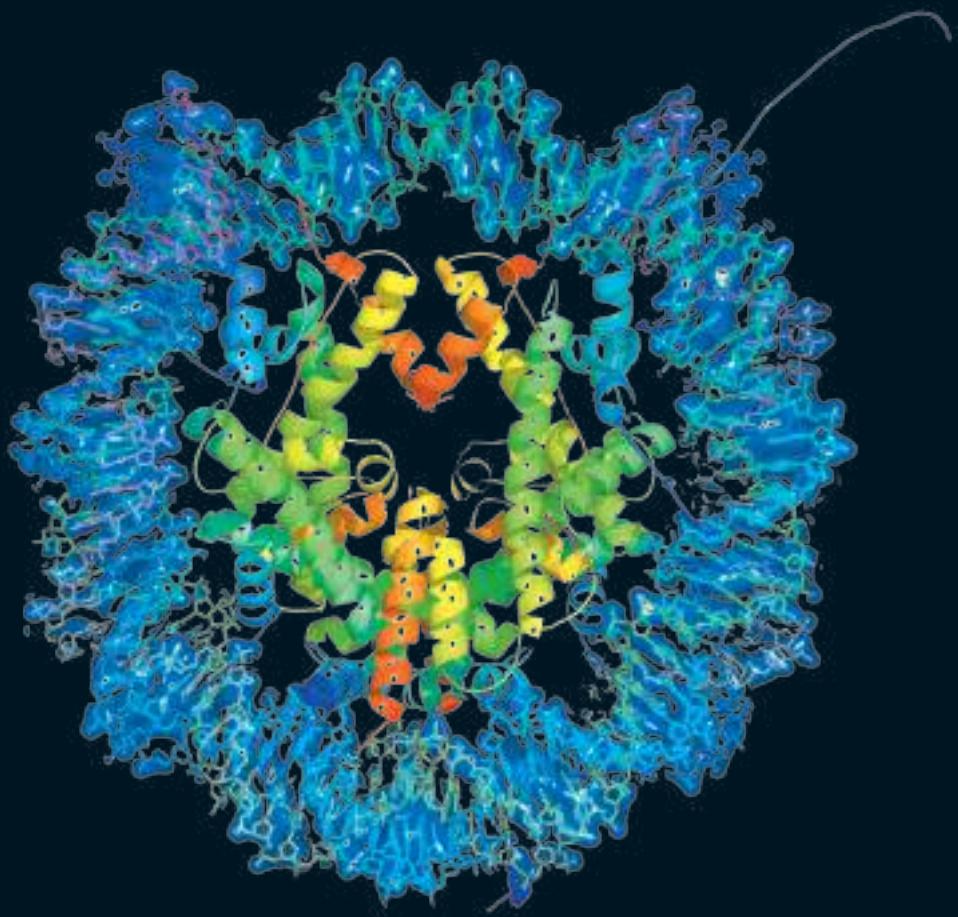


PACIFIC SYMPOSIUM ON BIOCOPUTING 2015



Edited by

**Russ B. Altman, A. Keith Dunker,
Lawrence Hunter, Marylyn D. Ritchie,
Tiffany Murray & Teri E. Klein**

PACIFIC SYMPOSIUM ON
BIOCOMPUTING 2015

PACIFIC SYMPOSIUM ON BIOCOMPUTING 2015

Kohala Coast, Hawaii, USA
4-8 January 2015

Edited by

Russ B. Altman
Stanford University, USA

A. Keith Dunker
Indiana University, USA

Lawrence Hunter
University of Colorado Health Sciences Center, USA

Marylyn D. Ritchie
The Pennsylvania State University, USA

Tiffany Murray
Stanford University, USA

Teri E. Klein
Stanford University, USA



NEW JERSEY • LONDON • SINGAPORE • BEIJING • SHANGHAI • HONG KONG • TAIPEI • CHENNAI

Published by

World Scientific Publishing Co. Pte. Ltd.

5 Toh Tuck Link, Singapore 596224

USA office: 27 Warren Street, Suite 401-402, Hackensack, NJ 07601

UK office: 57 Shelton Street, Covent Garden, London WC2H 9HE

ISSN: 2335-6936

British Library Cataloguing-in-Publication Data

A catalogue record for this book is available from the British Library.

BIOCOMPUTING 2015

Proceedings of the Pacific Symposium

Copyright © 2015 by World Scientific Publishing Co. Pte. Ltd.

All rights reserved. This book, or parts thereof, may not be reproduced in any form or by any means, electronic or mechanical, including photocopying, recording or any information storage and retrieval system now known or to be invented, without written permission from the publisher.

For photocopying of material in this volume, please pay a copying fee through the Copyright Clearance Center, Inc., 222 Rosewood Drive, Danvers, MA 01923, USA. In this case permission to photocopy is not required from the publisher.

ISBN 978-981-4644-73-0 (ebk)

Print-on-Demand in Singapore

Preface	vi
<i>A Twentieth Anniversary Tribute to PSB</i>	1
Darla Hewett, Michelle Whirl-Carrillo, Lawrence E. Hunter, Russ B. Altman, Teri E. Klein	
CANCER PANOMICS: COMPUTATIONAL METHODS AND INFRASTRUCTURE FOR INTEGRATIVE ANALYSIS OF CANCER HIGH-THROUGHPUT “OMICs” DATA	
<i>Session Introduction.....</i>	8
Søren Brunak, Francisco M. De La Vega, Adam Margolin, Benjamin J. Raphael, Gunnar Rätsch, Joshua M. Stuart	
<i>Cell Index Database (CELLX): A Web Tool for Cancer Precision Medicine</i>	10
Keith A. Ching, Kai Wang, Zhengyan Kan, Julio Fernandez, Wenyan Zhong, Jarek Kostrowicki, Tao Xie, Zhou Zhu, Jean-Francois Martini, Maria Koehler, Kim Arndt, Paul Rejto	
<i>Comparing Nonparametric Bayesian Tree Priors for Clonal Reconstruction of Tumors.....</i>	20
Amit G. Deshwar, Shankar Vembu, Quaid Morris	
<i>Stepwise Group Sparse Regression (SGSR): Gene-Set-Based Pharmacogenomic Predictive Models with Stepwise Selection of Functional Priors</i>	32
In Sock Jang, Rodrigo Dienstmann, Adam A. Margolin, Justin Guinney	
<i>Integrative Genome-wide Analysis of the Determinants of RNA Splicing in Kidney Renal Clear Cell Carcinoma</i>	44
Kjong-Van Lehmann, Andre Kahles, Cyriac Kandoth, William Lee, Nikolaus Schultz, Oliver Stegle, Gunnar Rätsch	
<i>An Integrated Framework for Reporting Clinically Relevant Biomarkers from Paired Tumor/Normal Genomic and Transcriptomic Sequencing Data In Support of Clinical Trials in Personalized Medicine.....</i>	56
Sara Nasser, Ahmet A. Kurdolgu, Tyler Izatt, Jessica Aldrich, Megan L. Russell, Alexis Christoforides, Wiabhav Tembe, Jeffery A. Keifer, Jason J. Corneveaux, Sara A. Byron, Karen M. Forman, Clarice Zuccaro, Jonathan J. Keats, Patricia M. LoRusso, John D. Carpten, Jeffrey M. Trent, David W. Craig	
<i>Characteristics of Drug Combination Therapy in Oncology by Analyzing Clinical Trial Data on Clinicaltrials.Gov.....</i>	68
Menghua Wu, Marina Sirota, Atul J Butte, Bin Chen	
CANCER PATHWAYS: AUTOMATIC EXTRACTION, REPRESENTATION, AND REASONING IN THE “BIG DATA” ERA	
<i>Session Introduction.....</i>	80
Graciela González, Chitta Baral, Jeff Kiefer, Suengchan Kim, Jieping Ye	
<i>Identifying Mutation Specific Cancer Pathways Using a Structurally Resolved Protein Interaction Network.....</i>	84
H. Billur Engin, Matan Hofree, Hannah Carter	

<i>Binning Somatic Mutations Based on Biological Knowledge for Predicting Survival: An Application in Renal Cell Carcinoma.....</i>	96
Dokyoon Kim, Ruowang Li, Scott M. Dudek, John R. Wallace, Marylyn D. Ritchie	
<i>Topological Features in Cancer Gene Expression Data.....</i>	108
Svetlana Lockwood, Bala Krishnamoorthy	
<i>Distant Supervision for Cancer Pathway Extraction from Text</i>	120
Hoifung Poon, Kristina Toutanova, Chris Quirk	
<i>Unsupervised Feature Construction and Knowledge Extraction from Genome-Wide Assays of Breast Cancer with Denoising Autoencoders.....</i>	132
Jie Tan, Matthew Ung, Chao Cheng, Casey S. Greene	
<i>Automated Gene Expression Pattern Annotation in the Mouse Brain.....</i>	144
Tao Yang, Xinlin Zhao, Binbin Lin, Tao Zeng, Shuiwang Ji, Jieping Ye	
CHARACTERIZING THE IMPORTANCE OF ENVIRONMENTAL EXPOSURES, INTERACTIONS BETWEEN THE ENVIRONMENT AND GENETIC ARCHITECTURE, AND GENETIC INTERACTIONS: NEW METHODS FOR UNDERSTANDING THE ETIOLOGY OF COMPLEX TRAITS AND DISEASE	
<i>Session Introduction.....</i>	156
Molly Hall, Shefali Setia Verma, Dennis P. Wall, Jason H. Moore, Brendan Keating, Daniel B. Campbell, Gregory Gibson, Folkert W. Asselbergs, Sarah A. Pendergrass	
<i>Measures of Exposure Impact Genetic Association Studies: An Example in Vitamin K Levels and VKORC1</i>	161
Dana C. Crawford, Kristin Brown-Gentry, Mark J. Rieder	
<i>A Bipartite Network Approach to Inferring Interactions Between Environmental Exposures and Human Diseases</i>	171
Christian Darabos, Emily D. Grussing, Maria E. Cricco, Kenzie A. Clark, Jason H. Moore	
<i>A Screening-Testing Approach for Detecting Gene-Environment Interactions Using Sequential Penalized and Unpenalized Multiple Logistic Regression</i>	183
H. Robert Frost, Angeline S. Andrew, Margaret R. Karagas, Jason H. Moore	
<i>Variable Selection Method for the Identification of Epistatic Models.....</i>	195
Emily Rose Holzinger, Silke Szymczak, Abhijit Dasgupta, James Malley, Qing Li, Joan E. Bailey-Wilson	
<i>Genome-Wide Genetic Interaction Analysis of Glaucoma Using Expert Knowledge Derived from Human Phenotype Networks.....</i>	207
Ting Hu, Christian Darabos, Maria E. Cricco, Emily Kong, Jason H. Moore	
<i>Identification of Gene-Gene and Gene-Environment Interactions Within the Fibrinogen Gene Cluster for Fibrinogen Levels in Three Ethnically Diverse Populations.....</i>	219
Janina M. Jeff, Kristin Brown-Gentry, Dana C. Crawford	

<i>Development of Exposome Correlations Globes to Map Out Environment-Wide Associations.....</i>	231
Chirag J. Patel, Arjun K. Manrai	

<i>Mitochondrial Variation and the Risk of Age-Related Macular Degeneration Across Diverse Populations</i>	243
--	-----

Nicole A. Restrepo, Sabrina L. Mitchell, Robert J. Goodloe, Deborah G. Murdock, Jonathan L. Haines, Dana C. Crawford

<i>iPINBPA: An Integrative Network-Based Functional Module Discovery Tool for Genome-Wide Association Studies</i>	255
---	-----

Lili Wang, Parvin Mousavi, Sergio E. Baranzini

CROWDSOURCING AND MINING CROWD DATA

<i>Session Introduction.....</i>	267
----------------------------------	-----

Robert Leaman, Benjamin M. Good, Andrew I. Su, Zhiyong Lu

<i>Reputation-Based Collaborative Network Biology.....</i>	270
--	-----

The sbv IMPROVER project team (in alphabetical order): Jean Binder, Stephanie Boue, Anselmo Di Fabio, R. Brett Fields, William Hayes, Julia Hoeng, Jennifer S. Park, Manuel C. Peitsch

<i>Microtask Crowdsourcing for Disease Mention Annotation in PubMed Abstracts.....</i>	282
--	-----

Benjamin M. Good, Max Nanis, Chunlei Wu, Andrew I. Su

<i>Crowdsourcing Image Annotation for Nucleus Detection and Segmentation in Computational Pathology: Evaluating Experts, Automated Methods, and the Crowd</i>	294
---	-----

Humayun Irshad, Laleh Montaser-Kouhsari, Gail Waltz, Octavian Bucur, Jonathan A Nowak, Fei Dong, Nicholas W Knoblauch, Andrew H Beck

<i>Analyzing Search Behavior of Healthcare Professionals for Drug Safety Surveillance</i>	306
---	-----

David J. Odgers, Rave Harpaz, Alison Callahan, Gregor Stiglic, Nigam H. Shah

<i>Refining Literature Curated Protein Interactions Using Expert Opinions.....</i>	318
--	-----

Oznur Tastan, Yanjun Qi, Jaime G. Carbonell, Judith Klein-Seetharaman

<i>Crowdsourcing RNA Structural Alignments with an Online Computer Game.....</i>	330
--	-----

Jérôme Waldspühl, Arthur Kam, Paul P. Gardner

PERSONALIZED MEDICINE: FROM GENOTYPES, MOLECULAR PHENOTYPES AND THE QUANTIFIED SELF, TOWARDS IMPROVED MEDICINE

<i>Session Introduction.....</i>	342
----------------------------------	-----

Joel T. Dudley, Jennifer Listgarten, Oliver Stegle, Steven E. Brenner, Leopold Parts

<i>KLEAT: Cleavage Site Analysis of Transcriptomes.....</i>	347
---	-----

Inanç Birol, Anthony Raymond, Readman Chiu, Ka Ming Nip, Shaun D Jackman, Maayan Kreitzman, T. Roderick Docking, Catherine A. Ennis, A. Gordon Robertson, Aly Karsan

<i>Causal Inference in Biology Networks with Integrated Belief Propagation.....</i>	359
---	-----

Rui Chang, Jonathan R. Karr, Eric E. Schadt

<i>Machine Learning from Concept to Clinic: Reliable Detection of BRAF V600E DNA Mutations in Thyroid Nodules Using High-Dimensional RNA Expression Data</i>	371
James Diggans, Su Yeon Kim, Zhanzhi Hu, Daniel Pankratz, Mei Wong, Jessica Reynolds, Ed Tom, Moraima Pagan, Robert Monroe, Juan Rosai, Virginia A. Livolsi, Richard B. Lanman, Richard T. Kloos, P. Sean Walsh, Giulia C. Kennedy	
<i>A Systematic Assessment of Linking Gene Expression with Genetic Variants for Prioritizing Candidate Targets</i>	383
Hua Fan-Minogue, Bin Chen, Weronika Sikora-Wohlfeld, Marina Sirota, Atul J. Butte	
<i>Drug-induced mRNA Signatures Are Enriched for the Minority of Genes that Are Highly Heritable</i>	395
Tianxiang Gao, Petter Brodin, Mark M Davis, Vladimir Jovic	
<i>An Integrative Pipeline for Multi-Modal Discovery of Disease Relationships</i>	407
Benjamin S. Glicksberg, Li Li, Wei-Yi Cheng, Khader Shameer, Jörg Hakenberg, Rafael Castellanos, Meng Ma, Lisong Shi, Hardik Shah, Joel T. Dudley, Rong Chen	
<i>PEAX: Interactive Visual Analysis and Exploration of Complex Clinical Phenotype and Gene Expression Association</i>	419
Michael A. Hinterberg, David P. Kao, Michael R. Bristow, Lawrence E. Hunter, J. David Port, Carsten Görg	
<i>T-ReCS: Stable Selection of Dynamically Formed Groups of Features with Application to Prediction of Clinical Outcomes</i>	431
Grace T. Huang, Ioannis Tsamardinos, Vineet Raghu, Naftali Kaminski, Panayiotis V. Benos	
<i>Meta-Analysis of Differential Gene Co-Expression: Application to Lupus</i>	443
Sumit B. Makashir, Leah C. Kottyan, Matthew T. Weirauch	
<i>Melancholic Depression Prediction by Identifying Representative Features in Metabolic and Microarray Profiles with Missing Values</i>	455
Zhi Nie, Tao Yang, Yashu Liu, Binbin Lin, Qingyang Li, Vaibhav A Narayan, Gayle Wittenberg, Jieping Ye	
<i>BayClone: Bayesian Nonparametric Inference of Tumor Subclones Using NGS Data</i>	467
Subhajit Sengupta, Jin Wang, Juhee Lee, Peter Müller, Kamalakar Gulukota, Arunava Banerjee, Yuan Ji	
WORKSHOPS	
<i>Human Evolutionary Genomics and the Search for the Genes that Made Us Human</i>	479
James M. Sikela	
<i>Discovery Informatics in Biological and Biomedical Sciences: Research Challenges and Opportunities</i>	482
Vasant Honavar	

<i>Inviting the Public: The Impact on Informatics Arising from Emerging Global Health Research Paradigms.....</i>	483
Richard Gayle, Mark Minie, Erik Nilsson	
<i>Training the Next Generation of Quantitative Biologists in the Era of Big Data.....</i>	488
Kristine A. Pattin, Anna C. Greene, Russ B. Altman, Kevin B. Cohen, Elizabeth Wethington, Carsten Görg, Lawrence E. Hunter, Spencer V. Muse, Predrag Radivojac, Jason H. Moore	
ERRATUM: <i>Next-Generation Analysis of Cataracts: Determining Knowledge Driven Gene-Gene Interactions Using Biofilter, and Gene-Environment Interactions Using the PhenX Toolkit.....</i>	493

PACIFIC SYMPOSIUM ON BIOCOMPUTING 2015

2015 marks the 20th Pacific Symposium on Biocomputing (PSB)! PSB was founded by Larry Hunter and Teri Klein, who had previously organized a Biocomputing session at HICSS (Hawaii International Conference on Systems Sciences) which grew too big for its host conference. They decided to split the Biocomputing track and created PSB—in the early years they recruited Russ Altman and Keith Dunker as co-organizers, and were pleased to add Marylyn Ritchie more recently. The mission of PSB is to provide a forum for the best emerging science in Biocomputing, providing both formal and informal mechanisms for scientific communication—with an emphasis on work in the pacific rim. In addition to being published by World Scientific and indexed in PubMED, the proceedings from all PSB meetings are available online at <http://psb.stanford.edu/psb-online/>. PSB has published more than 800 papers. These papers are often cited in journal articles that emerge early in the growth of a new subfield, and many papers have achieved hundreds of citations. The Twitter handle PSB 2015 is @PacSymBiocomp and the hashtag this year will be #psb15.

PSB depends on the community to define emerging areas in biomedical computation. Its sessions are usually conceived at the previous PSB meeting as people discuss trends and opportunities for new science. The typical program includes sessions that evolve over two to three years as well as entirely new sessions. This year we revisit cancer and personalized medicine (which continue to advance at dazzling speed), and add new sessions on crowdsourcing and gene-environment exposures. The efforts of a dedicated group of session organizers has produced an outstanding program, including introductory tutorials. The sessions of PSB 2015 and their hard-working organizers are as follows:

Cancer Panomics: Computational Methods and Infrastructure for Integrative Analysis of Cancer High-Throughput "Omics" Data to Enable Precision Oncology

Soren Brunak, Francisco M. De La Vega, Adam Margolin, Ben J. Raphael, Gunnar Raetsch, Joshua M. Stuart

Characterizing the Importance of Environmental Exposures, Interactions between the Environment and Genetic Architecture, and Genetic Interactions: New Methods for Understanding the Etiology of Complex Traits and Disease

Sarah A. Pendergrass, Shefali Setia Verma, Molly Hall, Jason H. Moore, Brendan Keating, Scott Selleck, Greg Gibson, Folkert Asselbergs, Heather Volk, Issac Pessah, Dennis Wall, Daniel B. Campbell

Crowdsourcing and Mining Crowd Data

Robert Leaman, Benjamin Good, Andrew Su, Zhiyong Lu

Personalized Medicine: From Genotypes, Molecular Phenotypes and the Quantified Self, Towards Improved Medicine

Joel Dudley, Steven E. Brenner, Jennifer Listgarten, Leopold Parts, Oliver Stegle

Cancer Pathways: Automatic Extraction, Representation, and Reasoning in the Big Data Era

Graciela H. González, Chitta Baral, Suengchan Kim, Jeff Kiefer, Jieping Ye

We are also pleased to present four workshops in which investigators with a common interest come together to exchange results and new ideas in a format that is more informal than the peer-reviewed sessions. For this year, the workshops and their organizers are:

Human Evolutionary Genomics and the Search for the Genes that Made Us Human

James M. Sikela

Training the Next Generation of Quantitative Biologists in the Era of Big Data
Kristine A. Pattin, Anna C. Greene, Jason Moore

Discovery Informatics in Biological and Biomedical Sciences: Research Challenges and Opportunities
Vasant G. Honavar

Including the Public in Research Projects: The impact on informatics arising from emerging health research paradigms

Richard Gayle, Mark Minie and Erik Nilsson

We thank our keynote speakers David Haussler (Science keynote) and Lucila Ohno-Machado (Ethical, Legal and Social Implications keynote).

Tiffany Murray has managed the peer review process and assembly of the proceedings since 2003, and also plays a key role in many other aspects of the meeting. We are grateful for the support of the Computational Genetics Laboratory at Dartmouth College; the Institute for Computational Biology, a collaborative effort of Case Western Reserve University, the Cleveland Clinic Foundation, and University Hospitals; and DNAexus for their support of PSB 2015. We also thank the National Institutes of Health, the National Science Foundation, and the International Society for Computational Biology (ISCB) for travel grant support. We are particularly grateful to the onsite PSB staff Al Conde, Brant Hansen, Georgia Hansen, BJ McKay-Morrison, Jackson Miller, Kasey Miller, and Paul Murray for their assistance. We also acknowledge the many busy researchers who reviewed the submitted manuscripts on a very tight schedule. The partial list following this preface does not include many who wished to remain anonymous, and of course we apologize to any who may have been left out by mistake.

We look forward to a great meeting once again.

Aloha!

Pacific Symposium on Biocomputing Co-Chairs,
October 12, 2014

Russ B. Altman

Departments of Bioengineering, Genetics & Medicine, Stanford University

A. Keith Dunker

Department of Biochemistry and Molecular Biology, Indiana University School of Medicine

Lawrence Hunter

Department of Pharmacology, University of Colorado Health Sciences Center

Teri E. Klein

Department of Genetics, Stanford University

Marylyn D. Ritchie

Department of Biochemistry and Molecular Biology, Pennsylvania State University

Thanks to the reviewers...

Finally, we wish to thank the scores of reviewers. PSB aims for every paper in this volume be reviewed by three independent referees. Since there is a large volume of submitted papers, paper reviews require a great deal of work from many people. We are grateful to all of you listed below and to anyone whose name we may have accidentally omitted or who wished to remain anonymous.

Zaky Adam	Christophe Giraud-Carrier	Ake Lu	Srikanth Srikari
Rehan Akbani	Anthony Gitter	Morgan Magnin	Oliver Stegle
Frank Albert	Anna Goldenberg	Matt Mahoney	Yan Sun
Chloe-Agathe Azencott	Graciela Gonzalez	Kimberly McAllister	Yasuo Tabei
Terri Beaty	Mehmet Gonen	Simon McCallum	Tasnia Tahsin
Gurkan Bebek	Assaf Gottlieb	Matthew McKenna	Haixu Tang
Brady Bernard	Casey Greene	Brett McKinney	Luis Tari
Theresa Marie Bernardo	Justin Guinney	Jeremy Miller	Duncan Thomas
Andreas Beyer	Melissa Haendel	Stephen Montgomery	Nam Tran
Sourav S. Bhowmick	Robin Haw	Sean Mooney	Lisa Tucker-Kellogg
Olivier Bodenreider	Jingrui He	Jason Moore	Tamir Tuller
Jeff Bond	Laura Heiser	Quaid Morris	Anna Tyler
David Andre	Katherine Hoadley	Jonathan Mortenson	Ryan Urbanowicz
Broniatowski	Gabriel Hoffman	Sara Mostafavi	Alfonso Valencia
Andrew Brown	Christopher Homan	Elias Neto	Charles Vaske
Marija Buljan	Jun Luke Huan	Azadeh Nikfarjam	Shankar Vembu
John Burger	Hailiang Huang	Olga Nikolova	Jean-Philippe Vert
William Bush	Heng Huang	Larsson Omberg	Bjarni Vilhalmsson
Kim Bussey	Vladimir Jojic	Aydogan Ozcan	Tomas Vinar
Daniel Campbell	Siddhartha Jonnalagadda	Qinxin Pan	Julia Vogt
Peter Carbonetto	Sungwon Jung	Leopold Parts	Heather Volk
Gregory Carter	Anne Justice	Vicente Pelechano	Yuliang Wang
Hannah Carter	Scott Kahn	Minoli Perera	Davy Weissenbacher
Su-Shing Chen	Konrad Karczewski	Alex Perez	John Wilbur
Rumi Chunara	Sarvnaz Karimi	Fernando Perez-Cruz	John Witte
Giovanni Ciriello	Brendan Keating	Nico Pfeifer	Denise Wolf
Kevin Cohen	Eimear Kenny	Alex Pico	Karin Verspoor
Courtney D. Corley	Shameer Khader	Jason Piper	Jinbo Xu
Eivind Coward	Jeff Kiefer	Snehit Prabhu	Jie Yang
Dana Crawford	Seungchan Kim	Gerald Quon	Pew-Thian Yap
Nabarun Dasgupta	Dokyoon Kim	Archana Ramesh	Meliha Yetisen
Munmun De Choudhury	Robert Kincaid	Oliver Ray	Elad Yom-Tov
Emek Demir	David Knowles	Mireille Regnier	Kristin Young
Amit Deshwar	Isaac Kohane	Marylyn Ritchie	Shipeng Yu
Zhihao Ding	Smita Krishnaswamy	Anna Ritz	Yuan Yuan
Joel Dudley	Christine Ladd-Acosta	Sushmita Roy	Judith Zaugg
Suzanne Fei	Hayan Lee	Wolfgang Sadee	Ping Zhang
Elana Fertig	Kjong-Van Lehmann	Kadeem Ho Sang	
Karen Fort	Christina Leslie	Abeed Sarker	
Nicolo Fusci	Sarah Yuqing Li	Richa Saxena	
Eric Gamazon	Hua Li	Michael Schubert	
Nils Gehlenborg	Jian-Liang Li	Andy Schwartz	
Olivier Gevaert	Han Liang	Jun Sese	
Debashis Ghosh	Christoph Lippert	Takehiko Soh	
Eugenia Giannopoulou	Mei Liu	Artem Sokolov	
Greg Gibson		Joe Song	

A TWENTIETH ANNIVERSARY TRIBUTE TO PSB

DARLA HEWETT¹, MICHELLE WHIRL-CARRILLO¹, LAWRENCE E HUNTER²,
RUSS B ALTMAN¹, TERI E KLEIN¹

¹*Stanford University, Shriram Center for Bioengineering and Chemical Engineering
443 Via Ortega, Stanford, CA 94305*

Email: teri.klein@stanford.edu

²*University of Colorado School of Medicine, Computational Bioscience Program,
Aurora CO 80045*

PSB brings together top researchers from around the world to exchange research results and address open issues in all aspects of computational biology. PSB 2015 marks the twentieth anniversary of PSB. Reaching a milestone year is an accomplishment well worth celebrating. It is long enough to have seen big changes occur, but recent enough to be relevant for today. As PSB celebrates twenty years of service, we would like to take this opportunity to congratulate the PSB community for your success. We would also like the community to join us in a time of celebration and reflection on this accomplishment.

1. PSB's Influence

PSB is one of the world's leading conferences in computational biology. It is where top researchers present and discuss current research in the theory and application of computational methods in problems of biological significance. The following facts, computed October 2014, highlight how PSB has impacted and supported our community:

- PSB has accepted and is tracking 887 papers.
- 81% of the papers submitted to PSB have been cited in Google Scholar.
- On the average, if a PSB paper has citations recorded in Google Scholar, it is cited 45 times.
- There are currently 32,504 citations of PSB Papers recorded in Google Scholar.
- The most highly cited PSB paper has 978 citations recorded in Google Scholar.
- 538 organizations have contributed to PSB.
- 2583 individuals have contributed 3671 times to PSB as an author, a session leader, or a workshop leader.

2. PSB Top 10 Papers

PSB has advanced science in many ways but these top ten highly cited papers, and the topics they represent, have been of particular importance to our community.

Session	Title	Authors	Description	Year
Genetic Relationships	REVEAL, A General Reverse Engineering Algorithm for Inference of Genetic Network Architectures	S Liang S Fuhrman R Somogyi	An algorithm for inferring genetic network architectures from state transition tables, which correspond to time series of gene expression patterns, using the Boolean network model.	1998
DNA Patterns	BioProspector: Discovering Conserved DNA Motifs in Upstream Regulatory Regions of Co-Expressed Genes	X Liu DL Brutlag JS Liu	BioProspector examines the upstream region of genes in the same gene expression pattern group and looks for regulatory sequence motifs.	2001
Protein Classification	The Spectrum Kernel: A String Kernel for SVM Protein Classification	C Leslie E Eskin WS Noble	String-based kernels, in conjunction with SVMs (support vector machines), offer a viable and computationally efficient alternative to other methods of protein classification and homology detection.	2002
Genetic Relationships	Principal Components Analysis to Summarize Microarray Experiments: Application To Sporulation Time Series	S Raychaudhuri JM Stuart RB Altman	Application of PCA to expression data provides a summary of the ways in which gene responses vary under different conditions. This analysis clarifies the relationship between previously reported clusters and is used to examine the relationships and differences between genes.	2000
Genetic Relationships	Mutual Information Relevance Networks: Functional Genomic Clustering Using Pairwise Entropy Measurements	AJ Butte IS Kohane	A technique that computes comprehensive pair-wise mutual information for all genes in RNA expression data to find functional genomic clusters.	2000
Gene Expression	Modeling Gene Expression with Differential Equations	T Chen HL He GM Church	A differential equation model for gene expression with two methods to construct the model from a set of temporal data.	1999
Genetic Relationships	Identification of Genetic Networks from a Small Number of Gene Expression Patterns Under the Boolean Network Model	T Akutsu S Miyano S Kuhara	REVEAL: improvement and mathematical proof.	1999

Genetic Relationships	Linear Modeling of mRNA Expression Levels During CNS Development and Injury	P D'Haeseleer X Wen S Furhman R Somogyi	A linear modeling approach that allows one to infer interactions between all the genes included in a data set. The resulting model can be used to generate interesting hypotheses to direct further experiments.	1999
Gene Expression	Modeling Regulatory Networks with Weight Matrices	DC Weaver CT Workman GD Stormo	Representing regulatory relationships between genes as linear coefficients or weights, with the “net” regulation influence on a gene’s expression being the mathematical summation of the independent regulatory inputs.	1999
Gene Expression	Using Graphical Models and Genomic Expression Data to Statistically Validate Models of Genetic Regulatory Networks.	AJ Hartemink DK Gifford T Jaakkola RA Young	A model-driven approach for analyzing genomic expression data that permits genetic regulatory networks to be represented in a biologically interpretable computational form.	2001

3. Twenty contributors to celebrate twenty years

Thousands of contributors have made significant and relevant contribution to PSB. However, we would like to highlight twenty people whom have recorded the most contributions to PSB. We believe these contributors should be counted on anyone's list of important contributors to the field of biological research. Since selection was based on contribution to PSB, not every important figure is included; and some people are included who would not describe themselves as important contributors despite their considerable impact on the field. However, based on the number and significance of their contributions to PSB, here is a list of twenty contributors who we would like to acknowledge as we celebrate the twentieth anniversary of PSB.

Russ Altman, MD, PhD

<http://helix-web.stanford.edu/>

Professor of Bioengineering, Genetics, and Medicine, (and Computer Science, by courtesy), Stanford University



Russ Altman directs the Helix group. The Helix group focuses on the creation and application of computational tools to solve problems in biology and medicine.

Francisco M. De La Vega, D.Sc.

[http://www.annaisystems.com/
management-team-3/](http://www.annaisystems.com/management-team-3/)

*Chief Scientific Officer,
Anni Systems*



Francisco De La Vega is a geneticist and computational biologist who has focused on analytical problems of large scale sequencing projects. Anni is a genomics management and analytics start-up, currently with a beachhead in cancer genomics.

Graciela Gonzalez, PhD

[https://webapp4.asu.edu/directory/
person/869038](https://webapp4.asu.edu/directory/person/869038)

Associate Professor and Data Core Director for NIH/NIA supported Alzheimer's Disease Centers, Arizona State University



Graciela Gonzalez leads the DIEGO (Discovery through Integration and Extraction of Genomic knOwledge) lab focusing her research on translational applications of information extraction using Natural Language Processing techniques.

Atul Butte, MD, PhD

<https://buttelab.stanford.edu/>

Associate Professor of Pediatrics & of Genetics and, by courtesy, of Computer Science, of Medicine & (BMIR) & of Pathology, Stanford University



The long-term research goal of the Butte Lab is to solve problems relevant to genomic medicine by developing new methodologies in translational bioinformatics.

Keith Dunker, PhD

[http://www.compbio.iupui.edu/group/5/
pages/about_us](http://www.compbio.iupui.edu/group/5/pages/about_us)

Director, Center for Computational Biology and Bioinformatics, Professor, Biochemistry and Molecular Biology, Professor, School of Informatics, Indiana University



Keith Dunker and his collaborators were the first to consider intrinsically disordered proteins as a distinct class with important biological functions. Ongoing work suggests that the original proteins on earth were intrinsically disordered and that protein evolution followed a disorder to order pathway.

Alexander J. Hartemink, PhD

<http://www.cs.duke.edu/~amink/>

*Professor of Computer Science, Statistical Science, and Biology
Duke University*



Alexander Hartemink's research interests are in computational systems biology and machine learning. Specifically, his work focuses on the development and application of new statistical learning algorithms to complex problems in systems biology.

David Haussler, PhD

<https://genomics.soe.ucsc.edu/haussler>
Investigator, Howard Hughes Medical Institute, Director and Distinguished Professor, Center for Biomolecular Science and Engineering, Director, Cancer Genomics Hub, Scientific Co-Director, California Institute for Quantitative Biosciences (QB3), University of California, Santa Cruz



David Haussler is known for his work leading the team that assembled the first human genome sequence. He is credited with pioneering the use of hidden Markov models (HMMs), stochastic context-free grammars, and the discriminative kernel method for analyzing DNA, RNA, and protein sequences. He was the first to apply the latter methods to the genome-wide search for gene expression biomarkers in cancer, now a major effort of his laboratory.

Teri E. Klein, PhD

<https://med.stanford.edu/profiles/teri-klein>
Director & Co-Principal Investigator, PharmGKB, Stanford University



Teri Klein's research interests extend over the broad spectrum of pharmacogenetics, computational biology and bioinformatics. Applications include the development of a pharmacogenetics knowledge base, structure-function relationships, de novo modeling and the structural basis of disease.

Satoru Kuhara, PhD

<http://hyoka.ofc.kyushu-u.ac.jp/search/details/K000030/english.html>

Professor, Bioscience & Biotechnology, Faculty of Agriculture, Kyushu University



Satoru Kuhara focus on the transcriptional regulation, the first step of the gene regulatory circuit and the study of the regulatory network in cells and in multi-cellular organization.

Lawrence Hunter, PhD

<http://compbio.ucdenver.edu/hunter/>

Director, Center for Computational Pharmacology, Computational Bioscience Program, School of Medicine at the University of Colorado



Larry Hunter's laboratory is focused on knowledge-driven extraction of information from the primary biomedical literature, the semantic integration of knowledge resources in molecular biology, and the use of knowledge in the analysis of high-throughput data.

Isaac S. Kohane, MD, PhD

<http://chip.org/zak/>

Professor of Pediatrics & Health Sciences Technology, Harvard Medical School



Isaac Kohane is transforming healthcare systems into discovery engines to accelerate innovation in clinical decision-support. He is working on neurodevelopmental diseases (NDD) and specifically autism spectrum disorder. Understanding the full landscape of its clinical manifestations, emphasizing early and systematic molecular-clinical diagnosis through the creation of an NDD information commons.

Shoudan Liang, PhD

<http://profiles.gulfcoastconsortia.org/profilesystem/editprofile.php?pid=3461>
Professor, Bioinformatics and Computational Biology, UT MD Anderson Cancer Center



Shoudan Liang creates algorithms to analyze microarray data with the aim of understanding the genetic circuitry that underlies diverse systems such as stem cell differentiation, development in animals, proliferation and apoptosis in immune and cancer cells.

Jun S. Liu, PhD

<http://www.people.fas.harvard.edu/~junliu/>

*Professor of Statistics
Harvard University*



Jun Liu thinks he can solve some of the mysteries about genes quicker with statistics than biologists can with laboratory experiments. Using a method that can be illustrated on a blackboard, he has successfully predicted the locations of on/off switches for genes in a bacterium.

Satoru Miyano, PhD

<http://dnagarden.hgc.jp/en/doku.php>

*Professor, Human Genome Center
Institute of Medical Science
The University of Tokyo*



Satoru Miyano's mission is to create computational strategy for systems biology and medicine towards translational bioinformatics. The supercomputer system is the indispensable infrastructure for his mission.

Marylyn Ritchie, PhD

<http://bmb.psu.edu/directory/mdr23>

*Director, Center for Systems Genomics
Professor of Biochemistry and Molecular Biology,
The Pennsylvania State University*



The mission of the Ritchie Lab is to improve our understanding of the underlying genetic architecture of common diseases. New methods focus on the detection of gene-gene interactions, gene-environment interactions, and network and/or pathway effects associated with human disease.

Gary Stormo, PhD

<http://stormo.wustl.edu/>

*Professor, Department of Genetics
and the Center for Genome Sciences
and Systems Biology, Washington
University School of Medicine in
St Louis*



The Stormo Lab's work is focused on understanding and modeling the regulation of gene expression, especially the basis of specificity in protein-nucleic acid interactions.

Xiaole Shirley Liu, PhD

<http://liulab.dfcf.harvard.edu/>

*Professor of Biostatistics
Harvard School of Public Health*



X. Shirley Liu has helped develop a number of widely used algorithms for transcription factor motif finding, ChIP-chip/Seq and DNase-seq data analysis, and is developing more algorithms and data integration approaches for high throughput data in transcriptional and epigenetic gene regulation.

William Stafford Noble, PhD

<http://noble.gs.washington.edu/>

*Professor of Genome Sciences
University of Washington*



Bill Noble's research focuses on the development and application of machine learning and statistical methods for interpreting complex biological data sets. Currently, his research can be divided into three areas: predicting protein properties, chromatin and gene regulation, and analysis of mass spectrometry data.

Roland Somogyi, PhD

<http://www.molecularmining.com/>

*Chief Scientific Officer,
Molecular Mining*



Roland Somogyi is developing the tools that will help us understand how genes work together in gene networks. He is known for his work in using gene expression data to reverse-engineer gene networks.

Richard A. Young, PhD

<http://younglab.wi.mit.edu/>

*Professor of Biology
Massachusetts Institute of Technology*



The Young laboratory is mapping the regulatory circuitry that controls cell state and differentiation in mice and humans. The Young lab uses experimental and computational technologies to determine how signaling pathways, transcription factors, chromatin regulators and small RNAs control gene expression programs in embryonic stem cells and differentiated cells.

4. Acknowledging PSB's Supporting Organizations

We would also like to acknowledge and thank the 538 organizations that have supported the 2583 authors, session leaders, and workshop leaders of PSB. We do understand that behind the contributors to PSB there are many organizations supporting their contribution. Organizational support of our community has been vital to PSB's success.

5. Looking Ahead

In contemplating a vision for the future of biological research, it is appropriate to consider the remarkable path that has brought us here and the significant role PSB has had in the journey. While we celebrate our progress, on this twentieth anniversary we also want to look ahead to the future. Looking ahead we have the opportunity to explore a new vision, of transformative new approaches to achieve health benefits. Although innovative biological analysis methods are rapidly permeating biomedical research, the challenge of establishing robust paths from knowledge and information to improved human health remains. We look forward to the future, hearing from you, collaborating with you, reading your papers, and being challenged by your continued innovation.

Congratulations PSB.

References

1. Reveal, a general reverse engineering algorithm for inference of genetic network architectures.
Liang S, Fuhrman S, Somogyi R. Pac Symp Biocomput. 1998; :18-29.
2. BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes.
Liu X, Brutlag DL, Liu JS. Pac Symp Biocomput. 2001 :127-38.
3. The spectrum kernel: a string kernel for SVM protein classification.
Leslie C, Eskin E, Noble WS. Pac Symp Biocomput. 2002 :564-75.
4. Principal components analysis to summarize microarray experiments: application to sporulation time series.
Raychaudhuri SI, Stuart JM, Altman RB. Pac Symp Biocomput. 2000 :455-66.
5. Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements.
Butte AJ, Kohane IS. Pac Symp Biocomput. 2000 :418-29.
6. Modeling gene expression with differential equations. *Chen T, He HL, Church GM.*
Pac Symp Biocomput. 1999 :29-40.
7. Identification of genetic networks from a small number of gene expression patterns under the Boolean network model. *Akutsu T, Miyano S, Kuhara S.* Pac Symp Biocomput. 1999 :17-28.
8. Linear modeling of mRNA expression levels during CNS development and injury.
D'haeseleer P, Wen X, Fuhrman S, Somogyi R. Pac Symp Biocomput. 1999 :41-52.
9. Modeling regulatory networks with weight matrices. *Weaver DC, Workman CT, Stormo GD.*
Pac Symp Biocomput. 1999 :112-23.
10. Using graphical models and genomic expression data to statistically validate models of genetic regulatory networks. *Hartemink AJ, Gifford DK, Jaakkola TS, Young RA.* Pac Symp Biocomput. 2001 :422-33.
11. A Vision for the Future of Genomics Research. *Francis S. Collins, Eric D. Green, Alan E. Guttmacher and Mark S. Guyer.* Nature, Vol. 422, No. 6934, April 24, 2003, p. 835-847.

CANCER PANOMICS: COMPUTATIONAL METHODS AND INFRASTRUCTURE FOR INTEGRATIVE ANALYSIS OF CANCER HIGH-THROUGHPUT “OMICS” DATA

SØREN BRUNAK

*Center for Biological Sequence Analysis Department of Systems Biology,
Technical University of Denmark, Copenhagen, Denmark
Email: brunak@cbs.dtu.dk*

FRANCISCO M. DE LA VEGA

*Annai Systems, Inc., Burlingame, CA, USA
Email: Francisco.dlv@gmail.com*

ADAM MARGOLIN

*Oregon Health & Science University, Portland, OR, USA
Email: margolin@ohsu.edu*

BENJAMIN J. RAPHAEL

*Brown University, Providence, RI, USA
Email: braphael@brown.edu*

GUNNAR RÄTSCH

*Computational Biology Center,
Memorial Sloan-Kettering Cancer Center, New York City, NY, USA
Email: raetsch@cbio.mskcc.org*

JOSHUA M. STUART

*Center for Biomolecular Science and Engineering,
University of California Santa Cruz, CA, USA
Email: jstuart@soe.ucsc.edu*

Targeted cancer treatment is becoming the goal of newly developed oncology medicines and has already shown promise in some spectacular cases such as the case of BRAF kinase inhibitors in BRAF-mutant (e.g. V600E) melanoma.¹ These developments are driven by the advent of high-throughput sequencing, which continues to drop in cost, and that has enabled the sequencing of the genome, transcriptome, and epigenome of the tumors of a large number of cancer patients in order to discover the molecular aberrations that drive the oncogenesis of several types of cancer.² Applying these technologies in the clinic promises to transform cancer treatment by identifying therapeutic vulnerabilities of each patient’s tumor.³ These approaches will need to address the panomics of cancer – the integration of the complex combination of patient-specific characteristics that drive the development of each person’s tumor and response to therapy. This in turn necessitates new computational methods to integrate large-scale “omics” data for each patient with their electronic medical records, and in the context of the results from large-scale pan-cancer research studies, to select the best therapy and/or clinical trial for the patient at hand.

This session of the Pacific Symposium on BioComputing 2015 features papers that address a range of issues from the analytical methods to discover molecular aberration from high-throughput sequencing

data, the prediction of pharmacogenetic effects in oncology drugs, the implications of clinical trials in personalized oncology, and the resources and IT infrastructure required to support the analysis of patient data.

The first set of contributions target novel analysis methodologies. One of the challenges of analyzing sequencing data from tumor samples is the fact that tumors are heterogeneous, and often composed of many sub-clones with different somatic mutations. The ability to deconvolute these mixtures and understand what somatic mutations co-occur in a given sub-clone, is important to understand the drivers of drug resistance of a particular tumor lineage. The work of Deshwar *et al.* aims to compare novel Bayesian methods for sub-clonal reconstruction of tumors that provides faster execution time and better resolution than other commonly used methods. On the subject of the analysis of transcriptome datasets, Lehmann *et al.* tackle the analysis and interpretation of the variation of alternative splicing in tumor samples, and are able to correlate this variability with QTLs that map to variants previously implicated in susceptibility to cancer and other traits, information which could be helpful when integrating a patient's germline with their tumor genome and corresponding transcriptome. Jang *et al.* focus on the development of pharmacogenetic predictive models utilizing domain-specific priors, and demonstrate that stepwise group sparse regression performs more accurately and provides better interpretability than purely data driven methods. On the other hand, Wu *et al.*, deal with the upcoming challenge of combination therapy - how to design and interpret trials where two drugs are provided in combination and reported in ClinicalTrials.gov.

On the practical side, performing panomic analysis of patient samples in the clinic would require IT infrastructures and resources that can deliver results in a fast time frame and can support the analyst to interpret and validate new molecular biomarkers. Nasser *et al.* present an integrated framework to analyze and present genome/transcriptome data of patients focused on clinical interpretation. They describe the challenges in developing a platform that can deliver results in a 24-hour timeframe. Ching *et al.* focus on the development of an online resource that integrates expression, copy number variation, mutation, compound activity, and meta data from cancer cells coming from publicly available projects, the Cell Index Database.

Precision oncology will require the analysis of data collected from a single patient. It will need to draw timely hypotheses about treatment in the context of prior knowledge of the specific form of cancer and based on information about the individual patient. This " $n=1$ " approach is significantly different from that of pattern discovery on large patient cohorts. Continued development of new methodology to meet the demand is clearly still an active area to pursue by the computational biology community.

References

1. Ribas, A. & Flaherty, K. T. BRAF targeted therapy changes the treatment paradigm in melanoma. *Nature Publishing Group* **8**, 426–433 (2011).
2. Omberg, L. *et al.* Enabling transparent and collaborative computational analysis of 12 tumor types within The Cancer Genome Atlas. *Nat. Genet.* **45**, 1121–1126 (2013).
3. Craig, D. W. *et al.* Genome and transcriptome sequencing in prospective refractory metastatic triple negative breast cancer uncovers therapeutic vulnerabilities. *Mol. Cancer Ther.* **12**, 104–116 (2012).

CELL INDEX DATABASE (CELLX): A WEB TOOL FOR CANCER PRECISION MEDICINE*

KEITH A. CHING¹, KAI WANG¹, ZHENGYAN KAN¹, JULIO FERNANDEZ¹, WENYAN ZHONG¹, JAREK KOSTROWICKI¹, TAO XIE¹, ZHOU ZHU¹, JEAN-FRANCOIS MARTINI², MARIA KOEHLER², KIM ARNDT¹, PAUL REJTO¹

¹*Oncology Research Unit, ²Oncology Business Unit, Pfizer Global Research & Development, Pfizer Inc., 10777 Science Center Drive San Diego, CA 92121, USA Email: keith.ching@pfizer.com*

The Cell Index Database, (CELLX) (<http://cellx.sourceforge.net>) provides a computational framework for integrating expression, copy number variation, mutation, compound activity, and meta data from cancer cells. CELLX provides the computational biologist a quick way to perform routine analyses as well as the means to rapidly integrate data for offline analysis. Data is accessible through a web interface which utilizes R to generate plots and perform clustering, correlations, and statistical tests for associations within and between data types for ~20,000 samples from TCGA, CCLE, Sanger, GSK, GEO, GTEx, and other public sources. We show how CELLX supports precision oncology through indications discovery, biomarker evaluation, and cell line screening analysis.

1. Introduction

To support precision medicine patient selection strategies, genomics data is used to identify oncogenic drivers or dysregulated pathways in cancer cells susceptible to therapeutic intervention. Notably, efforts by The Cancer Genome Atlas (TCGA, <http://cancergenome.nih.gov>), the Cancer Cell Line Encyclopedia (CCLE)[1], and Sanger Wellcome Trust Genomics of Drug Sensitivity in Cancer (GDSC)[2] have generated a plethora of data and datatypes that can be used for generating patient selection hypotheses. However, multiple genomics data types such as expression, copy number variation (CNV), and mutation are large and unwieldy to manage. For the computational biologist, much time and effort can be spent to assemble an up to date table of features which can be computed on because new data are often generated frequently and incrementally. Thus, there is a need for an infrastructure to perform simple, quick, and routine analyses on multi-dimensional genomics data as well as the automated assembly of data tables for offline computation using more sophisticated algorithms.

Currently, there exist several cancer genomics databases to access expression, CNV, mutation, and integrated data as reviewed in [3]. For example, BioGPS[4] provides expression data, Tumorscape[5] contains CNV measurements, the Sanger Catalog of Somatic Mutations in Cancer (COSMIC)[6] lists mutations, and the cBio Portal[7] integrates multiple TCGA data types. Additionally, databases with compound activity data include GDSC and CCLE. Here we present a publicly available web-based informatics tool to integrate data, perform analysis, and visualize results from public as well as private internal sources to support precision medicine activities.

* This work is supported by Pfizer, Inc.

2. Architecture

The underlying MySQL database consists of 22 tables for expression, CNV, mutation, compound, sample, meta data, RNAi, RPPA, and gene annotation data. The Perl CELLX application runs on an Apache web server. R-serve (<http://www.rforge.net/Rserve/>) instances generate plots and perform statistical analyses. An Apache Tomcat application server runs a custom Java servlet which bridges Perl and R by funneling Perl http requests to the R-serves and sends results back to the web server. A demo site, instructions, source code, database dumps, and data parsing / loading scripts are available at <http://cellx.sourceforge.net>.

3. Gene Based Search

A common starting point for indications discovery is asking where the target of interest is altered. CELLX can plot the relative expression or CNV of a gene within a dataset or across multiple compatible datasets. For instance, RNA-Seq data processed by RSEM[8] can be compared across tumors profiled not only by TCGA, but CCLE as well. CDK4 expression can be seen to have high outliers in Glioblastoma Multiforme (GBM), melanoma (SKCM), breast (BRCA), Lower Grade Glioma (LGG), and sarcomas (SARC) (Figure 1). A similar plot can be generated of CNV to identify datasets with amplifications or deletions. CELLX can chart the relationship between expression and CNV across datasets using scatter plots of expression versus CNV. A hallmark of amplification, CDK4 expression levels scale with CNV level in several datasets (Figure 2a,b).

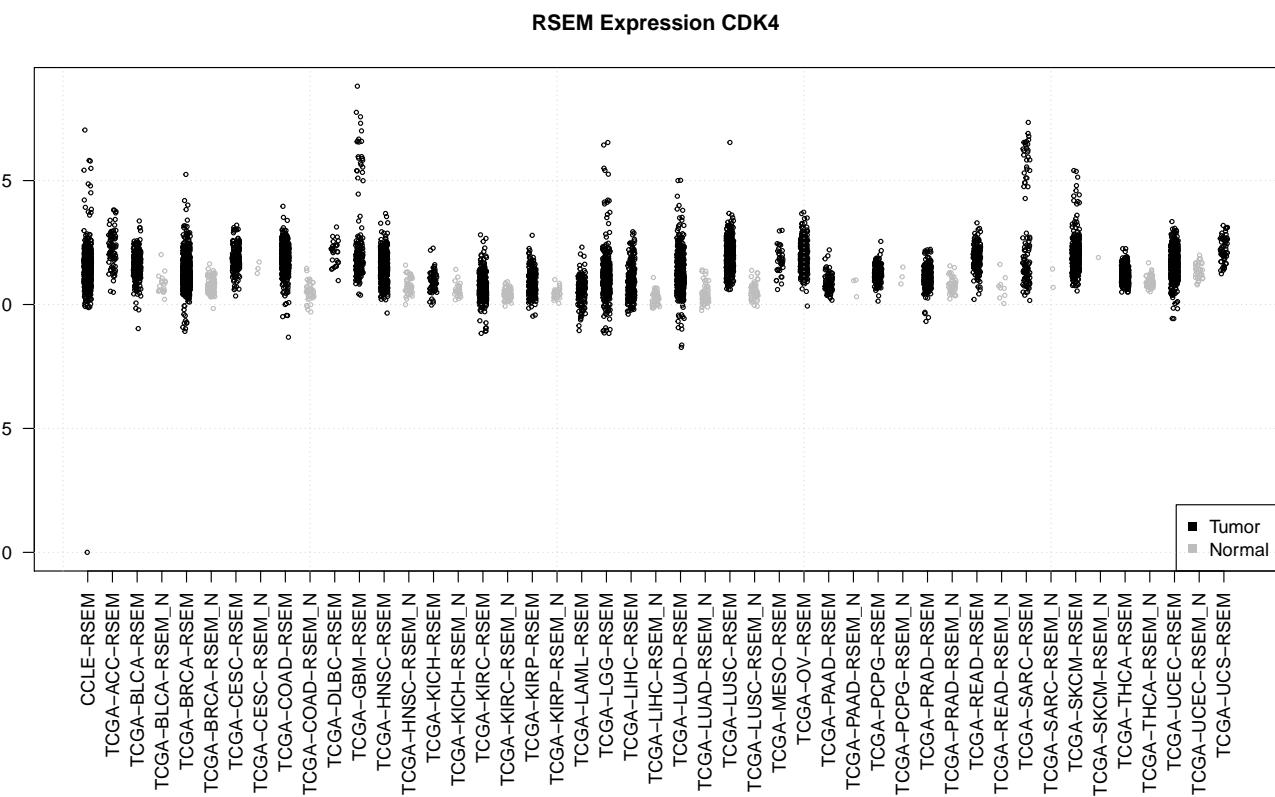


Figure 1. RNA-Seq RSEM gene expression of CDK4 (y-axis, log2) across datasets shows higher expression in tumor vs. adjacent normal tissue. Particular groups of outliers can be seen in GBM (glioblastoma multiforme), SARC (sarcoma), SKCM (skin cutaneous melanoma), LGG (brain lower grade glioma), and cell lines (CCLE).

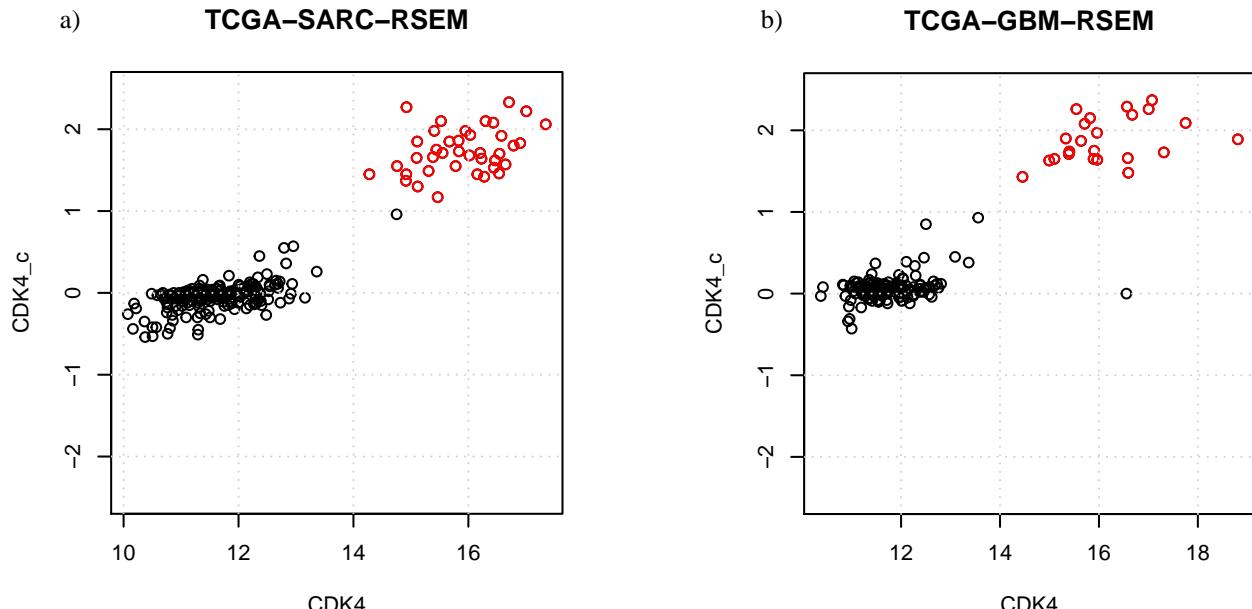


Figure 2. Correlation of expression and CNV. CNV (y-axis in log₂ diploid genome) vs. RSEM expression levels (log₂) for CDK4 show that a) SARC and b) GBM datasets have a sizable population of cells overexpressing CDK4 due to amplification of the locus. Additionally, expression levels scale with CNV levels. Clear outliers from the main distribution of CNV values can help determine appropriate CNV cut offs for amplification status. In this example, samples colored red have ≥ 1 log₂ diploid genomes (i.e. ≥ 4 copies).

4. Integrated Visualization

Mixed data types can be visualized in 2D scatter plots to look at the relationship between two datatypes on the same or different genes. For instance, expression of gene A on the x-axis can be plotted versus the CNV of gene B on the y-axis. Other plottable datatypes are protein levels for Reverse Phase Protein Arrays (RPPA), the mutation count per sample, the general amount of CNV per sample, IC50 values for compounds, and meta data. Multiple layers of data can be added to the plot to increase dimensionality. As a simple example, one can plot the expression of ERBB2 expression vs. ERBB2 CNV overlaid with ERBB2 mutations (Figure 3a) or breast cancer subtype meta data. (Figure 3b). The underlying data used to generate each plot is linked as a tab separated tsv file for downloading.

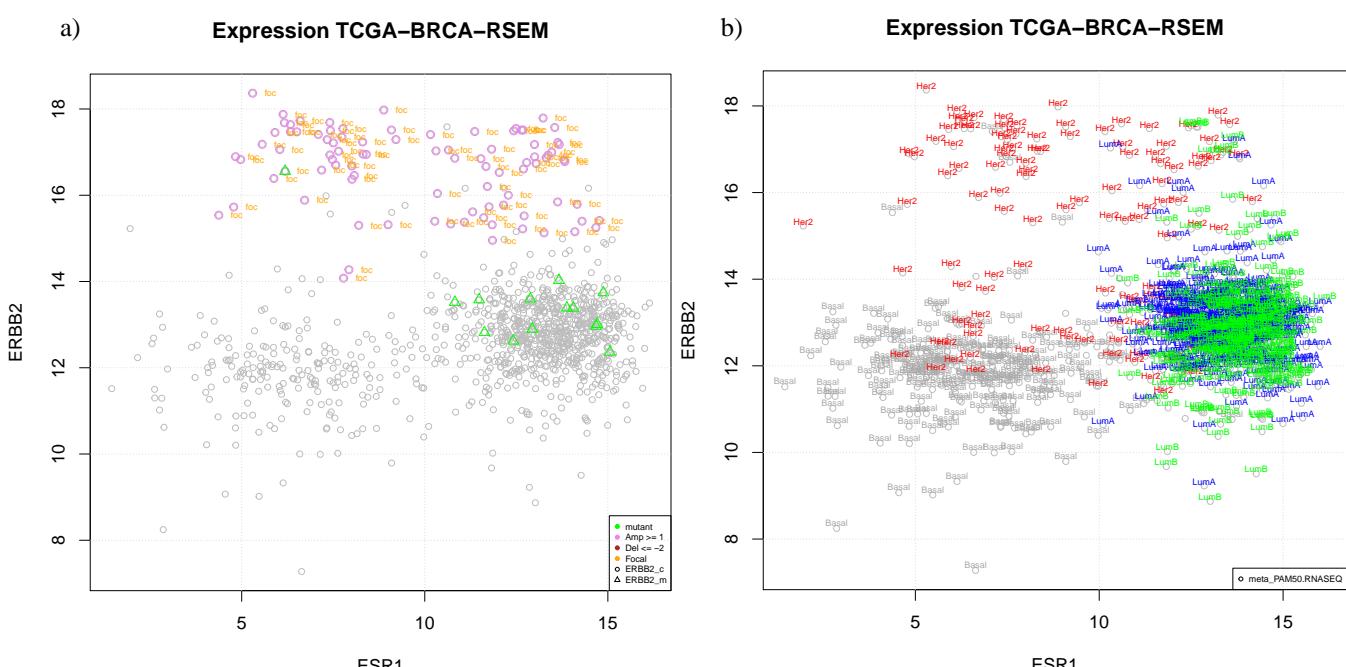


Figure 3. 2D scatter plots. a) Gene expression of ESR1 (x-axis, log₂) vs. ERBB2 (y-axis, log₂) gene expression. ERBB2 CNV over the selected threshold of 1 (log₂ diploid genome) is colored pink. Focal amplifications (≤ 10 MB) are denoted with 'foc'. Mutations in ERBB2 are colored green. c) Meta data for PAM50 subtype classification are colored and overlaid on the ESR1 vs. ERBB2 gene expression plot.

5. Biomarker Frequency Reports

Tables of the frequency of alterations across datasets can help to prioritize indications for therapies with known biomarkers. For instance, the venn report of the frequency of CDK4 biomarker alterations within datasets shows significant frequencies of CDK4 amplification in sarcoma, gliomas, and melanoma TCGA datasets (Table 1). Cutoffs can be defined by expression level, CNV level, and/or mutation status. The co-occurrence or exclusion of 2-4 biomarkers within the same sample can also be quantified.

Table 1. Frequency report for CDK4 alterations in TCGA. CDK4_c is the number of samples in which the CNV exceeds the set threshold, in this case ~4 copies. CDK4_m is the number of samples with a CDK4 mutation. The cells_c/m columns are the number of samples for which CNV or mutation data are available, respectively. Percentages are calculated as altered / total for each individual alteration type.

sourcename	CDK4_c	cells_c	CDK4_m	cells_m	cell_type	tumor_type	CNV%	MUT%
TCGA-SARC	35	171	0	0	soft_tissue	Sarcoma	20.47	NA
TCGA-GBM	73	607	0	150	neuronal	Glioblastoma multiforme	12.03	0
TCGA-LGG	14	471	1	612	neuronal	Brain Lower Grade Glioma	2.97	0.16
TCGA-ACC	2	90	0	91	adrenal_gland	Adrenocortical carcinoma	2.22	0
TCGA-SKCM	7	387	8	372	skin	Skin Cutaneous Melanoma	1.81	2.15
TCGA-LUAD	5	510	3	491	lung	Lung adenocarcinoma	0.98	0.61
TCGA-STAD	2	403	1	373	stomach	Stomach adenocarcinoma	0.5	0.27
TCGA-BRCA	5	1074	1	777	breast	Breast invasive carcinoma	0.47	0.13
TCGA-BLCA	1	255	2	242	urinary_tract	Bladder Urothelial Carcinoma	0.39	0.83
TCGA-OV	2	569	0	476	ovary	Ovarian serous cystadenocarcinoma	0.35	0
TCGA-LUSC	1	487	0	233	lung	Lung squamous cell carcinoma	0.21	0
TCGA-COAD	0	446	2	219	large_intestine	Colon adenocarcinoma	0	0.91
TCGA-PRAD	0	381	0	300	prostate	Prostate adenocarcinoma	0	0
TCGA-THCA	0	508	0	428	thyroid	Thyroid carcinoma	0	0
TCGA-PAAD	0	92	1	91	pancreas	Pancreatic adenocarcinoma	0	1.1
TCGA-PCPG	0	175	0	0	adrenal_gland	Pheochromocytoma and Paraganglioma	0	NA
TCGA-MESO	0	37	0	0	pleura	Mesothelioma	0	NA
TCGA-READ	0	164	0	1	rectum	Rectum adenocarcinoma	0	0
TCGA-UCEC	0	533	5	248	endometrium	Uterine Corpus Endometrial Carcinoma	0	2.02
TCGA-KIRC	0	521	6	328	kidney	Kidney renal clear cell carcinoma	0	1.83
TCGA-ESCA	0	126	0	0	oesophagus	Esophageal carcinoma	0	NA
TCGA-DLBC	0	28	0	79	haematopoietic	Lymphoid Neoplasm Diffuse Large B-cell	0	0
TCGA-KICH	0	66	0	66	kidney	Kidney Chromophobe	0	0
TCGA-UCS	0	57	0	57	uterus	Uterine Carcinosarcoma	0	0
TCGA-KIRP	0	212	0	169	kidney	Kidney renal papillary cell carcinoma	0	0
TCGA-LAML	0	194	0	118	haematopoietic	Acute Myeloid Leukemia	0	0
TCGA-LIHC	0	213	5	202	liver	Liver hepatocellular carcinoma	0	2.48
TCGA-HNSC	0	516	5	513	upper_aerodiges	Head and Neck squamous cell carcinoma	0	0.97
TCGA-CESC	0	206	0	41	cervix	Cervical squamous cell carcinoma and	0	0

6. Analysis

CELLX can identify genes whose expression correlates with a gene of interest and return a table of significant genes that can be visualized via a heat map with labelled metadata. For example, a search for genes correlated with CDK4 expression in the TCGA sarcoma dataset yields ACVR1L1 which is expressed by vascular endothelium and a potential anti-angiogenesis target. (Figure 4a)

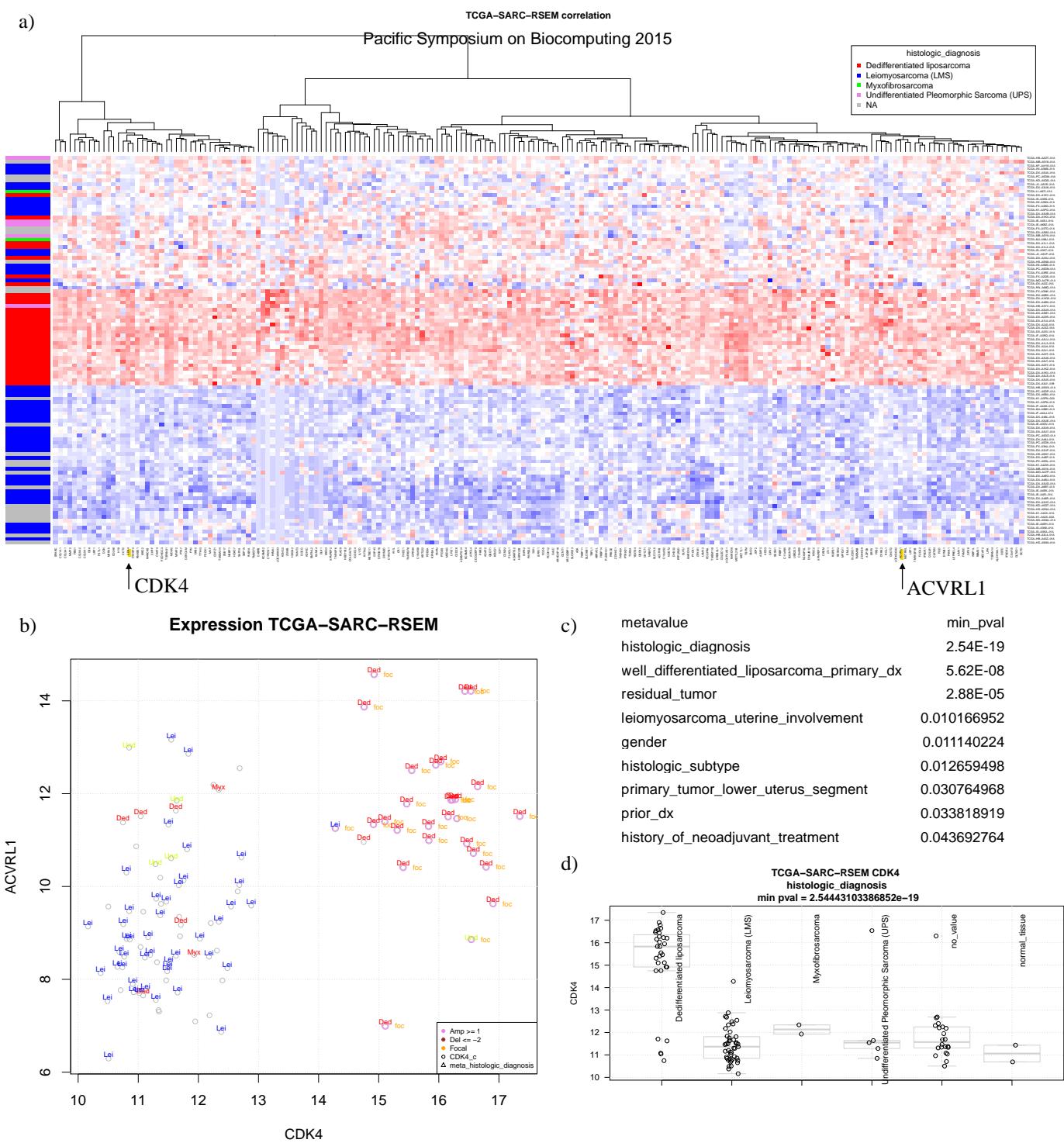


Figure 4. Analysis of features associated with CDK4 expression. a) Heatmap of top 200 genes (columns) correlated with CDK4 expression levels in samples (rows) from the TCGA sarcoma dataset showing ACVRL1 expression correlates with CDK4 (arrows). Meta data labels for histologic diagnosis are colored in a column on the left side of the plot. b) Scatter plot of CDK4 expression versus ACVRL1 expression showing high ACVRL1 expression in dedifferentiated liposarcomas. Metavalues from a) are colored and abbreviated by the first 3 letters. Amplification of CDK4 is denoted by a violet circle. foc=focal. c) Meta data with significantly different CDK4 expression levels. Min p-value is the lowest pairwise t-test score. d) Boxplot of histologic diagnosis by CDK4 expression data used in c).

A scatter plot of CDK4 vs. ACVRL1 shows higher ACVRL1 in Dedifferentiated Liposarcomas (DDPLS) vs. Leiomyosarcomas (Figure 4b). This is consistent with a study reporting immature and intermediate blood vessels in sarcomas and quantifying tumor microvessel density that is ~3X higher in DDLPS vs. Leiomyosarcomas. [9] The plot also shows that CDK4 expression is high in DDLPS and often focally amplified which is consistent with the literature.[10] CELLX can also

test for significant gene expression associated with meta data features by performing a t-test of a gene's expression grouped by a sample's meta data. As an example, a search for meta data with significantly different CDK4 expression in the TCGA sarcoma dataset reveals that the histologic diagnosis type has large differences in CDK4 expression levels (lowest p-val = 2.54e⁻¹⁹) as calculated by a pairwise t-test between all groups (Figure 4c). A box plot of the groups from histologic diagnosis shows that the CDK4 values from DDPLS are higher than other sarcomas (Figure 4d). Additional types of analyses include the identification of differentially expressed genes using t-tests of gene expression between groups defined by a gene's expression, a gene's mutation status, or a meta value label. For example, one could ask what genes are differentially expressed between samples with high CDK4 vs. low CDK4, samples with mutated EGFR vs. wild type EGFR, or samples annotated as male vs. female. Conversely, one can search for mutated genes which differentially express the query gene. e.g. which gene(s) mutations have higher or lower expression of EGFR than wild-type.

7. Precision Medicine

To support precision medicine, CELLX can be used to generate responder / non-responder hypotheses from cell line screening data. As a retrospective example, one can analyze the cell line sensitivity profile of Palbociclib, a CDK4/6 inhibitor under development for ER+ breast cancer. Published breast cell line IC50 values for Palbociclib[11] show a range of responses. (Figure 5a) CELLX can associate IC50 values with cell line expression, CNV, and mutation data from data sources such as CCLE. Samples divided into two groups by user defined cutoffs, in this case <1uM for responder cell lines (LOW IC50) and $\geq 1\mu\text{M}$ for non-responder cell lines (HIGH IC50) can be used to identify genes whose expression is significantly different between responder and non-responder cells by calculating t-tests on the expression of ~20,000 genes and displaying a p-value ranked table (Figure 5b). Hierarchical clustering on the top 100 most significant genes, ordering the samples from low to high IC50, and coloring the samples by intrinsic breast subtype as defined by PAM50[12] shows that luminal B and Her2 subtypes tend to be sensitive to Palbociclib whereas cells of the basal subtype tend to be resistant (Figure 5c). Luminal A cell line subtypes were not represented in the screening set. Additionally, CELLX can dynamically generate a combination CNV / mutation table for genes which meet user defined amplification / deletion thresholds or have annotated mutations. A ranked table of p-values from Fisher's exact test for all genes with either a CNV or mutation alteration (Table 2) highlights genes potentially associated with compound activity. While individually, the appearance of any one gene is not necessarily significant, together the combined results from the expression, CNV, and mutation associations highlight RB1, CCNE1, and to a lesser extent CDKN2A. Specifically, the expression of RB1 was low in resistant cells whereas CDKN2A and CCNE1 were high in resistant cells. Interestingly, unlike other targeted therapies where the small molecule target is often the biomarker of sensitivity (e.g. EGFR, MET, BRAF) the significant Palbociclib biomarkers represent markers of resistance. RB1 deficiency (CNV deletion, STOP mutations, and low expression) and concomitant high CDKN2A expression[13] are characteristics of the basal or triple negative breast subtype status (Figure 5c). Thus, if most of the RB1 deficient samples

belong to the triple negative subtype, the remaining luminal A/B (ER+/ERBB2+/-) and ERBB2+ segments would be enriched for possible CDK4i responders. In support of this notion, luminal B and Her2 breast subtype cell lines are mostly sensitive to CDK4i (Figure 5c).

CELLX can also confirm if the low RB1 expression found in triple negative breast cell lines also occurs in primary tissues by using the TCGA-BRCA breast invasive carcinoma dataset. CELLX can identify the genes that are most differentially expressed between RB1 high (≥ 9.5) vs. RB1 low (< 9.5) expressing cells using t-tests. Several of the top 100 ranking genes by p-value are related to cell cycle (RB1, CDKN2A, CCNE1) or DNA replication/repair (RFC2, RFC4, MCM5, MCM7, CDT1, NASP, POLK, POLD1, MUTYH, FANCE). Hierarchical clustering and labeling with the intrinsic subtype via PAM50[12] shows that similar to cell lines, we find that tumors with low RB1 and high CCNE1/CDKN2A expression are often of the basal subtype (Figure 6).

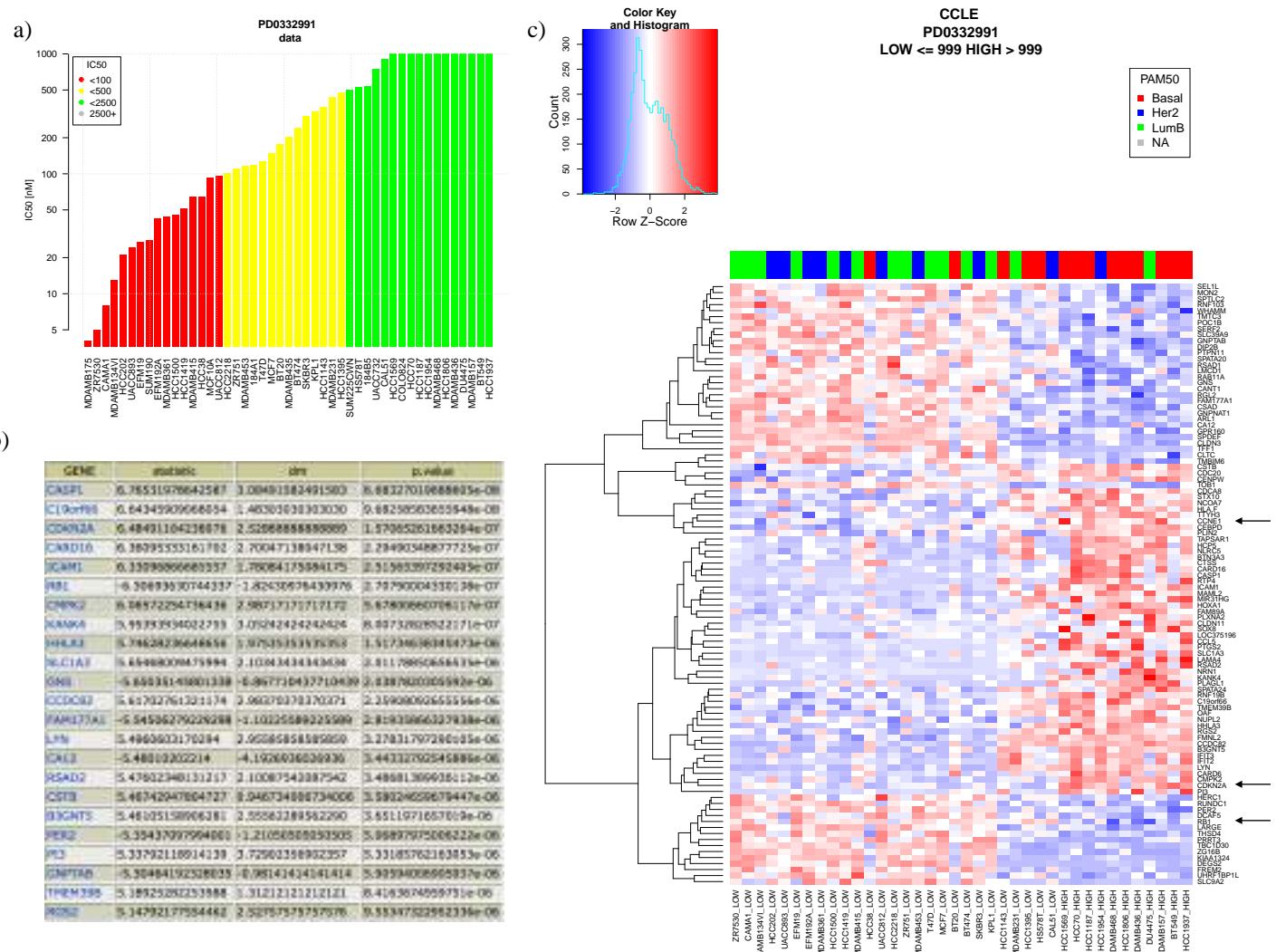


Figure 5. a) Waterfall plot of breast cell line responses to Palbociclib (PD0332991) colored by IC50 range. b) Example output listing the p-value of genes. dm = difference in group means, statistic = t-statistic (LOW-HIGH), p.value = uncorrected p-value of two-sided, two-class t-test with equal variances. Not shown: FDR and Hochberg adjusted p-values. c) Heatmap of gene expression of top 100 genes by t-test between sensitive (IC50 ≤ 999 nM, LOW) and resistant cell lines (IC50 > 999 nM, HIGH). The positions of RB1, CDKN2A, and CCNE1 are denoted with arrows. Cell lines are ordered by IC50 and colored by intrinsic breast subtype via PAM50.

Table 2. Association of mutations / CNV with response to Palbociclib (PD0332991). a) Ranking of genes by p-value for Fisher's Exact test. b) Breast cell line table of selected alterations. Breast cell lines are labeled LOW (sensitive) or HIGH (resistant) and marked altered or non-altered for mutation or CNV change in each gene. Cell lines are ordered by Palbociclib IC50 value. Genes with CNV values $\geq \text{abs}(1)$ or mutations from CCLE are marked as altered. CNV units are in log2 diploid genomes. (i.e. 1= \sim 4 copies) CCLE mutation nomenclature: del = deletion, p.0 = whole gene deletion, ? = unknown change, fs = frameshift, * = STOP codon

a)	GENE	pval	GENE	pval	b)	cell_name	PD0332991	RESPONSE	RB1	PIK3C2G	CCNE1	CDKN2A
	RB1	0.0004	ATP9B	0.0611		pvalue			0.0004	0.0048	0.0136	0.1362
	PIK3C2G	0.0048	CAPRIN1	0.0611		MDAMB175	4	LOW				
	C19orf12	0.0136	CTIF	0.0611		ZR7530	5	LOW				
	CCNE1	0.0136	DNM2	0.0611		CAMA1	8	LOW				
	LOC284395	0.0136	EHF	0.0611		MDAMB134VI	13	LOW				
	PLEKH1	0.0136	ELP2	0.0611		HCC202	21	LOW				
	POP4	0.0136	EPG5	0.0611		UACC893	24	LOW				
	URI1	0.0136	FANCI	0.0611		EFM19	27	LOW				
	VSTM2B	0.0136	HDLBP	0.0611		SUM190	28	LOW				
	DOCK3	0.0136	LRP6	0.0611		EFM192A	42	LOW				
	NCOA4	0.0136	MAPK4	0.0611		MDAMB361	44	LOW				
	ADRA1A	0.0136	MCPH1	0.0611		HCC1500	45	LOW				
	CTNNA1	0.0136	NKX6.3	0.0611		HCC1419	51	LOW				
	TCF12	0.0136	PDCD6	0.0611		HCC38	64	LOW				
	CDH1	0.0459	PEPB4	0.0611		MDAMB415	64	LOW				
	ANKS1B	0.0459	PTK2B	0.0611		MCF10A	92	LOW				
	DIP2C	0.0459	RP1L1	0.0611		UACC812	96	LOW	1.26	p.P129del	p.M52I	-2.24
	GSTT1	0.0595	SGK223	0.0611		HCC2218	100	LOW				
	GSTTP2	0.0595	SMAD4	0.0611		ZR751	110	LOW				
	LOC391322	0.0595	ZFYVE26	0.0611		MDAMB453	115	LOW				
	D2HGDH	0.0611	MTAP	0.0932		184A1	118	LOW				
	DHRS4L1	0.0611	USP32	0.0932		T47D	127	LOW				
	DHRS4L2	0.0611	BCAS1	0.0932		MCF7	148	LOW				
	ELAC1	0.0611	TRIM37	0.0932		BT20	177	LOW				
	GAL3ST2	0.0611	PIK3CA	0.0952		MDAMB435	201	LOW				
	LINC00906	0.0611	TP53	0.0952		BT474	240	LOW				
	LINC01029	0.0611	AUTS2	0.0971		SKBR3	300	LOW				
	LOC100420587	0.0611	LOC649352	0.0971		KPL1	327	LOW				
	LOC100505835	0.0611	MIR4650.1	0.0971		HCC1143	359	LOW				
	LOC102724958	0.0611	MIR4650.2	0.0971		MDAMB231	432	LOW				
	LOC439994	0.0611	SIGLEC14	0.0971		HCC1395	472	LOW				
	MIR6511B1	0.0611	FHIT	0.0971		SUM225CWN	503	LOW				
	NAALADL2	0.0611	PIK3C2B	0.0971		HS578T	524	LOW				
	NUTM2A.AS1	0.0611	PTEN	0.1176		184B5	538	LOW				
	RBFOX1	0.0611	CDKN2A	0.1362		UACC732	744	LOW				
	SALL3	0.0611	LOC284344	0.1560		CAL51	905	LOW				
	UGT2B28	0.0611	LPAR6	0.1560		MDAMB468	1000	HIGH				
	UQCRRFS1	0.0611	NRG1	0.1560		MDAMB436	1000	HIGH				
	APC	0.0611	PDE4D	0.1560		HCC1954	1000	HIGH				
	BTK	0.0611	EEF2K	0.1560		HCC1937	1000	HIGH				
	ELN	0.0611	EPHB3	0.1560		DU4475	1000	HIGH				
	EPHB6	0.0611	ITPR1	0.1560		HCC1569	1000	HIGH				
	GCNT2	0.0611	KIAA1549	0.1560		HCC1187	1000	HIGH				
	HIPK2	0.0611	MAP3K19	0.1560		BT549	1000	HIGH				
	KLK15	0.0611	MELK	0.1560		MDAMB157	1000	HIGH				
	NOS2	0.0611	MLKL	0.1560		COLO824	1000	HIGH				
	OMG	0.0611	MMP8	0.1560		HCC70	1000	HIGH				
	TBX22	0.0611	MYLK	0.1560		HCC1806	1000	HIGH				
	ZNF142	0.0611	PLCB2	0.1560								
	AGPAT5	0.0611	SPTA1	0.1560								

8. Summary

CELLX is an informatics infrastructure to manage multi-dimensional genomics datasets containing expression, copy number variation, mutation, and compound sensitivity information. A browser based web page enables an accessible way to visualize, analyze, and download the database data in a pre-formatted table suitable for offline computation. CELLX is presently

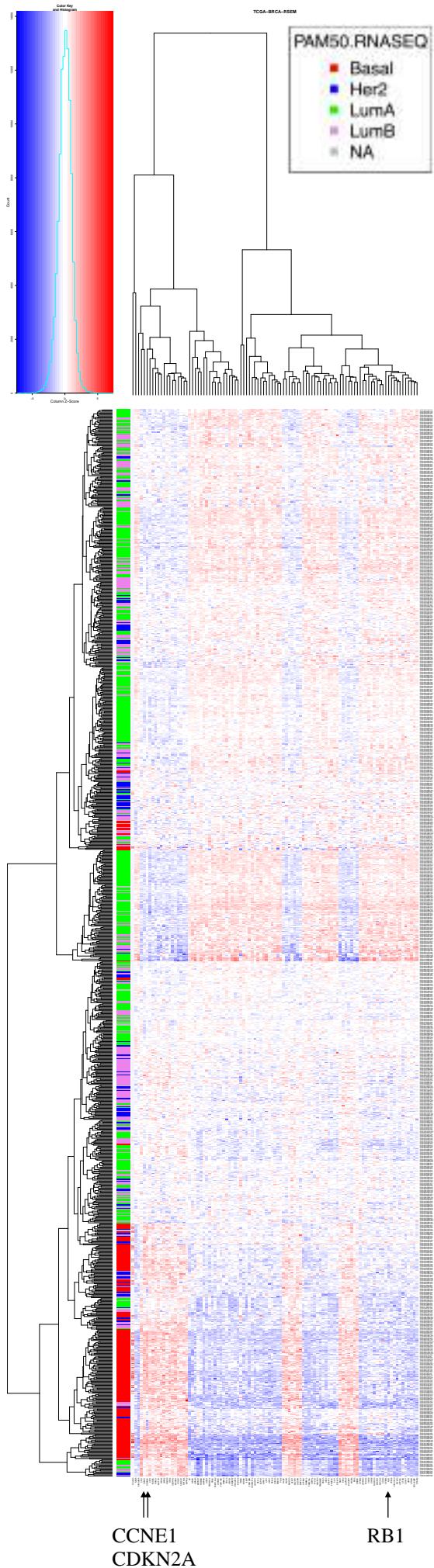


Figure 6. TCGA Breast differential gene expression between RB1 high and RB1 low expressing tumors. Hierarchical clustering of the top 100 genes in a heat map colored by breast subtype as determined by PAM50. Positions of CDKN2A, CCNE1, and RB1 are denoted by arrows.

focused on supporting oncology precision medicine through the evaluation of preconceived hypotheses as well as unbiased, data driven hypothesis generation. Though usable by the general user, CELLX is aimed at the computational biologist who desires more control over the data or wants to integrate custom data not available in public databases.

9. Data Processing

When available, summarized data from the source was used for TCGA, CCLE, and Tumorscape except for CNV calls. If Affymetrix SNP files were available, they were processed relative to the hg18 assembly using the *aroma.affymetrix* R package according to the methods of H. Bengtsson et al.[14] using the average baseline of 128 female HapMap samples[15] as the reference to maintain consistency and comparability across datasets. Microarray expression data from GEO, Sanger, and CCLE were GC Robust Multiarray Average normalized using R and the *gcrma*[16] library. Comparable to the TCGA RNA-Seq RSEM pipeline, CCLE RNA-Seq[17] data was processed using RSEM[8] on RefSeq sequences, quartile normalized to 1000, and log₂ transformed. The R library *genefu*[18] predicted PAM50 subtypes and *genefilter*[19] enabled fast t-tests, F-tests, and correlations. Plots were made using CELLX and edited using Preview and Pages.

Acknowledgements

The results published here are in whole or part based upon data generated by The Cancer Genome Atlas pilot project established by the NCI and NHGRI. Information about TCGA and the investigators and institutions who constitute the TCGA research network can be found at <http://cancergenome.nih.gov/> dbGaP Study Accession: phs000178.v8.p7 We thank Andy Futreal and the Wellcome Trust Sanger Institute for generously providing access to cell molecular profiling data. We also thank Adam Pavlicek and Shbing Deng for help with R and Heather Estrella for discussion and feedback.

References

1. Barretina J, et.al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*. 2012 Mar 28;483(7391):603-7. PMID: 22460905
2. Yang W, et.al. Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic Acids Res*. 2013 Jan;41:D955-61 PMID: 23180760
3. Chin L, Hahn WC, Getz G, Meyerson M. Making sense of cancer genomic data. *Genes Dev*. 2011 Mar 15;25(6):534-55. doi: 10.1101/gad.2017311. PMID:21406553
4. Wu C, Macleod I, Su AI. BioGPS and MyGene.info: organizing online, gene-centric information. *Nucleic Acids Res*. 2013 Jan;41:D561-5 PMID: 23175613
5. Beroukhim R, et.al. The landscape of somatic copy-number alteration across human cancers. *Nature*. 2010 Feb 18;463(7283):899-905. PMID: 20164920
6. Forbes SA, et.al. COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. *Nucleic Acids Res*. 2011 Jan;39:D945-50. PMID: 20952405
7. Cerami E, et.al. The cBio Cancer Genomics Portal: An Open Platform for Exploring Multidimensional Cancer Genomics Data. *Cancer Discov*. 2012 May;2(5):401-4. PMID: 22588877
8. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*. 2011 Aug 4;12:323. doi: 10.1186/1471-2105-12-323. PMID:21816040
9. Baneth V, Raica M, Cîmpean AM, Rom J. Assessment of angiogenesis in soft-tissue tumors. *Morphol Embryol*. 2005;46(4):323-7. PMID:16688371
10. Binh MB, Sastre-Garau X, Guillou L, de Pinieux G, Terrier P, Lagacé R, Aurias A, Hostein I, Coindre JM. MDM2 and CDK4 immunostainings are useful adjuncts in diagnosing well-differentiated and dedifferentiated liposarcoma subtypes: a comparative analysis of 559 soft tissue neoplasms with genetic data. *Am J Surg Pathol*. 2005 Oct;29(10):1340-7. PMID:16160477
11. Finn RS, Dering J, Conklin D, Kalous O, Cohen DJ, Desai AJ, Ginther C, Atefi M, Chen I, Fowst C, Los G, Slamon DJ. PD 0332991, a selective cyclin D kinase 4/6 inhibitor, preferentially inhibits proliferation of luminal estrogen receptor-positive human breast cancer cell lines in vitro. *Breast Cancer Res*. 2009;11(5):R77. PMID:19874578
12. Parker JS, Mullins M, Cheang MC, Leung S, Voduc D, Vickery T, Davies S, Fauron C, He X, Hu Z, Quackenbush JF, Stijleman IJ, Palazzo J, Marron JS, Nobel AB, Mardis E, Nielsen TO, Ellis MJ, Perou CM, Bernard PS. Supervised risk predictor of breast cancer based on intrinsic subtypes. *J Clin Oncol*. 2009 Mar 10;27(8):1160-7. doi: 10.1200/JCO.2008.18.1370. Epub 2009 Feb 9. PMID:19204204
13. Knudsen ES, Knudsen KE. Tailoring to RB: tumour suppressor status and therapeutic response. *Nat Rev Cancer*. 2008 Sep;8(9):714-24. doi: 10.1038/nrc2401 PMID:19143056
14. Bengtsson H, Irizarry R, Carvalho B, Speed TP (2008) Estimation and assessment of raw copy numbers at the single locus level. *Bioinformatics* 24: 759– 767. PMID:18204055
15. HapMap data available from http://hapmap.ncbi.nlm.nih.gov/downloads/raw_data/hapmap3_4_0/ Originally obtained from Affymetrix, but no longer available from that source.
16. Wu J and Gentry RIwcfJM. gcrma: Background Adjustment Using Sequence Information. R package version 2.36.0. <http://www.bioconductor.org/packages/release/bioc/html/gcrma.html>
17. CCLE RNA-Seq data obtained from The Cancer Genomics Hub (CGHub) <https://cgphub.ucsc.edu/>
18. Haibe-Kains B, Schroeder M, Bontempi G, Sotiriou C and Quackenbush J (2014). genefu: Relevant Functions for Gene Expression Analysis, Especially in Breast Cancer.. R package version 1.14.0, <http://compbio.dfci.harvard.edu>.
19. Gentleman R, Carey V, Huber W and Hahne F. genefilter: genefilter: methods for filtering genes from microarray experiments. R package version 1.46.1. <http://www.bioconductor.org/packages/release/bioc/html/genefilter.html>

COMPARING NONPARAMETRIC BAYESIAN TREE PRIORS FOR CLONAL RECONSTRUCTION OF TUMORS

AMIT G. DESHVAR

*Edward S. Rogers Sr. Department of Electrical and Computer Engineering,
University of Toronto, Toronto, ON, Canada
E-mail: amit.deshwar@utoronto.ca*

SHANKAR VEMBU

*Donnelly Center for Cellular and Biomolecular Research,
University of Toronto, Toronto, ON, Canada
E-mail: shankar.vembu@utoronto.ca*

QUAID MORRIS

*Donnelly Center for Cellular and Biomolecular Research,
Department of Molecular Genetics,
Edward S. Rogers Sr. Department of Electrical and Computer Engineering,
Department of Computer Science,
University of Toronto, Toronto, ON, Canada
E-mail: quaid.morris@utoronto.ca*

Statistical machine learning methods, especially nonparametric Bayesian methods, have become increasingly popular to infer clonal population structure of tumors. Here we describe the treeCRP, an extension of the Chinese restaurant process (CRP), a popular construction used in nonparametric mixture models, to infer the phylogeny and genotype of major subclonal lineages represented in the population of cancer cells. We also propose new split-merge updates tailored to the subclonal reconstruction problem that improve the mixing time of Markov chains. In comparisons with the tree-structured stick breaking prior used in PhyloSub, we demonstrate superior mixing and running time using the treeCRP with our new split-merge procedures. We also show that given the same number of samples, TSSB and treeCRP have similar ability to recover the subclonal structure of a tumor.

Keywords: Nonparametric Bayesian methods; Bayesian tree priors; Tumor phylogeny.

1. Background

The clonal theory of cancer posits that tumors contain multiple, genetically diverse subclonal populations of cells that evolved from a single progenitor population through successive waves of expansion and selection.¹ Recent genetic analyses of tumor subpopulations support this theory.^{2,3} These analyses also identify characteristic driver mutations involved in cancer development and progression⁴ and provide insight into understanding and predicting treatment response.⁵ Understanding this intratumor genotype heterogeneity is especially important because different subclonal populations have different abilities to metastasize and resist treatment.^{2,6} These somatic mutations are detected through high-throughput sequencing of tumor and normal tissue; and can be broadly divided into two types: Simple Somatic Mutations (SSMs) consisting of substitutions and small insertions / deletions, and Copy Number Variations (CNVs) resulting from larger structural changes.

Current, widely-used high-throughput sequencing (HTS) technology generates short reads that rarely span multiple SSM loci, so in almost all cases only the *variant allele frequency* (VAF), *i.e.*, the proportion of reads containing the variant, are available for individual SSMs. These VAFs have been used to partially reconstruct the tumor subpopulations.^{2,7–19} However, surprisingly, these VAFs can be used to completely reconstruct subpopulation genotypes in some cases, by reconstructing the evolutionary history of the subpopulations.^{14,15,19} SSM VAFs from multiple tumor samples improve this reconstruction.^{15,17,18}

These evolution-based subclonal reconstruction methods use a specific tree representation in which mutations are assigned to both internal and leaf nodes. This representation excludes tree inference methods, like hierarchical clustering or the nested Chinese restaurant process²⁰ that assign observations (mutations) only to leaf nodes. To our knowledge only two tree-based statistical models have been described that i) allow mutations to be assigned to internal nodes and ii) are *non-parametric*, *i.e.*, that do not require pre-specification of the number of nodes. PhyloSub¹⁵ has previously applied the tree-structured stick-breaking (TSSB) prior²¹ to this problem. Here, we derive a version of the tree-Chinese Restaurant Process (treeCRP)²² for subclonal reconstruction and new associated split-merge MCMC updates. We compare the two models in terms of their sampling efficiency and accuracy in subclonal reconstruction.

In the next section we provide an overview of the subclonal reconstruction problem. The remainder of this paper consists of a formal description of the treeCRP model and the results from a series of empirical comparisons of the TSSB model against several treeCRP variants.

2. Methods

2.1. Subclonal Reconstruction

Figure 1 provides an illustrative overview of the assumed process of tumor evolution and the task of subclonal reconstruction. Panel (i) of this figure shows a visualization of the evolution of a tumor over time as noncancerous cells (grey) are replaced by, at first, one clonal cancerous population (green) which then further develops into multiple cancerous subpopulations. Tumor cells define new subpopulations by acquiring new oncogenic mutations that allow their descendants to expand relative to the other tumor subpopulations. Each circle in Panel (i) refers to a subpopulation. We associate each subpopulation with the set of shared somatic mutations (shown as a diamond) that distinguish it from its parent subpopulation. However, each subpopulation also inherits all of its parent’s mutations; as such, mutations may be present in multiple subpopulations. We define the *subclonal lineage* of a mutation as the set of all subpopulations that contain it. For example, the subclonal lineage corresponding to the blue diamond includes the subpopulation (D) associated with that set of mutations and all decedent subpopulations (E,F,G). For clarity, and to highlight the link between subpopulations and their set of subpopulation-defining mutations, we will use the corresponding lower-case letter to refer to these mutations. For example, we will use *d* to refer to the set of mutations represented by the blue diamond.

Mutation sets, and their associated subpopulations, are defined by analyzing the population frequencies of somatic mutations detected in a tumor sample. In the simple case that we consider here, SSMs occur in one copy of diploid regions of the genome; allowing one to

estimate the *clonal frequency* (i.e. the proportion of the sampled cells with the mutation) of a mutation by simply doubling its variant allele frequency, i.e., the proportion of reads mapping to the mutated locus that contain the SSM. See Deshwar *et al.*²³ for the case when SSMs occur in non-diploid sections of the genome. Panel (ii) shows an example histogram of the SSM VAFs found in a heterogeneous tumor sample. Each subpopulation is defined by both the small number of oncogenic ‘driver’ mutations that cause rapid expansion but also a larger number of ‘passenger’ mutations acquired before the driver mutation(s) through errors in DNA replication (even noncancerous cells accumulate somatic mutations at a rate of 1.1 per cell division²⁴). When a subpopulation expands, both the driver and the passenger SSMs increase in clonal frequency, and so have essentially identical frequencies. Due to sampling noise in the measurement of the VAFs, these mutation sets correspond to clusters (or modes) in the VAF distribution. The central VAF of a particular cluster is determined by the population frequency of its subclonal lineage. It is important to note that a given VAF cluster need not correspond to a subpopulation that is currently present in the tumor. For example, in Panel (ii), there is a VAF mode corresponding to mutation set d even though subpopulation D has a population frequency of 0% in the tumor sample. Only methods that attempt to reconstruct phylogenies (shown as panel (iii)), such as PhyloSub¹⁵ and rec-BTP,¹⁹ can detect when ‘vestigial’ VAF clusters correspond to historical subpopulations that are no longer present in the sample.

2.2. Our Approach

We use a directed tree to represent the evolutionary relationship among the tumor subpopulations. Each node in the tree represents a subpopulation (either currently in the sample or that existed at some point in the tumor development) and the links connect parental subpopulations to their direct descendants. The set of SSMs assigned to a node are the defining set for the node’s associated subpopulation. The subclonal lineage of an SSM consists of the subpopulation it is assigned to and that population’s descendants. Each node i is also assigned a frequency $\phi_i \in [0, 1]$ which is the inferred clonal frequency of the SSMs in the node. The population frequency of the node’s subpopulation, η_i , is the difference between the node’s clonal frequency and the sum of the clonal frequencies of the node’s children, *i.e.*, $\eta_i = \phi_i - \sum_{j \in \mathcal{C}(i)} \phi_j$ where $\mathcal{C}(i)$ is the set of the indices of the children of node i . The complete set of SSMs present in a subpopulation is the union of the SSMs assigned to it and those of all its ancestral nodes.

2.3. Dirichlet process mixture models

The treeCRP is derived from the Dirichlet process mixture model (DPMM) which we introduce here. Consider the problem of clustering N data objects $\{x_i\}_{i=1}^N$ using a Bayesian finite mixture model of K components (clusters) with the following generative process:²⁵

$$\boldsymbol{\omega} \sim \text{Dirichlet}(\alpha/K, \dots, \alpha/K); \quad z_i \sim \text{Multinomial}(\boldsymbol{\omega}); \quad \phi_k \sim H; \quad x_i \sim F(\phi_{z_i}), \quad (1)$$

where $\boldsymbol{\omega}$ are the non-negative mixing weights such that $\sum_{k=1}^K \omega_k = 1$, α is the concentration parameter of the symmetric Dirichlet prior placed on the mixing weights, $z_i \in \{1, \dots, K\}$ is the cluster assignment variable, H is the prior distribution from which the component

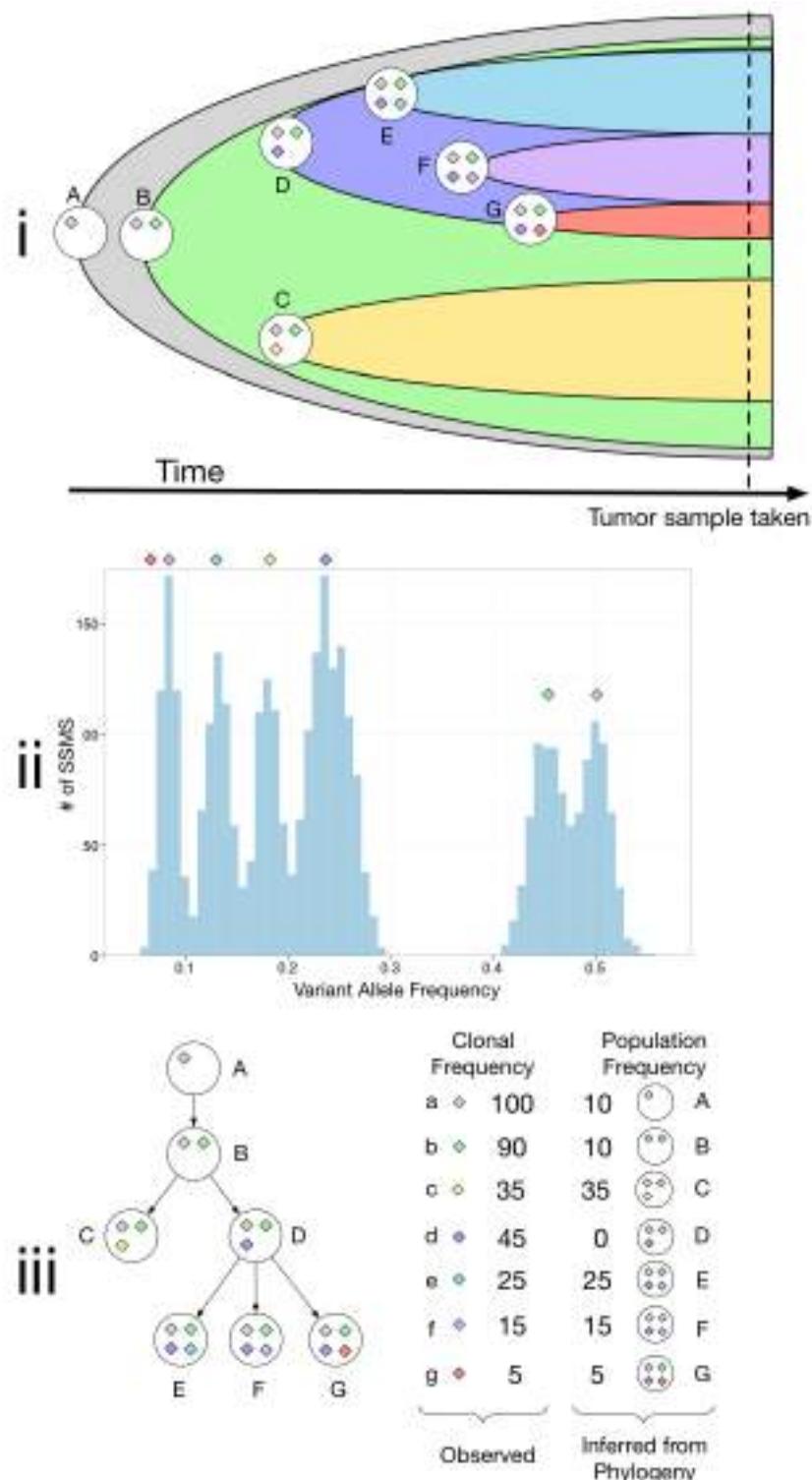


Fig. 1. The development of intratumor heterogeneity (i), the resulting distribution of variant allele frequencies (ii) and the desired output of subclonal inference (iii)

parameters $\{\phi_k\}$ are drawn, $F(\phi)$ is the component distribution parameterized by ϕ . The finite mixture model can be extended to a model with an infinite number of mixture components by replacing the Dirichlet prior with a Dirichlet process (DP) prior resulting in what is known as the DPMM.²⁶ Unlike finite mixture models, DPMMs automatically estimate the number of components from the data thereby circumventing the problem of fixing the number of components *a priori*. The Chinese Restaurant Process (CRP) provides a method to draw samples from a Dirichlet process. In this construction, an observation x_i is assigned to an existing cluster k with probability proportional to the number of objects N_k^{-i} in that cluster, excluding x_i . A new cluster $K+1$ is created with probability proportional to the concentration parameter. More formally,

$$\begin{aligned} p(z_i = k \mid \mathbf{z}_{\setminus i}, \alpha) &= \frac{N_k^{-i}}{N + \alpha - 1}, \forall k \in \{1, \dots, K\}; \\ p(z_i = K+1 \mid \mathbf{z}_{\setminus i}, \alpha) &= \frac{\alpha}{N + \alpha - 1}, \end{aligned} \quad (2)$$

where $\mathbf{z}_{\setminus i} = \{z_1, \dots, z_{i-1}, z_{i+1}, \dots, z_N\}$. The generative process for infinite mixture models using the Chinese restaurant process is:

$$z_i \sim \text{CRP}(\alpha; \mathbf{z}_{\setminus i}); \quad \phi_k \sim H; \quad x_i \sim F(\phi_{z_i}). \quad (3)$$

2.4. Tree-structured Chinese restaurant process

The Chinese restaurant process construction (2) described above can be used to produce a *flat* clustering of objects, where the clusters are independent of each other. Meeds *et al.*²² extended this construction for *relational* clustering that produces a clustering of objects where the clusters are connected to form a rooted tree.

In the tree-structured Chinese restaurant process (treeCRP), initially, a tree consists of a single root node (cluster) with all the data objects assigned to it. Subsequently, an object x_i is assigned to an existing node k with probability proportional to the number of objects N_k^{-i} in that node, excluding x_i . A new node $K+1$ is created as a child node to one of the K existing nodes in the tree with probability proportional to α/K . More formally,

$$\begin{aligned} p(z_i = k \mid \mathbf{z}_{\setminus i}, \alpha) &= \frac{N_k^{-i}}{N + \alpha - 1}, \forall k \in \{1, \dots, K\}; \\ p(z_i = K+1 \mid \mathbf{z}_{\setminus i}, \alpha) &= \frac{1}{K} \left(\frac{\alpha}{N + \alpha - 1} \right). \end{aligned} \quad (4)$$

2.5. Binomial observation model

Our probabilistic model for read count data is based on the one used by PhyloSub.¹⁵ Let a_i and b_i denote the number of reads matching the reference allele and the variant allele respectively at position i , and let $d_i = a_i + b_i$. This represents the total number of reads at locus i . Let η_k represent the population frequency of subpopulation k (node k in our tree). Let $\mu_i^r = 1 - \epsilon$ denote the probability of sampling a reference allele from the reference population where ϵ is the error rate of the sequencer. We set ϵ to 0.001 for all our experiments. Let μ_i^v denote the probability of sampling a reference allele from the variant population. For the purposes

of this paper we assume that all mutations are heterozygous and all loci have two copies, so we set μ_i^v to 0.5. Our model constrains the subpopulation frequencies η_i such that $\eta_i \geq 0 \forall i$ and that $\sum_i \eta_i = 1$. We recover the node clonal frequencies using the recursive equation: $\phi_i = \eta_i + \sum_{j \in \mathcal{C}(i)} \phi_j$. The observation model for read counts has the following likelihood:

$$a_i | z_i = k, d_i, \phi_k, \mu_i^r, \mu_i^v \sim \text{Binomial}(d_i, (1 - \phi_k)\mu_i^r + \phi_k\mu_i^v) . \quad (5)$$

2.6. Inference

Given this model and a set of N observations $\{(a_i, d_i, \mu_i^r, \mu_i^v)\}_{i=1}^N$, the tree structure and the subpopulation frequencies $\{\eta_k\}_{k=1}^K$ are inferred using Markov Chain Monte Carlo (MCMC) sampling. To sample the assignments of observations (SSMs) to nodes in the tree (z_i) we use Gibbs sampling where the probability of assigning a node is the prior probability (Equation (4)) multiplied by the likelihood of the data given that cluster identity (Equation (5)). After completing a single pass through the observations in random order we then use Metropolis-Hastings sampling for 100 iterations to sample the η_k variables. We use an asymmetric Dirichlet distribution as the proposal distribution where the concentration parameter is set during burn-in to achieve an acceptance ratio between 0.05 and 0.75.

For all experiments we sample for 2600 iterations and discard the first 100 samples as burn-in.

2.7. Split-merge updates

The Gibbs updates described previously only allow one cluster assignment to change at a time, which can result in slow mixing as described in the original treeCRP paper.²² Meeds *et al.*²² overcame this limitation by using split-merge updates in their implementation of the treeCRP. However, their updates relied on the partial conjugacy of the cluster parameters as described in Ref. 27, which is not the case in our observation model. In addition, the subclonal tree we are inferring has a natural ordering not present in the original treeCRP model. This natural ordering is that the clonal frequency of a node in our tree must always be greater than or equal to the sum of the clonal frequencies of its children. This natural ordering means that arbitrary split-merge moves are unlikely to be accepted. We therefore propose two “local” split-merge updates that are more likely to be accepted: the *parent-child split merge (PCSM)* and the *leaf-sibling-split-merge (LSSM)*.

The PCSM selects a node in the tree at random and then with equal probability either splits or merges that node. If merge is selected, then the node is merged with its parent (*i.e.*, its SSMs are assigned to its parent) and all its children become its parent’s children. If split is selected, a new node is added to the tree as the child of the selected node. The selected node’s children are split with the new node, assigning a given child to the new node with probability 0.5. The LSSM selects a leaf of the tree at random and, as the PCSM, selects with equal probability whether to add a new sibling node (*i.e.*, split) or merge the selected node with a randomly selected sibling node. Only leaf nodes are considered in this operation for implementation simplicity and because our subjective observation that leaf nodes exhibited slower mixing than the internal nodes. Whenever a new node is created through a split in either

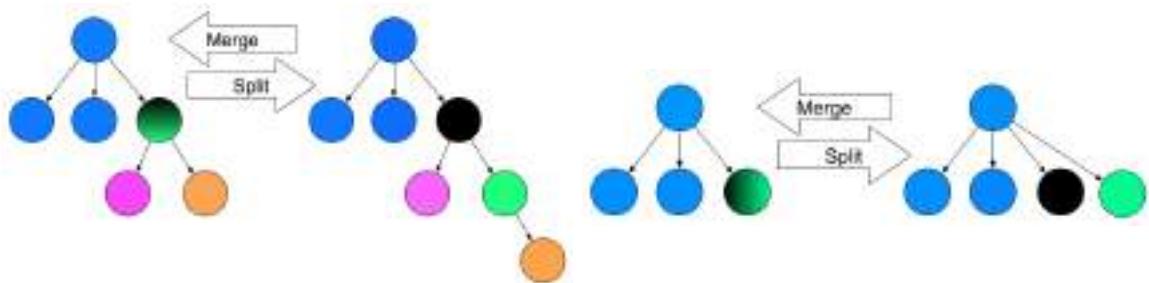


Fig. 2. Example of a Parent-Child Split-Merge move (*left*) and a Leaf-Sibling Split-Merge move (*right*)

Table 1. Table of subclonal lineage proportions used

Number of populations	ϕ values used
3	0.44, 0.11
4	0.56, 0.25, 0.06
5	0.64, 0.36, 0.16, 0.04
6	0.71, 0.44, 0.25, 0.11, 0.03

a PCSM or a LSSM update, the SSMs assigned to the split node are reassigned between the split and new node using restricted Gibbs updates. The population frequency of a new node (η_{new}) is selected uniformly at random between 0 and the population frequency of its parent, while the parents population frequency is decreased by η_{new} to maintain $\sum_i \eta_i = 1$. Figure 2 shows an example of PCMS and LSSM updates.

3. Results and Discussion

In order to compare our approaches we constructed a series of simulated datasets and applied PhyloSub¹⁵ (which uses the TSSB prior) and the treeCRP model with either Gibbs updates only (treeCRP-Gibbs), Gibbs updates and Parent Child Split-Merge moves (treeCRP-PCSM), Gibbs updates and Leaf-Sibling Split-Merge moves (treeCRP-LSSM) and all three types of updates (treeCRP-all). For treeCRP-all, we propose a PCSM update and then a LSSM update after each Gibbs update. Our simulations looked at a range of total population counts (3, 4, 5, 6), read depths (20, 30, 50, 70, 100, 200, 300) and number of SSMs per population (5, 10, 25, 50, 100, 200, 500). In each case, the first population is a normal population with no associated SSMs, while each subsequent population is a descendant of all previous populations (i.e. a chain phylogeny). For each simulated SSM k in population u , reference allele reads (a_k) were drawn as:

$$a_k \sim \text{Binomial}(d_k, 1 - \phi_u + 0.5\phi_u) ; \quad d_k \sim \text{Poisson}(r) ,$$

where ϕ_u is the clonal frequency of population u and r is the simulated read depth. A table of the ϕ values used can be found as Table 1.

First, we compared how quickly the Markov chain mixes for the different tree priors and MCMC sampling strategies. A chain that is fast-mixing requires fewer burn-in samples, is

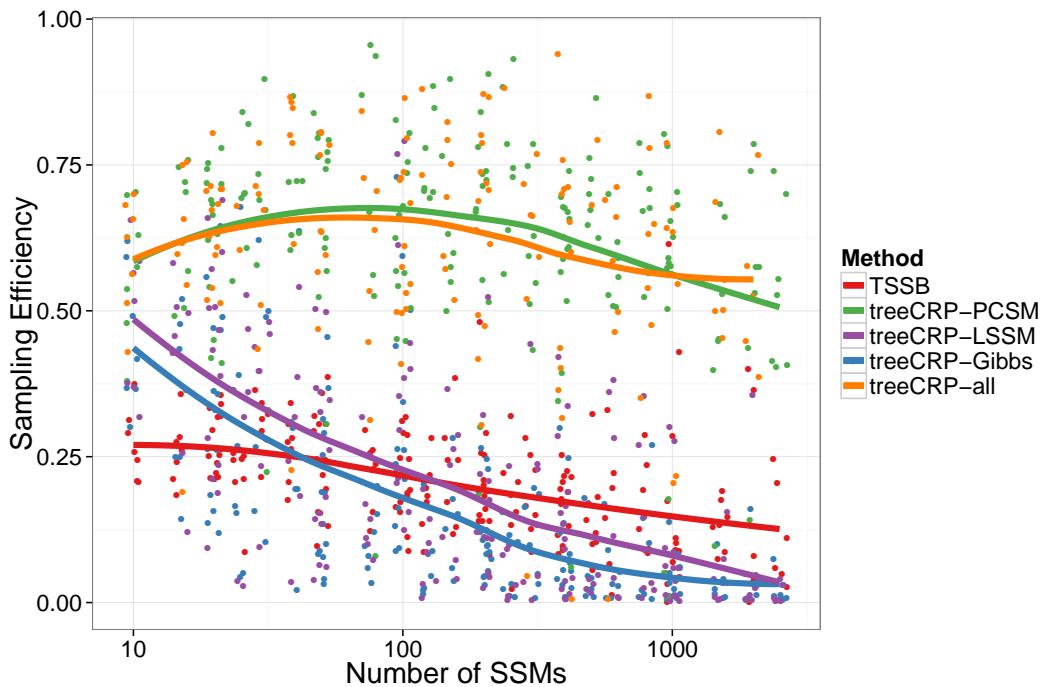


Fig. 3. The relationship between number of SSMs and sampling efficiency for the five algorithms. Lines are Loess curves. X-axis positions are jittered for clarity.

less likely to remain stuck in local modes and for a fixed desired accuracy requires fewer samples (or for the same number of samples delivers estimates with higher accuracy). We measure mixing by first computing the effective sample size of the chain and then dividing by the number of actual samples taken. By dividing the effective sample size by the number of actual samples we get a measure of sampling efficiency, where a higher value indicates better mixing. Effective sample size is calculated by the Coda package,²⁸ using the spectral density at frequency zero. Figure 3 shows the sampling efficiency for the five algorithms on our simulated datasets. The TSSB, the treeCRP-Gibbs and treeCRP-LSSM methods have consistently lower sampling efficiency than the treeCRP-PCSM and the treeCRP-all. Furthermore, for treeCRP-PCSM and treeCRP-all the sampling efficiency is not strongly affected by the number of SSMs, whereas the efficiency of the other three methods decreases with increasing numbers of SSMs.

Next, we assessed the accuracy of the mapping from population to SSM. To measure accuracy in a systematic way that accounts for class-imbalance, varying number of mutations and differing number of populations we used the Area Under the Precision-Recall Curve (AUPR) between the known true co-clustering matrix and the average co-clustering matrix over all samples. The co-clustering matrix M is a binary matrix where $M_{ij} = 1$ if SSM i and SSM j are in the same subclonal lineage. Figure 4 shows the distribution of AUPR values over all simulations. Although the absolute differences in AUPR are small, most of the pairwise differences are significant (7 out of 10, $P < 0.005$, Wilcoxon paired signed rank test, Bonferroni correction) and generally correspond to the differences in sampling efficiency in figure 3 except that there are no significant differences between the AUPRs for TSSB and those of treeCRP-all

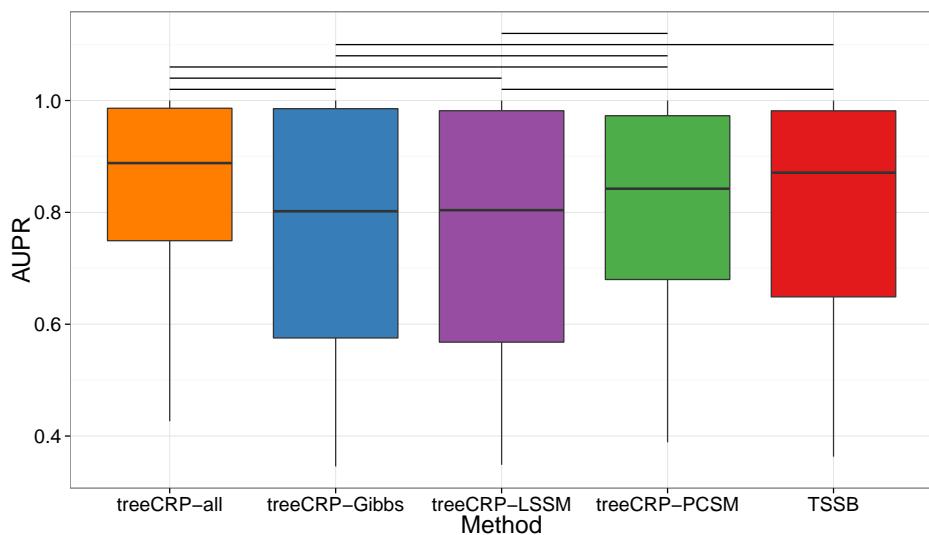


Fig. 4. Distribution of AUPR results for the five algorithms. Horizontal lines indicate statistically significant differences ($P < 0.005$, Wilcoxon paired signed rank test, Bonferroni correction)

and -PCSM. This suggests that the sampling efficiency differences have practical implications in reconstruction accuracy but neither prior is clearly superior.

Finally, we are interested in the relative time required to draw one sample. This is because in most situations a sampling budget is a given amount of CPU time, so greater efficiency per sample could be counteracted by greater computational effort to compute that sample. Because the cluster on which the experiments were run is composed of heterogeneous nodes it was meaningless to compare the runtimes of our experiments. Instead, we ran all five algorithms on the same computer using the simulated dataset with 5 subpopulations, read depth of 200 and 500 SSMs per subpopulation. Figure 5 (i) shows the average runtime per sampling iteration for the different algorithms, normalized to the runtime of the treeCRP-Gibbs algorithm while Figure 5 (ii) shows the runtime per effective sample. We observe that the TSSB algorithm is the slowest followed by the -all, -PCSM, -LSSM and finally the Gibbs only variant of the treeCRP. After adjusting for sampling efficiency, TSSB remains slower than the treeCRP methods but the treeCRP-PCSM and treeCRP-all are now about 5 times faster than the other treeCRP methods.

Chronic Lymphocytic Leukemia

To demonstrate the ability of the treeCRP family of algorithms to perform subclonal reconstruction on a real dataset we applied the four methods to a Chronic Lymphocytic Leukemia (CLL) dataset from Schuh *et al.*¹¹ The dataset consists of targeted sequencing data from three patients at five different time points; we reconstructed the tree using all five samples as input simultaneously. We examined the maximum likelihood tree found during sampling. All four treeCRP methods recovered the same tree structure and clustered the same SSMs together. Figure 6 shows the recovered tree structure along with the tree structure found in the original publication for the three patients CLL003, CLL006, and CLL077. The trees we recovered are

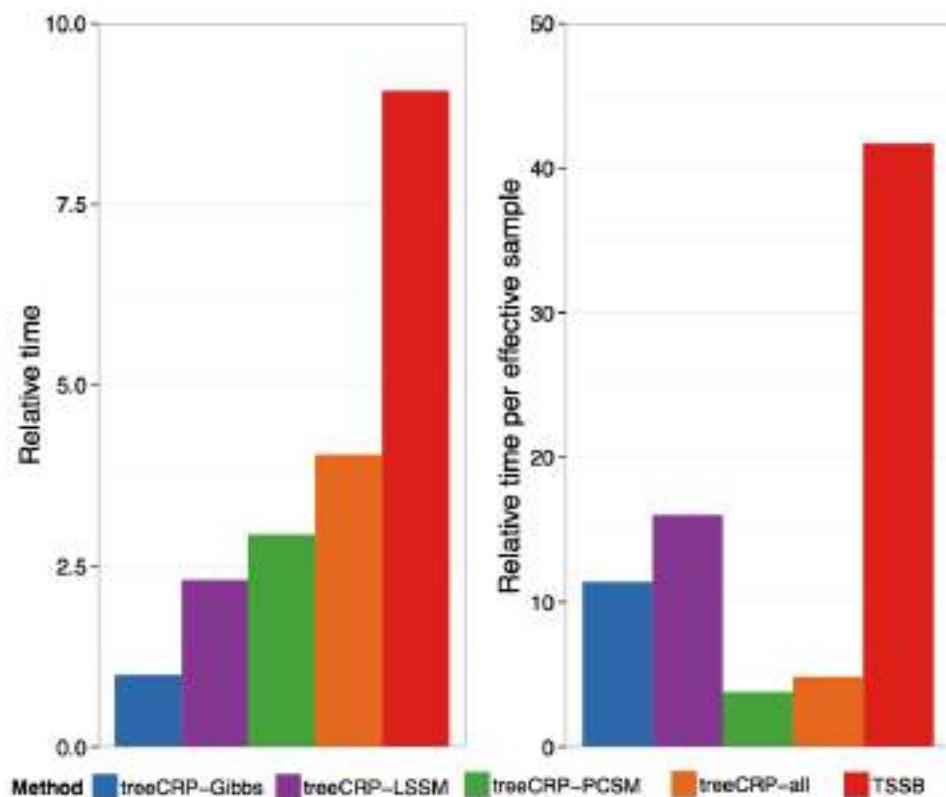


Fig. 5. Relative time per sample (i) and relative time per effective sample (ii) for the five methods

nearly identical to the expert derived trees, and those previously recovered by PhyloSub,¹⁵ with a small number of differences. For the top two rows in Figure 6, the differences are minor changes in clonal frequency estimates that result in reassignment of some SSMs to direct parents or children. The change in the bottom row (CLL077) is more substantial, as the treeCRP methods are predicting that the E2 population is a direct descendant of normal rather than of the B2 population, in other words, there are two independent cancerous lineages. This change occurred from our previous reconstruction¹⁵ because we no longer insist on a single cancer lineage in the new models. Although this reconstruction differs from the expert one, it is almost identical to one discovered by an independent, non-tree based method.¹⁷

We also compared the estimates of clonal frequencies from PhyloSub and treeCRP with the expert/baseline frequencies. The mean absolute difference between the baseline frequencies and frequencies estimated by Phylosub and treeCRP were (0.02, 0.018), (0.008, 0.028) and (0.012, 0.016) for CLL003, CLL006 and CLL077, respectively. The Pearson correlation between the baseline frequencies and frequencies estimated by Phylosub and treeCRP were (0.998, 0.999), (0.998, 0.984) and (0.999, 0.998) for CLL003, CLL006 and CLL077, respectively.

4. Conclusions

In our experiments with simulated data the treeCRP prior delivered similar subclonal reconstruction accuracy to the TSSB while having reduced runtime per sample and per effective sample. Among the treeCRP sampling strategies, treeCRP-all lead to the greatest subclonal

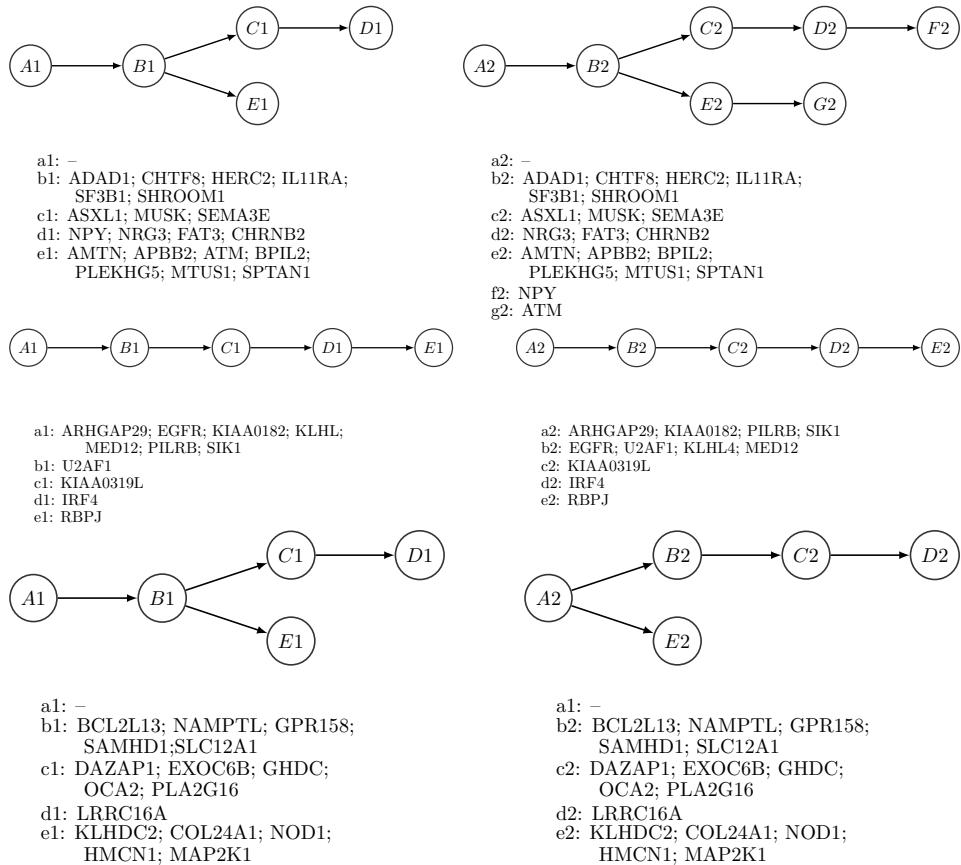


Fig. 6. Expert derived subclonal evolutionary trees (*left*) and trees found by the treeCRP methods (*right*) for patients CLL003 (top), CLL006 (middle) and CLL077 (bottom)

reconstruction accuracy and second highest sampling efficiency among all five tested methods. When compared to the TSSB method, for the same amount of CPU time, the treeCRP-all method could generate 10 times the number of effective samples thus permitting a 10-fold decrease in run time. Furthermore, treeCRP-all's sampling efficiency was independent of SSM number whereas TSSB's decreased with larger numbers of SSM. So, treeCRP-all appears much more suited to subclonal reconstruction using whole genome sequencing data with tens of thousands of SSMs. However, surprisingly, despite the increase in effective number of samples, there was not a significant difference in reconstruction accuracy between treeCRP-all and TSSB. Furthermore, the TSSB reconstruction was a better match to the expert reconstruction on the CLL dataset. As such, it remains an open question whether the decreased flexibility of the treeCRP prior (one hyperparameter versus two for TSSB) introduces a prior bias that interferes with subclonal reconstruction.

Acknowledgments

This work was funded by a National Science and Engineering Research Council (NSERC) operating grant and an Early Researcher Award from the Ontario Research Fund to QM. AGD is supported by an NSERC Vanier Canadian Graduate Scholarship.

References

1. P. C. Nowell, *Science* **194**, 23 (1976).
2. M. Gerlinger, A. J. Rowan, S. Horswell, J. Larkin, D. Endesfelder, E. Gronroos, P. Martinez, N. Matthews, A. Stewart, P. Tarpey *et al.*, *The New England Journal of Medicine* **366**, 883 (2012).
3. A. E. Hughes, V. Magrini, R. Demeter, C. A. Miller, R. Fulton, L. L. Fulton, W. C. Eades, K. Elliott, S. Heath, P. Westervelt *et al.*, *PLoS genetics* **10**, p. e1004462 (2014).
4. D. Hanahan and R. A. Weinberg, *Cell* **144**, 646 (2011).
5. S. Aparicio and C. Caldas, *New England Journal of Medicine* **368**, 842 (2013).
6. L. Ding, T. J. Ley, D. E. Larson, C. A. Miller, D. C. Koboldt, J. S. Welch, J. K. Ritchey, M. A. Young, T. Lamprecht, M. D. McLellan *et al.*, *Nature* **481**, 506 (2012).
7. C. G. Mullighan, L. A. Phillips, X. Su, J. Ma, C. B. Miller, S. A. Shurtleff and J. R. Downing, *Science* **322**, 1377 (2008).
8. N. E. Navin and J. Hicks, *Molecular Oncology* **4**, 267 (2010).
9. A. Marusyk and K. Polyak, *Biochimica et Biophysica Acta* **1805**, 105 (2010).
10. S. Nik-Zainal, P. V. Loo, D. C. Wedge, L. B. Alexandrov, C. D. Greenman, K. W. Lau, K. Raine, D. Jones, J. Marshall, M. Ramakrishna *et al.*, *Cell* **149**, 994 (2012).
11. A. Schuh, J. Becq, S. Humphray, A. Alexa, A. Burns, R. Clifford, S. M. Feller, R. Grocock, S. Henderson, I. Khrebtukova, Z. Kingsbury, S. Luo, D. McBride, L. Murray, T. Menju, A. Timbs, M. Ross, J. Taylor and D. Bentley, *Blood* **120**, 4191 (2012).
12. S. L. Carter, K. Cibulskis, E. Helman, A. McKenna, H. Shen, T. Zack, P. W. Laird, R. C. Onofrio, W. Winckler, B. A. Weir, R. Beroukhim, D. Pellman, D. A. Levine, E. S. Lander, M. Meyerson and G. Getz, *Nature Biotechnology* **30**, 413 (2012).
13. D. A. Landau, S. L. Carter, P. Stojanov, A. McKenna, K. Stevenson, M. S. Lawrence, C. Sougnez, C. Stewart, A. Sivachenko, L. Wang *et al.*, *Cell* **152**, 714 (2013).
14. F. Strino, F. Parisi, M. Micsinai and Y. Kluger, *Nucleic Acids Research* **41**, p. e165 (2013).
15. W. Jiao, S. Vembu, A. G. Deshwar, L. Stein and Q. Morris, *BMC Bioinformatics* **15**, p. 35 (2014).
16. A. Roth, J. Khattra, D. Yap, A. Wan, E. Laks, J. Biele, G. Ha, S. Aparicio, A. Bouchard-Côté and S. P. Shah, *Nature Methods* **11**, 396 (2014).
17. H. Zare, J. Wang, A. Hu, K. Weber, J. Smith, D. Nickerson, C. Song, D. Witten, C. A. Blau and W. S. Noble, *PLoS computational biology* **10**, p. e1003703 (2014).
18. A. Fischer, I. Vázquez-García, C. J. Illingworth and V. Mustonen, *Cell Reports* (2014).
19. I. Hajirasouliha, A. Mahmoody and B. J. Raphael, *Bioinformatics* **30**, i78 (2014).
20. D. M. Blei, T. L. Griffiths and M. I. Jordan, *Journal of the ACM (JACM)* **57**, p. 7 (2010).
21. R. P. Adams, Z. Ghahramani and M. I. Jordan, Tree-structured stick breaking for hierarchical data, in *Advances in Neural Information Processing Systems 23*, 2010.
22. E. Meeds, D. A. Ross, R. S. Zemel and S. T. Roweis, Learning stick-figure models using nonparametric bayesian priors over trees, in *In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2008.
23. A. G. Deshwar, S. Vembu, C. K. Yung, G. H. Jang, L. Stein and Q. Morris, *arXiv preprint arXiv:1406.7250* (2014).
24. S. Behjati, M. Huch, R. van Boxtel, W. Karthaus, D. C. Wedge, A. U. Tamuri, I. Martincorena, M. Petljak, L. B. Alexandrov, G. Gundem *et al.*, *Nature* (2014).
25. Y. W. Teh, Dirichlet processes, in *Encyclopedia of Machine Learning*, (Springer, 2010)
26. C. E. Antoniak, *Annals of Statistics* **2**, 1152 (1974).
27. S. Jain and R. M. Neal, *Journal of Computational and Graphical Statistics* **13** (2004).
28. M. Plummer, N. Best, K. Cowles and K. Vines, *R News* **6**, 7 (2006).

STEPWISE GROUP SPARSE REGRESSION (SGSR): GENE-SET-BASED PHARMACOGENOMIC PREDICTIVE MODELS WITH STEPWISE SELECTION OF FUNCTIONAL PRIORS¹

IN SOCK JANG, RODRIGO DIENSTMANN

Sage Bionetworks

1100 Fairview Ave. N Seattle, WA 98109, USA

Email: in.sock.jang@sagebase.org

Email: rodrigo.dienstmann@sagebase.org

ADAM A. MARGOLIN[†]

Oregon Health & Science University

3181 S.W. Sam Jackson Park Rd, Portland, OR 97239, USA

Email: margolin@ohsu.edu

JUSTIN GUINNEY[†]

Sage Bionetworks

1100 Fairview Ave. N Seattle, WA 98109, USA

Email: justin.guinney@sagebase.org

Complex mechanisms involving genomic aberrations in numerous proteins and pathways are believed to be a key cause of many diseases such as cancer. With recent advances in genomics, elucidating the molecular basis of cancer at a patient level is now feasible, and has led to personalized treatment strategies whereby a patient is treated according to his or her genomic profile. However, there is growing recognition that existing treatment modalities are overly simplistic, and do not fully account for the deep genomic complexity associated with sensitivity or resistance to cancer therapies. To overcome these limitations, large-scale pharmacogenomic screens of cancer cell lines – in conjunction with modern statistical learning approaches - have been used to explore the genetic underpinnings of drug response. While these analyses have demonstrated the ability to infer genetic predictors of compound sensitivity, to date most modeling approaches have been data-driven, i.e. they do not explicitly incorporate domain-specific knowledge (priors) in the process of learning a model. While a purely data-driven approach offers an unbiased perspective of the data – and may yield unexpected or novel insights - this strategy introduces challenges for both model interpretability and accuracy. In this study, we propose a novel prior-incorporated sparse regression model in which the choice of informative predictor sets is carried out by knowledge-driven priors (gene sets) in a stepwise fashion. Under regularization in a linear regression model, our algorithm is able to incorporate prior biological knowledge across the predictive variables thereby improving the interpretability of the final model with no loss – and often an improvement - in predictive performance. We evaluate the performance of our algorithm compared to well-known regularization methods such as LASSO, Ridge and Elastic net regression in the Cancer Cell Line Encyclopedia (CCLE) and Genomics of Drug Sensitivity in Cancer (Sanger) pharmacogenomics datasets, demonstrating that incorporation of the biological priors selected by our model confers improved predictability and interpretability, despite much fewer predictors, over existing state-of-the-art methods.

* This work is supported by grant U54CA149237 from the Integrative Cancer Biology Program and by grant U01CA176303 from the Cancer Target Discovery and Development of the National Cancer Institute

[†] Corresponding authors.

1. Introduction

High-throughput technologies such as microarray and deep sequencing have been extensively used to reveal that cancer subtypes can be molecularly defined based on their corresponding genomic alterations [1-4]. Moreover, two large-scale pharmacogenomics cell line screens have become available with genomic profiles and drug response of hundreds of clinical and preclinical anti-cancer compounds: the Cancer Cell Line Encyclopedia (CCLE) [5, 6] and the Genomics of Drug Sensitivity (Sanger) projects [7-9]. Both studies demonstrated that genomic features identified by modern machine learning algorithm could be a viable preclinical tool for identifying potential drug sensitivity or resistance markers, with the potential for guiding precision medicine applications and clinical trial design.

In contrast to data-driven pharmacogenomic modeling, decades of experimental molecular biology has produced a detailed (albeit incomplete) knowledge of gene-gene regulatory networks and pathways. The Kyoto Encyclopedia for Genes and Genomes (KEGG), for example, is a collection of comprehensive pathway information derived from experimental analyses and literature curation [10]. Pathway Commons is another rich resource that integrates biological pathway and molecular interaction information from many publicly available databases [11]. Importantly, pathway databases represent only the static regulatory relationships between genes or gene products and are typically context independent [12]. In addition, it is well known that pathways are not functionally independent but are highly coupled processes, with constitutive pathway genes playing multiple roles within different biological processes.

As computational approaches for modeling therapeutic response are being increasingly used in research and translational applications, systematic analyses and best practices recommendations have been recently published [13, 14]. However, these studies have primarily focused on computational or algorithmic improvements. Integrating prior knowledge in predictive algorithms may increase the biological interpretability of these models, and potentially mitigate issues of data over-fitting. Several analytical studies have already incorporated pathways or network information in the variable selection framework [15-21] or used network knowledge to identify differentially expressed genes [22, 23]. However, most of these studies considered only pre-selected pathways as “prior knowledge”, impeding an unbiased assessment of how each pathway is individually associated with model performance. In addition, group lasso algorithms [24-26] were proposed for solving the group sparsity problem. However, biological priors such as pathways are highly coupled and overlapping, and therefore do not optimally match the conditions required for group lasso.

In this study, we present the Stepwise Group Sparse Regression (SGSR) model, developed to leverage prior knowledge in order to improve predictive power and interpretability in the context of modeling drug response with genomic data. Specifically, we embedded a prior selection procedure into sparse regression, such that it could specify preferences for particular combination of priors in the model. The rationale is derived from forward stepwise selection, in which selection of gene-set-coherent features are encouraged through regularization, while the best combination of feature sets is determined by forward stepwise method. We first explored the effectiveness of the SGSR as compared to LASSO, Ridge and Elastic Net regression on the CCLE and Sanger cell line studies [5-7, 9, 13, 14], and then analyzed whether informative pathway priors

improved the selection of previously validated drug-targets in our model, e.g. MAPK pathway genes for MEK inhibitors. We also demonstrated and compared the effectiveness and power of SGSR using different genomic features as input variables, e.g. gene-expression (EXP) vs. copy-number alterations (COPY). For the public accessibility, we provide an R package at <https://github.com/Sage-Bionetworks/SGSR>, and share all results through <https://www.synapse.org/#!Synapse:syn2600070>.

2. Material and Methods

2.1. Materials: Datasets and Prior knowledge databases

Datasets: The CCLE and Sanger datasets contain anti-cancer compound screening data performed on large panels of molecularly characterized cancer cell lines. Both datasets contain high-throughput gene expression and copy number alterations, as well as mutation status on a subset of genes, summarized to gene-level features. Here, we utilize either EXP or COPY dataset to predict drug responses.

In Sanger we have 664 cell lines with EXP measurements on 12,024 genes (643 cell lines with COPY data on 12,082 genes), whereas CCLE has 491 cell lines with EXP measurements on 18,897 genes (488 cell lines with COPY data on 21,217 genes). All data was normalized as described in the original papers [5, 7]. Both studies provided multiple drug dose statistics such as IC50 and ActArea (or AUC) to summarize dose-response curves to compound sensitivity values for each cell line. We chose ActArea with CCLE and IC50 with Sanger, respectively, based on our previous analyses showing their predictive benefit [13]. In addition, we chose 28 out of 138 compounds in Sanger and all 24 compounds in CCLE: 14 overlapping drugs in both cell line studies, selected for cross-comparison. One of the main objectives of the proposed model is to improve interpretability by taking advantage of prior knowledge on pathways that may be implicated in sensitivity/resistance patterns to anti-cancer compounds. Sanger has drug response data to many agents that are not being investigated as anti-neoplastic drugs or that have multiple - and overlapping - targets, making interpretation of the results difficult. We decided to select for downstream analyses Sanger compounds for which there is substantial level of evidence in the literature in terms of preclinical or clinical oncological translation, making sure that we had at least one drug that inhibits relevant targets (known cancer drivers) included in the final list.

Prior knowledge databases: Curated pathway databases represent a valuable resource for scientists studying biological processes in cancer. We take advantage of this information accumulated over years of biomedical research and define a knowledge-driven prior as a set of genes that are mapped to a curated pathway. We anticipate that our model selects a set of pathways – and corresponding genes – that are most likely functionally important for drug sensitivity patterns, therefore increasing biological interpretability of the final set of features. Thus, our prior incorporated predictive model goes beyond traditional analyses by learning the complex structure of input variables and their functional relationships with response. As input to the SGSR model we used the GRAPH Interaction from pathway Topological Environment (graphite: R package built in Bioconductor [27]), providing access to publicly available canonical pathway databases such as KEGG (n=232), Biocarta (n=254), NCI/Nature (n=177) and gene ontology (GO) Biological Processes (n=825) and Molecular Functions (n=396) in MSigDB 3.0 [28].

2.2. Baseline regularized regression methods

A major challenge in the development of predictive models utilizing high-dimensional, genomic data is finding the optimal trade-off between predictive performance and model sparsity (often associated with model interpretability). In the context of drug sensitivity modeling, this trade-off is particularly acute as the selection of biomarkers for patient stratification is a primary goal. Simultaneously, model performance is used to evaluate the ultimate feasibility of drug prediction, and robustness of the biomarkers. Moreover, the incorporation of prior knowledge into data-driven models is a non-trivial task. Biological priors are highly coupled and oftentimes redundant, thereby complicating the process by which they might be included in model building.

To resolve these problems, we have implemented a predictive modeling framework that systematically incorporates prior biological knowledge. Here we present the prior incorporated sparse regression model and its internal prior selection procedure in terms of forward-stepwise selection. Throughout the text we consider the linear regression model $y = \mathbf{X}\beta + \varepsilon$, where y represents the $(n \times 1)$ vector of responses, \mathbf{X} corresponds to the $(n \times p)$ matrix of features, β corresponds to the $(p \times 1)$ vector of regression coefficients, and ε represents a $(n \times 1)$ noise vector. The original problem is to estimate vector of coefficients $\hat{\beta} = \text{argmin}(\|y - \mathbf{X}\beta\|^2)$ with least square criteria. In the “large p (features), small n (samples)” paradigm, the solution to the least-squared problem is undetermined and requires constraining the model space. Recent studies have shown that regularized regression can lead to practical solutions for modeling high-dimensional genomic data [13, 29-33]. Specifically, the LASSO model imposes an L1 penalty on β ($\hat{\beta} = \text{argmin}(\|y - \mathbf{X}\beta\|^2 + \lambda \|\beta\|_1)$) and typically results in sparse solutions where most coefficients are exactly zero. Conversely, the Ridge model imposes an L2 penalty on its model parameters ($\hat{\beta} = \text{argmin}(\|y - \mathbf{X}\beta\|^2 + \lambda \|\beta\|_2)$) and often produces a model where most coefficients are non-zero. However, in practice the use of these penalty functions have several limitations: the LASSO selects at most n variables before it saturates and if there is a group of highly correlated variables the LASSO tends to select one representative from a group and ignore the other components in the group. Meanwhile, models based on Ridge regression tend to perform well [13], but are hard to interpret due to lack of feature selection. To address these problems, Elastic Net regression linearly combines the L1 and L2 penalties of the LASSO and Ridge methods and optimizes two hyper-parameters (λ_1 and λ_2) $\hat{\beta} = \text{arg min}(\|y - \mathbf{X}\beta\|^2 + \lambda_1 \|\beta\|_1 + \lambda_2 \|\beta\|_2)$. Even though the Elastic Net regression method is able to select features that are not identified by LASSO because of high pairwise correlation – while still remaining parsimonious – there are intrinsic limitations of data-driven models: biological insights of model features can only be extracted after extensive post processing steps, including pathway enrichment analyses.

2.3. Stepwise Group Sparse Regression (SGSR)

The SGSR model is based on a stepwise forward “prior” selection procedure (see Figure 1 describing the workflow of the SGSR algorithm). We first define the following

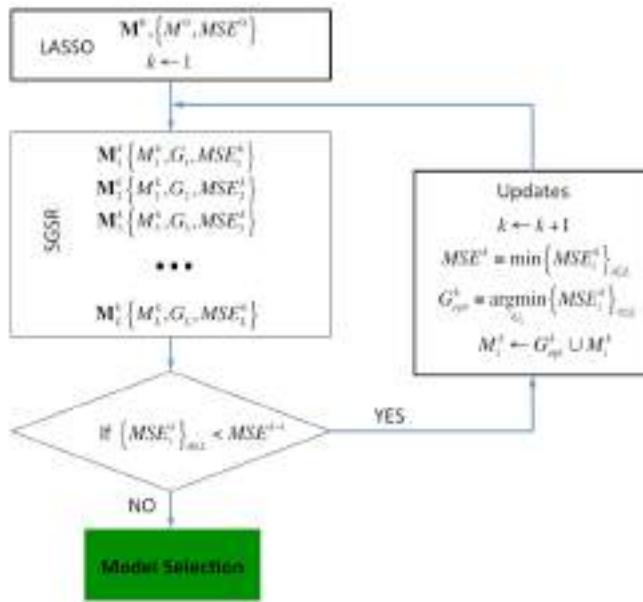


Figure 1. Workflow describing the SGSR algorithm. We define M_i^k, G_i , and MSE_i^k as the set of model features, genes in a gene set, and Mean Squared Error (MSE), respectively, corresponding to the i -th gene set and k -th round of the algorithm.

terms: M_i^k, G_i , and MSE_i^k correspond to the set of model features, genes in a gene set, and Mean Squared Error (MSE), respectively, corresponding to the i -th gene set and k -th round of the algorithm. We define L as the total number of pathways (gene sets) in the database. In the initialization step, we train a standard LASSO model without utilization of gene set priors, optimizing the LASSO's single hyper-parameter using 5-fold cross-validation. We define the set of selected features in this model as M^0 and its MSE as MSE^0 . The stepwise forward prior selection process begins by evaluating the addition of each gene set to the previous model (e.g. $M_i^k = M_i^{k-1} \cup G_i$) and the model that results in the largest reduction of MSE is selected as the model input for the next round (see Figure 1). Of note, the newly added genes from each gene set are unpenalized in the LASSO model, allowing them to enter into the model as a group. If none of the L models produces a lower MSE than the previous optimal model, then the iteration terminates and the previous M^{k-1} is returned.

More specifically, we have $\{\hat{\beta}^i = \text{argmin}(\|y - X\beta\|^2 + \lambda \|\beta_{\neq i}\|_1 + \|\beta_{\in i}\|_1)\}_{i \in L}^k$ and $\{MSE_i^k\}_{i \in L}$ in the k -th round, where $MSE_i^k = \frac{1}{N} \|y - X\hat{\beta}^i\|^2$, $\hat{\beta}^i$ are the coefficients trained by incorporating i -th prior's genes in its LASSO model, $\beta_{\neq i}$ are the coefficients for the predictors which do not belong to i -th prior's genes so that they should be penalized, $\beta_{\in i}$ are the coefficients that correspond to i -th prior's genes and should be always unpenalized in the model training, and $\{\bullet\}_L^k$ is the set of all L models in the k -th round of the stepwise selection procedure. When $\{MSE_i^k\}_L < MSE^{k-1}$ is satisfied, we select the k -

th best prior by $G_{opt}^k = \operatorname{argmin}_{G_i} \left\{ MSE_i^k \right\}_{i \in L}$. Finally, the algorithm's iteration is terminated either when no further MSE gain is achieved or when all pathways of given database are selected.

2.4. Assessment of model performance

For SGSR model running, we randomly split the input dataset into five non-overlapping sample groups: $4/5^{th}$ s of the samples are used for training, whereas $1/5^{th}$ of the samples are used for testing. The 5-fold cross validation scheme is once again applied within the $4/5^{th}$ s training samples so that we can tune the parameters and have an optimized set of priors. Afterwards, we apply the model in the remaining $1/5^{th}$ test samples and assess the final performance by summarizing the 5 sets of predicted drug responses with the Weighted Root Mean Squared Error (WRMSE) metric. The key reason for dividing the RMSE by the average of variance from observed and predicted values is that we can give proper weights to check whether or not the training procedure is successful. In the present analysis we discarded genomic features that have missing data in samples or that have a variance smaller than 0.02. At each split we obtained a prediction vector $\hat{\mathbf{y}}_j$, where $j \in \{1, 2, \dots, 5\}$, and we computed a single WRMSE between the concatenated predicted vector, $\hat{\mathbf{y}} = (\hat{\mathbf{y}}_1, \hat{\mathbf{y}}_2, \dots, \hat{\mathbf{y}}_5)$, and the full observed response vector, \mathbf{y} . $WRMSE = \sqrt{\frac{(\sum_i (y_i - \hat{y}_i)^2)}{N}} / \sqrt{mean(var(\mathbf{y}), var(\hat{\mathbf{y}}))}$

3. Results

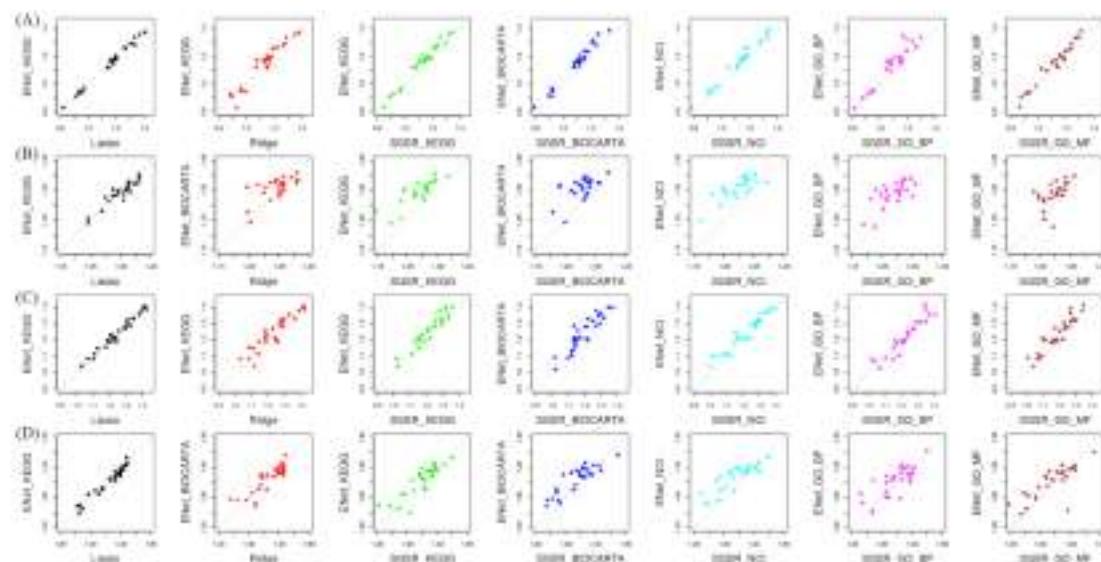


Figure 2. Comparison of model performance (weighted RMSE) between ElasticNet and the Ridge, LASSO, and SGSR algorithms. ElasticNet models are constrained to have a comparable number of features to the SGSR model. Each point corresponds to a single drug model. (A) CCLE with EXP (B) CCLE with COPY (C) Sanger with EXP (D) Sanger with COPY are applied for SGSR with 5 distinctive available pathway databases such as KEGG, BIOCARTA, Nature/NCI, GO_BP and GO_MF.

3.1. Model assessment with fixed sparsity

Using the SGSR framework, we are interested in generating models that are simultaneously sparse (i.e. have a minimal set of features in the model) and optimally predictive. As the Elastic Net regression framework was developed to optimize this trade-off, we compare the SGSR method with the Elastic Net model to determine whether incorporation of pathway knowledge can improve performance. Specifically, we compared overall model performance of SGSR and Elastic Net at comparable levels of model sparsity. Results using the CCLE and Sanger data sets are shown in Figure 2.

In general, we observed an overall improvement in predictive performance using the SGSR model over Elastic Net regression, in which the latter is constrained to have the same number of features as SGSR. This pattern is consistent, regardless of the pathway database selected, with the exception of the GO_BP pathways applied on the Sanger data set. Consistent with our previous work [13], we observed that models utilizing EXP data are more accurate. Interestingly, knowledge-driven priors significantly improved model performance when using COPY as input data, particularly in CCLE ($P<0.0001$ for all models, Wilcox rank sum test with all 5 corresponding pathway databases) while the performance improvement in Sanger with COPY depended on the type of pathway database that was utilized (see Table 1 (A)). Due to marginal gains of predictive performance with EXP, not all SGSR models were statistically significant. Overall, SGSR improved predictive accuracy over Elastic Net in the majority of comparisons (see “performance gain ratio” in Table 1(A)).

Benchmarked with Elastic Net regression									
(A)	CCLE				Sanger				
	EXP		COPY		EXP		COPY		
	wilcox test p-value	performance percentage							
SGSR_KEGG	0.19	31.1%	2.39E-05	99.99%	0.20	32.1%	0.23	46.4%	
SUSR_BIOCARTA	0.04	95.0%	1.88E-06	99.99%	0.03	95.0%	0.07	15.0%	
SGSR_NCI	0.14	35.0%	1.20E-06	99.99%	0.18	33.0%	0.07	10.0%	
SGSR_GO_BP	0.08	91.0%	1.59E-03	99.99%	0.84	32.1%	0.25	22.0%	
SGSR_GO_MF	0.29	35.0%	6.46E-05	91.7%	0.05	22.0%	0.31	22.0%	

Benchmarked with Lasso									
(B)	EXP				COPY				
	EXP		COPY		EXP		COPY		
	wilcox test p-value	performance percentage							
Ridge	0.14	35.0%	1.79E-03	92.0%	0.27	33.0%	0.57	46.4%	
SUSR_BIOCARTA	0.12	95.0%	2.61E-05	99.99%	0.89	95.0%	0.06	15.0%	
SGSR_NCI	0.10	91.0%	1.30E-06	99.99%	0.42	93.0%	0.01	10.0%	
SGSR_GO_BP	0.02	95.0%	1.10E-05	91.7%	0.94	33.0%	0.02	12.0%	
SUSR_GO_MF	0.10	35.0%	2.01E-05	91.7%	0.0001	22.0%	0.07	22.0%	

Benchmarked with Ridge									
(C)	EXP				COPY				
	EXP		COPY		EXP		COPY		
	wilcox test p-value	performance percentage							
Lasso	0.08	16.7%	0.03	57.9%	0.73	28.6%	0.45	50%	
SGSR_BIOCARTA	0.44	45.0%	0.002	99.99%	0.21	95.0%	0.10	15.0%	
SUSR_BIOCARTA	0.12	35.0%	0.002	97.9%	0.09	22.0%	0.03	12.0%	
SGSR_NCI	0.48	45.0%	0.001	97.9%	0.25	95.0%	0.06	15.0%	
SGSR_GO_BP	0.17	91.0%	0.0001	99.99%	0.43	33.0%	0.02	12.0%	
SGSR_GO_MF	0.34	35.0%	0.0001	99.99%	0.84	57.9%	0.03	50%	

Table 1. Performance assessment with Wilcox rank sum test and performance percentage. Orange are depicted for CCLE while light green are for Sanger (A) pairwise assessment table for fixed sparsity in Figure 2, (B) Pairwise assessment table for Figure 3 and 4 when LASSO is used for benchmarked model (C) Pairwise assessment table for Figure 3 and 4 when Ridge is used for benchmarked model. Performance percentage is computed by counting how many drug models of SGSR outperform the benchmark model. Red and green are depicted when SGSR shows better performance (>50%) than the benchmark model.

3.2 Assessment of data-driven model vs. knowledge-driven model

We also investigated the performance of the data-driven models and the SGSR knowledge-driven model, independent of sparsity constraints. Figures 3 and 4 summarize the results of the two data-driven models (Ridge & LASSO) with SGSR using several pathway databases. In general, we observed that Ridge outperforms LASSO, consistent with previous work [13]. The improvement of SGSR over LASSO was generally higher than what we observed with Ridge over LASSO. Using the CCLE data set, SGSR with COPY markedly outperformed the data-driven models while SGSR with EXP produced marginally better performance results (see Figure 3 and Table 1 (left orange panels of B and C)). Similarly, with the Sanger data, differences in favor of the SGSR algorithm showed consistent trends for both the COPY and EXP models (see Figure 4 and Table 1 (right light green panels of B and C)). Of note, the final number of predictors in SGSR models was on average only marginally increased as compared with the LASSO models (91.7%, 94.2%, 87.9% and 79.3% in CCLE EXP, CCLE COPY, Sanger EXP and Sanger COPY, respectively).

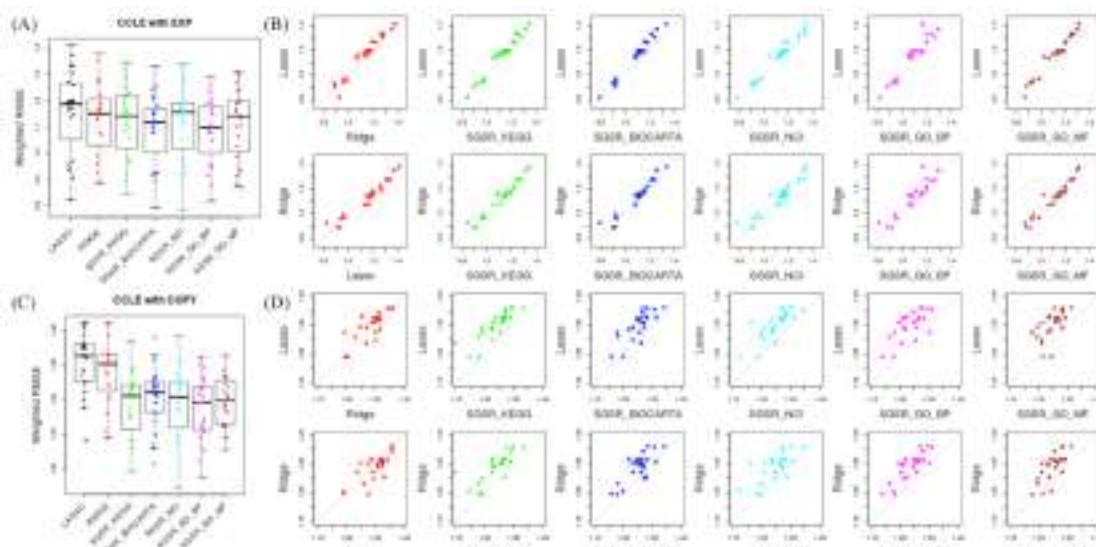


Figure 3. Performance comparison for CCLE pharmacogenomics data. (A) Predictability score with WRMSE metric of LASSO, Ridge, SGSR with KEGG, Biocarta, NCI/Nature, GO-BP and GO-MF pathways using EXP data across the 24 CCLE drugs, (B) Performance discrepancy between benchmarked LASSO, Ridge, and SGSR models with five available pathway databases with EXP; (C) Predictability score with WRMSE metric of LASSO, Ridge, SGSR with KEGG, Biocarta, NCI/Nature, GO-BP and GO-MF pathways with COPY across the 24 CCLE drugs, (D) Performance discrepancy between benchmarked LASSO, Ridge and SGSR models with five available pathway databases using COPY data.

3.3. Assessment of additional features identified by SGSR model

We next defined 2 tests to assess whether the improved performance by SGSR can be explained by factors other than the information contained in the gene-set priors. First, in order to check whether SGSR improves performance simply by adding additional features, we constructed a null distribution of predictive performance by generating 50 random models that had the same number of features added by SGSR. To do this, we preserved the original model fit by LASSO (M^0) and then randomly added

genes until we had a model with the same number of features as the SGSR model to which we are comparing. Second, to test whether similar performance could be obtained by incorporation of non-informative gene sets, we trained SGSR models using randomly permuted gene-set priors. Specifically, we preserved the input pathway database structure (i.e. maintained the same number of genes per gene set) but randomly shuffled the genes within each gene set. Figure 5 summarizes the predictive performance of SGSR models compared to the randomized models. In general, the WRMSE of SGSR models is significantly lower than that of both null models.

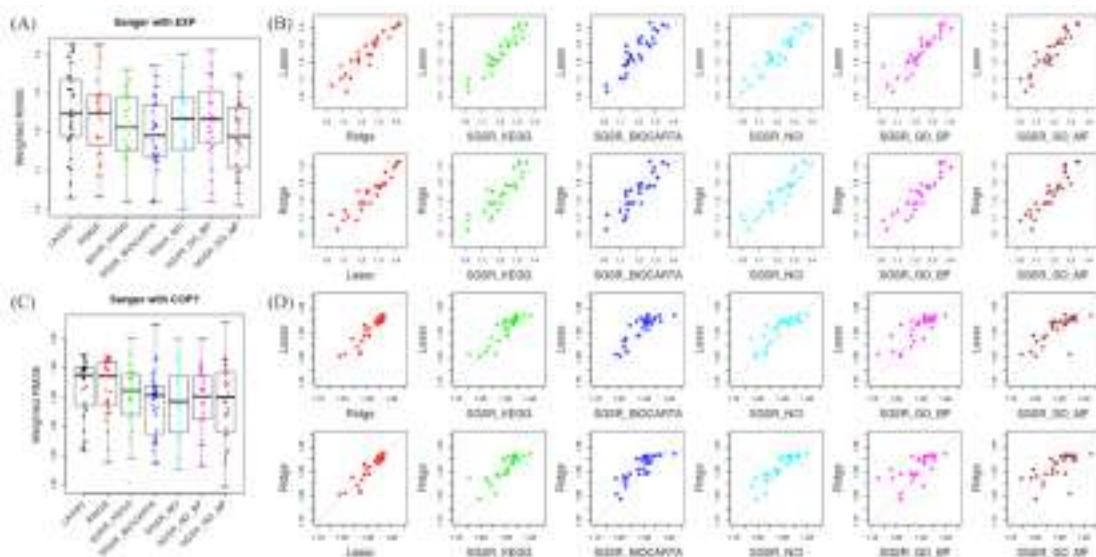


Figure 4. Performance comparison for Sanger pharmacogenomics data. (A) Predictability score with WRMSE metric of LASSO, Ridge, SGSR with KEGG, Biocarta, NCI/Nature, GO-BP and GO-MF pathways using EXP data across the preselected 28 Sanger drugs; (B) Performance discrepancy between benchmarked LASSO, Ridge and SGSR models with five available pathway databases with EXP (C) Predictability score with WRMSE metric of LASSO, Ridge, SGSR with KEGG, Biocarta, NCI/Nature, GO-BP and GO-MF pathways using COPY data across the preselected 28 Sanger drugs, (D) Performance discrepancy between benchmarked LASSO and Ridge and SGSR models with five available pathway databases with COPY.

3.4. Biological Interpretability from identified priors for anticancer compounds

One attractive characteristic of SGSR is the ability to perform feature selection with increased interpretability compared to state-of-the-art methods. To exemplify this, we analyzed the results of EXP-based SGSR models (with prior using NCI/Nature Cancer pathway database) of sensitivity/resistance to the MEK inhibitors AZD6244 and PD0325901, agents tested in both CCLE and Sanger. We then compared with the matching bootstrapped Elastic Net regression models. It is known that response to these agents correlates with mutation status of *KRAS/NRAS/BRAF* genes [5, 7]. However, we wanted to assess whether models built on gene expression measurements could give additional biologically meaningful information. Overall, predictive performance of SGSR models for AZD6244 and PD0325901 in both CCLE and Sanger data sets are comparable to the gold-standard method. In addition, top features (genes) identified in SGSR models for each agent significantly overlap both within and across data sets, underscoring the reproducibility and potential biological relevance of the findings. As shown in Table 2, overlapping genes of major interest include: (i) MAP2K1 (also known as MEK) and

MAPK1 (also known as ERK), important downstream effectors of the mitogen-activated protein kinase (MAPK) pathway; (ii) RHOA, a small GTPase known to interact with MAPK pathway to promote cell invasion [34]; (iii) AURKB, regulated by MAPK pathway to promote cell division [35]; (iv) Src family kinases SRC and FYN, which have a critical role in cell migration, proliferation and survival via the MAPK pathway [36, 37]; and (v) EDIL3 (EGF-like repeats and discoidin I-like domains 3), a stromal factor that is associated with angiogenic switch and poor prognosis in many cancers [38, 39]. By contrast, the genes described above were not inferred within the top 500 features by the bootstrapped Elastic Net regression models based on gene expression data. Although anecdotal, this analysis suggests that incorporating pathway information during the design of predictive models can identify functionally relevant biomarkers that would not be detected from a purely data-driven approach.

Biomarker	SGSR				Bootstrapped Elastic Net regression			
	CCLE		Sanger		CCLE		Sanger	
	AZD	PD	AZD	PD	AZD	PD	AZD	PD
EDIL3	Y	Y	Y	Y	7260	11170	5109	4361
RHOA	Y	Y	Y	Y	36833	18775	1335	5022
FYN	NA	NA	Y	Y	11962	9574	7932	10345
MAPK1	Y	Y	NA	Y	618	815	8914	10352
MAP2K1	Y	Y	NA	Y	30123	11896	11924	8338
AURKB	Y	Y	Y	Y	12979	16464	6772	8464
SRC	NA	NA	Y	Y	10820	16675	8501	8516

Table 2. For the SGSR model, the top 7 predictive features are displayed for AZD6244 (AZD) and PD0325901(PD). Cells highlighted in orange correspond to features with evidence of being functionally related to MEK inhibitor compounds, as described in the text. For comparison, the ranks of corresponding predictive features inferred by bootstrapped Elastic Net are displayed (18,897 and 12,024 features are considered in model building with CCLE and Sanger, respectively).

4. Discussion

The availability of large-scale pharmacogenomic screens on cancer cell line panels has begun to illuminate many of the genomic aberrations underlying compound sensitivity/resistance. The application of machine learning approaches optimized for feature selection on high-dimensional genomic data has been a critical tool in this analysis. Even though the tractability of penalized regression models has been proposed in earlier studies [5, 7, 13], the resultant models fail to incorporate well-known pathway characteristics that frequently underlie drug efficacy *in vitro* and in patients. In this study, we propose a novel SGSR algorithm that allows known pathway relationships to influence feature selection during model fitting, thereby enhancing interpretability of the final model without a concomitant decrease in model performance.

Our study benchmarks a statistically principled comparison with state-of-the-art machine learning algorithms - namely LASSO, ElasticNet and Ridge regression – to predict drug sensitivity using input features from gene expression or copy number. In general, we find that the SGSR model has better overall accuracy (smaller MSE) at comparable levels of model sparsity. Of note, we observed the highest gains in predictive performance in the models that originally gave weak predictions, such as those based on COPY data [13]. Moreover, we observe that the specific grouping of the pathways (gene sets) contributes meaningful information, demonstrated in our comparison of SGSR to

randomly constructed pathways. This is important, as we might expect that in aggregate the union of all genes from all pathways represent the set of genes/proteins that are more frequently studied, and therefore alone might explain the improved SGSR performance. However, the relevance of the specific gene set composition underscores the complex and pertinent information embedded in these gene sets. Finally, we consider the biological insights derived from our model (at the gene level) and interpretability of results (at the pathway level) as major advantages for cancer researchers.

In summary, SGSR provides a knowledge-incorporated sparse regression framework with significantly increased model interpretability without a trade-off of prediction accuracy. Notably, our modeling approach highlights the value of existing knowledge databases and their relevance in modeling disease phenotypes. Future directions might consider incorporation of even finer-grained relationships (dependence) embedded in these pathway databases, such as the protein interactions encoded in the Reactome pathways. We believe that SGSR advances current state-of-the art approaches for inferring molecular predictors of compound sensitivity, and may be used to identify functionally relevant gene sets used to guide translation of preclinical screens into precision medicine trials.

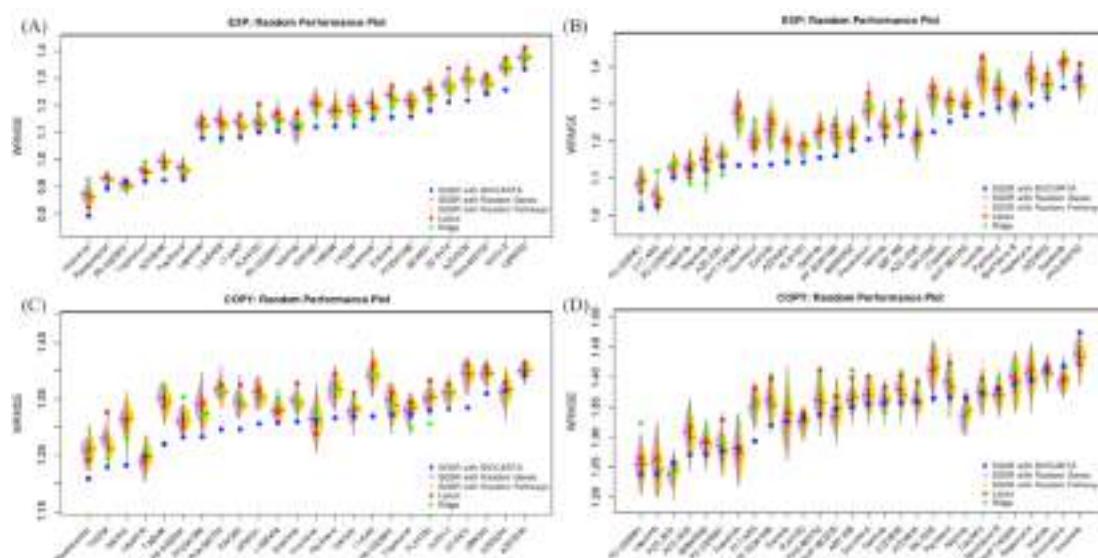


Figure 5. Model assessments per drug (WRMSE) including SGSR with BIOCARTA (blue); SGSR using random genes (purple, left distribution using 50 random models); SGSR using random pathways (yellow, right distribution using 50 random models); LASSO (red); RIDGE (green). (A) CCLE with EXP (B) Sanger with EXP (C) CCLE with COPY and (D) Sanger with COPY

5. References

1. Ferte, C., F. Andre, and J.C. Soria, Nat Rev Clin Oncol, 7(7)(2010).
2. Peggs, K. and S. Mackinnon, N Engl J Med, 348(11) (2003).
3. Roche-Lestienne, C., et al., N Engl J Med, 348(22) (2003).
4. Savage, D.G. and K.H. Antman, N Engl J Med, 2002. 346(9) (2002).
5. Barretina, J., et al., Nature, 483(7391) (2012).

6. Marum, L., Future Med Chem, 4(8) (2012).
7. Garnett, M.J., et al., Nature, 483(7391) (2012).
8. Yang, W., et al., Nucleic Acids Res, 41(Database issue) (2013).
9. Forbes, S.A., et al., Nucleic Acids Res, 39(Database issue) (2011).
10. Kanehisa, M., et al., Nucleic Acids Res, 40(Database issue) (2012).
11. Cerami, E.G., et al., Nucleic Acids Res, 39(Database issue) (2011).
12. Draghici, S., et al., Genome Res, 17(10) (2007).
13. Jang, I.S., et al., Pac Symp Biocomput, (2014).
14. Neto, E.C., et al., Pac Symp Biocomput, (2014).
15. Chen, L., et al., BMC Syst Biol, 5 (2011).
16. Li, C. and H. Li, Bioinformatics, 24(9) (2008).
17. Tai, F. and W. Pan, Bioinformatics, 23(23) (2007).
18. Tai, F. and W. Pan, 23(14) (2007).
19. Wang, Z., et al., Bioinformatics, 29(20) (2013).
20. Wei, P. and W. Pan, Bioinformatics, 24(3) (2008).
21. Jang, I.S., A. Margolin, and A. Califano, Interface Focus, 3(4) (2013).
22. Li, X., et al., Proteomics, 11(19) (2011).
23. Wei, Z. and H. Li, Bioinformatics, 23(12) (2007).
24. Yuan, M., et al. Journal of the Royal Statistical Society Series B-Statistical Methodology, 68 (2006)
25. Friedman, J., et al. arXiv:1001.0736 (2010)
26. Jacob, L., et al. In Proceedings of the 26th Annual International Conference on Machine Learning (2009)
27. Sales, G., et al., Bmc Bioinformatics, 13 (2012).
28. Liberzon, A., et al., Bioinformatics, 27(12) (2011).
29. Hoerl, A.E., R.W. Kennard, and R.W. Hoerl, Applied Statistics-Journal of the Royal Statistical Society Series C, 34(2) (1985).
30. Tibshirani, R., Journal of the Royal Statistical Society Series B-Methodological, 58(1) (1996).
31. Tibshirani, R., Journal of the Royal Statistical Society Series B-Statistical Methodology, 73 (2011).
32. Zou, H. and T. Hastie, Journal of the Royal Statistical Society Series B-Statistical Methodology, 67 (2005).
33. Fu, W.J.J., Journal of Computational and Graphical Statistics, 7(3) (1998).
34. Vilal, E., et al. Cancer Cell, 4(1) (2003).
35. Bonet, C., et al. J Biol Chem, 287(35) (2012).
36. Kim, L.C., et al. Nat Rev Clin Oncol, 6(10) (2009).
37. Yadav, V., et al. Mol Carcinog, 50(5) (2011).
38. Sun, J.C., et al. World J Gastroenterol, 16(36) (2010).
39. Damhofer, H., et al. Mol Oncol, 7(6) (2013).

INTEGRATIVE GENOME-WIDE ANALYSIS OF THE DETERMINANTS OF RNA SPLICING IN KIDNEY RENAL CLEAR CELL CARCINOMA

KJONG-VAN LEHMANN^{1,*‡}, ANDRÉ KAHLES^{1,‡} CYRIAC KANDOTH¹ WILLIAM LEE¹,
NIKOLAUS SCHULTZ¹ and OLIVER STEGLE² and GUNNAR RÄTSCH¹

1 Computational Biology Center, Memorial Kettering Cancer Center, New York, NY 10044, U.S.A

*2 European Molecular Biology Laboratory, European Bioinformatics Institute, Hinxton, Cambridge, CB10
1SD, United Kingdom*

‡ Both authors contributed equally.

We present a genome-wide analysis of splicing patterns of 282 kidney renal clear cell carcinoma patients in which we integrate data from whole-exome sequencing of tumor and normal samples, RNA-seq and copy number variation. We proposed a scoring mechanism to compare splicing patterns in tumor samples to normal samples in order to rank and detect tumor-specific isoforms that have a potential for new biomarkers. We identified a subset of genes that show introns only observable in tumor but not in normal samples, ENCODE and GEUVADIS samples. In order to improve our understanding of the underlying genetic mechanisms of splicing variation we performed a large-scale association analysis to find links between somatic or germline variants with alternative splicing events. We identified 915 *cis*- and *trans*-splicing quantitative trait loci (sQTL) associated with changes in splicing patterns. Some of these sQTL have previously been associated with being susceptibility loci for cancer and other diseases. Our analysis also allowed us to identify the function of several COSMIC variants showing significant association with changes in alternative splicing. This demonstrates the potential significance of variants affecting alternative splicing events and yields insights into the mechanisms related to an array of disease phenotypes.

Keywords: QTL, splicing, genomics, RNA-seq, Exome, CNV, statistical genetics

1. Introduction

The analysis of gene expression and the identification of expression quantitative trait loci (eQTLs) has become a standard part of the analyses performed in many population genetics studies. However, the variability in expression levels is only one of the factors shaping the complexity of the transcriptome. RNA-modifying processes, especially the process of alternative splicing, enable the formation of several RNA isoforms from a single gene locus and drastically increase transcriptome complexity. During splicing, specific parts are excised from the pre-mRNA (introns) and the remaining parts (exons) are re-connected. Through combinatorial choice of introns, different mRNAs can be generated. This tightly regulated process is also termed alternative splicing.^{1,2} The role of alternative splicing in cancer is being actively investigated,^{3,4} however it is often difficult to separate tissue specific effects from tumor specific changes. Defects in the splicing machinery or dysregulation of the process can lead to disease or play an active role in cancer progression.^{5–7} Interestingly, several strategies involving natural compounds or antisense oligonucleotides have been suggested to target aberrant splicing,^{8,9} making the detection of alternative splicing events as drug targets desirable. Yet, splicing efficiency has only recently been considered as a quantitative trait in genetic analysis. First studies describing systematically alternative splicing in the context for genomic variation have been conducted by Battle *et al.*¹⁰ and in context of the GEUVADIS project.¹¹

Also studies with a specific focus on *single* alterations that affect splicing have been published recently, for instance, the identification of somatic mutations in U2AF1 causative for altered splicing in acute myeloid leukemias¹² as well as identification of a somatic variant affecting SF3B1 function.¹³

In this work, we present the genome-wide analysis of alternative splicing events in 282 Kidney Renal Clear Cell Carcinoma (KIRC) samples generated in context of The Cancer Genome Atlas project (TCGA).¹⁴ We perform an integrative analysis of RNA-seq, whole-exome and copy number variation, in order to identify determinants of splicing variation caused by germline and somatic genetic variation. We first built a comprehensive inventory of alternative splicing events occurring in KIRC tumors and characterized tumor-specific splicing controlling for tissue specific effects. Encouraged by the presence of cancer-specific introns, we used a mixed model approach to systematically associate splicing alterations with germline and somatic genetic variants with the aim to identify splicing quantitative trait loci (sQTLs). This analysis enables us to shed some light on genetic mechanisms underlying alternative splicing patterns in cancer and normal cells.

2. Methods

We here provide an outline of the methodology taken. Please note that a detailed description of our methods can be found in the supplemental material.

2.1. Data Processing

Matching 282 whole exome and transcriptome samples have been downloaded from cgHub and were realigned using STAR. For comparison purposes we have also downloaded and re-aligned 140 RNA-Seq samples from the GEUVADIS project as well as 460 RNA-Seq samples from the ENCODE project. Expression counts have been generated based on the GENCODE annotation and splicing phenotypes have been generated using SplAdder.¹⁵

Germline variants have been called using the HaplotypeCaller in GATK and somatic variants have been identified using MuTect.

2.2. Tumor specific splicing analysis

Tumor-specific splicing has been identified by ranking all expressed genes by the ratio of the average number of samples that expressed a certain intron in the KIRC tumor samples over the average number of samples expressing the intron in KIRC normals, GEUVADIS and ENCODE combined. Functional enrichment analysis has been undertaken by making use of the GOrilla webserver.¹⁶

2.3. Quantitative Trait Analysis

Quantification of splicing measured in PSI has been performed using an inverse normal transform resolving ties randomly and variants have been encoded numerically under an additive genetic model (see Supplemental methods for details). A linear mixed model analysis has been used to find associations between germline mutations and splicing changes. We accounted for

population structure as well as possible hidden confounders using PANAMA and known confounders as in gene expression and copy number variation from Ciriello *et al.*¹⁷ Associations have been computed using LIMIX¹⁸ and Benjamini-Hochberg step-up procedure has been used for FDR estimation to correct for multiple testing.

3. Results

3.1. Identification of Tumor-specific Splicing

Based on the splicing graph constructed with SplAdder, we extracted 184,941 introns located in 15,387 genes that were part of alternative splicing events. Of these introns, 160,208 were confirmed by at least 10 spliced alignments in at least one of the samples from the KIRC, GEUVADIS or ENCODE sets. Interestingly, when ranked by exclusive occurrence in tumor samples (see methods), especially transmembrane proteins of the solute carrier family (SLC) comprising a family of roughly 450 genes, were significantly enriched amongst the top ranks showing a 12 fold enrichment (p-value $3.6 \cdot 10^{-5}$, hypergeom. test; compare Fig. 1, Panel A). Although single members of this family have been related to cancer biology, e.g., SLC28A1,¹⁹ in general not much is known about their function in context of cancer. Other top-ranked membrane-associated proteins show stronger known links to cancer, such as the transmembrane collagen COL23A1²⁰ or the transmembrane protein 176A.²¹ Encouragingly, the latter as well as its heterologous protein TMEM176B also appeared on top of the list of genes that show exclusive intron expression and harbor significant sQTL (Fig. 1, Panel B). It is notable that the accumulation of TMEM176A/B has been linked to several other cancer types²¹ previously. Moreover, we also found non-membrane associated genes with exclusive intron expression that are known to have links to cancer, such as secretagogin (SCGN)²² involved in cell proliferation, the cytochrome P450 epoxygenase CYP2J2²³ or the hypoxia-inducible factor EGLN3.²⁴ We observed that most exclusive introns were indeed result of splicing and not an artifact of lacking gene expression, although several genes show considerably less expression in normal samples. (see Supplemental Figure 1).

In agreement to these findings, a functional enrichment analysis on gene ontology (GO) categories showed significant enrichment of membrane transport processes but also in extracellular matrix organization and amino-acid metabolism — processes relevant for tumor growth and cancer progression. Interestingly, on the level of functional categories, we found significant enrichment for receptors in general (p-value $6.7 \cdot 10^{-14}$, 1.9 fold enrichment) and specifically G-protein coupled receptors (p-value $1.7 \cdot 10^{-6}$, 2.1 fold enrichment) as well as for substrate specific transmembrane transport (p-value $4.4 \cdot 10^{-9}$, 2 fold enrichment), pointing to a possible involvement in signaling. On the component level, we found significant enrichments of the plasma membrane (p-value $2.2 \cdot 10^{-20}$, 1.6 fold enrichment) and the extracellular region in general (p-value $4 \cdot 10^{-26}$, 2.2 fold enrichment). The interesting enrichments on the process level include ion transport (p-value $1.3 \cdot 10^{-13}$, 2 fold enrichment) and cell adhesion (p-value $1 \cdot 10^{-10}$, 1.9 fold enrichment). All these results are plausible in the light of what is known about cancer biology and will be further investigated. The identified cancer-specific isoforms have a potential use as diagnostic marker or as possible drug targets.

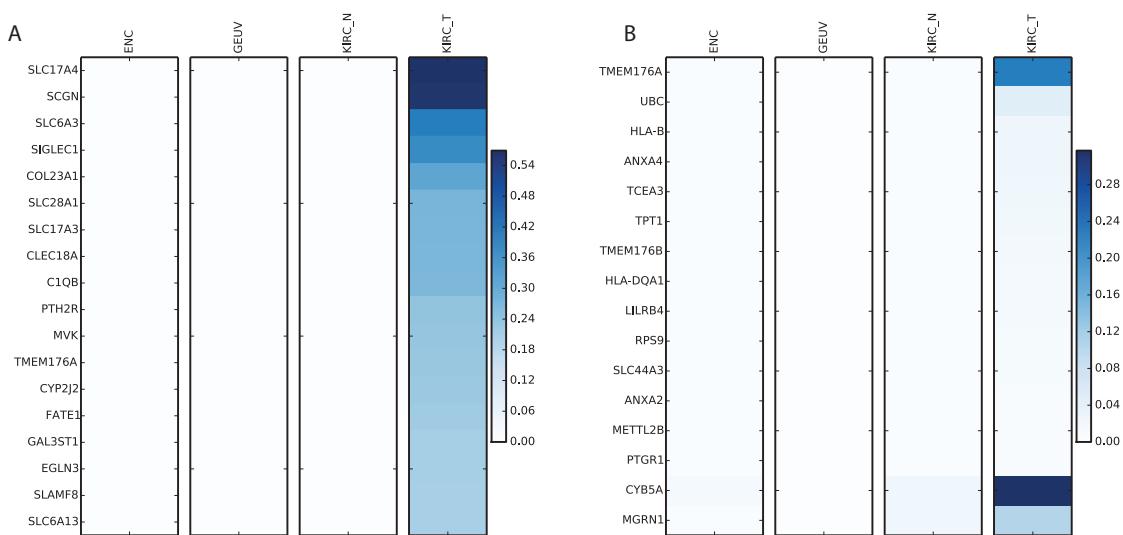


Fig. 1. Enrichment of introns exclusively present in tumor samples. *A*: List of genes that contain the top 20 most exclusive introns. Color represents fraction of samples that have confirmed this intron with ≥ 10 alignments. In all cases we observe no or little evidence of these introns in the control samples. *B*: List of genes with an sQTL that contain exclusively expressed introns. Color represents fraction of samples having confirmed the intron with ≥ 10 alignments.

3.2. Identification of SNVs Associated with Splicing Changes

After preprocessing and filtering, we have analyzed 11,383 exon skip events as well as 3,961 alternative 5'- and 5,038 3'-end events. All events have been associated with 458,266 variants. Since each event can represent the same transcript structure, these events are certainly not independent leading to exon skip events in 5,623 genes, 3,703 genes with alternative 3'-ends and 3,278 genes showing alternative 5' ends. After a very conservative correction (p -value $< 5 \cdot 10^{-9}$ and 5% effect size), we find 251 polymorphic sQTLs of which 228 are *cis*-QTLs and 23 are *trans*-QTLs. Full break down in Table 1 and an overview of all sQTLs can be seen in Figure 2. It clearly demonstrates the generally higher power in detecting *cis*-QTLs due to the more direct nature of their effects and thus are easier to detect.

Table 1. Break down of sQTL associations. This table shows how many sQTLs with more than 10% effect size and p -value $< 5 \cdot 10^{-7}$ p -value and $< 5 \cdot 10^{-9}$ (Filtered) are found to be annotated in various functional databases. The top two rows sum to all 915 sQTLs detected and subsequent subsets are shown below.

Category	Exon Skip (Filtered)	Alt 3' (Filtered)	Alt 5' (Filtered)
<i>cis</i>	228(127)	101(62)	62(39)
<i>trans</i>	312(12)	104(8)	108(3)
<i>cis</i> ClinVar	2(1)	3(2)	0(0)
<i>trans</i> ClinVar	2(0)	0(0)	1(0)
<i>cis</i> COSMIC	18(11)	8(4)	1(0)
<i>trans</i> COSMIC	6(1)	1(0)	2(0)
<i>cis</i> GWAS	1(1)	2(1)	0(0)
<i>trans</i> GWAS	1(0)	0(0)	0(0)

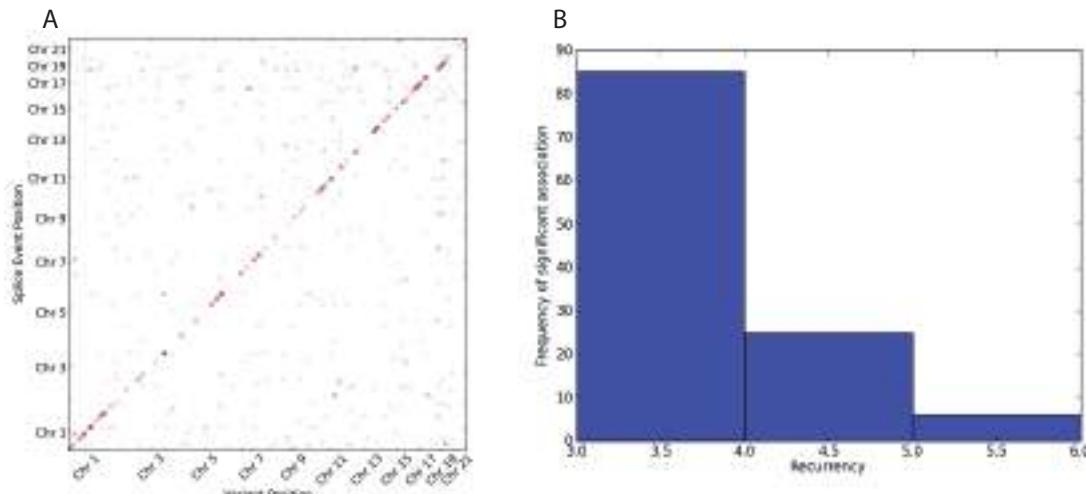


Fig. 2. Summary plots of associations found in somatic and germline calls. *A*: Germline sQTL overview. Every dot represents a sQTL. Results have been filtered for 5% effect size. Red dots indicate sQTLs with $p\text{-value} < 5 \cdot 10^{-9}$ and blue dots indicate sQTLs with $p\text{-value} < 5 \cdot 10^{-7}$. *B*: Histogram indicating how many variants are found recurring (x-axis) in a given number of samples (y-axis).

3.2.1. Associating Somatic Variants

We also considered a small set of 128 recurrent somatic mutations that are expected to be highly enriched with functional variants. Surprisingly, we found a large fraction of those somatic variants to be associated in *trans* and none in *cis* ($p\text{-value} < 5 \cdot 10^{-4}$).

Out of these associations, five are annotated COSMIC variants for which we have found none of them to be significant in our previous analysis on the tumor germline calls. While it could be reasoned that most somatic variants may have a functional effect, it is notable that a high fraction is significant after Bonferroni correction and that most of them are rare (see Figure 2 B). While we are confident in our statistical analysis, technical and biological validation will ultimately be able to separate false positive from biological meaningful results.

3.2.2. ClinVar Annotated sQTL Suggest Functional Mechanisms

In an effort to establish links of interest between sQTLs and existing variants of interest we have compared our results to variants annotated in ClinVar. This analysis revealed that two polymorphic sites both associated ($p\text{-value} < 2.6 \cdot 10^{-8}$ and $p\text{-value} < 1.1 \cdot 10^{-9}$) with the same alternative 3' splicing event within the Paraoxonase 2 gene have previously been associated with risk for coronary heart disease.²⁵ We found another variant associated with three different alternative splice events in the ACP1 gene. While this variant is of benign clinical importance, we have confirmed previous mechanistic insights.²⁶ Another polymorphic site in the SOD 2 gene is of clinical interest and is associated with an exon skip variation within the same gene. This variant is associated with increased risk of nephropathy in diabetics. This gene is of particular interests since variants in this gene have been associated with various diseases as in idiopathic cardiomyopathy (IDC), premature aging, sporadic motor neuron disease, and cancer (genecards source). A type 1 diabetes risk missense variant has also been

associated with an alternate 3'-end in the OAS1 gene and it is known that variants influencing alternative splicing in this gene are of functional importance. It is notable that this region is also spanned by a long non-coding RNA (ENSG00000257452).²⁷ We have also linked a risk variant for Myocardial infarction as well as an autoimmune disease susceptibility variant as sQTL variants with an exon skip event and an alternate 5'-event, respectively. This analysis demonstrates successfully how our sQTL analysis can confirm and suggest mechanistic insights into clinically and molecularly significant phenotypes.

3.2.3. COSMIC Annotated sQTL

We identified 16 *cis*-sQTLs under very conservative thresholds ($p\text{-value} < 5 \cdot 10^{-9}$ and 0.25% effect size), which are also annotated in COSMIC suggesting their potential effect and involvement in cancer. Of particular interest are those variants annotated as sQTL in commonly mutated cancer genes.

Fig. 3 demonstrates an example where the most significant variant is annotated in COSMIC and shows a large difference in the splicing index across the different alleles in gene PMF1. This gene is known to be associated with bladder carcinoma and thus is of specific interest. While this somatic variant is rare, it overlaps a more common germline variant and thus did allow us to identify it as an sQTL in this study.

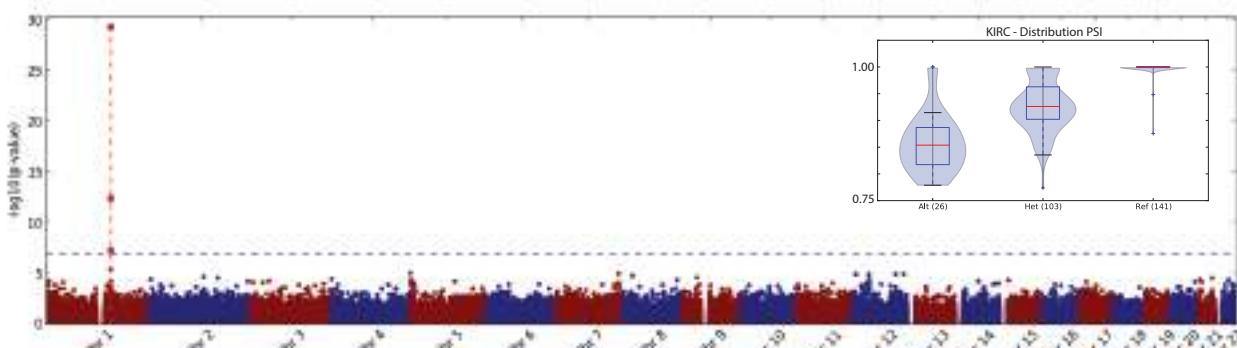


Fig. 3. An example of a COSMIC annotated variant with alternative splicing. The dotted red line indicates the position of the alternative splice event. The figure on the top right shows a violin plot of PSI demonstrating the shift of distribution for each of the three possible genotype across all samples.

We found that the tumor necrosis factor related protein encoded by the C1QTNF3 gene harbors a recurrent COSMIC variant (COSM449566) which also appears to be a *cis* sQTL. This suggests potential functional effects with an effect size of $\sim 30.25\%$. Further analysis of such variants may be promising in understanding the effect of some commonly observed somatic variants.

3.2.4. sQTL as Susceptibility loci for Cancer

A comparison of the set of sQTL with a known catalogue of genome-wide association study (GWAS) hits did yield four loci which have been linked to previous studies.²⁸ Interestingly,

three out of these loci are susceptibility loci for different types of cancers while the other one has been previously linked to multiple sclerosis. A variant introducing an alternate splice junction which then leads to a change of the alternate 3'-end of exon two and subsequently causing a truncation of that exon has previous been identified in two association studies. This variant is being highly associated with changes in serum magnesium levels (PMID) and it has also been identified as a susceptibility locus for gastric adenocarcinoma and esophageal squamous cell carcinoma. An sQTL in a mitochondrial carrier protein (SLC25) has recently been identified as a potential new susceptibility locus for testicular cancer²⁹ which is of particular interest with respect to our finding of SLC enrichment. The well-studied BABAM1 gene appears to host an sQTL which is also an annotated COSMIC variant which is associated in several publications with hormone receptor-negative breast cancer and is also a susceptibility locus for ovarian cancer^{30313233,34}. These results are encouraging and may be suggestive of alternative splicing patterns as underlying mechanism for cancer progression rather than as functional consequence.

4. Discussion

We have completed an extensive analysis on RNA-Seq samples originating from patients with Kidney Renal Clear Cell Carcinoma and present the first systematic sQTL analysis on kidney cancer to our knowledge. State-of-the-art methods are being used to systematically investigate splicing patterns in KIRC samples. We have developed a ranking mechanism to identify splicing events specific to tumor samples in comparison with normal samples. Our analysis demonstrates that we do not only find sQTL for tumor specific splicing events 1 B but are also able to identify functionally annotated variants providing potential new mechanistic insights.

Our analysis revealed a subset of genes that showed large differences in splicing events between tumor and normal samples. Although we tried to control for tissue-specific expression by taking the normal samples into account, this effect may still be partially confounded by cancer specific gene expression (see supplemental Fig. 1). While these genes may certainly be interesting with regards to being cancer markers, it is not absolutely clear whether altered functions associated with the observed events in these genes are driving tumor progression or are mere passenger events. However, considering that several top ranked genes have been previously linked to cancer metabolism and treatment outcome encourages further molecular analyses of these findings. TMEM176 may be a new interesting target due to the identification of tumor specific intron expression as well as the identification of sQTL associated with this event.

The analysis of the variants found in the matched whole exome sequencing data and their association with patterns of alternative splicing revealed various germline and COSMIC variants either directly causal or in linkage disequilibrium with causal variants. Some of the *cis*-sQTLs have previously been annotated as being susceptibility loci for cancer or other diseases and our analysis suggests a potential mechanistic involvement of splicing aberrations. We were further able to link various sQTLs to variants annotated in ClinVar suggesting splicing related mechanisms. In order to understand to what extent somatic mutations may be driving splicing

changes, we have analyzed a small set of somatic variants identified by matching tumor normal pairs. After Bonferroni correction across the events and variants tested, most of the identified somatic variants remain significant which causes some concerns with regard to the amount of false positives in this set. Further analysis demonstrates that most of the associations are seen in less than four samples. While this might be an intrinsic property of somatic mutations, we suggest to explore also different approaches to take the lower frequency of somatic variants into account, specifically addressing false positive results. However, we are encouraged that many of the associations we found are indeed correct and meaningful in the context of cancer biology.

Our analysis is a step forward towards gaining further insight into the involvement of splicing patterns in cancer. It still remains to be seen to what extent alternative splicing patterns can create large changes in phenotypic outcome. Studies of natural population suggest that the effect of germline variants is generally small, however our results suggest that several germline and somatic variants may contribute towards functional changes. While we can confirm previously known alternative splicing events and the genetic markers driving these splicing changes, the functional role of many of these genes and changes in splicing patterns in cancer is unknown. We believe that new approaches and larger sample sizes are needed to gain further insight into the role of somatic variants. Our future work will involve addressing these issues and including samples of larger sizes from the TCGA project to gain more power to study the effect of rare somatic variants in cancer. This will involve the integration of rare-variant analysis approaches and integration of whole-genome data generated by the TCGA and ICGC project.

Acknowledgements: We gratefully acknowledge discussions with Ari Hakimi, Ed Reznik and Chris Sander. This work is funded by the Sloan Kettering Institute (to G.R.) and was relying on the Beckman genomic data storage facility (Geoffrey Beene grant to G.R. & N.S.). Partial funding under NIH grant 1R01CA176785-01A1.

Supplemental materials Additional material including a detailed method description can be found under the following url: <http://www.raetschlab.org/suppl/kirc-splicing>

References

1. M. Green, *Annual Review of Genetics* **20**, 671 (1986).
2. A. R. Kornblihtt, I. E. Schor, M. Alló, G. Dujardin, E. Petrillo and M. J. Muñoz, *Nature Reviews Molecular Cell Biology* **14**, 153 (March 2013).
3. A. Valletti, O. Palumbo, M. D'Antonio, M. Gigante, M. Carella, E. Picardi, A. M. D'Erchia, R. Elena, G. Pesole and others, *Cancer Epidemiology Biomarkers & Prevention* **21**, IA06 (2012).
4. D. Ramsköld, S. Luo, Y.-C. Wang, R. Li, Q. Deng, O. R. Faridani, G. A. Daniels, I. Khrebtukova, J. F. Loring, L. C. Laurent and others, *Nature biotechnology* **30**, 777 (2012).
5. J. Tazi, N. Bakkour and S. Stamm, *Biochimica et Biophysica Acta* **1792**, 14 (January 2009).
6. A. Srebrow and A. R. Kornblihtt, *Journal of Cell Science* **119**, 2635 (2006).
7. V. Quesada, L. Conde, N. Villamor, G. R. Ordóñez, P. Jares, L. Bassaganyas, A. J. Ramsay, S. Beà, M. Pinyol, A. Martínez-Trillos, M. López-Guerra, D. Colomer, A. Navarro, T. Baumann, M. Aymerich, M. Rozman, J. Delgado, E. Giné, J. M. Hernández, M. González-Díaz, D. a.

- Puente, G. Velasco, J. M. P. Freije, J. M. C. Tubío, R. Royo, J. L. Gelpí, M. Orozco, D. G. Pisano, J. Zamora, M. Vázquez, A. Valencia, H. Himmelbauer, M. Bayés, S. Heath, M. Gut, I. Gut, X. Estivill, A. López-Guillermo, X. S. Puente, E. Campo and C. López-Otín, *Nature Genetics* **44**, 47 (January 2012).
8. S. Bonnal, L. Vigevani and J. Valcárcel, *Nature Reviews Drug Discovery* **11**, 847 (November 2012).
 9. B. Khoo and A. Krainer, *Current Opinion in Molecular Therapeutics* **11**, 108 (2009).
 10. A. Battle, S. Mostafavi, X. Zhu, J. B. Potash, M. M. Weissman, C. McCormick, C. D. Haudenschild, K. B. Beckman, J. Shi, R. Mei, A. E. Urban, S. B. Montgomery, D. F. Levinson and D. Koller, *Genome Research* **24**, 14 (January 2014).
 11. T. Lappalainen, M. Sammeth, M. R. Friedlander, P. A. t Hoen, J. Monlong, M. A. Rivas, M. Gonzalez-Porta, N. Kurbatova, T. Griebel, P. G. Ferreira, M. Barann, T. Wieland, L. Greger, M. van Iterson, J. Almlöf, P. Ribeca, I. Pulyakhina, D. Esser, T. Giger, A. Tikhonov, M. Sultan, G. Bertier, D. G. MacArthur, M. Lek, E. Lizano, H. P. Buermans, I. Padialleau, T. Schwarzmayr, O. Karlberg, H. Ongen, H. Kilpinen, S. Beltran, M. Gut, K. Kahlem, V. Amstislavskiy, O. Stegle, M. Pirinen, S. B. Montgomery, P. Donnelly, M. I. McCarthy, P. Flückeck, T. M. Strom, C. Geuvadis, H. Lehrach, S. Schreiber, R. Sudbrak, A. Carracedo, S. E. Antonarakis, R. Hasler, A. C. Syvanen, G. J. van Ommen, A. Brazma, T. Meitinger, P. Rosenstiel, R. Guigo, I. G. Gut, X. Estivill and E. T. Dermitzakis, *Nature* **501**, 506 (2013).
 12. A. N. Brooks, P. S. Choi, L. de Waal, T. Sharifnia, M. Imielinski, G. Saksena, C. S. Pedamallu, A. Sivachenko, M. Rosenberg, J. Chmielecki, M. S. Lawrence, D. S. DeLuca, G. Getz and M. Meyerson, *PloS One* **9**, p. e87361 (January 2014).
 13. V. Quesada, L. Conde, N. Villamor, G. R. Ordóñez, P. Jares, L. Bassaganyas, A. J. Ramsay, S. Beà, M. Pinyol, A. Martínez-Trillos and others, *Nature genetics* **44**, 47 (2012).
 14. J. Kaiser, *Science* **310**, p. 1751 (2005).
 15. X. Gan, O. Stegle, J. Behr, J. G. Steffen, P. Drewe, K. L. Hildebrand, R. Lyngsoe, S. J. Schultheiss, E. J. Osborne, V. T. Sreedharan, A. Kahles, R. Bohnert, G. Jean, P. Derwent, P. Kersey, E. J. Belfield, N. P. Harberd, E. Kemen, C. Toomajian, P. X. Kover, R. M. Clark, G. Rätsch and R. Mott, *Nature* **108**, 10249 (August 2011).
 16. E. Eden, R. Navon, I. Steinfield, D. Lipson and Z. Yakhini, *BMC Bioinformatics* **10**, p. 48 (2009).
 17. G. Ciriello, M. L. Miller, B. A. Aksoy, Y. Senbabaoğlu, N. Schultz and C. Sander, *Nature Genetics* **45**, 1127 (September 2013).
 18. C. Lippert, F. Casale, B. Rakitsch and O. Stegle, *bioRxiv* , 0 (2014).
 19. Y. D. Bhutia, S. W. Hung, B. Patel, D. Lovin and R. Govindarajan, *Cancer Research* **71**, 1825 (March 2011).
 20. K. A. Spivey, J. Banyard, L. M. Solis, I. I. Wistuba, J. A. Barletta, L. Gandhi, H. A. Feldman, S. J. Rodig, L. R. Chirieac and B. R. Zetter, *Cancer Epidemiology, Biomarkers & Prevention* **19**, 1362 (May 2010).
 21. M. P. Cuajungco, W. Podevin, V. K. Valluri, Q. Bui, V. H. Nguyen and K. Taylor, *Acta Histochemical* **114**, 705 (November 2012).
 22. X. Xing, M. Lai, W. Gartner, E. Xu, Q. Huang, H. Li and G. Chen, *Proteomics* **6**, 2916 (May 2006).
 23. C. Chen, G. Li, W. Liao, J. Wu, L. Liu, D. Ma, J. Zhou, R. H. Elbekai, M. L. Edin, D. C. Zeldin and D. W. Wang, *The Journal of Pharmacology and Experimental Therapeutics* **329**, 908 (June 2009).
 24. V. A. Sciorra, M. A. Sanchez, A. Kunibe and A. E. Wurmser, *PloS One* **7**, p. e40053 (January 2012).
 25. D. K. Sanghera, C. E. Aston, N. Saha and M. I. Kamboh, *American Journal of Human Genetics* **62**, 36 (January 1998).

26. J. Dissing and A. H. Johnsen, *Biochimica et Biophysica Acta* **1121**, 261 (June 1992).
27. M.-C. Tessier, H.-Q. Qu, R. Fréchette, F. Bacot, R. Grabs, S. P. Taback, M. L. Lawson, S. E. Kirsch, T. J. Hudson and C. Polychronakos, *Journal of Medical Genetics* **43**, 129 (February 2006).
28. L. A. Hindorff, P. Sethupathy, H. A. Junkins, E. M. Ramos, J. P. Mehta, F. S. Collins and T. A. Manolio, *Proceedings of the National Academy of Sciences of the United States of America* **106**, 9362 (June 2009).
29. E. Ruark, S. Seal, H. McDonald, F. Zhang, A. Elliot, K. Lau, E. Perdeaux, E. Rapley, R. Eeles, J. Peto, Z. Kote-Jarai, K. Muir, J. Nsengimana, J. Shipley, D. T. Bishop, M. R. Stratton, D. F. Easton, R. A. Huddart, N. Rahman and C. Turnbull, *Nature Genetics* **45**, 686 (June 2013).
30. F. J. Couch, X. Wang, L. McGuffog, A. Lee, C. Olswold, K. B. Kuchenbaecker, P. Soucy, Z. Fredericksen, D. Barrowdale, J. Dennis, M. M. Gaudet, E. Dicks, M. Kosel, S. Healey, O. M. Sinilnikova, A. Lee, F. Bacot, D. Vincent, F. B. L. Hogervorst, S. Peock, D. Stoppa-Lyonnet, A. Jakubowska, kConFab Investigators, P. Radice, R. K. Schmutzler, SWE-BRCA, S. M. Domchek, M. Piedmonte, C. F. Singer, E. Friedman, M. Thomassen, Ontario Cancer Genetics Network, T. V. O. Hansen, S. L. Neuhausen, C. I. Szabo, I. Blanco, M. H. Greene, B. Y. Karlan, J. Garber, C. M. Phelan, J. N. Weitzel, M. Montagna, E. Olah, I. L. Andrulitis, A. K. Godwin, D. Yannoukakos, D. E. Goldgar, T. Caldes, H. Nevanlinna, A. Osorio, M.-B. Terry, M. B. Daly, E. J. van Rensburg, U. Hamann, S. J. Ramus, A. E. Toland, M. A. Caligo, O. I. Olopade, N. Tung, K. Claes, M. S. Beattie, M. C. Southey, E. N. Imyanitov, M. Tischkowitz, R. Janavicius, E. M. John, A. Kwong, O. Diez, J. Balmaña, R. B. Barkardottir, B. K. Arun, G. Rennert, S.-H. Teo, P. A. Ganz, I. Campbell, A. H. van der Hout, C. H. M. van Deurzen, C. Seynaeve, E. B. Gómez García, F. E. van Leeuwen, H. E. J. Meijers-Heijboer, J. J. P. Gille, M. G. E. M. Ausems, M. J. Blok, M. J. L. Ligtenberg, M. A. Rookus, P. Devilee, S. Verhoef, T. A. M. van Os, J. T. Wijnen, HEBON, EMBRACE, D. Frost, S. Ellis, E. Fineberg, R. Platte, D. G. Evans, L. Izatt, R. A. Eeles, J. Adlard, D. M. Eccles, J. Cook, C. Brewer, F. Douglas, S. Hodgson, P. J. Morrison, L. E. Side, A. Donaldson, C. Houghton, M. T. Rogers, H. Dorkins, J. Eason, H. Gregory, E. McCann, A. Murray, A. Calender, A. Hardouin, P. Berthet, C. Delnatte, C. Nogues, C. Lasset, C. Houdayer, D. Leroux, E. Rouleau, F. Prieur, F. Damiola, H. Sobol, I. Coupier, L. Venat-Bouvet, L. Castera, M. Gauthier-Villars, M. Léoné, P. Pujol, S. Mazoyer, Y.-J. Bignon, GEMO Study Collaborators, E. Złowocka-Perłowska, J. Gronwald, J. Lubinski, K. Durda, K. Jaworska, T. Huzarski, A. B. Spurdle, A. Viel, B. Peissel, B. Bonanni, G. Melloni, L. Ottini, L. Papi, L. Varesco, M. G. Tibiletti, P. Peterlongo, S. Volorio, S. Manoukian, V. Pensotti, N. Arnold, C. Engel, H. Deissler, D. Gadzicki, A. Gehrig, K. Kast, K. Rhiem, A. Meindl, D. Niederacher, N. Ditsch, H. Plendl, S. Preisler-Adams, S. Engert, C. Sutter, R. Varon-Mateeva, B. Wappenschmidt, B. H. F. Weber, B. Arver, M. Stenmark-Askmalm, N. Loman, R. Rosenquist, Z. Einbeigi, K. L. Nathanson, T. R. Rebbeck, S. V. Blank, D. E. Cohn, G. C. Rodriguez, L. Small, M. Friedlander, V. L. Bae-Jump, A. Fink-Retter, C. Rappaport, D. Gschwantler-Kaulich, G. Pfeiler, M.-K. Tea, N. M. Lindor, B. Kaufman, S. Shimon Paluch, Y. Laitman, A.-B. Skytte, A.-M. Gerdes, I. S. Pedersen, S. T. Moeller, T. A. Kruse, U. B. Jensen, J. Vijai, K. Sarrel, M. Robson, N. Kauff, A. M. Mulligan, G. Glendon, H. Ozcelik, B. Ejlertsen, F. C. Nielsen, L. Jønson, M. K. Andersen, Y. C. Ding, L. Steele, L. Foretova, A. Teulé, C. Lazaro, J. Brunet, M. A. Pujana, P. L. Mai, J. T. Loud and C. a. Walsh, *PLoS genetics* **9**, p. e1003212 (2013).
31. P. D. P. Pharoah, Y.-Y. Tsai, S. J. Ramus, C. M. Phelan, E. L. Goode, K. Lawrenson, M. Buckley, B. L. Fridley, J. P. Tyrer, H. Shen, R. Weber, R. Karevan, M. C. Larson, H. Song, D. C. Tessier, F. Bacot, D. Vincent, J. M. Cunningham, J. Dennis, E. Dicks, Australian Cancer Study, Australian Ovarian Cancer Study Group, K. K. Aben, H. Anton-Culver, N. Antonenkova, S. M. Armasu, L. Baglietto, E. V. Bandera, M. W. Beckmann, M. J. Birrer, G. Bloom, N. Bogdanova,

- J. D. Brenton, L. A. Brinton, A. Brooks-Wilson, R. Brown, R. Butzow, I. Campbell, M. E. Carney, R. S. Carvalho, J. Chang-Claude, Y. A. Chen, Z. Chen, W.-H. Chow, M. S. Cicek, G. Coetzee, L. S. Cook, D. W. Cramer, C. Cybulski, A. Dansonka-Mieszkowska, E. Despierre, J. A. Doherty, T. Dörk, A. du Bois, M. Dürst, D. Eccles, R. Edwards, A. B. Ekici, P. A. Fasching, D. Fenstermacher, J. Flanagan, Y.-T. Gao, M. Garcia-Closas, A. Gentry-Maharaj, G. Giles, A. Gjyshi, M. Gore, J. Gronwald, Q. Guo, M. K. Halle, P. Harter, A. Hein, F. Heitz, P. Hillemanns, M. Hoatlin, E. Hogdall, C. K. Høgdall, S. Hosono, A. Jakubowska, A. Jensen, K. R. Kalli, B. Y. Karlan, L. E. Kelemen, L. A. Kiemeney, S. K. Kjaer, G. E. Konecny, C. Krakstad, J. Kupryjanczyk, D. Lambrechts, S. Lambrechts, N. D. Le, N. Lee, J. Lee, A. Leminen, B. K. Lim, J. Lissowska, J. Lubinski, L. Lundvall, G. Lurie, L. F. A. G. Massuger, K. Matsuo, V. McGuire, J. R. McLaughlin, U. Menon, F. Modugno, K. B. Moysich, T. Nakanishi, S. A. Narod, R. B. Ness, H. Nevanlinna, S. Nickels, H. Noushmehr, K. Odunsi, S. Olson, I. Orlow, J. Paul, T. Pejovic, L. M. Pelttari, J. Permuth-Wey, M. C. Pike, E. M. Poole, X. Qu, H. A. Risch, L. Rodriguez-Rodriguez, M. A. Rossing, A. Rudolph, I. Runnebaum, I. K. Rzepecka, H. B. Salvesen, I. Schwaab, G. Severi, H. Shen, V. Shridhar, X.-O. Shu, W. Sieh, M. C. Southe, P. Spellman, K. Tajima, S.-H. Teo, K. L. Terry, P. J. Thompson, A. Timorek, S. S. Tworoger, A. M. van Altena, D. van den Berg, I. Vergote, R. A. Vierkant, A. F. Vitonis, S. Wang-Gohrke, N. Wentzensen, A. S. Whittemore, E. Wik, B. Winterhoff, Y. L. Woo, A. H. Wu, H. P. Yang, W. Zheng, A. Ziogas, F. Zulkifli, M. T. Goodman, P. Hall, D. F. Easton, C. L. Pearce, A. Berchuck, G. Chenevix-Trench, E. Iversen, A. N. A. Monteiro, S. A. Gayther, J. M. Schildkraut and T. A. Sellers, *Nature genetics* **45**, 362 (April 2013).
32. A. C. Antoniou, X. Wang, Z. S. Fredericksen, L. McGuffog, R. Tarrell, O. M. Sinilnikova, S. Healey, J. Morrison, C. Kartsonaki, T. Lesnick, M. Ghoussaini, D. Barrowdale, EMBRACE, S. Peock, M. Cook, C. Oliver, D. Frost, D. Eccles, D. G. Evans, R. Eeles, L. Izatt, C. Chu, F. Douglas, J. Paterson, D. Stoppa-Lyonnet, C. Houdayer, S. Mazoyer, S. Giraud, C. Lasset, A. Remenieras, O. Caron, A. Hardouin, P. Berthet, GEMO Study Collaborators, F. B. L. Hogervorst, M. A. Rookus, A. Jager, A. van den Ouwehand, N. Hoogerbrugge, R. B. van der Luijt, H. Meijers-Heijboer, E. B. Gómez García, HEBON, P. Devilee, M. P. G. Vreeswijk, J. Lubinski, A. Jakubowska, J. Gronwald, T. Huzarski, T. Byrski, B. Górska, C. Cybulski, A. B. Spurdle, H. Holland, kConFab, D. E. Goldgar, E. M. John, J. L. Hopper, M. Southe, S. S. Buys, M. B. Daly, M.-B. Terry, R. K. Schmutzler, B. Wappenschmidt, C. Engel, A. Meindl, S. Preisler-Adams, N. Arnold, D. Niederacher, C. Sutter, S. M. Domchek, K. L. Nathanson, T. Rebbeck, J. L. Blum, M. Piedmonte, G. C. Rodriguez, K. Wakeley, J. F. Boggess, J. Basil, S. V. Blank, E. Friedman, B. Kaufman, Y. Laitman, R. Milgrom, I. L. Andrulis, G. Glendon, H. Ozcelik, T. Kirchhoff, J. Vija, M. M. Gaudet, D. Altshuler, C. Guiducci, SWE-BRCA, N. Loman, K. Harbst, J. Rantala, H. Ehrencrona, A.-M. Gerdes, M. Thomassen, L. Sunde, P. Peterlongo, S. Manoukian, B. Bonanni, A. Viel, P. Radice, T. Caldes, M. de la Hoya, C. F. Singer, A. Fink-Retter, M. H. Greene, P. L. Mai, J. T. Loud, L. Guidugli, N. M. Lindor, T. V. O. Hansen, F. C. Nielsen, I. Blanco, C. Lazaro, J. Garber, S. J. Ramus, S. A. Gayther, C. Phelan, S. Narod, C. I. Szabo, MOD SQUAD, J. Benitez, A. Osorio, H. Nevanlinna, T. Heikkinen, M. A. Caligo, M. S. Beattie, U. Hamann, A. K. Godwin, M. Montagna, C. Casella, S. L. Neuhausen, B. Y. Karlan, N. Tung, A. E. Toland, J. Weitzel, O. Olopade, J. Simard, P. Soucy, W. S. Rubinstein, A. Arason, G. Renner, N. G. Martin, G. W. Montgomery, J. Chang-Claude, D. Flesch-Janys, H. Brauch, GENICA, G. Severi, L. Baglietto, A. Cox, S. S. Cross, P. Miron, S. M. Gerty, W. Tapper, D. Yannoukakos, G. Fountzilas, P. A. Fasching, M. W. Beckmann, I. Dos Santos Silva, J. Peto, D. Lambrechts, R. Paridaens, T. Rüdiger, A. Försti, R. Winqvist, K. Pylkäs, R. B. Diasio, A. M. Lee, J. Eckel-Passow, C. Vachon, F. Blows, K. Driver, A. Dunning, P. P. D. Pharoah, K. Offit, V. S. Pankratz, H. Hakonarson, G. Chenevix-Trench, D. F. Easton and F. J. Couch, *Nature genetics* **42**, 885 (October 2010).

33. K. L. Bolton, J. Tyrer, H. Song, S. J. Ramus, M. Notaridou, C. Jones, T. Sher, A. Gentry-Maharaj, E. Wozniak, Y.-Y. Tsai, J. Weidhaas, D. Paik, D. J. Van Den Berg, D. O. Stram, C. L. Pearce, A. H. Wu, W. Brewster, H. Anton-Culver, A. Ziogas, S. A. Narod, D. A. Levine, S. B. Kaye, R. Brown, J. Paul, J. Flanagan, W. Sieh, V. McGuire, A. S. Whittemore, I. Campbell, M. E. Gore, J. Lissowska, H. P. Yang, K. Medrek, J. Gronwald, J. Lubinski, A. Jakubowska, N. D. Le, L. S. Cook, L. E. Kelemen, A. Brook-Wilson, L. F. A. G. Massuger, L. A. Kiemeney, K. K. H. Aben, A. M. van Altena, R. Houlston, I. Tomlinson, R. T. Palmieri, P. G. Moorman, J. Schildkraut, E. S. Iversen, C. Phelan, R. A. Vierkant, J. M. Cunningham, E. L. Goode, B. L. Fridley, S. Kruger-Kjaer, J. Blaeker, E. Hogdall, C. Hogdall, J. Gross, B. Y. Karlan, R. B. Ness, R. P. Edwards, K. Odunsi, K. B. Moyisch, J. A. Baker, F. Modugno, T. Heikkinnen, R. Butzow, H. Nevanlinna, A. Leminen, N. Bogdanova, N. Antonenkova, T. Doerk, P. Hillemanns, M. Dürst, I. Runnebaum, P. J. Thompson, M. E. Carney, M. T. Goodman, G. Lurie, S. Wang-Gohrke, R. Hein, J. Chang-Claude, M. A. Rossing, K. L. Cushing-Haugen, J. Doherty, C. Chen, T. Rafnar, S. Besenbacher, P. Sulem, K. Stefansson, M. J. Birrer, K. L. Terry, D. Hernandez, D. W. Cramer, I. Vergote, F. Amant, D. Lambrechts, E. Despierre, P. A. Fasching, M. W. Beckmann, F. C. Thiel, A. B. Ekici, X. Chen, Australian Ovarian Cancer Study Group, Australian Cancer Study (Ovarian Cancer), Ovarian Cancer Association Consortium, S. E. Johnatty, P. M. Webb, J. Beesley, S. Chanock, M. Garcia-Closas, T. Sellers, D. F. Easton, A. Berchuck, G. Chenevix-Trench, P. D. P. Pharoah and S. A. Gayther, *Nature genetics* **42**, 880 (October 2010).
34. M. Garcia-Closas, F. J. Couch, S. Lindstrom, K. Michailidou and M. K. e. a. Schmidt, *Nature genetics* **45**, 392 (April 2013).

AN INTEGRATED FRAMEWORK FOR REPORTING CLINICALLY RELEVANT BIOMARKERS FROM PAIRED TUMOR/NORMAL GENOMIC AND TRANSCRIPTOMIC SEQUENCING DATA IN SUPPORT OF CLINICAL TRIALS IN PERSONALIZED MEDICINE

SARA NASSER¹, AHMET A. KURDOGLU¹, TYLER IZATT¹, JESSICA ALDRICH¹, MEGAN L. RUSSELL¹, ALEXIS CHRISTOFORIDES¹, WIABHAV TEMBE¹, JEFFERY A. KIEFER², JASON J. CORNEVEAUX¹, SARA A. BYRON², KAREN M. FORMAN³, CLARICE ZUCCARO³, JONATHAN J. KEATS¹, PATRICIA M. LORUSSO⁴, JOHN D. CARPTEN², JEFFREY M. TRENT² AND DAVID W.

CRAIG^{1*}

¹*Neurogenomics Division, Translational Genomics Research Institute, Phoenix, AZ 85004, USA*

²*Integrated Cancer Division, Translational Genomics Research Institute, Phoenix, AZ 85004, USA*

⁴*Yale Cancer Center, Yale School of Medicine, 333 Cedar Street, New Haven, CT 06510*

³*Barbara Ann Karmanos Cancer Institute, Wayne State University, Detroit, MI*

*E-mail: dcraig@tgen.org

The ability to rapidly sequence the tumor and germline DNA of an individual holds the eventual promise of revolutionizing our ability to match targeted therapies to tumors harboring the associated genetic biomarkers. Analyzing high throughput genomic data consisting of millions of base pairs and discovering alterations in clinically actionable genes in a structured and real time manner is at the crux of personalized testing. This requires a computational architecture that can monitor and track a system within a regulated environment as terabytes of data are reduced to a small number of therapeutically relevant variants, delivered as a diagnostic laboratory developed test. These high complexity assays require data structures that enable real-time and retrospective ad-hoc analysis, with a capability of updating to keep up with the rapidly changing genomic and therapeutic options, all under a regulated environment that is relevant under both CMS and FDA depending on application. We describe a flexible computational framework that uses a paired tumor/normal sample allowing for complete analysis and reporting in approximately 24 hours, providing identification of single nucleotide changes, small insertions and deletions, chromosomal rearrangements, gene fusions and gene expression with positive predictive values over 90%. In this paper we present the challenges in integrating clinical, genomic and annotation databases to provide interpreted draft reports which we utilize within ongoing clinical research protocols. We demonstrate the need to retire from existing performance measurements of accuracy and specificity and measure metrics that are meaningful to a genomic diagnostic environment. This paper presents a three-tier infrastructure that is currently being used to analyze an individual genome and provide available therapeutic options via a clinical report. Our framework utilizes a non-relational variant-centric database that is scaleable to a large amount of data and addresses the challenges and limitations of a relational database system. Our system is continuously monitored via multiple trackers each catering differently to the diversity of users involved in this process. These trackers designed in analytics web-app framework provide status updates for an individual sample accurate to a few minutes. In this paper, we also present our outcome delivery process that is designed and delivered adhering to the standards defined by various regulation agencies involved in clinical genomic testing.

Keywords: Genomic Testing; Next-Gen Sequencing Analysis; Personalized Medicine

1. Introduction

Cancer onset and progression leads to a variety of genomic events such as chromosomal aberrations and genomic mutations. Cancer develops through a stochastic process that produces cellular heterogeneity and structural complexity of the cellular genome. New advances in the depth and dimensionality of tumor profiling through the use of Next Generation Sequencing (NGS) have emphasized that there are in fact an under-appreciated number of acquired genetic changes and aberrations in cancer genomes that show clonal evidence of being selected. While the function of many of the mutations within a cancer are unknown and may represent passenger events, a subset of these are possibly biologically relevant and therapeutically actionable. Coupled with this knowledge and the ability to sequence a patient's genome within a clinically relevant timeframe, there is an increasing desire to utilize this vast amount data for patient care at an individual level.

National Cancer Institute lists hundred's of FDA approved targeted therapies which is further supplemented as an even larger number of therapies still under clinical investigation. Most notably, the BRAF inhibitors approved by FDA have demonstrated clinical response in patients with BRAF mutations.¹ More recently within our and collaborating research protocols, therapies targeting genomic events like FGFR fusions in Cholangiocarcinoma² and EGFR mutations in Lung cancer³ have shown promising results and in some cases suggested that these targeted therapies instead of chemotherapy may be the best choice of treatment.

Studies to date are largely limited in scope, suffer from small numbers of patients, or therapeutic options, or lack the randomization or prospective design necessary to provide an unbiased assessment. One could critique that inability to access therapeutic drugs recommended by a group of experts (often termed a tumor board), or may indicate investigational agents only available through other trials. Determining the effectiveness of using genomics to inform therapy selection is the subject of numerous research studies, including several studies at the Translational Genomics Research Institute that have driven the development of data analytics pipeline achieving both technical, regulatory, and clinical goals. The pipelines we describe largely originate from two multi-year studies. The first is a study of the feasibility of using molecular-guided therapy for patients with BRAF wild-type metastatic melanoma (BRAFwt MM) as part of the Stand Up To Cancer Melanoma Dream Team. Detailed in other publications in preparation, the inclusion of investigational agents within the pharmacopeia (or compendium of drugs that could be indicated within a report), our pipeline was developed both to satisfy requirements within a laboratory developed test (LDT), regulated under the Center for Medicaid Services(CMS), as well as the Food and Drug Administration (FDA).⁴ The second is a study of glioblastoma funded under the Ivy Foundation facilitating the development of the engine of rules identifying therapeutic options within this study.

As these examples show, the framework and mindset in code development for flexible platforms is regulated under multiple agencies. CMS regulates all laboratory testing (except research) performed on humans in the U.S. through the Clinical Laboratory Improvement Amendments (CLIA).⁵ Particularly in the case of clinical research trials that impact care of patients, the FDA also provides regulatory oversite. Providing a framework that provides analytical validity for both FDA and CMS, requires understanding that precision, specificity,

sensitivity, reproducibility, repeatability be characterized with datasets that are often generated under a variety of conditions and truly characterizes the limits of detection. Understanding the regulatory environment is changing two involve multiple agencies, requires additional diligence with pipeline reporting, such critically as the ability to provide negative calls. For example traditional uses of the VCF format do not provide negative calls, and the ability to understand both false positives and false negatives is an increasing requirement for a field where the VCF specification could fall short. For example, there should not be ambiguity about where the lack of a variant is a 'no call' or a 'negative for BRAF 600V/E' below 0.01

The challenges of conducting trials and research in personalized genomics is elevated by the fact that next-generation sequencing data analysis requires intensive computational processing.⁶ Effectiveness requires proper versioning, rapid turnaround, and reasonable disk requirements. A major aspect is whether the platform allows for implementing improvements and proper versioning within a clinical validated setting. Thus a CLIA certified lab needs to be equipped with high performance computers with additional security layers that can expedite genomic analysis. Additionally, interpretation of the findings requires accessing several genomic databases for annotations, pharmaceutical databases to gather gene-to-drug relations and tie in these databases to provide a clear and concise report.

Annotations and drug-gene matching process needs to be continually updated to include newly discovered variants and drug-gene matches. Further, tumor specific transcript variants exist and may require additional methods for detections and reporting. For instance, the EGFRvIII variant that has been reported in 30-40% of highly aggressive glioblastomas⁷ is not available in most annotations and thus cannot be detected via usual methods.

In the following sections we provide an overview of the TGen Personal Genomics System and describe the framework that is being used for genomic testing and data delivery.

2. System Layout for Facilitating Clinical Research Trails in Personalized Genomics

The system developed to support a wide variety of clinical research trails for biopsy to report monitoring, tracking, and clinical reporting is a dynamic and flexible analysis framework which integrates genome and transcriptome sequencing data to inform interpretation of next-generation sequencing data on samples, specifically developed for cancer genomics. Our current platform is being utilized within a CLIA lab and has also been utilized under FDA within larger protocols. An overview is provided in Figure 1, which builds on stages supported by layers.

2.1. Overview of Stages Going From Consent to Report

The stages all begin with the patient and the doctor, who enroll and consent patients and provide a biopsy and/or blood draw to our clinical laboratory. This first stage is supported by a patient-centric custom-built clinical data portal. These portals and database described later in section 2.3 collect encoded, predefined non-public health information (PHI) for each patient. The second stage is specimen processing and accessioning whereby the specimens are evaluated for suitability and DNA/RNA analyte is isolated. Each step is managed by a

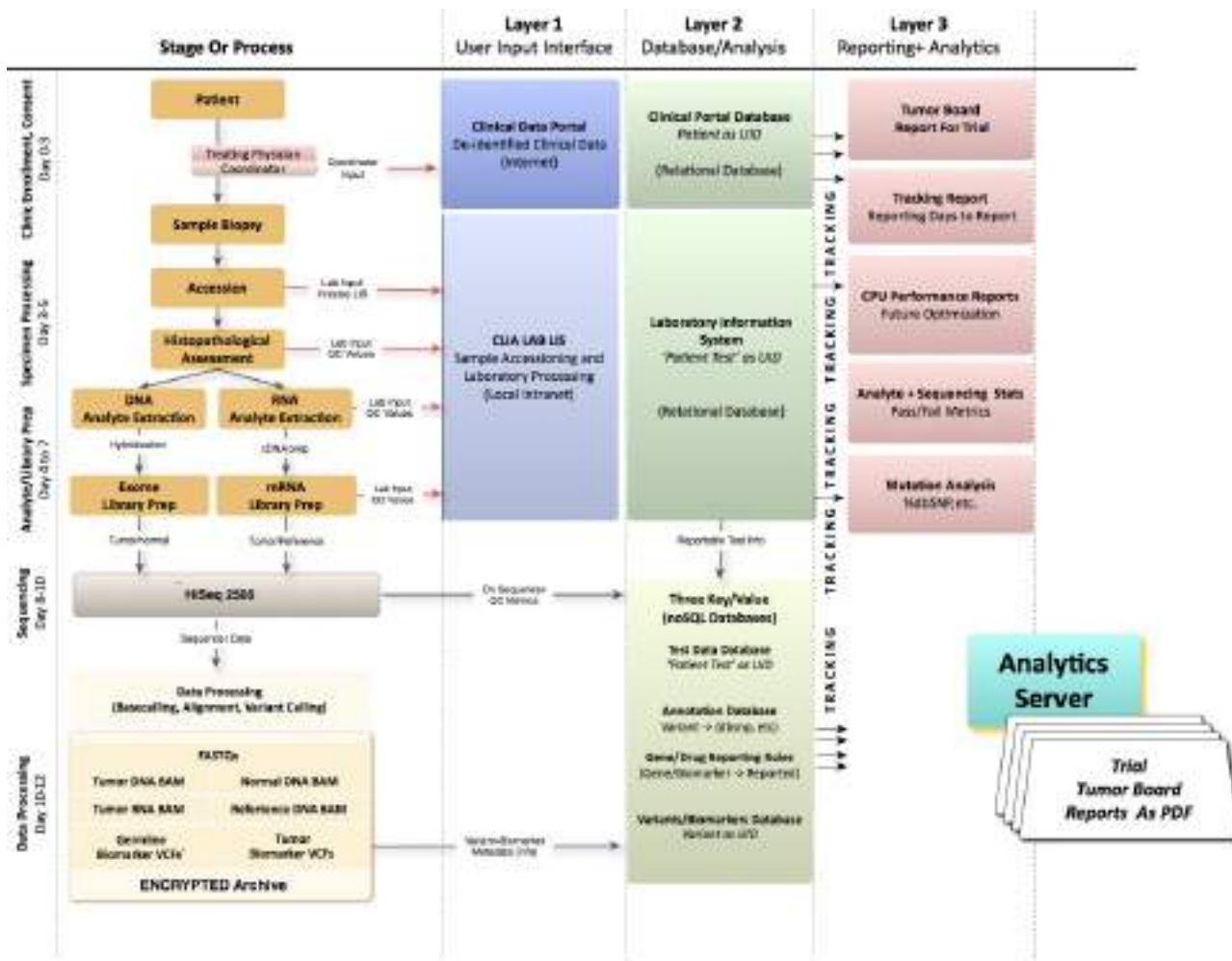


Fig. 1. Data Flow depicting the three layers of our system showcase the flow of samples from the clinic to sequencing and analysis. Finally, an interpretive report is sent back to the clinic

FileMaker Pro¹² relational database. The third stage is library preparation and sequencing, again supported by a FileMaker Pro database which collects both images and data. As the Illumina 2500 sequencer writes to a shared disk system, data is collected into the relational database. The fourth stage is analysis and reporting, which is a framework built around (1) standardized file formats, (2) multiple quality control checks, (3) automated processing, (4) scheduled releases of sequence data, sequencing alignments, and variant calls, and (5) centralized primary data processing. As variants are generated, they are placed into a NoSQL database utilizing a document primary key as a 'biomarker' .

Our framework is compatible with most external programs by plugging into allowing us to use the best tools and never be obligated to one tool, or feel required to develop an entire pipeline from scratch when components may be added according to their license agreements. Currently, we use external open-source style tools for mapping whereas the rest of the pipeline is largely built on internally developed software. As other groups develop muta-

tion detection tools, our framework will be compatible because of design. Our personalized medicine framework is a modular-based standard driven framework built to allow flexibility of adding/swapping out component analysis programs rapidly. It functions currently on Dell blade systems or appliances though are portable to other Unix-based environments. Fundamentally, many aspects are coded with map-reduce in mind allowing eventual porting to other frameworks. Component programs do not have to be proprietary, and thus other tools may be added if shown to be effective at identifying variants. It is optimized for oncology or cancer, though expandable to other areas such as neurological diseases. Current implementation is within a Torque-style queuing system common to many computing environments whereby jobs are monitored using background processes such that samples flow in an automated fashion to generation of reports.

2.2. User Input Layers

User interfaces are developed with the understanding they support trials at early stages which require prototyping by lab staff and clinical teams in an iterative manner before fixing for validation in use in a trial. User interaction happens at two levels: i) Clinical Data Portal and ii) Laboratory Data.

- **Clinical Portal:** The User Layer provides user interfaces to interact with the databases, extract reports and track information. To provide the flexibility for the research trials, pragmatic open source solutions are used, recognizing they are not necessarily appropriate for production clinical environments that are in open networks. Pragmatically, portals for clinical data are designed in WordPressTM, a PHP management system tied in with an additional layer of security. WordPress utilizes extensive plugin framework that allows for easy addition and removal of features. The Clinical Data Portal (fields as shown in Table 1) are paired with their genomic counterpart for tumor board presentation.
- **Laboratory Data Frontend:** Laboratory and Sequencing data is entered in a user friendly FileMaker Pro database⁸ described in section 2.3. FileMaker Pro's flexible and user friendly framework provides ease to handle the massive amount of data that is generated during sequencing.

2.3. Database/Analysis Layer

NGS technologies provide a high-resolution and high- throughput approach to identify individual nucleotide bases from DNA samples. The goal of the NGS bioinformatics pipeline is to identify germline and somatic genetic variants events from tumor/normal pairs at the genomic (DNA) level, including coding point mutations and small insertions/deletions, copy number changes, and structural events (intra-chromosomal rearrangements and translocations).

An overview of the Analysis Workflow is provided in Figure 1. Briefly, each flowcell contains up to 4 tumor/normal pairs with an obligate reference control barcoded according to Illumina specifications (the control is described in detail in the Protocol section). Data is written from the HiSeq2500 to the scratch portion of a server in the form of BCL folders within the

Table 1. Clinical Data Portal: An example list of fields collected in the clinical data portal. Columns indicate sections in the portal and rows correspond to the fields within each section.*This is collected for all professions of the disease.

PatientSummary	PrimaryDx	ProgressiveDisease*	CurrentPresentation
PatientIdentifier	SpecimenSize	SiteOfRecurrence	ComorbidConditions
Date of Consent	SatelliteNodules	DateofRecurrance	MenopausalStatus
PatientAge	StageT ,N,M	Surgery	PreviousCancers
PatientRace	Ulceration	SurgeryResponse	PriorTreatments
PatientEthnicity	Mitoses	SurgeryDate	DrugAllergies
PatientGender	ClarksLevel	SurgeryType	Medications
PatientSummary	MutationBRAF/NRAS/CKIT	RadiationLocation	Physical Exam
	LymphNodeInvolvement	RadiationType	Imaging/Radiology
	IFNTType/Cycle	RadiationResponse	WBC/ANC/AlkPhos
	ClinicalTrialVaccine	TreatmentofRecurrance	Proteinuria

IlluminaRunFolders directory. An analysis run is triggered by the Clinical Laboratory Information System (DCLIS); depositing files within the ConversionArea folder that is processed into MergeSheets and SampleSheets. Using a queuing system and write FAIL/COMPLETED system BCL folders/files are converted to FASTQ files (raw sequence) and aligned to the genome using BWA-MEM⁹ followed by a standard best-practice cascade of variant calling software tools.

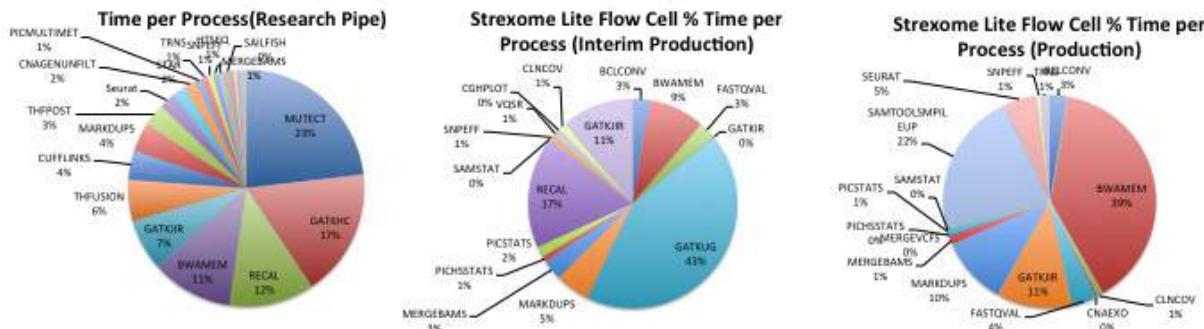


Fig. 2. Percentage of CPU Times utilization listed for three different pipelines highlight the importance of optimization. In the research pipeline, "Mutect" which was not optimized took the largest chunk of CPU hours. The middle chart display the CPU time distribution for the Interim Pipeline: germline variant caller GATKUG caller was taking a large chunk of CPU hours, replacing it with Samtools significantly improved turn-around time.

2.3.1. Databases

At the backbone of our infrastructure sits a database layer, which consists of annotation databases, drug rule matching databases and clinical databases. Although, each database is

built in a different environment to cater to its users, we aggregate the relative information so it can be used for reporting and tracking.

- **Sequencing Database** Sequencing data is built in a user friendly FileMaker Pro database.⁸ This relational database captures relations between "Patients", "Samples", "Orders" and "Sequencing Statistics". FileMakerPro's audit trail allows realtime tracking for document changes which are critical for a CLIA lab.
- **Annotations** Annotation is handled by committing sources of annotation into the database (typically by first exporting a text copy of the public database), and then performing an annotate action on the variant collection. The annotate action appends additional fields to all variants that are in annotated regions. We chose to handle annotation this way for two reasons: It allows for the natural retention of a "snapshot" of the annotation version that was used for the analysis. Many public variation databases are frequently modified, and some do not follow a strict versioning scheme that uniquely identifies a copy of the data. This can help with maintaining the repeatability of an analysis. Databases of many forms can be represented easily. Inserting an annotation as a field on a variant record allows us to create queries that use it. Solving the data representation and access problem of modern sequencing helps sustains the progress of genomics by a) using expert and analysis time more efficiently and b) by allowing even small labs to perform complicated knowledge extraction from the abundance of genotypic information that is available.
- **Biomarker Database** This database will integrate the disparate gene state annotation data and compile all available genomic information to facilitate the efficient and effective access for knowledge mining of the various dimensions of gene states. We will also include genes that become evident only from integrated analysis of genome and transcriptome analysis, such as a gene mapping to a large hemizygous deletion region and contains an obviously inactivating frameshift or nonsense mutation in the retained allele. In cases, where RNA data is present we integrate RNA Allele specific expression for the variants.
- **Statistics Database** This is a project centric noSQL database that holds a sample's sequencing information and collects all statistics that are used in determining the quality of the run. Statistics such as total bases aligned, mean target coverage help decide the accuracy of the alignment. In addition, performance metrics are also collected for each run, thus allowing analytics report generation (discussed in section /citerep).
- **Drug Rule Matching** We developed a conceptual framework for annotating the relationships between genomic alterations and drug response. This drug-rule matching algorithm identifies drug candidates for individual genomic alterations, including somatic mutations, indels, gene fusions, DNA copy number changes, and RNA expression changes, based on literature-curated evidence within a structured framework. The primary annotation source is PubMed publications, with other information sources captured when appropriate. New drug-gene associations from the literature will be added to the drug rule matching database through versioning.

2.3.2. Genetic Variant Calling

Our framework uses tumor and constitutional sample for somatic variants. Several tools have been developed to identify somatic events including Mutect,¹⁰ Strelka¹¹ to identify somatic mutations, GATK,¹² samtools¹³ callers to identify germline variants. Figure 2.3 compares our research pipeline, the interim production pipeline and production pipeline. Some software that become bottlenecks were identified in the research and interim-production settings and were optimized and/or replaced for realtime production.

2.4. Reporting and Analytics Layer

This layer allow end-users to interact with the framework at any given time point and generate interim reports. This layer is also responsible of sample tracking, delivery and maintenance.

- **Test Tracking:** A Clinical Genomic test consists of multiple processes that start with receiving the sample, followed by sample isolation, library preparation, sequencing, analysis and report generation. Users/Clients often are interested in tracking the status of their samples to estimate progress or get a priori information. We provide multiple trackers which are designed in JasperSoftTM, one such tracker depicted in Figure 3 can only be visualized by authorized personnel authenticated using JasperSoft's authentication. *JasperReports Server* uses the Spring sub-project, Acegi Security, for authentication and authorization.¹⁴

Patient Status	Initial Date	Test assigned	Sample Received	Sample Received Started	Sample Received Complete	Library Prep Started	Library Prep Complete	Report Generated (Household Client)	Analysis Started	Analysis Completed
Sarcoma - Tumor-normal CBT17_8802_20140810_T2_A18TS	2014-08-10	✓100	Received	AcqBio-Tumor	-	-	-	-	-	-
				AcqBio-Tumor	AcqBio-Normal	✓	AcqBio-Tumor	✓	✓	-
				AcqBio-Normal	AcqBio-Normal	✓	AcqBio-Normal	✓	✓	-
LUSC-FED-SUCCINA - FWD/Rev+ & PNCD/Rev- CBT17_8801_20140810_T3_TSMBU	2014-08-10	✓14	Received	Karyo-Normal	AcqBio-Normal	✓	Karyo-Normal	✓	✓	-
				AcqBio-Tumor	AcqBio-Tumor	✓	AcqBio-Tumor	✓	✓	-

Fig. 3. An Internal Tracker provides information from sample receipt to final report generation. Each step provides an internal ID and the number of days utilized. Total Active Time is color coded providing quick information on the number of days since the sample receipt.

- **Archiving:** Every sample analyzed in the CLIA lab is encrypted and archived. Encryption is a slow process, thus archiving is run as a maintenance task. We use a two-level *gpg* encryption using asymmetric keys, the top level encrypts the entire package and a second layer of encryption is provided for documents that may contain patient-specific data.

3. Validation

FDA and CMS require extensive testing for repeatability and reproducibility. Thus each flow-cell contains tumor/normal pairs with a barcoded reference control COLO829.¹⁵ This reference control is used to validate a flowcell by generating performance measures for a run. Reporting

metrics is a challenge as we need to carefully consider the fact that traditional performance metrics might not apply in a marker-positive framework.¹⁶ In Table 2, pre-production performance metrics are reported on the full range. Genomic tests conducted on approximately 3 billion base pairs of the human genome return only a small number of variants. Thus for these rare-events, the number of true negatives in the test will always be much larger than the true positives or false negatives. Taking into consideration only the reportable range addresses the issues in production. Table 2 also bring into light that Accuracy and Sensitivity (which are traditionally reported) might not be the best indicators of performance of a test in a genomic framework. This is illustrated by two examples where specificity is > 99% and sensitivity is at 50% in cases when the false positives are greater than true positives, whereas Positive Predictive Values(PPV) reports more reliable numbers.

Table 2. Performance Metrics on preproduction COLO829, hypothetical examples and production COLO829. Hypothetical examples indicate that sensitivity and specificity are not the best indicators of performance.

Sample	<i>FN</i>	<i>FP</i>	<i>TN</i>	<i>TP</i>	<i>FDR</i>	<i>Sensitivity</i>	<i>Specificity</i>	<i>PPV</i>
PreCOLO829_1	14	119	53535119	116	50.64%	89.23%	99.99%	49.36%
PreCOLO829_2	27	101	53535129	101	50.00%	78.91%	100.00%	50.00%
PreCOLO829_3	2	17	12716974	14	54.84%	87.50%	100.00%	45.16%
PreCOLO829_4	2	18	12716973	14	56.25%	87.50%	100.00%	43.75%
PreCOLO829_5	2	19	12716972	14	57.58%	87.50%	100.00%	42.42%
PreCOLO829_6	2	16	12716975	14	53.33%	87.50%	100.00%	46.67%
Hypothetical_1	15	17	12716975	1	94.44%	6.25	99.99	5.56
Hypothetical_2	1	20	12716975	1	95.24%	50.00	99.99	54.31
ProdCOLO829_1	23	25	1000000	258	8.83%	91.81%	99.97%	91.17%
ProdCOLO829_2	33	33	1000000	248	11.74%	88.26%	99.98%	88.26%
ProdCOLO829_3	26	11	1000000	255	4.14%	90.75%	99.97%	95.86%
ProdCOLO829_4	28	16	1000000	253	5.95%	90.04%	99.97%	94.05%
ProdCOLO829_5	26	20	1000000	255	7.27%	90.75%	99.97%	92.73%
ProdCOLO829_6	25	22	1000000	256	7.91%	91.10%	99.97%	92.09%

3.1. Targeted Variant Detection

Personalized Genome Testing is a growing field with emergent need for tools that are more focussed on actionable events. In this section, we illustrate with an example the need for development of such new tools. Analyses indicated that 92 – 94% of human genes undergo alternative splicing, 86% with a minor isoform frequency of 15% or more.¹⁷ EGFRvIII is a functional and permanently activated mutation of the epidermal growth factor receptor EGFR, a protein that contributes to cell growth and has been well validated as a target for cancer therapy.^{18,19} Existing tools to detect isoforms presence using differential expression of a test and control sample,²⁰ these tools are quite promising but require much difficult to acquire RNA-Seq control. Additionally, Cuflinks²¹ quantifies the transcript by read alignment and

Sailfish²² provide expression levels for an isoform using an alignment-free approach. Table 3) contains FPKM/RPKM values reported by Cufflinks and Sailfish. Intuitively, a high R/FPKM value for this variant would imply its presence. In Table 3, GBM13 an experimentally validated sample reported an FPKM 165.354, whereas GBM 7 with no evidence of EGFRvIII reported an FPKM value of 820.109. Since these tools rely on the presence of reads in individual exons, they lack to provide evidence of a contiguous segment that defines this particular isoform. Thus R/FPKM values indicate the presence of reads across the whole region, as this region is a subset of the wild-type EGFR, the R/FPKM values might be misleading. Following the objective to detect the presence of variants (such as fusions, isoforms) we use a targeted de Novo assembly approach focusing only on certain region of the genome. In a clinical setting where only certain genes have actionable drugs, an approach to assemble and detect clinically actionable variants seems suitable. We have used this approach in past to correctly detect FGFR fusion.² A Denovo assembler Trinity²³ is used to assemble a region (a region is acquired by initial alignment), which are subsequently aligned to the reference genome for validation check using BWA-MEM.⁹ Guided Assembly for this variant provides a clear response (Yes/No) which is required in clinical testing.

Table 3. EGFRvIII detection using Cufflinks, Sailfish and Guided assembly approach. GBM 6,7,12 all have a high FPKM value but upon examination evidence of EGFRvIII variant was not found for these samples. *EGFRvIII was experimentally validated for this sample.

Samples	Cufflinks (FPKM)		Sailfish (RPKM)		Guided Assembly
	EGFRvIII variant	Wild-type	EGFRvIII	Wild-type	EGFRvIII
GBM3	148.281	6.44284	71.1637	3.38737	Yes
GBM4	171.925	152.204	101.339	77.7003	No
GBM5	485.282	284.416	120.435	232.007	No
GBM6	550.976	432.228	238.849	175.429	Yes
GBM7	820.109	9.90112	382.24	6.24513	No
GBM8	150.059	43.8713	83.3531	21.5956	No
GBM10	266.228	31.7377	136.81	15.5846	Yes
GBM11	18.3666	0.321978	3.91242	0	No
GBM12	653.088	0.000121561	136.032	20.8175	No
GBM13*	165.354	0.000114743	12.5984	0.210588	Yes

4. Discussion

We have presented a three layer system, whereby inputs at layer 1 are through established prototyping solutions are supported by a second layer of automated interpretive engine utilizing intense knowledge mining genes that are known to be directly altered cancer, or who play critical roles in molecular mechanisms that are the targets of pharmaceutical agents. This organization allows for rapid prototyping, implementing, and analytically validating data analysis pipelines in support of open protocols and clinical research trials of personalized medicine built around a three layer design supporting data collection, analysis, and report delivery from

consent to reporting. The first layer leveraging of prototyping environments whereby experiment and clinical collaborators can design interfaces, fitting in with a structured relational and non-relational databases provide a capability of data collection and tracking to create a development cycle that is both agile and rationale.

In a second layer, we describe a framework where both relational and non-relational databases are used to collect all information linking to a patient as it moves through various stages to the identification of biomarkers. Use of non-relational key-value document stores through non-relational databases provides design flexibility of the first layer. Overall, utilization of both relation and non-relational document store databases is based on integration around two these two key concepts 'patients' and 'biomarkers', for which all other concepts depend. Moving across multiple studies with this 'variant' or 'patient' centric conceptually works with most bioinformatics pipelines that are focused on identifying tumor specific (somatic) mutations or biomarkers that compare or germline variants utilizing a variety of tools. A key is that two samples such as tumor and normal tracking to a single parent object. Overall, the mindset that our processes identify biomarkers' from 'patients' in a multi-step linear workflow, define joining as always building around 'patients' and 'biomarkers'. Utilization standardized reporting frameworks such as Jaspersoft at the third level provide environments that provide consistent and flexible reporting consistent with industry standards, fitting over the second database layer. Reports may be of the type that represent 'tumor board reports' or 'sensitivity/specificity' reports for supporting regulatory agencies. Reports may be tracking of where a specimen is in a system, or may be CPU utilization at a particular point in time.

Our assay captures coding mutations and structural events within cancer genes. The final output, an interpretive Cancer Panel Report, provides the physician with a list of agents that are associated with tumor specific DNA mutations. Importantly, mutations can have positive or negative correlations to drugs and our system highlights both. This unbiased report includes all relevant patient related information along with both basic and detailed information related to the tumor's mutational spectrum, and candidate relationships with known therapies.

The framework and data structures we use as part of trials in personalized medicine are conceptually fitting into either three layers supporting a multi-step linear process moving from patient to biomarker sets the mechanism for how data is integrated and analyzed in support of patient of care. The modular-based standard driven framework allows flexibility of adding/swapping out component analysis programs rapidly, thus is not constraint by the tools used. Staying with goals of Bench to Bedside, our future direction is in developing and improving tools that focus towards clinical applications and integrating state-of-the-art software within the system.

5. Acknowledgements

This research was supported by Stand Up To Cancer: Melanoma Research Alliance Melanoma Dream Team Translational Cancer Research Grant (SU2C-AACR-DT0612). Stand Up To Cancer is a program of the Entertainment Industry Foundation administered by the American Association for Cancer Research. Research reported in this publication was also supported by generous philanthropic contributions from the Dorrance Family Foundation and Ben and

Catherine Ivy Foundation.

References

1. P. B. Chapman, A. Hauschild, C. Robert, J. B. Haanen, P. Ascierto, J. Larkin, R. Dummer, C. Garbe, A. Testori, M. Maio, D. Hogg, P. Lorigan, C. Lebbe, T. Jouary, D. Schadendorf, A. Ribas, S. J. O'Day, J. A. Sosman, J. M. Kirkwood, A. M. M. Eggermont, B. Dreno, K. Nolop, J. Li, B. Nelson, J. Hou, R. J. Lee, K. T. Flaherty, G. A. McArthur and BRIM-3 Study Group, *N Engl J Med* **364**, 2507 (Jun 2011).
2. M. J. Borad and et al, *PLoS Genet* **10**, p. e1004135 (Feb 2014).
3. G. J. Weiss, W. S. Liang, M. J. Demeure, J. A. Kiefer, G. Hostetter, T. Izatt, S. Sinari, A. Christoforides, J. Aldrich, A. Kurdoglu, L. Phillips, H. Benson, R. Reiman, A. Baker, V. Marsh, D. D. Von Hoff, J. D. Carpten and D. W. Craig, *PLoS One* **8**, p. e76438 (2013).
4. U. Food and D. Administration, Paving the way for personalized medicine: Fda's role in a new era of medical product development:fda's role in a new era of medical product development.
5. C. for Medicare and M. Services, Clinical laboratory improvement amendments.
6. M.-P. Schapranow, *Analyze Genomes*, tech. rep., Hasso Plattner Institute.
7. C. A. Del Vecchio, C. P. Giacomini, H. Vogel, K. C. Jensen, T. Florio, A. Merlo, J. R. Pollack and A. J. Wong, *Oncogene* **32**, 2670 (May 2013).
8. Filemaker pro (2013).
9. H. Li, *arXiv* **1303**, p. 3997 (2013).
10. K. Cibulskis, M. S. Lawrence, S. L. Carter, A. Sivachenko, D. Jaffe, C. Sougnez, S. Gabriel, M. Meyerson, E. S. Lander and G. Getz, *Nat Biotechnol* **31**, 213 (Mar 2013).
11. C. T. Saunders, W. S. W. Wong, S. Swamy, J. Becq, L. J. Murray and R. K. Cheetham, *Bioinformatics* **28**, 1811 (Jul 2012).
12. A. McKenna, M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly and M. A. DePristo, *Genome Res* **20**, 1297 (Sep 2010).
13. H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin and 1000 Genome Project Data Processing Subgroup, *Bioinformatics* **25**, 2078 (Aug 2009).
14. L. T. Ben Alex Ben Alex, Spring security (July 2014).
15. E. D. Pleasance and et al, *Nature* **463**, 191 (Jan 2010).
16. L. Tang and X.-H. Zhou, *Stat Med* **32**, 620 (Feb 2013).
17. E. T. Wang, R. Sandberg, S. Luo, I. Khrebtukova, L. Zhang, C. Mayr, S. F. Kingsmore, G. P. Schroth and C. B. Burge, *Nature* **456**, 470 (Nov 2008).
18. M. W. Pedersen, M. Meltorn, L. Damstrup and H. S. Poulsen, *Ann Oncol* **12**, 745 (Jun 2001).
19. J. L. Munoz, V. Rodriguez-Cruz, S. J. Greco, V. Nagula, K. W. Scotto and P. Rameshwar, *Mol Cancer Ther* (Jul 2014).
20. Y. Katz, E. T. Wang, E. M. Airoldi and C. B. Burge, *Nat Methods* **7**, 1009 (Dec 2010).
21. C. Trapnell, B. A. Williams, G. Pertea, A. Mortazavi, G. Kwan, M. J. van Baren, S. L. Salzberg, B. J. Wold and L. Pachter, *Nat Biotechnol* **28**, 511 (May 2010).
22. R. Patro, S. M. Mount and C. Kingsford, *Nat Biotechnol* **32**, 462 (May 2014).
23. B. J. Haas, A. Papanicolaou, M. Yassour, M. Grabherr, P. D. Blood, J. Bowden, M. B. Couger, D. Eccles, B. Li, M. Lieber, M. D. Macmanes, M. Ott, J. Orvis, N. Pochet, F. Strozzi, N. Weeks, R. Westerman, T. William, C. N. Dewey, R. Henschel, R. D. Leduc, N. Friedman and A. Regev, *Nat Protoc* **8**, 1494 (Aug 2013).

CHARACTERISTICS OF DRUG COMBINATION THERAPY IN ONCOLOGY BY ANALYZING CLINICAL TRIAL DATA ON CLINICALTRIALS.GOV

MENGHUA WU

1. *Division of Systems Medicine, Department of Pediatrics, Stanford University School of Medicine, 1265 Welch Road, Stanford, CA, 94305, USA*
2. *The Harker High School, 500 Saratoga Ave, San Jose, CA, 95129, USA.*
Email: rachel.wu@gmail.com

MARINA SIROTA

- Division of Systems Medicine, Department of Pediatrics, Stanford University School of Medicine, 1265 Welch Road, Stanford, CA, 94305, USA.*
Email: msirota@stanford.edu

ATUL J. BUTTE

- Division of Systems Medicine, Department of Pediatrics, Stanford University School of Medicine, 1265 Welch Road, Stanford, CA, 94305, USA.*
Email: abutte@stanford.edu

BIN CHEN*

- Division of Systems Medicine, Department of Pediatrics, Stanford University School of Medicine, 1265 Welch Road, Stanford, CA, 94305, USA.*
Corresponding Email: binchen1@stanford.edu

Within the past few decades, drug combination therapy has been intensively studied in oncology and other complex disease areas, especially during the early drug discovery stage, as drug combinations have the potential to improve treatment response, minimize development of resistance or minimize adverse events. In the present, designing combination trials relies mainly on clinical and empirical experience. While empirical experience has indeed crafted efficacious combination therapy clinical trials (combination trials), however, garnering experience with patients can take a lifetime. The preliminary step to eliminating this barrier of time, then, is to understand the current state of combination trials. Thus, we present the first large-scale study of clinical trials (2008-2013) from ClinicalTrials.gov to compare combination trials to non-combination trials, with a focus on oncology. In this work, we developed a classifier to identify combination trials and oncology trials through natural language processing techniques. After clustering trials, we categorized them based on selected characteristics and observed trends present. Among the characteristics studied were primary purpose, funding source, endpoint measurement, allocation, and trial phase. We observe a higher prevalence of combination therapy in oncology (25.6% use combination trials) in comparison to other disease trials (6.9%). However, surprisingly the prevalence of combinations does not increase over the years. In addition, the trials supported by the NIH are significantly more likely to use combinations of drugs than those supported by industry. Our preliminary study of current combination trials may facilitate future trial design and move more preclinical combination studies to the clinical trial stage.

1. Introduction

Since many diseases including cancer are driven by complex molecular and environmental interactions, targeting a single component may not be sufficient to disrupt those mechanisms [1, 2]. Interest in early drug discovery stages has increasingly evolved to target multiple molecules, pathways, or networks [3-6].

Due to the myriad of potential targets and causes of disease, the rate-limiting step to making a meaningful difference with personalized and precision medicine will be the availability of therapeutic options, which are only validated in the context of clinical trials. Increasingly, panomics technologies are being used to identify novel therapeutic options (e.g. target discovery, repositioned drugs) among existing conventional sets of treatments. Of particular note, the visionary document of The American Society of Clinical Oncology stated that combination therapy is critical to developing prevention strategies and curative therapies [7]. Successful combination therapies in oncology include trastuzumab in combination with paclitaxel for breast cancer, and cetuximab in combination with irinotecan for metastatic colorectal cancer. In addition, combinational antiretroviral therapy has become an effective treatment for HIV infections. Despite the promises of combination therapy, however, its success requires the precise optimization of effective doses, the prevention of adverse drug-drug interactions, and many other factors [8, 9]. Currently, combination design is still based primarily on empirical clinical studies [10]. Thus, systematic examination of current combination trials could facilitate more rational design of clinical trials and guide preclinical tests in the early drug discovery stage. To our knowledge, the characteristics of drug combinations in clinical trials remain elusive.

ClinicalTrials.gov, the most robust of the international clinical trial registries, provides a unique opportunity to take a snapshot of all drug combination trial therapy. In September 2007, the federal law required sponsors or designees to register trials and record key elements in this registry. In addition, many journals require the registration of clinical trials before publication. This registry currently (as of July 2014) contains 171,527 studies in 187 countries and increases at a rate of approximately 350 studies per week. A recent effort on the creation of the database for Aggregate Analysis of ClinicalTrials.gov (AACT) facilitates systematic analysis of clinical trials in this registry [11]. Several recent studies have been conducted to examine the characteristics of all the trials [12] or of individual disease areas, including oncology [13-16]. However, neither AACT nor the ClinicalTrials.gov website explicitly annotates combination trials, as their free text data delimits only between individual treatments; combinations, on the other hand, are reported in various ways, including multiple delimited drugs or strings of natural language drug combinations. This inconsistency renders identification and analysis of these trials to be impossible without mining the free text.

In this work, we aimed to learn more about combination trial design in the United States, focusing on basic characteristics such as funding source, primary purpose, trial phase, and prevalence of such trials over time. We first developed a classifier to identify combination trials and oncology trials. By leveraging the information from AACT, we present an initial view of combination trials in oncology. By systematically identifying combinatorial studies within the

database, we make it possible to answer future questions regarding combination therapy and further guide the design of combination trials and preclinical studies.

2. Methods

The clinical trial data used in this study were downloaded from ClinicalTrials.gov on July 15, 2014. The AACT database referenced here reflects ClinicalTrials.gov as of March 27, 2014. We restricted our analysis of clinical trials to interventional trials between 2008 and 2013, as those trials are a complete and unbiased sample following the legislation passed in 2007 requiring all ongoing clinical trials to be registered. The overall workflow is shown in Figure 1.

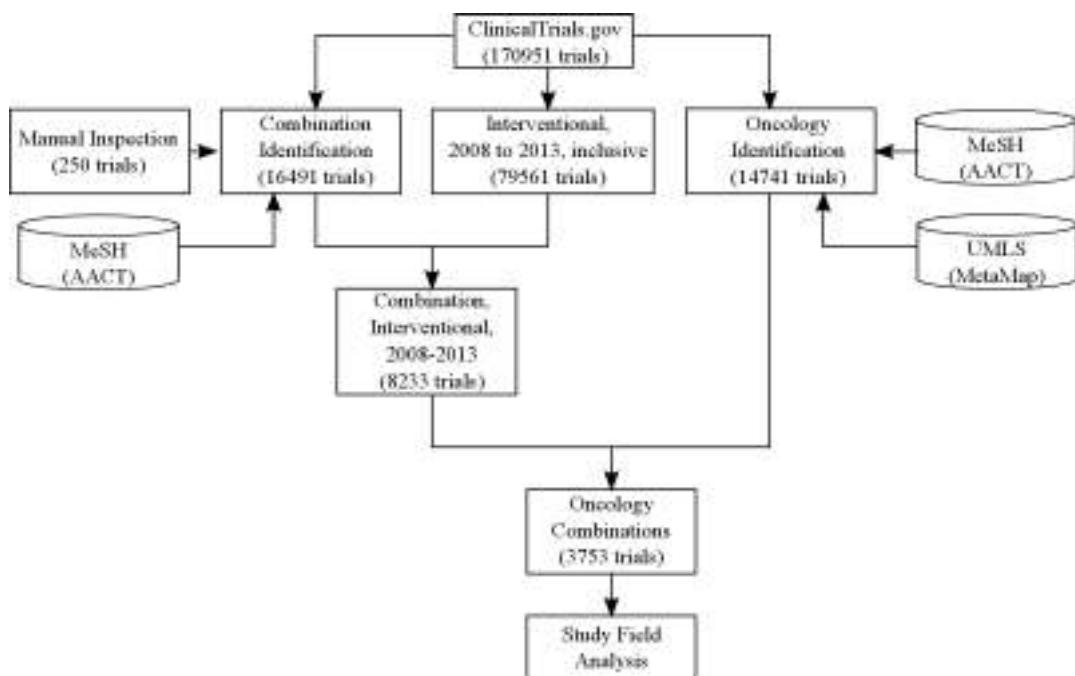


Figure 1: Overview of study workflow. This flowchart documents our process of standardizing and subsetting clinical trial data based on type and content of study.

2.1 Identifying drug combination trials

Though neither ClinicalTrials.gov nor AACT expressly annotates drug combination trials, some study fields may indicate if a combination is used. We developed a scoring system to identify combination trials based on the information extracted from these fields. A combination trial is defined as a trial in which at least two drugs are administered to a group of patients. Without specific note, a drug can be a small molecule or a biological agent in our study. We summarized discriminating features and assigned different scores to each feature, depending on how confidently that feature could identify combinations. The scores were first assigned based on empirical experience and later adjusted using a training set consisting of 300 manually annotated trials. For example, if the word, “combination,” appeared in the title, then the trial was highly probable to contain combinations, so we considered the occurrence of “combination” in the title a

highly weighted feature. On the other hand, if the title included the word “and,” it may indicate that two drugs are used together, but it also may refer to unrelated words, so we weighed this feature lower. Such features were also extracted manually from a list of interventions that contained non-standard, free text descriptions of combinations. These strings often did not map to the standardized vocabulary for drugs as defined by PubChem and DrugBank. For example, trial NCT01121575 specifies its intervention as “PF-00299804 followed by combined PF-02341066 and PF-00299804.” A complete list of our features and study fields can be found in Table 1. Sub-scores were added together to generate the final score for the trial. Trials with scores greater than 2 were classified as combination therapy trials. We validated the scoring system using independent 250 manually annotated trials, with which the system could achieve an impressive performance (with precision 0.94 and recall 0.85). Due to the small feature set, other classifiers such as random forest did not produce a better performance (data not shown), so we decided to choose this simpler scoring system for the classification.

Table 1: Features for combination detection. Each feature is labeled with its field in AACT, the exact words that were looked for, and the sub-score assigned to it for each occurrence. Sub-scores were added together to generate the final score for the trial. Trials with scores greater than 2 were classified as combination therapy trials.

Field	Term	Score
title	“combine”, “combination”, “combining”	2
title	“with/without”, “and/or”	2
title	“plus”	2
title	“interaction”, “co-administered”	1
title	“and”, “with”	0.5
intervention	“plus”, “and”, “+”, “/”	1
intervention	“placebo”, “vehicle”	-1
summary	“combine”, “combination”, “combining”	2
summary	“together with”, “interaction with”, “alone or with”, “co-administered”, “parallel assignment”	1
arm	“combine”, “combination”, “combining”	2
arm	“together with”, “interaction with”, “alone or with”, “parallel assignment”, “plus”	1
arm	“Drug:” (frequency per individual arm)	1
keyword	“combination”	1

2.2 Identifying oncology trials

Oncology trials were inferred using disease condition terms (including both Medical Subject Heading [MeSH] and non-MeSH terms) provided by the data submitters and additional condition MeSH terms annotated by a National Library of Medicine (NLM) algorithm [11]. The average number of MeSH terms for the interventional trials between 2008 and 2013 was 2.3. If the trial had at least one MeSH term that started with C04 (Neoplasms), it was considered an oncology trial; otherwise, it was considered a non-oncology trial. For example, C04.557.337 represents

leukemia, while C02.839.040 represents acquired immunodeficiency disorder and does not start with “C04”. The trials were also grouped into other disease categories (e.g., Cardiovascular Diseases with MeSH ID starting with C14, Nervous System Diseases with MeSH ID starting with C10, and others) based on their MeSH terms. The MeSH IDs associated with each trial were provided by AACT.

Unfortunately, 15,306 out of 79,561 trials were not associated with any MeSH terms, but provided conditions using free text. To include these trials in our analysis, we used MetaMap [17] to annotate the conditions that they studied. MetaMap uses a knowledge-intensive approach based on symbolic, NLP and computational-linguistic techniques to map free text into Unified Medical Language System (UMLS) concepts. This tool also provides confidence scores and concept categories. Trials with at least one term categorized as “Neoplastic Process” with the maximum confidence score of 1000 were considered oncology trials. Trials categorized otherwise, or with lower confidence scores, were considered non-oncology trials. For example, trial NCT00546247 is not associated any MeSH terms, even though it studied “Advanced Solid Tumors” according to the condition description; this trial was as a Neoplastic Process successfully recognized by MetaMap.

Both the MeSH-based approach and the UMLS-based approach used text mining to process clinical trial conditions. We further compared their predictions using the trials where MeSH and non-MeSH terms were provided. MeSH identified 35,144 oncology trials, MetaMap identified 21,292 oncology trials, and 18,960 of them were common. This shows that 89% of oncology trials identified by the UMLS-based approach were corroborated by the MeSH-based approach.

2.3 Characterization of drug combination clinical trials in oncology

We examined combination trials in oncology in the context of: (1) start year, (2) primary purpose, (3) endpoint measurements (type of study), (4) phases, (5) allocation (randomization), and (6) funding sources. These were extracted from the trials’ study design, start date, phase, and funded by fields. Trials with missing features were ignored for analysis of that particular feature. For funded sources, some trials may have multiple sources. The results were compared across combination trials in oncology, combination trials in non-oncology, and non-combination trials in oncology.

2.4 Statistical analysis

Pearson’s Chi-Squared test was used to compare frequencies of trials associated with different features in Tables 2 and 3, for oncology combinations vs. oncology non-combinations and oncology combinations vs. non-oncology combinations, respectively. Linear regression and analysis of variance (ANOVA) were used to identify significant factors associated with trial start year. P-value < 0.001 was the default significance cutoff.

3. Results

We examined 170,951 clinical trials available through ClinicalTrials.gov. We further identified 16,491 combination trials and 36,430 oncology trials. After restricting the trials to interventional

studies from 2008 to 2013 to obtain the most comprehensive and unbiased set of trials for our analysis, 8,233 combination trials and 14,652 oncology trials remained. Among the 8,233 combination trials, 45.6% (3,753 trials) are related to oncology. In addition, among the 14,652 oncology trials, 25.6% contain combinations, while only 6.9% of non-oncology trials contain combinations, indicating that oncology trials have significantly more combination trials than non-oncology trials (Chi-Squared test, P-value < 0.001).

When looking at other specific disease types, we found that viral diseases and digestive diseases were also more likely to contain combinations (22.3% and 18.6%, respectively), while cardiovascular conditions, pathological conditions, and nervous systems diseases were less likely to contain combinations (8.5%, 5.6%, and 5.4%, respectively).

Table 2. Combination trials across different disease types

Disease Type	Oncology	Viral Diseases	Digestive Diseases	Cardiovascular Diseases	Pathological Conditions	Neurological Diseases
Combination (number of trials)	3753	797	1295	722	963	463
Non-Combination (number of trials)	10899	2771	5662	7782	16056	8056
% Combination	25.6	22.3	18.6	8.5	5.7	5.4

3.1 Trend of oncology combination trials across years

The number of total oncology trials and combination oncology trials stayed constant from 2008 to 2013 (Figure 2a). Surprisingly, the ratio of oncology combination trials to all oncology trials decreased significantly (P value < 0.05; Figure 2b). In 2008, there are 2,339 oncology trials in total, and 691 of these trials contained combinations (29.5%). The ratio decreases to 22.7% in 2012. As shown in Figure 2c, the primary decrease is caused by the steady decrease in phase 2 trials. In addition, over the years, industry has consistently shown less interest in combinations than the NIH, but the NIH has recently decreased involvement in combinations as well (Figure 2b, P value < 0.05 by ANOVA).

3.2 Comparison of combination and non-combination trials in oncology

Table 3 lists the comparison across combination trials in non-oncology, combination trials in oncology, and non-combination trials in oncology. Oncology combinations are more likely to be funded by the NIH than non-oncology combinations (22.7% vs. 5.1%; P-value < 0.001). Combinations in oncology are less likely to be used for prevention (0.9% in oncology vs. 10.9% in non-oncology; P-value < 0.001) and basic science (0.4% in oncology vs. 5.6% in non-oncology; P value < 0.001). Combination trials in oncology are less likely to be reported as phase 4 than those in non-oncology (1.5% vs. 19.4%; P value < 0.001) and are more likely to be non-randomized (33.0% vs. 13.5%; P value < 0.001).

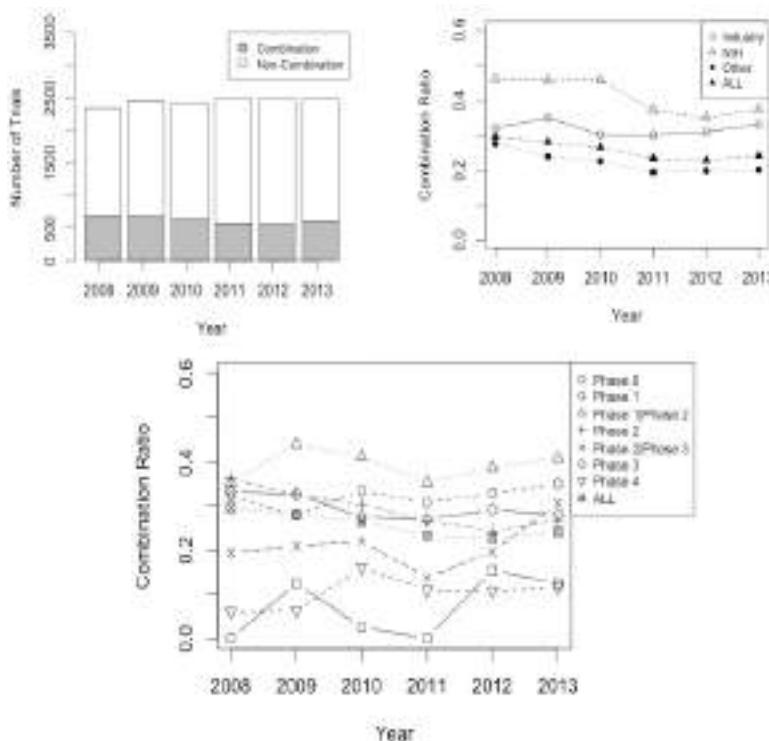


Figure 2. a) Total oncology trials and oncology combination trials over the years, b) The ratio of oncology combination trials vs. all oncology trials over the years, grouped by funding sources, c) The ratio of oncology combination trials vs. all oncology trials over the years, grouped by phases.

Then within oncology trials, the primary funding source of combination trials is “Other” (61.7%), which includes universities, independent institutes, etc., followed by industry (49.0%), the NIH (22.7%), and the U.S. Federal Government (0.4%). Of the 2,058 oncology trials supported by the NIH, 853 trials are combinations (22.7% of all combination trials) and 1,205 trials are non-combinations (11.1% of all non-combination trials), indicating that combination trials are more likely to be supported by the NIH than non-combination trials (P -value < 0.001).

As expected, the primary purpose of combination trials is to test treatments (97.5%). Very few combination trials are conducted for prevention (0.9%), screening (0.08%), supportive care (0.8%), basic science (0.3%), diagnosis (0.3%), or health services research (0.03%). On the other hand, larger portions of non-combination trials are conducted for diagnosis (8.1%), supportive care (5.7%) and prevention (4.8%). Prevention is less likely to appear as a primary purpose in combination trials (4.8% in non-combination trials vs. 0.9% in combination trials; P value < 0.001).

Nearly half of the combination trials are reported as phase 2 (45.5%), and very few are reported as phase 0 (0.4%) or phase 4 (1.5%). Of the 526 phase 4 trials in oncology, only 10.1% (53 trials) include combinations, while 26.5% of all oncology trials include combinations.

Table 3. Comparison of combination trials and non-combination trials in oncology and non-oncology

	Non-Oncology Combination	Oncology Combination	Oncology Non-Combination
Funded By ^(a)	n=4480	n=3753	n=10899
Industry	3086 (68.9%)	1839 (49.0%)	3946 (36.2%)
NIH	230 (5.1%)	853 (22.7%)	1205 (11.1%)
Other	1654 (36.9%)	2318 (61.8%)	8195 (75.2%)
U.S. Fed	39 (0.9%)	14 (0.4%)	90 (0.8%)
Primary Purpose	n=4217	n=3732	n=10545
Basic Science	236 (5.6%)	13 (0.3%)	148 (1.4%)
Diagnostic	44 (1.0%)	12 (0.3%)	858 (8.1%)
Health Services Research	9 (0.2%)	1 (0.03%)	97 (0.9%)
Prevention	458 (10.9%)	34 (0.9%)	508 (4.8%)
Screening	6 (0.1%)	3 (0.08%)	145 (1.4%)
Supportive Care	43 (1.0%)	30 (0.8%)	597 (5.7%)
Treatment	3421 (81.1%)	3639 (97.5%)	8192 (77.7%)
Endpoint Measures	n=4179	n=3311	n=9202
Bio-availability Study	63 (1.5%)	4 (0.1%)	24 (0.3%)
Bio-equivalence Study	95 (2.3%)	0 (0%)	52 (0.6%)
Efficacy Study	809 (19.4%)	795 (24.0%)	3051 (33.2%)
Pharmacodynamics	78 (1.9%)	7 (0.2%)	86 (0.9%)
Pharmacokinetics	425 (10.2%)	24 (0.7%)	103 (1.1%)
Pharmacokinetics/Dynamics	137 (3.3%)	12 (0.4%)	71 (0.8%)
Safety Study	437 (10.7%)	492 (14.9%)	947 (10.3%)
Safety/Efficacy Study	2135 (51.1%)	1977 (59.7%)	4868 (52.9%)
Allocation	n=4176	n=2413	n=6142
Non-Randomized	563 (13.5%)	795 (33.0%)	2189 (35.6%)
Randomized	3613 (86.5%)	1618 (67.0%)	3953 (64.4%)
Phase	n=4134	n=3634	n=8530
Phase 0	12 (0.3%)	15 (0.4%)	180 (2.1%)
Phase 1	1114 (27.0%)	778 (21.4%)	1657 (19.4%)
Phase 1/Phase 2	133 (3.2%)	582 (16.0%)	896 (10.5%)
Phase 2	851 (20.6%)	1655 (45.5%)	3950 (46.3%)
Phase 2/Phase 3	88 (2.1%)	52 (1.4%)	201 (2.4%)
Phase 3	1136 (27.5%)	499 (13.7%)	1183 (13.9%)
Phase 4	800 (19.4%)	53 (1.5%)	463 (5.4%)

(a) The sum of percentages exceed 100 because trials may be funded by more than agencies

4. Discussion

The advances of genomics and high-throughput technologies enable identifying molecular aberrations of individual tumors and other molecular features that could guide individualized treatment. As tumors are consequences of defects in a complex network comprising a multitude of

environmental factors, genetic mutations, and polymorphisms, increasingly more preclinical combinatory studies suggest that targeting multiple components in the tumors may be necessary [4-6]. However, the parameters in clinical trials are very different with those in preclinical settings, while in the present, designing combination trials relies mainly on clinical and empirical experience. Therefore, understanding existing combination trials is of critical importance for future design of clinical trials and preclinical studies.

To our knowledge, there has not yet been any systematic study of combination clinical trials conducted. One critical barrier is that no combination clinical trial dataset is publicly available. Hence, we took the first step towards collecting combination trials and extracting useful quantitative data about these combination trial characteristics from a massive data repository. We developed a simple, yet precise, classifier to identify combination trials, and we also leveraged public natural language processing tools (e.g., MetaMap) and datasets (e.g., MeSH) to identify oncology trials. The dataset we collected paves the way for future drug combination studies.

Our analysis shows that nearly half of all combination trials are conducted in oncology, and a quarter of oncology trials use combination therapies, indicating drug combination is indeed, prevalent in oncology. According to FDA guidance, combinations are intended to treat serious diseases or conditions associated with morbidity that have substantial impact on day-to-day functioning. Oncology and infectious diseases are among the most popular severe diseases for which combination therapy is highly desired, while less severe diseases such as pathological conditions and neurological diseases do not demand combinatory therapy [18].

We found that the trials supported by the NIH are significantly more likely to use combinations of drugs than those supported by industry in the last few years. This phenomenon may be caused by companies' tendencies to focus on developing their own specific drugs rather than testing drugs from possible outside sources; conversely, academic labs tend to focus less on the drug vendors and more on finding efficacious combinations. Over the years, industry's interest in combination usage remains constant. Surprisingly, academic interest in combinations appears to be declining from 2010 to 2012 and then starts increasing in 2012. Notably, the interest drops significantly from 2010 to 2011; this coincides the release of the draft guidance on combination therapy codevelopment issued by Food and Drug Administration (FDA) in December 2010 [18]. In the guidance, FDA suggests all of the following criteria should be met for the consideration of co-development of combination therapy: 1) the combination is intended to treat a serious disease or condition, 2) there is a strong biological rationale for use of the combination, 3) a full nonclinical or a short-term clinical study suggests the combination is superior to the individual agents, and 4) there is a compelling reason why the new investigational drugs cannot be developed independently. This requirement may lend increases in costs and time to combination trials, leading a deceasing interest to initiate combination trials. Another possible explanation is that industry and academia are increasingly linked, and conflicts of interest may exist [19-21]. The success of many single-target drugs in oncology may partially explain why industry still prefers to single agent therapy [22,23], and why academia has followed this trend in recent years. In addition, this requirement explains why we found more safety and safety/efficacy studies within oncology combination trials, and why the main primary purpose is treatment.

We also found that combination trials are more likely conducted for treatment and happen in phase1/phase2 and phase 2, signifying that most combinations are still being tested for safety and/or efficacy. FDA suggests whenever possible, the safety profile and dose response of individual new drugs should be characterized in phase 1, resulting in a fewer number of phase 1 combination trials in phase 1 than phase 2 trials [18].

Our current work has several limitations. First, ClinicalTrials.gov does not capture all the studies performed in the USA, especially phase 0-1 trials, which are not required to be registered in the repository. Nonetheless, it still covers over 80% of all studies [12]. Second, some information is missing or misrepresented in the database. For example, 3,872 trials in our data set do not specify a start date. As we only retained trials between 2008 and 2013 for final analysis, the trials without start dates were unable to be used. Third, although our classifier has good performance, some trials are still misclassified. Much effort is required for manual inspection of trials, but even human error in annotating trials cannot be avoided due to the ambiguous information provided. In the future, we will incorporate a greater number of features and keywords to improve classifier performance and potentially introduce more sophisticated algorithms to identify combination trials. For this study, however, due to the low number of falsely identified trials and our classifier's satisfactory precision, the overall trend is unlikely to change.

Finally, we only assessed the fundamental characteristics of combination trials in this work. In order to facilitate the design of combination trials, it would be important to assess other features as specific disease types and interventions applied. The data contained within ClinicalTrials.gov is free text submitted without strict guidelines for sponsors, so much manual effort is required to process these data. In the future, we will stratify our data into different cancer types (e.g., lung cancer, leukemia, etc.) and seek characteristics within the different types. In addition, we will identify patterns within specific drug combinations (e.g., which two drugs are used together frequently, which drugs reach further stages of clinical trials) by analyzing various drug features. We will aim to extract more information from our datasets regarding the interactions between drugs listed in clinical trials, especially regarding the exact combinations formed from the individual drugs. Many drugs that are tested in clinical trials are unapproved compounds or biological agents, which renders them hard to standardize and extract from the clinical trial data. In addition, many combination trials only list all the interventions applied without specifying the exact relationships between them. We believe by analyzing other fields and integrating with other sources (e.g., drug-target relationships [24]), we can better understand the nature of drug combinations in these trials.

5. Conclusion

Cancer is a heterogeneous disease that involves a multitude of genetic and environment factors. Its complexity also accounts for its resistance against many current targeted therapies. Increasing numbers of studies aim to target multiple factors and have proven successful in many cases, especially in preclinical models. However, moving from preclinical models to clinical trials requires consideration of many more clinical factors. Thus, understanding the characteristics of current combination trials may facilitate the design of future trials. In this work, we developed

methods to identify combination trials and oncology trials from ClinicalTrials.gov, and we took the first step to exploring the fundamental characteristic of drug combination trials. Surprisingly, we found that interest in drug combinations does not increase. Understanding the barriers that prevent combinations from reaching the clinical stage may help advance more combinational therapy studies. In the future, we will integrate our clinical data with other molecular features to extract more patterns regarding drug combinations.

6. Acknowledgements

This research was funded in part by the Lucile Packard Foundation for Children's Health, the Stanford Child Health Research Institute, the Hewlett Packard Foundation, and the National Institute of General Medical Sciences (R01 GM079719). Menghua Wu was supported by Stanford Institutes of Medicine Summer Research Program (SIMR). We thank Jenna Bollyky for providing valuable comments.

References

1. Chen B, Butte AJ. Network medicine in disease analysis and therapeutics. *Clin Pharmacol Ther.* 2013 Dec;94(6):627-9.
2. Barabasi AL, Gulbahce N, Loscalzo J. Network medicine: a network-based approach to human disease. *Nature reviews Genetics.* 2011 Jan;12(1):56-68.
3. Hopkins AL. Network pharmacology: the next paradigm in drug discovery. *Nature chemical biology.* 2008 Nov;4(11):682-90.
4. Lee MJ, Ye AS, Gardino AK, et al. Sequential application of anticancer drugs enhances cell death by rewiring apoptotic signaling networks. *Cell.* 2012 May 11;149(4):780-94.
5. Miller ML, Molinelli EJ, Nair JS, et al. Drug synergy screen and network modeling in dedifferentiated liposarcoma identifies CDK4 and IGF1R as synergistic drug targets. *Science signaling.* 2013 Sep 24;6(294):ra85.
6. Mathews Griner LA, Guha R, Shinn P, et al. High-throughput combinatorial screening identifies drugs that cooperate with ibrutinib to kill activated B-cell-like diffuse large B-cell lymphoma cells. *Proceedings of the National Academy of Sciences of the United States of America.* 2014 Feb 11;111(6):2349-54.
7. Oncology ASoC. Shaping the Future of Oncology: Envisioning Cancer Care in 2030. 2014 Available from: <http://www.asco.org/about-asco/asco-vision>
8. Verweij J, Disis ML, Cannistra SA. Phase I studies of drug combinations. *J Clin Oncol.* 2010 Oct 20;28(30):4545-6.
9. Ananthakrishnan R, Menon S. Design of oncology clinical trials: a review. *Critical reviews in oncology/hematology.* 2013 Oct;88(1):144-53.
10. Al-Lazikani B, Banerji U, Workman P. Combinatorial drug therapy for cancer in the post-genomic era. *Nat Biotechnol.* 2012 Jul;30(7):679-92.
11. Tasneem A, Aberle L, Ananth H, et al. The database for aggregate analysis of ClinicalTrials.gov (AACT) and subsequent regrouping by clinical specialty. *PloS one.* 2012;7(3):e33677.

12. Califf RM, Zarin DA, Kramer JM, Sherman RE, Aberle LH, Tasneem A. Characteristics of clinical trials registered in ClinicalTrials.gov, 2007-2010. *JAMA*. 2012 May 2;307(17):1838-47.
13. Goswami ND, Pfeiffer CD, Horton JR, Chiswell K, Tasneem A, Tsalik EL. The state of infectious diseases clinical trials: a systematic review of ClinicalTrials.gov. *PloS one*. 2013;8(10):e77086.
14. Hill KD, Chiswell K, Califf RM, Pearson G, Li JS. Characteristics of pediatric cardiovascular clinical trials registered on ClinicalTrials.gov. *American heart journal*. 2014 Jun;167(6):921-9 e2.
15. Hirsch BR, Califf RM, Cheng SK, et al. Characteristics of oncology clinical trials: insights from a systematic analysis of ClinicalTrials.gov. *JAMA internal medicine*. 2013 Jun 10;173(11):972-9.
16. Stockmann C, Sherwin CM, Ampofo K, et al. Characteristics of antimicrobial studies registered in the USA through ClinicalTrials.Gov. *International journal of antimicrobial agents*. 2013 Aug;42(2):161-6.
17. Aronson AR, Lang FM. An overview of MetaMap: historical perspective and recent advances. *Journal of the American Medical Informatics Association : JAMIA*. 2010 May-Jun;17(3):229-36.
18. U.S. Department of Health and Human Services. Guidance for Industry: Codevelopment of Two or More New Investigational Drugs for Use in Combination. *Clinical Medical*: 2013 June
19. Lexchin J, Bero LA, Djulbegovic Benjamin, et al. Pharmaceutical industry sponsorship and research outcome and quality: systematic review. *BMJ* 2003;326:1167
20. Bodenheimer T. Uneasy alliance: Clinical investigators and the Pharmaceutical Industry. *Health Policy Report*: 2000 May;342(20):1539-1544.
21. Angell M. Industry-sponsored clinical research: A broken system. *JAMA*. 2008;300(9):1069-1071.
22. DiMasi JA, Grabowski HG. Economics of new oncology drug development. *Journal of Clinical Oncology*: 2007 January;25(2): 209-216
23. Kamb A, Wee S, Lengauer C. Why is cancer drug discovery so difficult? *Nature Reviews Drug Discovery*: 2007 February;6:115-120
24. Rask-Andersen M, Masuram S, Schiøth HB. The druggable genome: Evaluation of drug targets in clinical trials suggests major shifts in molecular class and indication. *Annual review of pharmacology and toxicology*. 2014;54:9-26.

CANCER PATHWAYS: AUTOMATIC EXTRACTION, REPRESENTATION, AND REASONING IN THE ‘BIG DATA’ ERA

GRACIELA GONZALEZ

*Department of Biomedical Informatics,
Arizona State University, Scottsdale, AZ 85259, USA
Email: ggonzalez@asu.edu*

CHITTA BARAL

*School of Computing, Informatics, and Decision Systems Engineering,
Arizona State University, Tempe, AZ 85287, USA
Email: chitta@asu.edu*

JEFF KIEFER

*Knowledge Mining Laboratory,
Translational Genomics Research Institute (TGen), Scottsdale, AZ, 85259, USA
Email: jkiefer@tgen.org*

SEUNGCHAN KIM

*Integrated Cancer Genomics Division, Biocomputing Unit
Translational Genomics Research Institute (TGen), Phoenix, AZ 85004, USA
Email: skim@tgen.org*

JIEPING YE

*School of Computing, Informatics, and Decision Systems Engineering,
Arizona State University, Tempe, AZ 85287, USA
Email: jieping.ye@asu.edu*

There has been great interest and research initiatives in the biomedical community around harnessing “big data”, including data from the literature, high-throughput gene expression experiments, array CGH and high-throughput siRNA and many other types of data to generate novel hypothesis to address the most crucial biomedical questions and aid in the discovery of more effective and improved therapeutic options for the treatment of complex and pervasive diseases such as cancer. Cancer research has progressed rapidly in the last decade with the implementation of high-dimensional genomic technologies. The large amount of data generated over the years has enabled a systems-based approach to uncovering and elucidating the complex signaling networks associated with cancer. However, even though new technologies have advanced our understanding of cancer biology beyond what could be imagined even a decade ago, there still exist unique challenges associated precisely with the amount of data that is now routinely generated from even a single patient. The data must be stored and processed, with novel analysis strategies called for to uncover new insights into cancer biology that are literally hidden in ‘big data’. Interest in taming ‘big data’ through methods and systems to extract, represent, and transform it into knowledge that can effectively be used for reasoning and question answering will only increase over time,

enabling scientists to finally use the data for personalized treatment, discovery and validation.

Work presented in this session includes novel approaches to explore cancer gene expression data, applying algebraic topology (Lockwood and Krishnamoorthy) and Denoising autoencoders (Tan et al) to identify significant properties of genomic data that cannot be found by traditional algorithms. There is also a novel methodology for leveraging somatic mutation data for predicting survival in cancer samples (Kim et al), a computational system for automated gene expression pattern annotation on mouse brain images that could prove to be key to understanding the pathogenesis of brain tumors and their early detection (Yang et al). With respect to knowledge extraction, this session includes work on a weakly supervised machine learning approach for automatic pathway extraction from PubMed abstracts (Poon et al), and on the use protein interaction data from multiple sources to investigate mutations in 125 genes that were earlier identified as driving tumorigenesis when mutated (Engin et al).

1. Introduction

This session brings together researchers in text and data mining, knowledge representation and reasoning with bench scientists, geneticists and translational scientists. It serves as a unique forum to discuss novel approaches to the complex text extraction and knowledge representation and reasoning problems that come with dealing with big data. Given that computational approaches often draw upon disease-specific resources and expertise, we focus the session on disseminating approaches to extract, represent, or reason with knowledge derived from the literature, databases, or experimental data that respond to biological questions about the signaling pathways in cancer pathophysiology.

2. Challenges

Improving text and data mining methods for any task requires careful consideration and evaluation. The biomedical domain presents specific challenges given the diversity, complexity and volume of the information being mined. This section presents a brief overview of the fundamental challenges faced by researchers in these areas.

2.1. Extraction, Representation, and Reasoning

While there exists significant research in information extraction from biomedical text, little has been done on the extraction of more general knowledge that requires the integrated use of information from many different publications, databases, and experimental results into a coherent story, a (proven or hypothetical) biological pathway that can be used to drive new research directions. For this task, accurate named entity recognition and named entity identification (also referred to as normalization) are important, as well as the extraction of (binary) relationships or events involving any two of these entities. However, when the goal is to go beyond atomic events and into establishing an ordered sequence of these events to reconstruct the “biological stories” they are telling us (pathways), the problem has strong similarity with the classical planning and scheduling problem in Artificial Intelligence (AI) and shares similar challenges. Furthermore, to answer questions about such pathways, one needs to start with a query representation that is equally comprehensible to human as to machines.

This session includes specialized examples that make some advances into the mostly uncharted territory of pathway extraction, showing the trend towards finer granularity in the type of information needed for meaningful advances, requiring a tighter collaboration between the text mining community and domain experts.

2.2. Data Mining

Recent advance in technologies has generated a large amount of high-dimensional genomic data. These data are a key to illuminate fundamental principles of cancer biology. However, analysis of these data poses unique challenges for data mining. One of the key issues is the curse of dimensionality, i.e., an enormous number of samples is required to perform accurate prediction on problems with large numbers of features. Feature selection, which selects a small number of features by removing the irrelevant, redundant, and noisy information, is an effective way to overcome the curse of dimensionality. However, existing methods for feature selection are less effective when there are strong empirical correlations among the features. Furthermore, the presence of missing values which are ubiquitous in biomedical data further complicates the problem. Several recent works apply unsupervised feature learning algorithms to address these challenges. The goal of these methods is to identify highly correlated features and construct representative features from the data.

Included in this session are works that address some of these challenges.

3. Overview of Contributions

Lockwood and Krishnamoorthy apply tools from algebraic topology to explore cancer gene expression data. Specifically, the proposed method selects a small relevant subset from tens of thousands of genes while simultaneously identifying higher order topological features. By employing tools from algebraic topology, the proposed method is capable of identifying geometric properties of the data that cannot be found by traditional algorithms such as clustering.

Poon et al applied distant supervision, a state-of-the-art weakly supervised machine learning approach, for automatic pathway extraction from PubMed abstracts. From 22 million PubMed abstracts, they extracted 1.5 million pathway interactions at a precision of 25%. More than 10% of interactions are mentioned in the context of one or more cancer types. The paper reports interesting results, exploring the effectiveness of utilizing the information available in the pathway database (PID) for automatic annotation of the training data. The applied machine learning approach does not rely on the manually annotated sentences and can potentially be applied to other problems.

Tan et al present a method based on Denoising autoencoders (DAs) for feature extraction from biological data. DAs aim to learn compact and efficient representations from input data. DAs serve as building blocks for deep networks, which have been applied successfully for image and audio processing. The authors present an interesting application of DAs to, for the first time, identify and extract complex patterns from genomic data.

Kim et al touches on a number of themes appropriate for this session. They present a novel methodology for leveraging somatic mutation data for predicting survival in cancer samples. As a test case, they apply their methodology to Renal Cell Carcinoma data from the TCGA. The authors take advantage of their previous tool and methods, BioBin, ATHENA, and Grammatical Evolution Neural Networks (GENN) in the current study based on prior knowledge resources.

Yang et al present a computational system for automated gene expression pattern annotation on mouse brain images. The annotation system aims to provide an accurate description of the locations where the genes are active. Since genetic factor is one of the significant risk factors for brain cancer, such a description is crucial for understanding the pathogenesis of brain tumors and for early detection.

Engin et al. use protein interaction data from multiple sources to investigate mutations in 125 genes that were earlier identified as driving tumorigenesis when mutated. They do structural enrichment of driver protein-protein interactions (PPIs) and map chromosomal coordinates to PDB coordinates. They use TCGA mutation data and RNAseq data as source of patient mutation and expression profiles. Thus by integrating protein interaction networks with protein structure, and using patient related mutation and gene expression data, they observe “patient specific network wiring” and analyze the HRAS subnetwork. Overall, this paper is a good example of cancer network analysis and the integration and mining of various different kinds of data.

IDENTIFYING MUTATION SPECIFIC CANCER PATHWAYS USING A STRUCTURALLY RESOLVED PROTEIN INTERACTION NETWORK

H. BILLUR ENGIN

*School of Medicine, University of California San Diego, 9500 Gilman Dr.
San Diego, CA 92093, USA
Email: hengin@ucsd.edu*

MATAN HOFREE

*Department of Computer Science and Engineering, University of California San Diego, 9500 Gilman Dr.
San Diego, CA 92093, USA
Email: mhofree@ucsd.edu*

HANNAH CARTER*

*School of Medicine, University of California San Diego, 9500 Gilman Dr.
San Diego, CA 92093, USA
Email: hkarter@ucsd.edu*

Here we present a method for extracting candidate cancer pathways from tumor ‘omics data while explicitly accounting for diverse consequences of mutations for protein interactions. Disease-causing mutations are frequently observed at either core or interface residues mediating protein interactions. Mutations at core residues frequently destabilize protein structure while mutations at interface residues can specifically affect the binding energies of protein-protein interactions. As a result, mutations in a protein may result in distinct interaction profiles and thus have different phenotypic consequences. We describe a protein structure-guided pipeline for extracting interacting protein sets specific to a particular mutation. Of 59 cancer genes with 3D co-complexed structures in the Protein Data Bank, 43 showed evidence of mutations with different functional consequences. Literature survey reciprocated functional predictions specific to distinct mutations on APC, ATRX, BRCA1, CBL and HRAS. Our analysis suggests that accounting for mutation-specific perturbations to cancer pathways will be essential for personalized cancer therapy.

1. Introduction

Cancer is a complex genetic disease in which the genomes of normal cells accumulate somatic mutations. A subset of these mutations confer neoplastic behaviors to cells through disregulation of a small number of common pathways¹. Identifying the genes that participate in these pathways is an important objective in cancer genomics. However, linking somatically altered genes to perturbed pathways remains an open problem².

Individual proteins rarely mediate cellular behaviors; instead molecular machines comprising multiple proteins arbitrate various intracellular processes. As a result, proteins that interact physically within the cell are frequently involved in the same biological activities. This phenomenon, sometimes called guilt-by-association, has motivated the development of a variety of computational methods to identify disease-specific regions on the human Protein-Protein Interaction (PPI) network from molecular measurement data. Ideker *et al.*³ integrate PPIs with mRNA expression data to detect differentially expressed sub-networks of genes, while, NIMMI⁴ combines PPI networks with GWAS data to produce sub-networks that are functionally related and enriched for genetic variants linked with a trait. HotNet⁵ maps mutation data onto PPI networks to identify sub-networks significantly enriched with cancer causing mutations, and NetBox⁶ detects oncogenic network modules from DNA Copy Number Variants (CNVs), PPIs and signaling

* This work is supported by NIH grant DP5 OD017937-01.

pathways. Additionally, Hofree *et al.*⁷ combine patient mutation profiles, gene interaction networks and PPIs to find network regions that are specific to subtypes of cancer.

Missense mutations, a class of non-synonymous single nucleotide variants (nsSNVs), cause an amino acid substitution, resulting in a subtly different protein sequence. These sequence changes can alter protein structure. The resulting consequences for protein activity span the spectrum from neutral to completely disruptive. To date, methods for automated pathway extraction have treated missense mutations either as disruptive or neutral to protein activity, however it is well established that distinct amino acid sites within a protein mediate different functions. Simply modeling proteins as active or not may detract from the biological relevance of extracted pathways.

Recently, several groups have published high-resolution three-dimensional (3D) PPI networks⁸⁻¹⁰ that include the molecular details of binding interfaces. Applications of these to investigate inherited disease mutations¹¹⁻¹⁵ have suggested that a) nsSNVs located at protein interfaces result in distinct phenotypes from those located in the protein core^{9,10}, b) known disease associated variants outside of the core are enriched at residues participating in protein interaction interfaces¹⁰ c) in particular, in-frame disease mutations are enriched at interface regions of interacting proteins⁹ and d) disease mutations at distinct interfaces of the same protein can be associated with distinct disease phenotypes⁹. In cancer, 3D location of mutations at an interface has served as evidence that protein interactions may be important for metastasis site determination.¹¹ These observations suggest that distinct changes to the network topology of protein interaction networks will result in different phenotypes. Thus efforts to identify disease-associated pathways may need to account for mutation-specific effects to the PPI network.

Here, we investigate the extent to which distinct somatic mutations observed in known cancer genes have distinct phenotypic consequences. We present a structure-guided sub-network extraction pipeline (**Figure 1**) that identifies protein sets associated with specific missense mutations. We divide mutations observed in tumor exome sequencing data from The Cancer Genome Atlas (TCGA) into two categories: core and interface, then use structurally resolved protein interaction data to model the effects of mutations on PPI network topology. Then using a diffusion-based approach, we identify distinct sets of interacting proteins in the global interaction network associated with different residue alterations of the same cancer gene and show that in many cases, these protein sets are implicated in distinct biological activities.

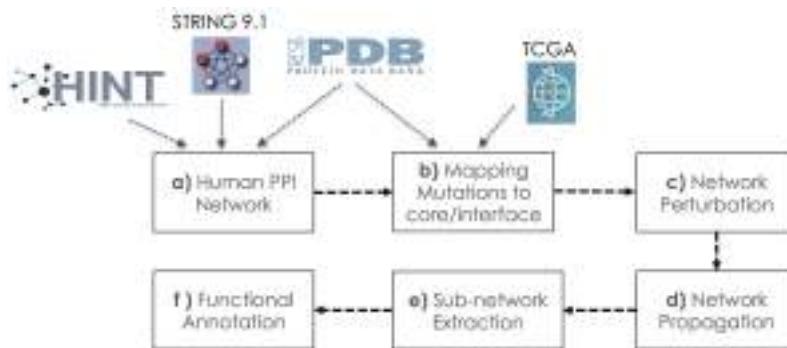


Figure 1. A pipeline for mining molecular cancer related sub-networks accounting for different effects of distinct mutations. Steps include network assembly (a), mapping of mutations to interface versus core residues of cancer genes (b), removal of affected edges (c), extraction of associated protein sets (d & e) and functional annotation (f).

2. Materials and Methods

2.1 Sources of Protein Interaction Data

We assembled a highly reliable set of human PPIs from STRING¹⁶, HINT¹⁷ and the Protein Data Bank (PDB)¹⁸ (**Figure 1a**). The respective contributions were 57985 interactions among 10211 proteins from STRING v.9.1 with experimental support and a confidence score higher than 0.4, 24523 interactions involving 8126 proteins from HINT, and 11583 physical interactions between 1653 proteins from the PDB. After eliminating self-interactions, our network comprises of 74699 interactions among 11951 proteins.

2.2 Cancer Genes

We investigated mutations in 125 genes implicated by Vogelstein *et al.*'s¹ as driving tumorigenesis. Of these genes, 123 were present in our human PPI, participating in 7503 interactions. Ninety-seven of the encoded proteins had structural entries in the PDB, and 59 had PDB structures in complex with one or more binding partners, resulting in a total of 169 structurally resolved interaction interfaces (**Figure 1a**).

2.3 Source of Mutation Data

The TCGA mutation data (merged results of MutSig v2.0 and MutSigCV v0.9) was downloaded from the 01/15/2014 Firehose release (<http://gdac.broadinstitute.org>). Only missense mutations were considered in this analysis.

2.4 Mapping Mutations to Protein Structure

2.4.1 From DNA Sequence Position to Structural Position

In order to determine the three-dimensional location of mutated residues, chromosomal coordinates of nsSNVs were mapped onto to PDB coordinates. Chromosomal coordinates were mapped to transcripts in Gencode19¹⁹ using psl format files downloaded from the UCSC Genome Browser²⁰. UniProt proteins were aligned to transcripts in Gencode19 using tblastn²¹ software. Lastly, we performed the Uniprot to PDB mapping with the PDBSWS²² server.

2.4.2 Designating Interface and Core Residues

We classified nsSNVs into two groups depending on their structural location: core or interface. To designate a residue as participating in a protein interaction interface, we used the consensus of interface predictions made by the HotPoint²³ and KFC2²⁴ servers to identify residues in physical contact. We removed incomplete interfaces by discarding interactions with fewer than 5 residues in at least one of the interacting chains. We used NACCESS²⁵ to calculate the accessible surface area (ASA) of all protein residues. Residues with an ASA of 0 were classified as core.

2.4.3 Positioning Mutation Data on Protein Structure

We obtained core residue positions for 97 of the proteins encoded by cancer genes and positions at

interfaces for 169 interactions between these proteins and their partners. In total there were 398 mutations located at either core or interface residues of cancer genes (**Figure 1b**). We made the simplifying assumption that mutations mapping to the same interface are likely to affect it in the same way. Thus we select a single representative mutated residue for each interface. Because we are interested in differential functional consequences of mutations in the same protein, we focused on genes with at least 2 mutations in different locations (i.e. different interfaces or core and interface mutations). This reduced our list to 137 distinct events in 43 cancer genes (**Supplementary Table 1**).

2.5 Network Perturbation

If a mutation occurs in the protein core we assume that all of the protein's interactions are affected (**Figure 2b**), and if it occurs at an interface, only the interactions mediated by the interface are affected (**Figure 2c**). To implement these perturbations, we removed edges from our structurally resolved PPI network corresponding to the affected interactions (**Figure 1c**).

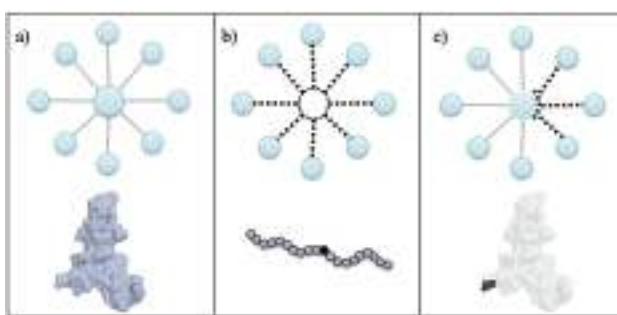


Figure 2. Modeling mutations as network perturbations. a) The unaltered protein-protein interactions of a wild type protein, b) a core mutation has the tendency to destabilize the protein. We depict this phenomenon by removing all edges involving the protein c) an interface mutation may affect some of the interactions of a protein. In this case we remove the potentially affected edges of the protein from the network.

2.6 Network Propagation Algorithm

We used network propagation²⁶ to implicate protein sets most likely to be affected by each mutation (**Figure 1d**). This method has been applied to the related problem of clustering of patients based on somatic mutation profiles by Hofree *et al.*⁷, and uses a random walk (with restarts) according to the function in Eq.(1).

$$\mathbf{F}_{t+1} = \alpha \mathbf{F}_t \mathbf{A} + (1 - \alpha) \mathbf{F}_0 \quad (1)$$

\mathbf{F}_0 is a binary vector with size equal to the number of proteins in the network. Mutated cancer genes are set to 1, representing 'heat sources', while other proteins are initialized to 0. The \mathbf{A} matrix is the degree normalized adjacency matrix of the PPI network. The α parameter affects the distance that the heat signal propagates during the diffusion. The distribution of the propagated values was similar for different α values and the choice of this parameter had limited impact on the results within the range of [0.4-0.7], as was previously reported.²⁶ We used 0.4 as the α parameter.

In order to avoid numerical inaccuracy issues, the propagation algorithm is solved by iterative use of equation (1) until convergence (i.e. the sum of absolute differences between elements of

F_{t+1} and F_t is smaller than 10^{-6}). The algorithm returns F_t , which contains a value for each node in the network proportional to the expected number of times the node is visited during a random walk originating from the heat source, and restarting at the heat source with probability α .

2.7 The Differential Heat Profiles

We performed network propagation for each of the 137 representative mutations separately. For each mutation, we calculated F_t vectors for the unaltered network and the perturbed network. For subsequent analysis steps (protein module detection and functional annotation), we used the differential heat profiles, obtained by subtracting the F_t values for each gene in the unaltered and perturbed networks.

As methods used in this analysis are sensitive to differences in scale differential heat profiles were aggregated into a mutation x gene matrix and quantile normalized using the “preprocessCore” package of Bioconductor²⁷ for R²⁸.

2.8 Sub-network Extraction

We used an approach similar to that used by the HotNet⁵ method to identify altered sub-networks in our global PPI from the differential heat profiles for the 137 mutations (**Figure 1e**). First, each edge was assigned the minimum heat value of the corresponding protein pair. Edges were then sorted by heat value and the top 10th percentile of edges were extracted. Next, we executed our pipeline for 1000 random mutations with similar consequences to those observed in the TCGA data (390 core and 610 interface affecting 1-10 edges). We removed edges that had differential heat scores in the top 10th percentile in over 5% of the random runs as these edges likely resulted from the underlying topology of our PPI network rather than the perturbation of interest. This procedure resulted in a set of connected components for each of the 137 mutations, representing mutation-specific candidate cancer pathway genes.

2.9 Functional Annotation

We used David²⁹ to annotate the gene sets in the mutation-specific connected components from the GO Biological Process data set³⁰. For each cancer gene, functional annotations were divided into those common to all mutations and those specific to particular mutations (**Figure 1f**).

3. Results and Discussions

3.1 A Pipeline to Extract Mutation-Specific Pathways

We constructed a pipeline (**Figure 1**) for mining and annotating cancer related protein sets from somatic mutation data while accounting for mutation-specific network perturbations. We applied this pipeline to analyze mutations observed in 125 frequently mutated cancer genes, where the vast majority of observed mutations are likely to be cancer causing driver mutations. Our pipeline can be applied to mutations in any gene, however for genes not known to drive tumorigenesis, efforts should be made to discriminate between causal driver and non-causal passenger events.

3.2 Mutational Distribution in Cancer Genes

We investigated the spatial distribution of somatic missense mutations on 125 cancer genes using individual crystal structures and co-crystallization of the encoded proteins with interaction partners. We then incorporated these structural data into a larger network of experimentally validated PPIs. We note that because these 125 genes are well studied, there is a positive bias towards data availability relative to other genes in the human interactome. Nonetheless, only 59 of the 125 proteins had structural complexes in the PDB that included an interaction partner. Despite this, co-crystallization of the 59 driver genes with interaction partners covered fully 6% of the experimentally validated protein interactions (**Supplementary Table 2**).

In order to assess the extent of structural diversity of missense mutations observed across known cancer genes, we mapped mutations to core and interface regions. We observed that cancer gene encoded proteins for which co-crystallization structures are available harbor mutations at an average of 2.5 distinct sites (**Supplementary Table 2**). Of core and interface sites observed to harbor mutations, 21% demonstrated tissue specificity (Fisher's Exact test, Bonferroni corrected p-value < 0.05) (**Supplementary Table 3**). In particular, 4 cancer genes showed significant differences in mutation counts at distinct sites in different cancer types. These observations suggest that the physical location of mutations in known cancer genes may have functional significance.

3.3 Determining the Altered Sub-networks and Their Functionality

To identify perturbed network modules in the global network, we applied a HotNet-like method (section 2.8) to the differential heat profiles obtained for missense mutations at protein core or interface residues. We focused on cancer genes with mutations mapping to multiple distinct locations likely to have different functional consequences. Filtering redundant mutations (those occurring at residues in the same protein core or interface), we retained 137 mutations for 43 cancer genes. These 137 events returned an average of 56 altered sub-networks derived from an average of 686 proteins (**Supplementary Table 4**). We annotated the resulting sub-networks from the GO Biological Process database, and found that all 43 cancer genes harbored events that implicated specific functional consequences. Published events were consistent with our functional annotations for sites in APC, ATRX, BRCA1, CBL and HRAS via literature search (**Section 4** and **Supplementary Table 5**). For this purpose we assumed mutagenesis experiments reported for other residues at the same interface or core would be equivalent to the events we modeled.

3.4 HRAS Case Study: Implicated Sub-Networks and Functions

The RAS family oncogenes, KRAS, HRAS and NRAS were among the first discovered oncogenes, and are frequently mutated across a variety of human cancers. These genes regulate cell proliferation, differentiation and survival³¹ via interaction with a number of different protein targets. Amino acid substitutions occurring on these 3 genes disturb signaling through these pathways and lead to tumorigenesis.

In our current network, we have structurally resolved interfaces for HRAS binding to RASA1 and SOS1, but not for KRAS and NRAS. Given the high degree of similarity among RAS proteins, and various experimental findings that support similar functional capabilities³², the model we present here for HRAS likely generalizes to KRAS and NRAS as well.

Among TCGA patients we observe four amino acid substitutions localizing to protein-binding interfaces on RAS proteins (G12 in 6 patients, G13 in 8 patients, A59 in 1 patient and Q61 in 21 patients) and no mutations affecting the protein core (**Figure 3**). By superimposing the RASA1 – HRAS (PDB ID: 1WQ1) and SOS1 – HRAS (PDB ID: 1BKD) complexes, we observed that these two interactors utilize the same binding site on HRAS. Physical distance between residues in co-crystallized structures implicated three residues in interactions with RASA1 (residues 12, 13 and 61) and SOS1 (residues 13, 59 and 61) respectively. This suggests that neighboring residues 12-13 and 59-61 participate in different interactions. Mutations at residues 12 and 13 have been observed to have prognostic and therapeutic differences. For example, KRAS G13 mutated colorectal cancers show some response to cetuximab, while G12 mutated cancers do not respond or may even progress more rapidly³³. We applied our pipeline to each mutated interface, resulting in predictions for mutations altering signaling through RASA1(HRAS G12), signaling through SOS1 (HRAS A59) or both simultaneously(HRAS G13/Q61).

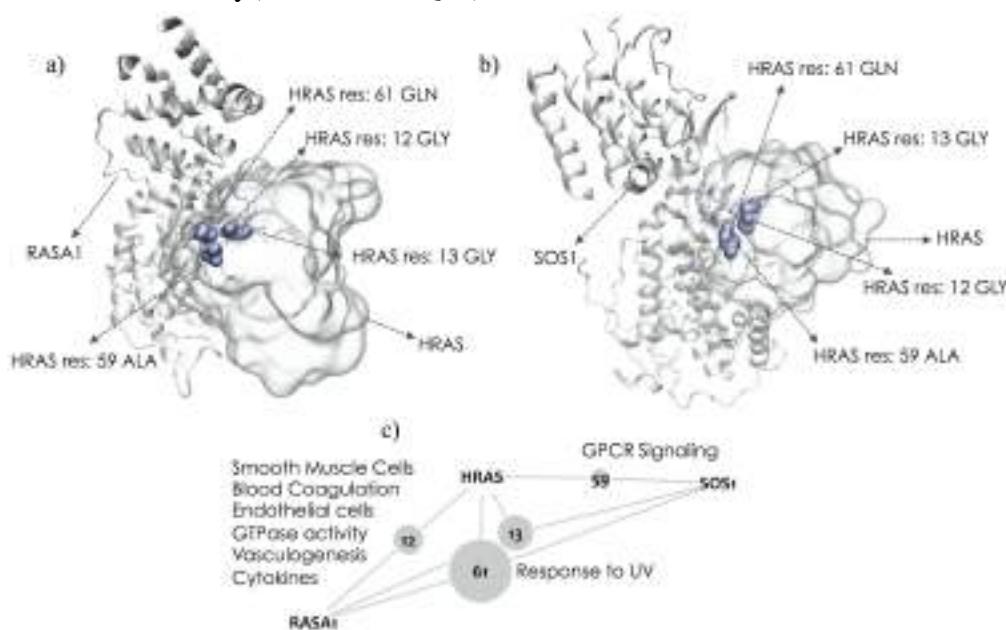


Figure 3. The RASA1–HRAS complex (a) and the SOS1–HRAS complex (b) show that both interactors utilize the same binding site on HRAS. Residues 12, 13, 59 and 61 on HRAS that participate in the interface region of these interactions are highlighted in blue. Residue 12 mediates the HRAS-RASA1, residue 59 mediates the HRAS-SOS1 interaction, and 13 and 61 participate in both interactions. The small network (c) provides a schematic of the residues mediating particular interactions. The grey nodes represent the residues. Their sizes are proportional to the frequency with which they are mutated in the TCGA cohort. Functions predicted to be specifically altered by mutations at each interface are listed on the edges.

3.4.1 HRAS G12 alterations

G12 mutations in HRAS returned protein modules involved in GTPase activity, cytokine production, vasculogenesis, blood coagulation, endothelial cell differentiation/proliferation and smooth muscle cell proliferation/migration (**Supplementary Table 6**). Evidence suggests that these functions may be linked. Inappropriate blood coagulation is frequently observed in cancer

patients and is closely related to tumor growth³⁴. Vasculogenesis, which involves proliferation, migration and remodeling of endothelial cells, have been found to be related with tumor recurrence³⁵. Chemokines play an important role in the behavior of endothelial cells during vessel formation³⁶. Vascular smooth muscle cells provide homeostatic control and protect newly formed vessels against rupture and regression via inhibition of endothelial cell proliferation and migration³⁷. Oncogenic HRAS G12 is known to stimulate chemokine secretion³⁸, cause VGEF activation and endothelial cell apoptosis³⁹, and is thought to be essential for solid tumor maintainence³⁷. Furthermore, VEGF (Entrez Gene ID: 7422), one of the prominent molecules that control vasculogenesis, is present in the protein set (**Supplementary Table 7**) implicated by the HRAS G12 mutation. Amino acid changes at HRAS G12 have been shown to affect the strength of GTPase activity and binding to GTP⁴⁰.

3.4.2 HRAS G13/Q61 alterations

Residues G13 and Q61, which participate in a common binding site on HRAS, mediate interactions with both RASA1 and SOS1 and are frequently mutated in cancer. GO analysis of the protein modules implicated by these residues found functional enrichment for response to UV light (**Supplementary Table 8-9**). Even though there is an extensive body of evidence supporting the connection between UV radiation and melanoma⁴¹⁻⁴³, the exact mechanism remains unclear. Yang *et al*⁴⁴ proposed a possible mechanism that drives melanoma photocarcinogenesis through KRAS Q61 mutagenesis. Besides, it's shown that UV-radiation has a bias towards targeting pyrimidine dimers that more frequently lead to RAS Q61 mutations⁴⁵.

3.4.3 HRAS A59 alterations

Mutations at HRAS A59 exclusively affect the SOS1 interaction. When analyzed with our model, the implicated proteins were involved in GPCR signaling. (**Supplementary Table 10-11**). RAS activation is catalyzed by guanine nucleotide exchange factors (GEFs) which include GPCRs. On the other hand SOS1⁴⁶ acts as a GEF for HRAS. SOS1 forms a complex with GRB2 that is obligatory for GPCR-mediated RAS activation. Since these proteins are tightly bound to each other, when the interaction between SOS1 and HRAS is hindered, it is not unexpected to see GPCR signaling as an altered pathway.

4. Additional Validation

We assessed two mutated residues affecting interactions between APC and KHDRBS1 (res640) and CTNNB1 (res1527) respectively. Consistent with the published finding that the R640G of APC causes exon 14 skipping by disrupting ASF/SF2 binding⁴⁷, our functional annotations included mRNA splice related activities and the implicated protein set included ASF/SF2. In contrast, functional annotations specific to the res1527 perturbation included a number of nervous system related activities, such as “neuron apoptosis” and “negative regulation of neuron differentiation”. In the literature, mutations at codon 1495 which is also at the CTNNB1 interface have been observed in Medulloblastoma⁴⁸.

Mutated residues at positions 220 and 263 map to the core of ATRX and an interface that mediates binding to members of the Histone H3 family, H3F3A and HIST1H3A, respectively.

ATRX has been implicated in chromatin remodeling and regulation of gene expression. The ATRX ADD domain and HP1 are required for ATRX localization to heterochromatin. Mutation E218A reduces pericentromeric localization of ATRX without disturbing the stability of the ADD domain⁴⁹. H3 tails bearing tri-methylated Lys9 (H3-K9) are also required for ATRX localization via the ADD domain⁵⁰. Annotations for residue 263 were specific to histone H4-K acetylation, while the ATRX core mutations, which presumably destabilize the protein, returned histone H3-K9 methylation. This is consistent with the ADD domain being unaffected by the interface mutations.

We evaluated residues affecting BRCA1's interactions with BRIP1 (res1813,res1699) and BARD1 (res96). The BRCA1–BRIP1–TOPBP1 complex is associated with DNA repair during replication and is essential for the S-phase checkpoint in response to collapsed replication forks.⁵¹ The mutated residues affecting BRCA1 and BRIP1 interactions returned related terms including “DNA repair”, “maintenance of fidelity during DNA-dependent DNA replication”, and “DNA replication checkpoint and replication fork protection”. Weakening of the BRCA1–BARD1 interaction due to ionizing radiation leads to the induction of p21 and initiation of the G1/S checkpoint⁵². Our annotations for this included “DNA damage response, signal transduction by p53 class mediator resulting in transcription of p21 class mediator”. In addition, of 10 BRCA1 mutations found to be functional by Carvalho *et al.*⁵³, 7 mapped to interface or core residues with our pipeline.

CBL's interfaces with UBE2D2 (res418, res417, res384) and EGFR (res322) were observed to harbor mutations in TCGA samples. CBL residue R420 is involved in ubiquitination⁵⁴ and the G298E mutation was shown to abolish NFAT activation⁵⁵. Consistent with these findings, our annotations specific to the UBE2D2 interface included terms related to ubiquitination, while annotations specific to the EGFR interaction included T-cell activation and, NFAT activation molecule 1 was present in the protein set.

5. Conclusions

We describe here our efforts to incorporate information about the differential consequences of somatic mutations in the same protein for extracting cancer pathways from large tumor ‘omics data sets. Although there is limited structural knowledge available for PPIs, what exists provides strong evidence for specific functional consequences of mutations at distinct sites within the same protein. Among 59 proteins with sufficient structural data, 43 had mutations with specific functional annotations. Despite the paucity of functionally characterized missense mutations in databases and literature, we were able to find supporting evidence in the literature for mutated sites on 6 genes of the 43 genes. A case study investigating the mutation consequences for distinct interactions of HRAS further highlights that biological processes associated with each event can be specific and may have phenotypic relevance to the patient. For more systematic validation, experimental assays could be designed to validate predictions of our method, guided by the implicated protein sub-networks and the associated annotations. In aggregate, our findings suggest that perturbation to cancer pathways may in fact be mutation-specific and point to the need for analysis methods aware of tumor-specific network topologies.

6. Supplementary Material

http://chianti.ucsd.edu/perm/hcarter/psb2015/Supplementary_Material.docx

References

- 1 Vogelstein, B. *et al.* Cancer genome landscapes. *Science* **339**, 1546-1558, doi:10.1126/science.1235122 (2013).
- 2 Garraway, L. A. & Lander, E. S. Lessons from the cancer genome. *Cell* **153**, 17-37, doi:10.1016/j.cell.2013.03.002 (2013).
- 3 Ideker, T., Ozier, O., Schwikowski, B. & Siegel, A. F. Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics* **18 Suppl 1**, S233-240 (2002).
- 4 Akula, N. *et al.* A network-based approach to prioritize results from genome-wide association studies. *PloS one* **6**, e24220, doi:10.1371/journal.pone.0024220 (2011).
- 5 Vandin, F., Upfal, E. & Raphael, B. J. Algorithms for detecting significantly mutated pathways in cancer. *Journal of computational biology : a journal of computational molecular cell biology* **18**, 507-522, doi:10.1089/cmb.2010.0265 (2011).
- 6 Cerami, E., Demir, E., Schultz, N., Taylor, B. S. & Sander, C. Automated network analysis identifies core pathways in glioblastoma. *PloS one* **5**, e8918, doi:10.1371/journal.pone.0008918 (2010).
- 7 Hofree, M., Shen, J. P., Carter, H., Gross, A. & Ideker, T. Network-based stratification of tumor mutations. *Nature methods* **10**, 1108-1115, doi:10.1038/nmeth.2651 (2013).
- 8 Mosca, R., Ceol, A. & Aloy, P. Interactome3D: adding structural details to protein networks. *Nature methods* **10**, 47-53, doi:10.1038/nmeth.2289 (2013).
- 9 Wang, X. *et al.* Three-dimensional reconstruction of protein networks provides insight into human genetic disease. *Nature biotechnology* **30**, 159-164, doi:10.1038/nbt.2106 (2012).
- 10 David, A., Razali, R., Wass, M. N. & Sternberg, M. J. Protein-protein interaction sites are hot spots for disease-associated nonsynonymous SNPs. *Human mutation* **33**, 359-363, doi:10.1002/humu.21656 (2012).
- 11 Engin, H. B., Guney, E., Keskin, O., Oliva, B. & Gursoy, A. Integrating structure to protein-protein interaction networks that drive metastasis to brain and lung in breast cancer. *PloS one* **8**, e81035, doi:10.1371/journal.pone.0081035 (2013).
- 12 Acuner-Ozbabacan , S. E. *et al.* The structural network of Interleukin-10 and its implications in inflammation and cancer. *BMC Genomics* **15**, S2 (2014).
- 13 Acuner Ozbabacan, S. E., Gursoy, A., Nussinov, R. & Keskin, O. The structural pathway of interleukin 1 (IL-1) initiated signaling reveals mechanisms of oncogenic mutations and SNPs in inflammation and cancer. *PLoS computational biology* **10**, e1003470, doi:10.1371/journal.pcbi.1003470 (2014).
- 14 Zhong, Q. *et al.* Edgetic perturbation models of human inherited disorders. *Molecular systems biology* **5**, 321, doi:10.1038/msb.2009.80 (2009).
- 15 Schuster-Bockler, B. & Bateman, A. Protein interactions in human genetic diseases. *Genome biology* **9**, R9, doi:10.1186/gb-2008-9-1-r9 (2008).
- 16 Franceschini, A. *et al.* STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic acids research* **41**, D808-815, doi:10.1093/nar/gks1094 (2013).
- 17 Das, J. & Yu, H. HINT: High-quality protein interactomes and their applications in understanding human disease. *BMC systems biology* **6**, 92, doi:10.1186/1752-0509-6-92 (2012).
- 18 Bernstein, F. C. *et al.* The Protein Data Bank. A computer-based archival file for macromolecular structures. *European journal of biochemistry / FEBS* **80**, 319-324 (1977).

- 19 Harrow, J. *et al.* GENCODE: the reference human genome annotation for The ENCODE Project. *Genome research* **22**, 1760-1774, doi:10.1101/gr.135350.111 (2012).
- 20 Karolchik, D. *et al.* The UCSC Genome Browser Database. *Nucleic acids research* **31**, 51-54 (2003).
- 21 Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research* **25**, 3389-3402 (1997).
- 22 Martin, A. C. Mapping PDB chains to UniProtKB entries. *Bioinformatics* **21**, 4297-4301, doi:10.1093/bioinformatics/bti694 (2005).
- 23 Tuncbag, N., Keskin, O. & Gursoy, A. HotPoint: hot spot prediction server for protein interfaces. *Nucleic acids research* **38**, W402-406, doi:10.1093/nar/gkq323 (2010).
- 24 Zhu, X. & Mitchell, J. C. KFC2: a knowledge-based hot spot prediction method based on interface solvation, atomic density, and plasticity features. *Proteins* **79**, 2671-2683, doi:10.1002/prot.23094 (2011).
- 25 Hubbard SJ, T. J. NACCESS. *Department of Biochemistry and Molecular Biology, University College London* (1993).
- 26 Vanunu, O., Magger, O., Ruppin, E., Shlomi, T. & Sharan, R. Associating genes and protein complexes with disease via network propagation. *PLoS computational biology* **6**, e1000641, doi:10.1371/journal.pcbi.1000641 (2010).
- 27 Gentleman, R. C. *et al.* Bioconductor: open software development for computational biology and bioinformatics. *Genome biology* **5**, R80, doi:10.1186/gb-2004-5-10-r80 (2004).
- 28 R: A Language and Environment for Statistical Computing (R Foundation for Statistical Computing, Vienna, Austria, 2014).
- 29 Huang da, W., Sherman, B. T. & Lempicki, R. A. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic acids research* **37**, 1-13, doi:10.1093/nar/gkn923 (2009).
- 30 Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature genetics* **25**, 25-29, doi:10.1038/75556 (2000).
- 31 Fernandez-Medarde, A. & Santos, E. Ras in cancer and developmental diseases. *Genes & cancer* **2**, 344-358, doi:10.1177/1947601911411084 (2011).
- 32 Potenza, N. *et al.* Replacement of K-Ras with H-Ras supports normal embryonic development despite inducing cardiovascular pathology in adult mice. *EMBO reports* **6**, 432-437, doi:10.1038/sj.embo.7400397 (2005).
- 33 De Roock, W. *et al.* Effects of KRAS, BRAF, NRAS, and PIK3CA mutations on the efficacy of cetuximab plus chemotherapy in chemotherapy-refractory metastatic colorectal cancer: a retrospective consortium analysis. *The lancet oncology* **11**, 753-762, doi:10.1016/S1470-2045(10)70130-3 (2010).
- 34 Nash, G. F., Walsh, D. C. & Kakkar, A. K. The role of the coagulation system in tumour angiogenesis. *The lancet oncology* **2**, 608-613 (2001).
- 35 Kioi, M. *et al.* Inhibition of vasculogenesis, but not angiogenesis, prevents the recurrence of glioblastoma after irradiation in mice. *The Journal of clinical investigation* **120**, 694-705, doi:10.1172/JCI40283 (2010).
- 36 Gupta, S. K., Lysko, P. G., Pillarisetti, K., Ohlstein, E. & Stadel, J. M. Chemokine receptors in human endothelial cells. Functional expression of CXCR4 and its transcriptional regulation by inflammatory cytokines. *The Journal of biological chemistry* **273**, 4282-4287 (1998).
- 37 Carmeliet, P. Mechanisms of angiogenesis and arteriogenesis. *Nature medicine* **6**, 389-395, doi:10.1038/74651 (2000).

- 38 O'Hayer, K. M., Brady, D. C. & Counter, C. M. ELR+ CXC chemokines and oncogenic Ras-mediated tumorigenesis. *Carcinogenesis* **30**, 1841-1847, doi:10.1093/carcin/bgp198 (2009).
- 39 Chin, L. *et al.* Essential role for oncogenic Ras in tumour maintenance. *Nature* **400**, 468-472, doi:10.1038/22788 (1999).
- 40 Pylayeva-Gupta, Y., Grabocka, E. & Bar-Sagi, D. RAS oncogenes: weaving a tumorigenic web. *Nature reviews. Cancer* **11**, 761-774, doi:10.1038/nrc3106 (2011).
- 41 Tronnier, M., Smolle, J. & Wolff, H. H. Ultraviolet irradiation induces acute changes in melanocytic nevi. *The Journal of investigative dermatology* **104**, 475-478 (1995).
- 42 Zaidi, M. R. *et al.* Interferon-gamma links ultraviolet radiation to melanomagenesis in mice. *Nature* **469**, 548-553, doi:10.1038/nature09666 (2011).
- 43 Bald, T. *et al.* Ultraviolet-radiation-induced inflammation promotes angiotropism and metastasis in melanoma. *Nature* **507**, 109-113, doi:10.1038/nature13111 (2014).
- 44 Yang, G., Curley, D., Bosenberg, M. W. & Tsao, H. Loss of xeroderma pigmentosum C (Xpc) enhances melanoma photocarcinogenesis in Ink4a-Arf-deficient mice. *Cancer research* **67**, 5649-5657, doi:10.1158/0008-5472.CAN-06-3806 (2007).
- 45 Tormanen, V. T. & Pfeifer, G. P. Mapping of UV photoproducts within ras proto-oncogenes in UV-irradiated cells: correlation with mutations in human skin cancer. *Oncogene* **7**, 1729-1736 (1992).
- 46 Buday, L. & Downward, J. Many faces of Ras activation. *Biochimica et biophysica acta* **1786**, 178-187, doi:10.1016/j.bbcan.2008.05.001 (2008).
- 47 Goncalves, V. *et al.* A missense mutation in the APC tumor suppressor gene disrupts an ASF/SF2 splicing enhancer motif and causes pathogenic skipping of exon 14. *Mutation research* **662**, 33-36, doi:10.1016/j.mrfmmm.2008.12.001 (2009).
- 48 Huang, H. *et al.* APC mutations in sporadic medulloblastomas. *The American journal of pathology* **156**, 433-437, doi:10.1016/S0002-9440(10)64747-5 (2000).
- 49 Estermann, S. *et al.* Combinatorial readout of histone H3 modifications specifies localization of ATRX to heterochromatin. *Nature structural & molecular biology* **18**, 777-782, doi:10.1038/nsmb.2070 (2011).
- 50 Kourmouli, N., Sun, Y. M., van der Sar, S., Singh, P. B. & Brown, J. P. Epigenetic regulation of mammalian pericentric heterochromatin in vivo by HP1. *Biochemical and biophysical research communications* **337**, 901-907, doi:10.1016/j.bbrc.2005.09.132 (2005).
- 51 Levran, O. *et al.* The BRCA1-interacting helicase BRIP1 is deficient in Fanconi anemia. *Nature genetics* **37**, 931-933, doi:10.1038/ng1624 (2005).
- 52 Roy, R., Chun, J. & Powell, S. N. BRCA1 and BRCA2: different roles in a common pathway of genome protection. *Nature reviews. Cancer* **12**, 68-78, doi:10.1038/nrc3181 (2012).
- 53 Carvalho, M. A. *et al.* Determination of cancer risk associated with germ line BRCA1 missense variants by functional analysis. *Cancer research* **67**, 1494-1501, doi:10.1158/0008-5472.CAN-06-3297 (2007).
- 54 Makishima, H. *et al.* CBL mutation-related patterns of phosphorylation and sensitivity to tyrosine kinase inhibitors. *Leukemia* **26**, 1547-1554, doi:10.1038/leu.2012.7 (2012).
- 55 Zhang, Z., Elly, C., Qiu, L., Altman, A. & Liu, Y. C. A direct interaction between the adaptor protein Cbl-b and the kinase zap-70 induces a positive signal in T cells. *Current biology : CB* **9**, 203-206 (1999).

BINNING SOMATIC MUTATIONS BASED ON BIOLOGICAL KNOWLEDGE FOR PREDICTING SURVIVAL: AN APPLICATION IN RENAL CELL CARCINOMA

DOKYOON KIM, RUOWANG LI, SCOTT M. DUDEK, JOHN R. WALLACE, MARYLYN D. RITCHIE

*Center for Systems Genomics, Department of Biochemistry and Molecular Biology, Pennsylvania State University, University Park, Pennsylvania, USA
Email: marylyn.ritchie@psu.edu*

Enormous efforts of whole exome and genome sequencing from hundreds to thousands of patients have provided the landscape of somatic genomic alterations in many cancer types to distinguish between driver mutations and passenger mutations. Driver mutations show strong associations with cancer clinical outcomes such as survival. However, due to the heterogeneity of tumors, somatic mutation profiles are exceptionally sparse whereas other types of genomic data such as miRNA or gene expression contain much more complete data for all genomic features with quantitative values measured in each patient. To overcome the extreme sparseness of somatic mutation profiles and allow for the discovery of combinations of somatic mutations that may predict cancer clinical outcomes, here we propose a new approach for binning somatic mutations based on existing biological knowledge. Through the analysis using renal cell carcinoma dataset from The Cancer Genome Atlas (TCGA), we identified combinations of somatic mutation burden based on pathways, protein families, evolutionary conversed regions, and regulatory regions associated with survival. Due to the nature of heterogeneity in cancer, using a binning strategy for somatic mutation profiles based on biological knowledge will be valuable for improved prognostic biomarkers and potentially for tailoring therapeutic strategies by identifying combinations of driver mutations.

Keywords: Somatic mutation, pathway, somatic mutation burden, survival analysis, renal cell carcinoma

1. Introduction

Cancer is a complex and heterogeneous disease, many of which are caused by somatic mutations or structural alterations. Recent meta-dimensional omics data from The Cancer Genome Atlas (TCGA) or the International Cancer Genome Consortium (ICGC) have provided exceptional opportunities to investigate the complex genetic basis of disease for improving the ability to diagnose, treat, and prevent cancer [1,2]. Multiple alterations affecting cancer can be observed directly as somatic mutations or copy number changes, and indirectly as changes in epigenomic, transcriptomic, and proteomic dimensions. In particular, one of the main issues in cancer research is to distinguish between driver mutations and passenger mutations based on somatic mutation profiles. Massive efforts of whole exome/genome sequencing from hundreds to thousands of patients have provided the landscape of somatic genomic alterations in each specific cancer or across several cancer types [3,4].

Driver mutations show strong associations with survival in several different types of cancer [5]. Thus, evaluation of survival models to predict the disease trajectory of cancer patients based on somatic mutation profiles is one of the most imperative foci in the development of prediction models for cancer prognosis. However, due to the heterogeneity of tumors, somatic mutation profiles are exceptionally sparse whereas other types of genomic data such as miRNA or gene

expression contain nearly complete data for all genomic features with quantitative values measured in each patient. Somatic mutations, in contrast, occur with low to intermediate frequency among cancer patients (2-20%). Thus, it is common that patients do not share any somatic mutations even though they have same clinical features such as prognosis [6].

Previously, we proposed a framework for data integration to predict clinical outcomes in ovarian cancer [7]. However, somatic mutation profiles were not appropriate for predicting outcomes due to the sparseness. To overcome this challenge, we developed a new strategy to use somatic mutation profiles by performing a biologically based collapsing/binning of the mutations to look for an accumulation of somatic mutations in specific types of features based on biological knowledge such as pathways. Then, these somatic mutation burden features in specific pathways can be tested for association with cancer outcome such as survival. It may be desirable to focus on identifying driver pathways instead of driver mutations associated with survival because the patterns of altered pathways may be similar even though patients within a cancer subtype might have diverse mutations [8]. The hypothesis is that rather than looking for the shared mutation to be the driver, it is important to look for the shared pathway (or other biological feature) to be the driver. We applied grammatical evolution neural networks to identify not only specific pathways associated with cancer survival, but also interactions/combinations of pathways. In addition, we tested not only pathways as biological features, but also protein families, regulatory regions, and evolutionary conversed regions to test the association between different types of knowledge-based somatic mutation burden and survival. To test the utility of the proposed strategy, we applied our approach on somatic mutation profiles from renal cell carcinoma from TCGA, which is the most common type of kidney cancer.

2. Methods

2.1. Data

Somatic mutations from renal cell carcinoma patients were retrieved from the TCGA (<http://tcga-data.nci.nih.gov/>) on 1 July 2014. Due to the extreme sparseness of somatic mutation profile, we used all classes of somatic mutations generated from the mutation calling conducted by Baylor College of Medicine (BCM) as a mutation annotation format (MAF) and extracted patient mutation profiles where the analyte was a DNA sample from the tumor. Somatic mutations from 417 patients with renal cell carcinoma were retained for subsequent analysis. Based on chromosomal positions, there were 27,194 unique somatic mutations across all patients. As a clinical outcome, survival information was downloaded for 417 patients as well as sex and age for adjusting potential confounding factors when modeling.

2.2. BioBin

BioBin is a flexible collapsing or binning method using biological knowledge to automate the binning of low frequency variants for association tests [9,10]. The main function of BioBin

provides access to comprehensive knowledge-guided multi-level binning. For example, bin boundaries can be formed using genomic locations from: genes, regulatory regions, evolutionary conserved regions, and/or pathways. BioBin uses a built-in database called the Library of Knowledge Integration (LOKI), which is a repository of data assembled from public databases. LOKI contains multiple data resources [11].

Similar to germline low frequency variants, somatic mutations tend to occur with low or intermediate frequency among cancer patients. Thus, we used BioBin for binning somatic mutations based on biological knowledge, such as pathway, in order to overcome the sparseness of somatic mutation profiles. First, we converted MAF to variant call format (VCF) as an input for BioBin (Fig. 1). Then, we applied BioBin to generate KEGG pathway, Pfam, evolutionary conversed region (ECR), and regulatory bin profiles by accumulating somatic mutations in a specific bin. BioBin is open source and available at <http://ritchielab.psu.edu>.

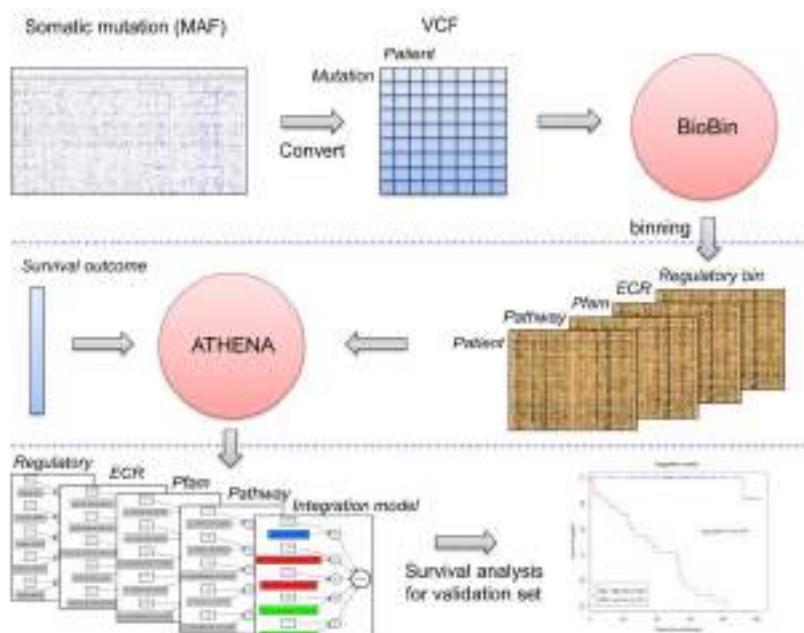


Fig. 1. Illustration of somatic mutation analysis using BioBin and ATHENA for predicting survival.

2.3. ATHENA

The Analysis Tool for Heritable and Environmental Network Associations (ATHENA) is a multi-functional software package designed to perform three essential functions to determine the meta-dimensional models of complex disease: (1) performing feature selections from categorical or continuous independent features; (2) building additive and interaction models that explain or predict categorical or continuous clinical outcomes; (3) interpreting the candidate models for use in further translational bioinformatics [7,12]. For this analysis, we used Grammatical Evolution Neural Networks (GENN) as the modeling component. ATHENA is open source and available at <http://ritchielab.psu.edu>.

2.4. Grammatical Evolution Neural Networks (GENN)

Various computational methods have been developed to identify non-linear interactions between genomic variables that have small or large main effects such as a multi-factor dimensionality reduction (MDR) [13]. However, MDR performs an exhaustive search of all possible combination of interacting loci to generate multi-locus predictor models. The search space increases exponentially with the number of variables and became infeasible when integrating meta-dimensional genomics data. Thus, stochastic methods using evolutionary algorithm have been developed and shown to utilize the full dimensionality of the data without exhaustively evaluating all possible combinations of variables [14,15].

Artificial Neural Network (ANN) is a flexible and robust machine learning technique designed to imitate neurons in the brain to solve complex problems. ANN is a good candidate for identifying complex and non-linear interactions that influence variance in an outcome of interest. Generally, the method for applying ANN to a classification problem is to use gradient descent algorithm such as backpropagation to fit the weights of the network given input variables and network architecture. However, the variables and network architecture are not known a priori. In order to simultaneously optimize the input variables, weights, and network structures, evolutionary algorithm approaches have been developed and applied [15,16]. Genetic programming, a specialization of genetic algorithms, is an evolutionary algorithm-based method that uses “survival of the fittest” to evolve optimal solutions from a population of random solutions [17]. Grammatical evolution is a more flexible version of genetic programming since it can also evolve functional solutions, or computer program, via grammar rules [15]. The details of the grammar rules were described in a previous study [15]. The GENN algorithm is briefly described as follows:

- (1) The data is divided into five parts for five cross validation with 4/5 for training and 1/5 for testing.
- (2) A random population of binary strings is generated to be ANNs using a Backus-Naur grammar. The total population is divided into demes as sub-populations across a user-defined number of CPUs for parallelization.
- (3) All ANNs are evaluated with training data, and the solutions with the lowest prediction errors are selected for crossover and reproduction. The new population is composed of mutated original solutions and new random solutions.
- (4) Step 3 is repeated for a set number of generations. Migration of best solutions also occurs between demes during evolution for a pre-defined number of times.
- (5) The best solution at the final generation is tested using the remaining 1/5 test dataset and fitness is recorded.
- (6) Steps 2-5 are repeated four more times, each time using a different 4/5 of the training data and 1/5 of testing data.

2.5. Survival fitness function

The goal of this study is to predict censored survival outcome based on somatic mutation burden generated from BioBin. In general, it is difficult to directly predict raw survival data via measuring of goodness-of-fit, such as R^2 , due to the censored observations. Thus, an appropriate measure of goodness-of-fit should be required for predicting censored survival data. Martingale residuals are defined as the difference between the cumulative hazards assigned to an individual i with failure time t_i and its observed status, $\delta_i = 0$ censored, $\delta_i = 1$ event [18]. Thus, martingale residuals could be intuitively interpreted as the surplus deaths. Martingale residuals are calculated from the fitted Cox model as

$$M_i = \delta_i - \Lambda(t_i) \quad (1)$$

where Λ is a cumulative hazard function [18].

According to the model, the result of cumulative hazard function reflects the number of expected death events per individuals failing at t_i . The range of martingale residuals is between negative infinity and 1 because the cumulative hazard function does not have upper limit. However, the sum of all martingale residuals is zero. Each patient with a negative martingale residual is interpreted as a good prognosis, whereas one with a positive martingale residual is interpreted as a poor prognosis. The martingale residual of each patient is obtained from the reduced model with no genomic effects from somatic mutations. Thus, martingale residuals can be used as a new continuous outcome since they reflect the unexplained portion beyond what is explained by the adjusted clinical covariates excluding the genomic features [18]. In addition, another advantage of martingale residual is that the model can be adjusted by potential confounders such as sex and age. As a proof of concept, we adjusted for age and sex when calculating martingale residuals using *survival* R package.

Since the distribution of martingale residuals is more exponentially shaped, the assumption of R^2 , which has normally distributed residuals, is not satisfied. Thus, a new fitness function was proposed for measuring the mean absolute differences (MAD) between observed martingale residuals (M_i) and predicted martingale residuals ($M_{i|x}$) from GENN with genomic covariate vector x [19]. The new fitness function is formulated as follows:

$$MAD = \frac{\sum_i |M_i - M_{i|x}|}{\sum_i |M_i|} \quad (2)$$

$$Fitness\ function = 1 - MAD \quad (3)$$

The output of the MAD fitness function will be from 0 to 1. The model with 1 fitness score represents the best predictive model whereas the one with 0 fitness score means the worst predictive model. We used MAD for the subsequent experiments.

2.6. Experiment setup

Figure 1 shows the overview of the analysis pipeline, which consists of a binning step using

BioBin and a modeling step using ATHENA. After converting MAF to VCF, BioBin was used to generate pathway, Pfam, ECR, regulatory bin profiles. Then, we used ATHENA to build additive/interaction models associated with survival. Martingale residuals and each bin profile can be used as an input for ATHENA. For building GENN models, we randomly split the input dataset into two groups, 4/5 dataset ($n=333$) for learning models and 1/5 dataset ($n=84$) for the validation. This is independent of the cross-validation (CV) procedure. The CV procedure was performed on the learning dataset which is 4/5 of the total dataset. Based on GENN results from 5-fold CV, the features from each model across 5 CVs were selected, and then, we reran GENN using selected features to generate the final model from entire training dataset. Lastly, the final GENN model can be used to predict survival from the validation dataset. To avoid over-fitting, the validation dataset was not used for the entire learning step. Table 1 shows the GENN parameters for the analysis. Based on the output of the final GENN model as predicted martingale residuals, the validation dataset was divided into two sub groups, low-risk group and high-risk group, by the median threshold of predicted martingale residuals. Then, survival analysis was performed using *survival* R package.

Table 1. GENN parameter settings

Parameter	Value
Number of demes (CPUs)	20
Population size/ Deme	5,000
Number of generations	1,000
Number of migrations	20
Probability of crossover	0.9
Probability of mutation	0.01
Fitness function	1 – MAD

3. Results and Discussion

3.1. Binning somatic mutations using BioBin

To predict survival based on somatic mutation burden, BioBin was used to generate KEGG pathway, Pfam, ECR, and regulatory bin profiles. Somatic mutation burden analysis can be biased when using bins consisting of extremely small number of mutations, thus bins from KEGG pathway, Pfam, ECR, and regulatory regions with more than 10 mutations were selected for the further study. The total number of KEGG pathway, Pfam, ECR, and regulatory bins were 272, 922, 250, and 41, respectively. Since somatic mutation profiles were conducted by whole-exome sequencing, regulatory bin profiles had a relatively small number of bins compared to other bins. Figure 2 shows the difference of sparseness between raw somatic mutation profiles and pathway bin profiles.

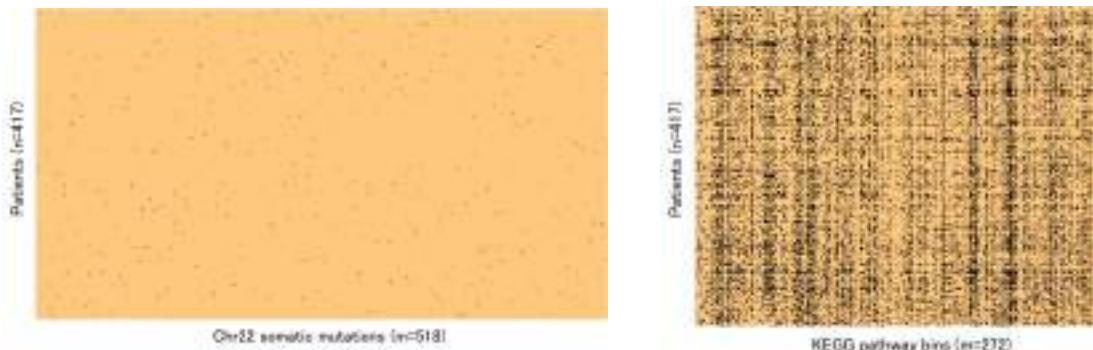


Fig. 2. Difference of sparseness between raw somatic mutation profiles and KEGG pathway bin profiles. For somatic mutation profiles, mutations from chromosome 22 were extracted for generating the heatmap figure. Each black dot represents the presence of either somatic mutation in the left heatmap or mutation burden in a pathway in the right heatmap.

3.2. GENN modeling for somatic mutation burden

A simulation study was conducted to demonstrate the validity of the proposed survival fitness function and martingale residuals as a new outcome for predicting survival (data not shown) [Kim et al., submitted]. According to the results from the simulation data, martingale residuals performed well as a new outcome in terms of finding true survival genes and limited false positives using GENN. Next, somatic mutation profiles in renal cell carcinoma were analyzed to identify additive/interaction models based on knowledge-based somatic mutation burden. After generating pathway, Pfam, ECR, and regulatory bins using BioBin, GENN models were trained to predict survival from the validation dataset. The final model of GENN is the evolved neural network with optimized input variables, weights, and network structure to identify additive or interaction models that predict survival outcome. Figure 3 shows the best GENN models from each bin profile: KEGG pathway, Pfam, ECR, and regulatory bins, respectively. Finally, the final GENN model was used to predict survival from the validation dataset, which consisted of 84 patients. The fitness scores from the validation dataset for each of the best models with pathway, Pfam, ECR, and regulatory bin profiles were 0.641, 0.67, 0.665, and 0.654, respectively (Fig 3 and Table 2). Among four different bin profiles, Pfam bin profiles showed the best performance for predicting survival.

Table 2. Performance comparison between different types of bin profiles. Performance was measured from the validation dataset.

GENN model	1 - MAD	Permutation p-value
KEGG pathway bins	0.641	0.602
Pfam bins	0.67	0.108
ECR bins	0.665	0.2
Regulatory bins	0.654	0.423
Integration	0.685	0.026

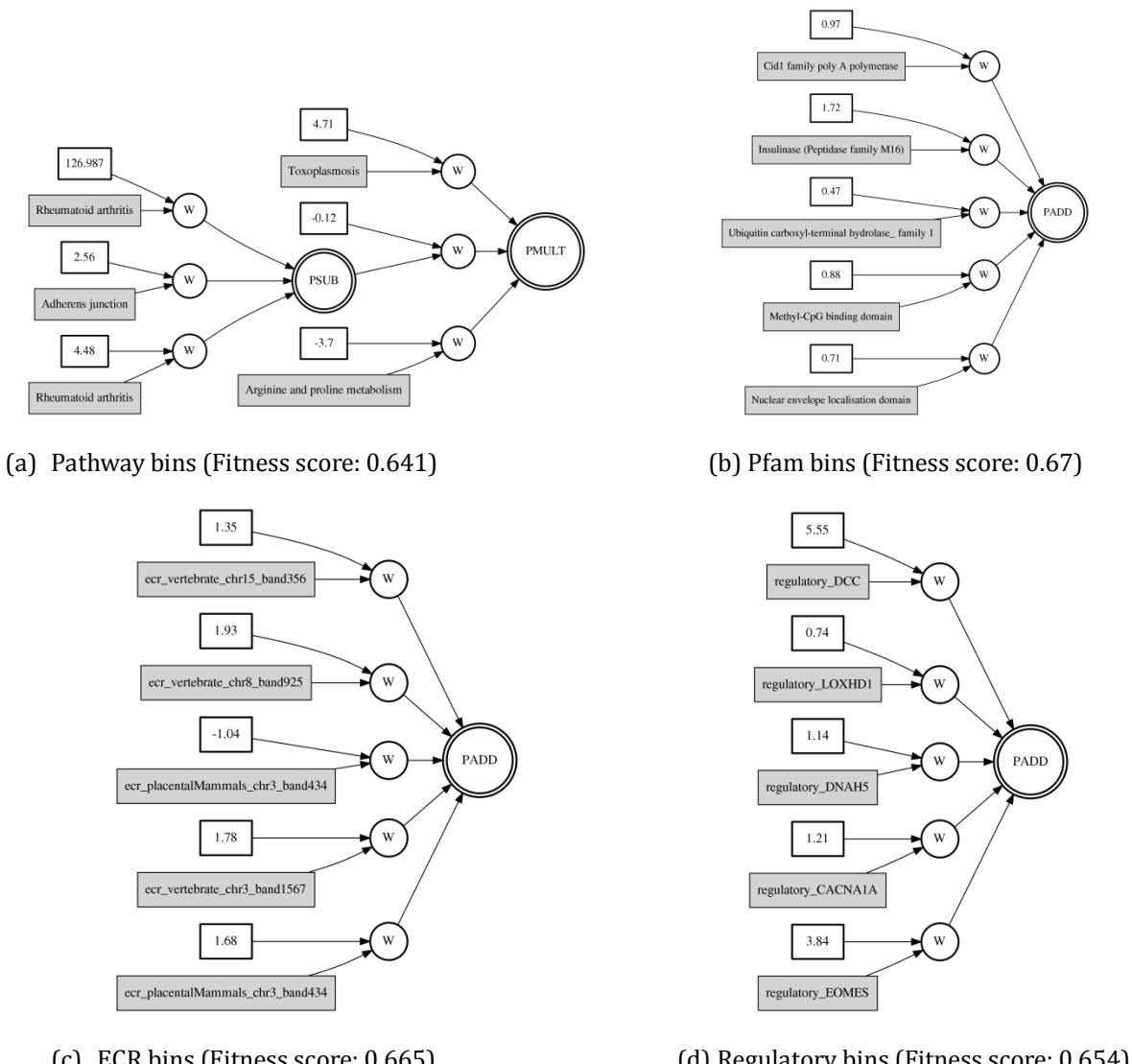


Fig. 3. Best GENN models from each knowledge-based somatic mutation profiles. PSUB, PMUL, and PADD are a subtraction, multiplication, and addition activation node, respectively. Knowledge features such as pathway, Pfam, ECR, regulatory regions are shown in the gray boxes. (a) pathway-based mutation profiles (b) Pfam-based mutation profiles (c) ECR-based mutation profiles (d) regulatory-based mutation profiles

To build an interaction model between different knowledge-guided bins associated with survival in renal cancer, we integrated pathway, Pfam, ECR, and regulatory bin profiles. The final integration model was generated using GENN with variables from the best models of each bin profile. The final integration model was also used to predict survival for the validation dataset. In terms of predictive power, the integration model showed the best performance with a fitness score of 0.685 (Table 2). The selected features in the final integration model are methyl-CpG binding domain and insulinase (Peptidase family M16) from Pfam bin profile, ecr_vertebrate_chr3_band1567 and

ecr_vertebrate_chr15_band356 from ECR bin profiles, and regulatory_EOMES from regulatory bin profiles (Fig. 4). To test the statistical significance of each GENN model, permutation testing was performed. The survival outcome for the validation dataset was randomly permuted 1000 times and permutation p-values of each GENN model were obtained from the 1000 random validation sets (Table 2). The integration model showed a significant result ($P = 0.026$) while other GENN models were not significant based on permutation testing. In addition, survival analysis was performed for two sub groups, low-risk and high-risk groups, which were divided by a median threshold of predicted martingale residuals for the validation dataset. Kaplan-Meier analysis showed that the two groups generated from the integration model were significantly different based on survival (Fig. 5).

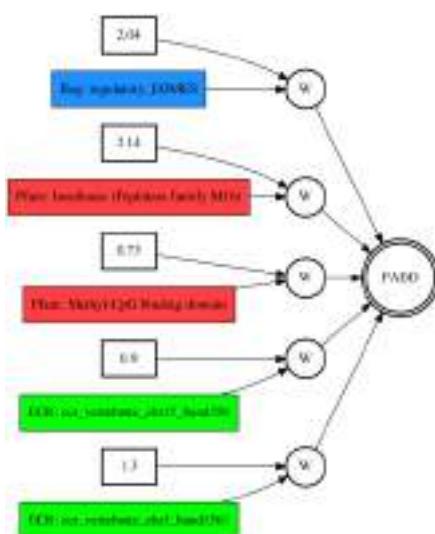


Fig. 4. Integration model containing variables from different knowledge-based mutation profiles. Red, green, and blue boxes represent Pfam, ECR, and regulatory region features, respectively. PADD represents an addition activation node. A fitness score of the integration model was 0.685.

3.3. Biological interpretation

Four pathways, arginine and proline metabolism, adherens junction, toxoplasmosis, and rheumatoid arthritis, were found in the GENN models. Adherens junctions are one of the most relevant junctional complexes in the kidney epithelium and adherens junction disruption is associated with cell proliferation, invasion, and angiogenesis in renal cell carcinoma [20]. In addition, arginine and proline metabolism has been shown to be important in renal cell carcinoma using proteomic and metabolic profiles [21]. Interestingly, rheumatoid arthritis was identified as one of the features in the final model. Associations between rheumatoid arthritis and kidney cancer have been reported in many studies. For example, infliximab, anti-tumor necrosis factor α (TNF- α) antibody, is licensed for use in rheumatoid arthritis and TNF- α might be a therapeutic target in renal cell carcinoma [22]. Notably, the top pathway model showed complex and non-linear interactions between pathways associated with survival. This suggests it is important to consider

interactions associated with survival, which might have crucial roles in molecular pathogenesis, progression, and prognosis of renal cell carcinoma, that are not easily detected by traditional pathway analysis approaches. For Pfam model, a combination of methyl-CpG binding domain, cid1 family poly A polymerase, ubiquitin carboxyl-terminal hydrolase family 1, insulinase (peptidase family M16), and nuclear envelope localization domain was found to be associated with survival. In particular, the epigenetic silencing of cancer-related genes such as *ABCG2* has been shown to be associated with renal cell carcinoma via being mediated through recruitment of a group of proteins, called methyl-CpG binding domain (*MBD*) [23]. Epigenetic control of the ubiquitin carboxyl terminal hydrolase family 1, which plays an important role in cell growth and differentiation, can be disturbed in renal cell carcinoma [24]. Several evolutionary conversed regions were also selected from ECR model. Many genes located in selected evolutionary conversed regions such as *BAP1*, *EIF4G1*, *EBAG9*, or *FBN1* were found as import genes involved in several cancers. In particular, *BAP1* loss defines a new class of renal cell carcinoma [25]. In addition, somatic mutation burden in the regulatory regions of *CACNA1A*, *LOXHD1*, *DNAH5*, *DCC*, or *EOMES* might play a functionally significant role in renal cell carcinoma survival. In the integration model, where we used multiple knowledge sources, methyl-CpG binding domain and insulinase (Peptidase family M16) from Pfam model, chr3_band1567 and chr15_band356 from ECR model, and *EOMES* from regulatory model were selected. Combination of somatic mutation burden based on multiple biological knowledge sources might reflect the complex molecular pathogenesis and progression of renal cell carcinoma.

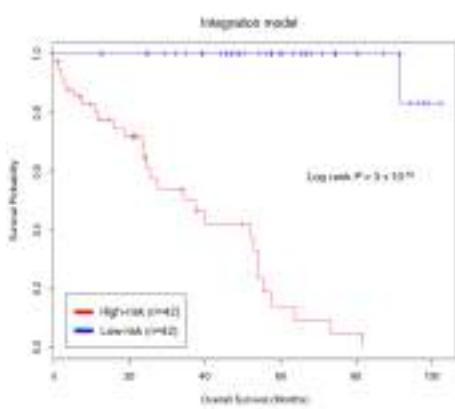


Fig. 5. Kaplan-Meier survival plots for the validation dataset. Validation dataset was divided into high-risk and low-risk groups based on a median value of predicted outputs from the integration model.

4. Conclusions

In this study, we proposed a new approach of binning somatic mutation based on biological knowledge for predicting survival in order to overcome the extreme sparseness of somatic mutation profiles. Through the analysis using renal cell carcinoma dataset, we identified interaction/combinations of somatic mutation burden based on pathway, protein families,

evolutionary conversed regions, and regulatory regions associated with survival. Knowledge guided binning/collapsing somatic mutations dramatically reduce the sparseness of profiles as well as search space, from 27,194 mutations to 272 pathways. In terms of predictive power, some of the GENN models were not significant based on the permutation test. However, it might be due to the fitness function based on mean absolute difference. Even though survival outcome of 84 patients were shuffled, a mean difference between observed martingale residuals and predicted martingale residuals could not be too large. In addition, a small number of patients in the validation dataset limit our power as sample size is often a limiting factor. Improving the methodology by incorporating directionality of the pathway when binning somatic mutations could also potentially increase the predictive power of pathway-based approach. Notably, the predictive power of the integration model outperformed other models from single types of biological knowledge sources. These results suggest that each biological knowledge source can be complementary to the prediction power of survival because each knowledge source has its specific biological context. Furthermore, the survival analysis for the validation dataset demonstrated that somatic mutation burden based on biological knowledge showed significant associations with cancer prognosis in renal cell carcinoma.

The present study underpins our on-going work. First, not only somatic mutations but also germline mutations can be regarded as important genomic features that are associated with cancer outcomes [5]. Thus, as one of promising future works, it would be valuable to combine both types of mutations to investigate the associations with cancer outcomes. It would be also interesting to investigate whether known somatic mutations influence the models. In addition, the proposed approach could be applied to explore associations with other cancer clinical outcomes such as stage, grade, recurrence, or metastasis. Furthermore, the current approach by biological-based bins such as pathways may unnecessarily inject noise into the model. It would be intriguing to apply an adequate filtering step to the mutations in future studies. Due to the nature of heterogeneity in cancer, using a binning strategy for somatic mutation profiles based on biological knowledge will be valuable for improved prognostic biomarkers and tailoring therapeutic strategies by identifying interactions/combinations of driver mutations.

Acknowledgments

This work was funded by NIH grant R01 LM010040, NHLBI grant U01 HL065962, and CTSI: UL1 RR033184-01. This work is also supported by a grant with the Pennsylvania Department of Health using Tobacco CURE Funds.

References

1. Cancer Genome Atlas Research N (2013) Comprehensive molecular characterization of clear cell renal cell carcinoma. *Nature* 499: 43-49.
2. International Cancer Genome Consortium (2010) International network of cancer genome projects. *Nature* 464: 993-998.
3. Cancer Genome Atlas Research N, Weinstein JN, Collisson EA, Mills GB, Shaw KR, et al. (2013) The Cancer Genome Atlas Pan-Cancer analysis project. *Nat Genet* 45: 1113-1120.

4. Brennan CW, Verhaak RG, McKenna A, Campos B, Noushmehr H, et al. (2013) The somatic genomic landscape of glioblastoma. *Cell* 155: 462-477.
5. Yang D, Khan S, Sun Y, Hess K, Shmulevich I, et al. (2011) Association of BRCA1 and BRCA2 mutations with survival, chemotherapy sensitivity, and gene mutator phenotype in patients with ovarian cancer. *JAMA* 306: 1557-1565.
6. Lawrence MS, Stojanov P, Polak P, Kryukov GV, Cibulskis K, et al. (2013) Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* 499: 214-218.
7. Kim D, Li R, Dudek SM, Ritchie MD (2013) ATHENA: Identifying interactions between different levels of genomic data associated with cancer clinical outcomes using grammatical evolution neural network. *BioData Min* 6: 23.
8. Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA, Jr., et al. (2013) Cancer genome landscapes. *Science* 339: 1546-1558.
9. Moore CB, Wallace JR, Frase AT, Pendergrass SA, Ritchie MD (2013) BioBin: a bioinformatics tool for automating the binning of rare variants using publicly available biological knowledge. *BMC Med Genomics* 6 Suppl 2: S6.
10. Moore CB, Wallace JR, Wolfe DJ, Frase AT, Pendergrass SA, et al. (2013) Low frequency variants, collapsed based on biological knowledge, uncover complexity of population stratification in 1000 genomes project data. *PLoS Genet* 9: e1003959.
11. Pendergrass SA, Frase A, Wallace J, Wolfe D, Katiyar N, et al. (2013) Genomic analyses with biofilter 2.0: knowledge driven filtering, annotation, and model development. *BioData Min* 6: 25.
12. Holzinger ER, Dudek SM, Frase AT, Pendergrass SA, Ritchie MD (2013) ATHENA: the analysis tool for heritable and environmental network associations. *Bioinformatics*.
13. Ritchie MD, Hahn LW, Roodi N, Bailey LR, Dupont WD, et al. (2001) Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *Am J Hum Genet* 69: 138-147.
14. Ritchie MD, White BC, Parker JS, Hahn LW, Moore JH (2003) Optimization of neural network architecture using genetic programming improves detection and modeling of gene-gene interactions in studies of human diseases. *BMC Bioinformatics* 4: 28.
15. Motsinger-Reif AA, Dudek SM, Hahn LW, Ritchie MD (2008) Comparison of approaches for machine-learning optimization of neural networks for detecting gene-gene interactions in genetic epidemiology. *Genet Epidemiol* 32: 325-340.
16. Turner SD, Dudek SM, Ritchie MD (2010) ATHENA: A knowledge-based hybrid backpropagation-grammatical evolution neural network algorithm for discovering epistasis among quantitative trait Loci. *BioData Min* 3: 5.
17. Ritchie MD, Motsinger AA, Bush WS, Coffey CS, Moore JH (2007) Genetic Programming Neural Networks: A Powerful Bioinformatics Tool for Human Genetics. *Appl Soft Comput* 7: 471-479.
18. Therneau TM, Grambsch PM, Fleming TR (1990) Martingale-Based Residuals for Survival Models. *Biometrika* 77: 147-160.
19. Müller M (2004) Goodness-of-fit criteria for survival data. *Sonderforschungsbereich Paper* 382.
20. Peruzzi B, Athauda G, Bottaro DP (2006) The von Hippel-Lindau tumor suppressor gene product represses oncogenic beta-catenin signaling in renal carcinoma cells. *Proc Natl Acad Sci U S A* 103: 14531-14536.
21. Perroud B, Lee J, Valkova N, Dhirapong A, Lin PY, et al. (2006) Pathway analysis of kidney cancer using proteomics and metabolic profiling. *Mol Cancer* 5: 64.
22. Harrison ML, Obermueller E, Maisey NR, Hoare S, Edmonds K, et al. (2007) Tumor necrosis factor alpha as a new target for renal cell carcinoma: two sequential phase II trials of infliximab at standard and high dose. *J Clin Oncol* 25: 4542-4549.
23. To KK, Zhan Z, Bates SE (2006) Aberrant promoter methylation of the ABCG2 gene in renal carcinoma. *Mol Cell Biol* 26: 8572-8585.
24. Seliger B, Handke D, Schabel E, Bukur J, Lichtenfels R, et al. (2009) Epigenetic control of the ubiquitin carboxyl terminal hydrolase 1 in renal cell carcinoma. *J Transl Med* 7: 90.
25. Pena-Llopis S, Vega-Rubin-de-Celis S, Liao A, Leng N, Pavia-Jimenez A, et al. (2012) BAP1 loss defines a new class of renal cell carcinoma. *Nat Genet* 44: 751-759.

TOPOLOGICAL FEATURES IN CANCER GENE EXPRESSION DATA

S. LOCKWOOD

*School of Electrical Engineering and Computer Science, Washington State University,
Pullman, WA 99164, U.S.A.
E-mail: svetlana.lockwood@email.wsu.edu*

B. KRISHNAMOORTHY

*Department of Mathematics, Washington State University,
Pullman, WA 99164, U.S.A.
E-mail: bkrishna@math.wsu.edu*

We present a new method for exploring cancer gene expression data based on tools from algebraic topology. Our method selects a small relevant subset from tens of thousands of genes while *simultaneously* identifying nontrivial higher order topological features, i.e., holes, in the data. We first circumvent the problem of high dimensionality by *dualizing* the data, i.e., by studying genes as points in the sample space. Then we select a small subset of the genes as landmarks to construct topological structures that capture persistent, i.e., topologically significant, features of the data set in its first homology group. Furthermore, we demonstrate that many members of these loops have been implicated for cancer biogenesis in scientific literature. We illustrate our method on five different data sets belonging to brain, breast, leukemia, and ovarian cancers.

Keywords: Persistent homology; Cancer; High-dimensional data.

1. Introduction

During the past few decades, science has made great progress in understanding the biology of cancer [1, 2]. The latest technological tools allow assaying tens of thousands of genes simultaneously, providing large volumes of data to search for cancer biomarkers [3, 4]. Ideally, scientists would like to extract some qualitative signal from this data in the hope to better understand the underlying biological processes. At the same time it is desirable that the extracted signal is robust in the presence of noise and errors while effectively describing the dataset [5, 6].

One suite of methods which have enjoyed increased level of success in recent years is based on concepts from the mathematical field of algebraic topology, in particular, *persistent homology* [7–9]. The key benefits of using topological features to describe data include coordinate-free description of shape, robustness in the presence of noise and invariance under many transformations, as well as highly compressed representations of structures [10]. Analysis of the homology of data allows detection of high-dimensional features based on *connectivity*, such as loops and voids, which could not be detected using traditional methods such as clustering [8, 11]. Further, identifying the most *persistent* of such features promises to pick up the significant shapes while ignoring noise [8]. Such analysis of the stable topological features of the data could provide helpful insights as demonstrated by several studies, including some recent ones on cancer data [12, 13].

1.1. Our contributions

We present a new method of topological analysis of various cancer gene expression data sets. Our method belongs to the category of exploratory data analysis. In order to more efficiently handle the huge number of genes whose expressions are recorded in such data sets (typically in the order of tens of thousands), we transpose the data and analyze it in its *dual space*, i.e., with each gene represented in the much lower dimensional (in the order of a few hundred) sample space. We then sample critical genes as guided by the topological analysis. In particular, we choose a small subset (typically 120–200) of genes as *landmarks* [14], and construct a family of nested simplicial complexes, indexed by a proximity parameter. We observe topological features (loops) in the first homology group (H_1) that remain significant over a large range of values of the proximity parameter (we consider small loops as topological noise). By repeating the procedure for different numbers of landmarks, we select stable features that persist over large ranges of both the number of landmarks and the proximity parameter. We then further analyze these loops with respect to their membership, working under the hypothesis that their topological connectivity could reveal functional connectivity. Through the search of scientific literature, we establish that many loop members have been implicated in cancer biogenesis. We applied our methodology to five different data sets from a variety of cancers (brain, breast, ovarian, and acute myeloid leukemia (AML)), and observed that in each of the five cases, many members of the significant loops in H_1 have been identified in the literature as having connections to cancer.

Our method is capable of identifying geometric properties of the data that cannot be found by traditional algorithms such as clustering [15, 16]. By employing tools from algebraic topology, our method goes beyond clustering and detects connected components around holes (loops) in the data space. The shown methodology is also different from techniques such as graph [17, 18] or manifold learning [19–23]. Graph algorithms, while identifying connectivity, miss wealth of information beyond clustering. Manifold learning algorithms assume that the data comes from an intrinsically low-dimensional space and their goal is to find a low-dimensional embedding. We do not make such assumptions about the data.

1.2. Related work

Several applications of tools from algebraic topology to analyze complex biological data from the domain of cancer research have been reported recently. DeWoshkin et al. [12] used computational homology for analysis of comparative genomic hybridization (CGH) arrays of breast cancer patients. They analyzed DNA copy numbers by looking at the characteristics of H_0 group. Using *Betti₀* (β_0) numbers, which are the ranks of the zeroth homology groups (H_0), their method was able to distinguish between recurrent and non-recurrent patient groups.

Likewise, Seeman et al. [13] applied persistent homology tools to analyze cancer data. Their algorithm starts with a set of genes that are preselected using the *nondimensionalized standard deviation* metric [13]. Then, by applying persistent homology analysis to the H_0 group, the patient set is recursively subdivided to yield three subgroups with distinct cancer types. By inspecting the cluster membership, a core subset of genes is selected which allows sharper differentiation between the cancer subtypes.

Another example of topological data analysis is the work of Nicolau et al. [24]. Their method termed *progression analysis of disease* (PAD) is applied to differentiate three subgroups of breast cancer patients. PAD is a combination of two algorithms – disease-specific genome analysis (DSGA) [25] and the topology-based algorithm termed *Mapper* [26]. First, DSGA transforms the data by decomposing it into two components – the disease component and the healthy state component, where the disease component is a vector of residuals from a *Healthy State Model* (HSM) linear fit. A small subset of genes that show a significant deviation from the healthy state are retained and passed on to Mapper, which applies a specified filter function to reveal the topology of the data. Mapper identified three clusters corresponding to ER+, ER-, and normal-like subgroups of breast cancer. This work is somewhat different from the previous two papers mentioned above because it does not explicitly analyze features of any of the homology groups.

All studies mentioned above utilized β_0 numbers, thus performing analyses that are topologically equivalent to clustering. In contrast, our method relies on β_1 numbers (ranks of H_1 groups). One can think of β_1 numbers characterizing the loops constructed from connected components (genes) around “holes” in the data. The underlying idea is that connections around holes may imply connections between the participating genes and biological functions. Also, most of other methods use some data preprocessing to limit the initial pool of candidate genes. Our method selects the optimal number of genes as part of the analysis itself.

2. Mathematical background

We review some basic definitions from algebraic topology used in our work. For details, refer one of the standard textbooks [27, 28]. Illustrations of simplices, persistent homology, and identification of topological features from landmarks are available in the literature [8, 14, 34].

2.1. Simplices and simplicial complexes

Topology represents the shape of point sets using combinatorial objects called simplicial complexes. Consider a finite set of points in \mathbb{R}^n . More generally, the space need not be Euclidean. We just need a unique pairwise distance be defined for every pair of points. The building blocks of the combinatorial objects are *simplices*, which are made of collections of these points.

Formally, the convex hull of $k+1$ affinely independent points $\{v_0, v_1, \dots, v_k\}$ is a k -simplex. The dimension of the simplex is k , and v_j s are its vertices. Thus, a vertex is a 0-simplex, a line segment connecting two vertices is a 1-simplex, a triangle is a 2-simplex, and so on. Observe that each p -simplex σ is made of lower dimensional simplices, i.e., k -simplices τ with $k \leq p$. Here, τ is called a *face* of σ , denoted $\tau \subset \sigma$. A collection of simplices satisfying two regularity properties forms a *simplicial complex*. The first property is that each face of every simplex in a simplicial complex K is also in K . Second, each pair of simplices in K intersect in a face of both, or not at all. Due to these properties, algorithms to study shape and topology run much more efficiently on the simplicial complex than on the original point set.

To construct a simplicial complex on a given point set, one typically considers balls of a given diameter ϵ (called ϵ -ball) centered at each point. The two widely studied complexes of this form are the Čech and the Vietoris-Rips complexes. A k -simplex is included in the Čech

complex if there exists an ϵ -ball containing all its $k + 1$ vertices. Such a simplex is included in the Vietoris-Rips complex R_ϵ if each pair of its vertices is within a distance ϵ . As such, Vietoris-Rips complexes are somewhat easier to construct, since we only need to inspect pairwise, and not higher order, distances.

However, both the Čech and the Vietoris-Rips complexes have as vertex set all of the points in the data. Such complexes are computationally intensive for datasets of tens of thousands of points. The feasible option is to work with an approximation of the topological space of interest [14]. The key idea is to select only a small subset of points (landmarks), while the rest of points serve as *witnesses* to the existence of simplices. Termed *witness complexes*, such complexes have a number of advantages. They are easily computed, adaptable to arbitrary metrics, and do not suffer from the curse of dimensionality. They also provide a less noisy picture of the topological space. We use the *lazy Witness complex*, in which conditions for inclusion are checked only for pairs and not for higher order groups of points [14], analogous to the distinction between the constructions of Vietoris-Rips and Čech complexes.

We employ the heuristic landmark selection procedure called *sequential maxmin* to select a representative set of landmark points [14, 29, 30]. The first landmark is selected randomly from the point set S . Then the algorithm proceeds inductively. If L_{i-1} is the set of the first $i - 1$ landmarks, then the i -th landmark is the point of S which maximizes the function $d(x, L_{i-1})$, the distance between the point x and the set L_{i-1} . We vary the total number of landmarks, exploring each of the resulting lazy witness complexes. The final number of landmarks is chosen so that the resulting witness complex maximally exposes topological features.

2.2. Persistent homology

Homology is the concept from algebraic topology which captures how space is *connected*. Thus, homology can be used to characterize interesting features of a simplicial complex such as connected clusters, holes, enclosed voids, etc., which could reveal underlying relationships and behavior of the data set. Homology of a space can be described by its *Betti numbers*. The k -th Betti number β_k of a simplicial complex is the rank of its k -th homology group. For $k = 0, 1, 2$, the β_k have intuitive interpretation. β_0 represents a number of connected components, β_1 the number of holes, and β_2 the number of enclosed voids. For example, a sphere has $\beta_0 = 1, \beta_1 = 0, \beta_2 = 1$, as it has one component, no holes, and one enclosed void.

Consider the formation of a simplicial complex using balls of diameter ϵ centered on points in a set. For small ϵ , the simplicial complex is just a set of disjoint vertices. For sufficiently large ϵ , the simplicial complex becomes one big cluster. What value of ϵ reveals the “correct” structure? Persistent homology [7, 8] gives a rigorous response to this question. By increasing ϵ , a sequence of nested simplicial complexes called a filtration is created, which is examined for attributes of connectivity and their robustness. Topological features appear and disappear as ϵ increases. The features which exist over a longer range of the parameter ϵ are considered as signal, and short-lived features as noise [7, 31]. This formulation allows a visualization as a collection of barcodes (one in each dimension), with each feature represented by a bar. The longer the life span of a feature, the longer its bar. In the example barcodes in Figs. 1–7, the x-axis represents the ϵ parameter, and the bars of persistent loops of interest are circled.

Research questions

Our approach could address several critical questions in the context of cancer data analysis. First, could we select a small subset of relevant genes while *simultaneously* identifying robust nontrivial structure, i.e., topology, of the data? Most previous approaches require the selection of a subset of genes *before* exploring the resulting structure, and hence limiting the generality. Second, could we elucidate higher order interactions (than clusters) between genes that could have potential implications for cancer biogenesis? Higher order structures such as loops could reveal critical subsets of genes with relevant nontrivial interactions, which together have implications to the cancer. Third, could this method work even when data is available from only a *subset* of patients?

3. Data

We analyzed five publicly available microarray datasets of gene expression from four different types of cancer – breast, ovarian, brain, and acute myeloid leukemia (AML). Four of the datasets have the same protocol, GPL570 (HG_U133_Plus_2, Affymetrix Human Genome U133 Plus 2.0 Array). The fifth dataset has a different protocol, HG_U95Av2, which has a fewer number of genes (see Table 1). By including data sets from different protocols, we could verify that the topological features identified are not just artifacts of a particular protocol.

The number of genes represents the number of unique gene id tags defined by a protocol excluding controls. While the brain dataset is of the same protocol as the breast and ovarian datasets, the former one has fewer genes – 46201 vs. 54613. This variability, however, did not affect our procedure to find topological features.

All datasets, except for AML170, were obtained from NCBI Gene Expression Omnibus <http://www.ncbi.nlm.nih.gov/geo/> in October 2013. AML170 was retrieved at the same time from the National Cancer Institute caArray Database at <https://array.nci.nih.gov/caarray/project/willm-00119>.

4. Methods

We work with the raw gene expression values. In particular, we do not log-transform them.

4.1. Dual space of data

Traditionally, gene expression data is viewed in its gene space, i.e., the expression profile of patient i with m genes is a point in \mathbb{R}^m , $\mathbf{x}_i = [x_{i_1} \ x_{i_2} \ \dots \ x_{i_m}]$. Each x_{i_j} is an expression of gene j in the patient i . For example, each patient in the Brain dataset is a point in \mathbb{R}^{46201} .

We analyze expression data in its dual space, i.e., in its *sample space*. Hence a gene j is represented as a point in \mathbb{R}^n , $\mathbf{x}_j = [x_{j_1} \ x_{j_2} \ \dots \ x_{j_n}]$, where n is the number of samples or

Table 1. Datasets used in the study.

Dataset	Series	Protocol	# Genes	# Samples
Brain	GSE36245	GPL570	46201	46
Breast	GSE3744	GPL570	54613	47
Ovarian	GSE51373	GPL570	54613	28
AML188	GSE10358	GPL570	54613	188
AML170	willm-00119	HG_U95Av2	12558	170

patients. Each x_{j_i} is the expression of the gene j in the patient i . For example, in the same Brain dataset, the points now sit in \mathbb{R}^{46} space. Hence we study gene expression across the span of all patients.

The key motivation for this approach is to handle the high dimensionality in a meaningful way for analyzing the *shape* of data. Given the small size of patients, one could efficiently construct a Vietoris-Rips complex using the set of pairwise distances between patients in the gene space (no need to choose landmarks). But such distances become less discriminatory when the number of genes is large [32, 33]. Working in the dual space, we let our topological method select a manageable number of genes as landmarks to construct the witness complexes, which potentially capture interesting topology of the data. Hence we do not preselect a small number of genes before the topological analysis, as is done by some previous studies [13, 24].

4.2. Choosing the number of landmarks

For construction of the witness complexes, the number of landmarks has to be defined *a priori* (Sec. 2.1). Hence the question becomes how many landmarks do we select? We let the data itself guide the selection of genes used as landmarks. If there is a significant loop feature in the data, it would persist through a range of landmarks in H_1 of the complexes. We reconstruct the topological space incrementing the number of landmarks while observing appearance and disappearance of the topological features. Initially, there would be very few small, noisy features because of insufficient number of points. As the number of landmarks increases, some features stabilize, i.e., do not change much either in size or membership. Then they reach their maximal size, and start to diminish once some critical number of landmarks is exceeded (when the “holes” are all filled in). The “optimal” number of landmarks is chosen when the length of the bar representing the topological feature is maximal.

A typical example of such behavior is seen in the Breast dataset (see Fig. 1). A small loop appears when the number of landmarks is $L = 50$. It stabilizes around $L = 90$, reaches its maximum span at $L = 110$, and then decreases as L grows.

4.3. Composition of loops

One of the goals of our method is to determine the genes which participate in H_1 features, which could indicate potential implications for cancer biogenesis. Since the first landmark is chosen randomly in the sequential maxmin procedure, the composition of the loops identified may differ based on this first choice. To circumvent this effect, we do 20 different runs in each case to collect possible variations in loop formation. Members of the loops are then pooled together for further analysis. Due to the almost deterministic nature of sequential maxmin selection (apart from the first landmark being selected randomly), we observed very little variation over the 20 runs in most cases. The recovered members of loops are then queried in scientific literature for cancer-related reports.

Table 2. Numbers of landmarks selected in each dataset.

Dataset	# landmarks
Brain	120
Breast	110
Ovarian	200
AML188	150
AML170	130

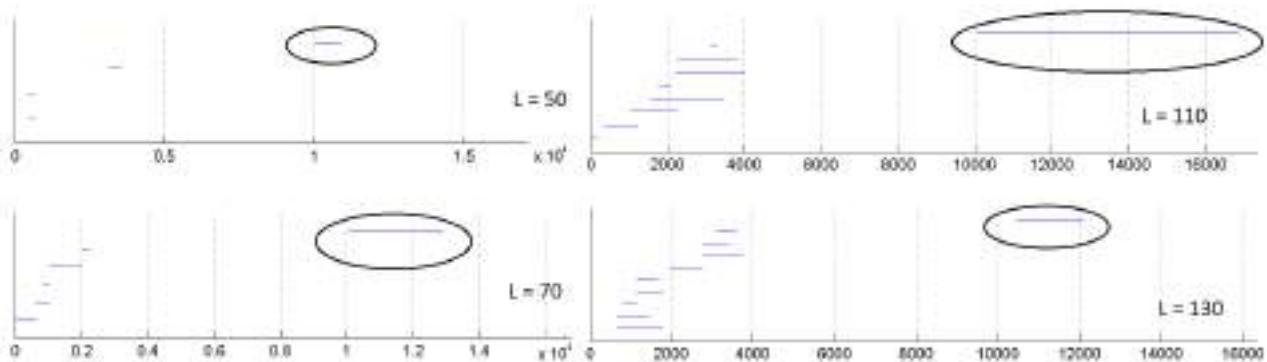


Fig. 1. Evolution of the loop of interest (circled) for varying number of landmarks from $L=50$ to $L=130$ in the breast dataset. Here and in Figs. 2–7, the x-axis represents the ϵ parameter.

We implemented our computations using the package JavaPlex [34]. We explored the barcodes for H_0 , H_1 , and H_2 , but interesting persistent features with members related to cancer biogenesis were detected only for H_1 .

5. Results

Persistent topological features in the homology group H_1 were observed in every cancer dataset we analyzed. Representative examples are shown in Figs. 2–6. The AML datasets both had two persistent loops, while the other datasets have one loop each. The Ovarian dataset had a few medium length bars in the H_1 barcode, but we investigated only the longest loop. Once the persistent loops were identified, they were inspected with respect to their composition and relation to cancer through search of scientific literature. Below is a brief description of results for each of the datasets (the full list of all loop members is also available [35]).

Table 3. Selected representatives of loops in different datasets.

Gene	Dataset	Description	References
CAV1	Brain	tumor suppressor gene	[36]
RPL36	Brain	prognostic marker in hepatocellular carcinoma	[37]
RPS11	Breast	downregulation in breast carcinoma cells	[38]
FTL	Breast	prognostic biomarkers in breast cancer	[39]
LDHA	Ovarian	overexpressed in tumors, important for cell growth	[40]
GNAS	Ovarian	biomarker for survival in ovarian cancer	[41]
LAMP1	AML170	regulation of melanoma metastasis	[42]
PABPC1	AML170	correlation with tumor progression	[43]
HLF	AML188	promotes resistance to cell death	[44]
DTNA	AML188	induces apoptosis in leukemia cells	[45]

5.1. Brain dataset

The lifespan of the longest loop in this dataset was about 820 (bar between $\approx [480, 1300]$). The loop was consistent over different choices of the first landmark. We identified

13 loop members, out of which 9 were found in cancer literature. Some cancer-related members include EGR1 and CAV1, which have genes been characterized as cancer suppressor genes [36, 46], A2M, which has been identified as a predictor for bone metastases [47], and RPL36 which has been found to be a prognostic marker in hepatocellular carcinoma [37].

5.2. Breast dataset

The lifespan of the longest loop in this dataset is in the range [10080.0, 16684.2]. As with the brain dataset, this loop is very consistent. However, there were only 10 members of this loop, and 8 of which were found in cancer literature. An interesting feature of this loop is that it had five ribosomal proteins which are known to play a critical role in tightly coordinating p53 signaling with ribosomal biogenesis [48].

5.3. Ovarian dataset

The Ovarian dataset had the most variable features in H_1 . However, we investigated the loop corresponding to the most consistent and longest bar, which ranged from about 4000 to over 7000. This loop consisted of 17 members, and 9 were mentioned in the cancer-related literature. Among cancer-related members were GNAS, which was identified as “an independent, qualitative, and reproducible biomarker to predict progression-free survival in epithelial ovarian cancer” [41], and HNRNPA1, which has been identified as a potential biomarker for colorectal cancer [49].

5.4. AML188 dataset

Acute myeloid leukemia 188 (AML188) had two significant loops (as did AML170). The first one occurred at [25200.0, 102200.0] and the second one at [78400.0, 146219.24]. The first loop has 27 members while the second one only 6. Altogether, only 14 of these 33 genes were mentioned in cancer literature. Some cancer-related representatives were hepatic leukemia

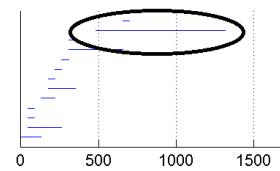


Fig. 2. Representative loop in brain dataset.

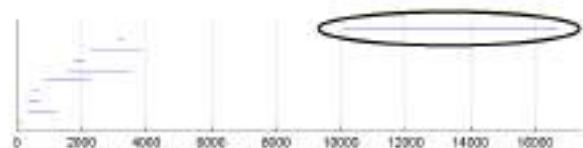


Fig. 3. Representative loop in breast dataset.

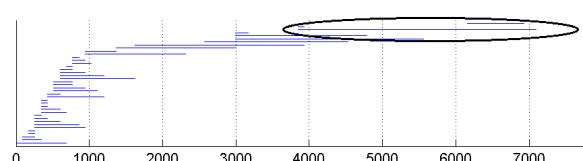


Fig. 4. Representative loop in ovarian dataset.

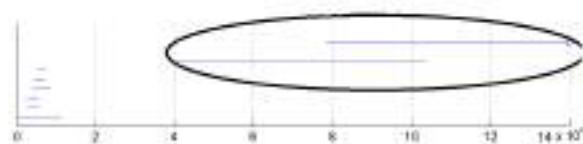


Fig. 5. Representative loops in AML188 dataset.

factor (HLF) which promotes resistance to cell death [44], RPL35A known for inhibition of cell death [50], and GRK5 which regulates tumor growth [51]. A group of zinc finger proteins were present in the first loop, some of which have been reported as novel biomarkers for detection of squamous cell carcinoma [52, 53].

5.5. *AML170 dataset*

AML170 comes from caArray database and its protocol HG_U95Av2 has only 12558 genes. Even though this protocol had a smaller (about 1/4) number of genes compared to the other data sets, we still detected two loops in this dataset. They were relatively shorter than the loops in AML188, and occurred at [5800.0, 11400.0] and [11000.0, 17600.0]. The two loops comprised of 19 members, of which 10 were found in cancer-related literature. These relevant members included ubiquitin C (UBC), which was recently identified as a novel cancer-related protein [54], PABPC1 whose positive expression is correlated with tumor progression in esophageal cancer [43], and LAMP1 which facilitates lung metastasis [42].

6. Discussion

The Breast, Brain, and Ovarian datasets had only one persistent loop, while AML170 and AML188 had two. Also, the AML datasets had a higher number of patients (samples) than the other three sets (see Table 1). Is this fact just a coincidence or, indeed, does the number of H_1 features (loops) correlate with the number of dimensions? To address this question, we chose samples of random and progressively larger (25–175) subsets of patients from AML170 and AML188 while also increasing the number of landmarks, and studied the evolution of H_1 features. In other words, we repeated our method on smaller subsets of patients from these datasets. Both the AML datasets contained two loops even with only 25 dimensions (see Fig. 7 for AML170), and continued to do so for the progressively larger subsets. Thus, the number of significant H_1 features appears to depend on intrinsic qualities of the data rather than the number of dimensions, demonstrating the robustness of our method to the number of patients in the dataset.

An important property of a loop is its lifespan [55]. One may note that the life span of loops for different datasets vary significantly. For example, the lifespan of a significant H_1 feature in the brain dataset is only 820, while for AML188 the lifespan of the first loop is 77,000. This difference is not only due to the increase in the actual size of a loop as indicated by the number of points comprising the loop (13 vs. 27 in this case), but also in part because

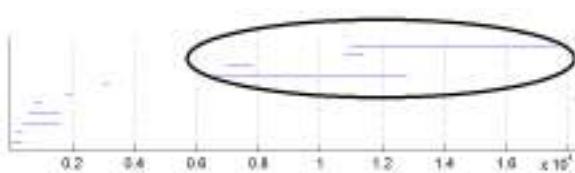


Fig. 6. Representative loops in AML170 dataset.

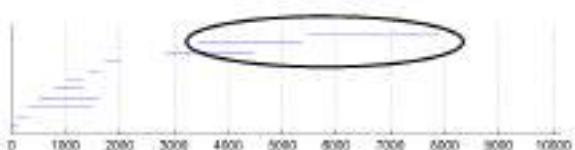


Fig. 7. Two persistent loops in the AML170 dataset detected using only 25 dimensions at the number of landmarks $L=130$.

of the different absolute values in microarray expression data. The maximum value for the brain dataset is 24×10^3 , while for AML188 is 3×10^6 . Therefore, the absolute length of an H_1 feature is not as important as its length relative to other H_1 features.

The crucial step of our method is the choice of landmarks. The goal here is the efficient inference of the topology of data, while selecting a small subset of potentially relevant genes. Landmarks chosen using sequential maxmin tend to cover the dataset better and are spread apart from each other, but the procedure is also known to pick outliers [14]. In our datasets, outliers are typically identified by extreme expression values [56]. We examined the expression values of the chosen landmarks, and found that very few of them had extreme values (Figs. 8 and 9). Similarly, the expressions of the genes implicated in cancer biogenesis (among the loop members) did not have any extreme values, and in fact appear to follow normal distributions. We infer that this observation results because sequential maxmin indeed picks points on the outskirts of the topological features. Further, the distribution of expressions of the loop members suggests that the group as a whole could have potential implications for the disease. More interestingly, the “hole” structure means such groups could not potentially be identified by traditional coexpression or even differential expression analyses [57].

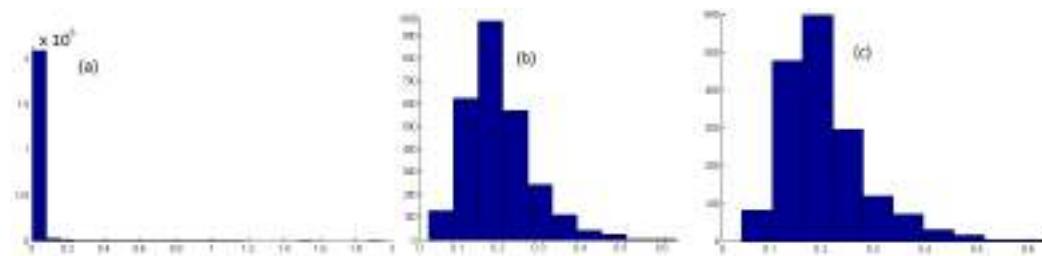


Fig. 8. Histograms for AML170 dataset, x-axis represents gene expression level of scale 10^4 . (a) distribution of gene expressions for the whole set; (b) distribution of gene expressions for only the loop members; and (c) distribution of gene expressions for cancer-related loop members.

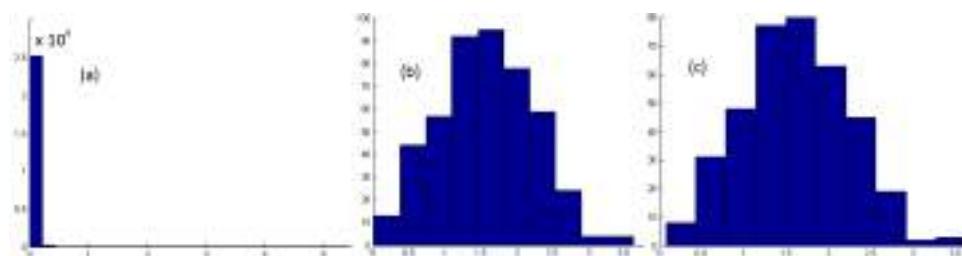


Fig. 9. Histograms for Breast dataset, x-axis represents gene expression level of scale 10^5 . (a) distribution of gene expressions for the whole set, (b) distribution of gene expressions for the loop members only, (c) distribution of gene expressions for cancer-related loop members.

The computational complexity of our method is based on the current implementation of JavaPlex [34], where the two main steps are building and filtering simplicial complexes, and computing homology. Up to dimension 2 (β_2), homology could be computed in $O(n^3)$ time [28]. However, building simplicial complexes relies on clique enumeration, which is NP-complete, and has a complexity of $O(3^{n/3})$ [58, 59]. Further, JavaPlex requires explicit enumeration

of simplicial facets appearing at each filtration step, implying the need for large memory resources [34].

The cancer-related loop members identified from the two AML datasets were distinct apart from the prominent group of ribosomal proteins. This observation could be explained by two main reasons. First, AML188 has four times the number of genes as AML170 (see Table 1). Second, JavaPlex identifies only *one* representative of each homology class. That is, if a significant topological feature (hole) exists, it will be identified, but only one loop will be found around that hole. There could be other relevant points in proximity to the hole, but are not guaranteed to be included in the loop. If we have prior knowledge of some genes being relevant, we could try to identify loops around the holes that include these genes as members. In this case, the other members of the identified loops could also have potential implications for the cancer biogenesis. Methods to find a member of a homology class that includes specific points could be of independent interest in the context of optimal homology problems [60, 61].

7. Conclusion

We have presented a method to look at cancer data from a different angle. Unlike previous methods, we look at characteristics of the first homology group (H_1). We identify the persistent H_1 features (which are loops, rather than connected components) and inspect their membership. Importantly, our approach finds potentially interesting connections among genes which cannot be found otherwise using traditional methods. This geometric connectedness may imply functional connectedness, however, this is yet to be investigated by oncologists. If such connections are indeed implied, then the genes in the loops could together form a characteristic “signature” for the cancer in question.

Acknowledgment: Krishnamoorthy acknowledges support from the National Science Foundation through grant #1064600.

References

- [1] X. Yu, S. Narayanan, A. Vazquez and D. R. Carpizo, *Apoptosis* **19**, 1055 (2014).
- [2] J. D. Patel, L. Krilov, S. Adams, C. Aghajanian, *et al.*, *J. Clin. Oncology* **32**, 129 (2014).
- [3] A. Singh and N. Kumar, *International Journal of Current Research and Review* **5**, 1 (2013).
- [4] J. C. Venter, M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, *et al.*, *Science* **291**, 1304 (2001).
- [5] S. Valarmathi, A. Sulthana, K. Latha, R. Rathan, R. Sridhar and S. Balasubramanian, *Proc. 2nd Intl. Conf. Soft Comput. Prob. Solv (SocProS 2012)* , 1451 (2014).
- [6] A. E. Pozhitkov, P. A. Noble, J. Bryk and D. Tautz, *PloS One* **9**, p. e91295 (2014).
- [7] H. Edelsbrunner, D. Letscher and A. Zomorodian, *Discret. Comput. Geom.* **28**, 511 (2002).
- [8] R. Ghrist, *Bulletin of the American Mathematical Society* **45**, 61 (2008).
- [9] G. Carlsson, A. Zomorodian, A. Collins and L. Guibas, *Proc. Eurogr./ACM SIGGRAPH Symp. Geom. Proc. SGP '04*, 124 (2004).
- [10] P. Y. Lum, G. Singh, A. Lehman, T. Ishkanov, *et al.*, *Scientific Reports* **3** (2013).
- [11] G. Carlsson, *Bulletin of the American Mathematical Society* **46**, 255 (January 2009).
- [12] D. DeWoskin, J. Climent, I. Cruz-White, M. Vazquez, C. Park and J. Arsuaga, *Topology and its Applications* **157**, 157 (2010).
- [13] L. Seemann, J. Shulman and G. H. Gunaratne, *ISRN Bioinformatics* (2012).
- [14] V. de Silva and G. Carlsson, *SPG '04: Proc. Sympos. Point-Based Graphics*, 157 (2004).

- [15] R. Fa, A. K. Nandi and L.-Y. Gong, *Comm. Control. Sig. Proces. (ISCCSP)*, 1 (2012).
- [16] A. K. Jain, *Pattern Recognition Letters* **31**, 651 (2010).
- [17] D. A. Spielman and S.-H. Teng, *SIAM Journal on Computing* **42**, 1 (2013).
- [18] D. J. Cook and L. B. Holder, *Mining graph data* (John Wiley & Sons, 2006).
- [19] J. B. Tenenbaum, V. D. Silva and J. C. Langford, *Science* **290**, 2319 (2000).
- [20] S. T. Roweis and K. S. Lawrence, *Science* **290**, 2323 (2000).
- [21] D. L. Donoho and C. Grimes, *Proceedings of the National Academy of Sciences* **100**, 5591 (2003).
- [22] M. Belkin and P. Niyogi, *NIPS* **14**, 585 (2001).
- [23] R. R. Coifman and S. Lafon, *Applied and computational harmonic analysis* **21**, 5 (2006).
- [24] M. Nicolau, A. J. Levine and G. Carlsson, *PNAS* **108**, 7265 (2011).
- [25] M. Nicolau, R. Tibshirani, A.-L. Børresen-Dale and S. S. Jeffrey, *Bioinformatics* **23**, 957 (2007).
- [26] G. Singh, F. Mémoli and G. E. Carlsson, *Symp. Point Based Graphics*, 91 (2007).
- [27] J. R. Munkres, *Elements of Algebraic Topology* (Addison-Wesley, 1984).
- [28] H. Edelsbrunner and J. L. Harer, *Computational Topology* (American Math. Society, 2009).
- [29] H. Adams and G. Carlsson, *SIAM J. Img. Sci.* **2**, 110 (2009).
- [30] G. Carlsson, T. Ishkhanov, V. de Silva and A. Zomorodian, *Intl. Jnl. Comp. Vis.* **76**, 1 (2008).
- [31] A. Zomorodian and G. Carlsson, *Discrete Computational Geometry* **33**, 249 (2005).
- [32] K. Beyer, J. Goldstein, R. Ramakrishnan and U. Shaft, *Database Theory—ICDT99*, 217 (1999).
- [33] M. Ledoux, *The concentration of measure phenomenon* (American Mathematical Soc., 2005).
- [34] A. Tausz, M. Vejdemo-Johansson and H. Adams, Javaplex: A software package for computing persistent topological invariants (2011), Available at <http://comptop.stanford.edu/programs/>.
- [35] Supplement: <http://math.wsu.edu/faculty/bkrishna/Topo/Cancer/Supplement.pdf>.
- [36] Cav1 caveolin 1. NCBI Gene Database <http://www.ncbi.nlm.nih.gov/gene/857>.
- [37] M. J. Song, C. K. Jung, C. H. Park, W. Hur, J. E. Choi, et al., *Pathol. Int.* **61**, 638 (2011).
- [38] D. Nadano, C. Aoki, T. Yoshinaka, S. Irie and T. A. Sato, *Biochemistry* **40**, 15184 (2001).
- [39] G. Ricolleau, C. Charbonnel, L. Lod, D. Loussouarn, et al., *Proteomics* **6**, 1963 (2006).
- [40] D. Zhao, S. W. Zou, Y. Liu, X. Zhou, Y. Mo, et al., *Cancer Cell* **23**, 464 (2013).
- [41] E. Tominaga, H. Tsuda, T. Arao, S. Nishimura, et al., *Gynecol. Oncol.* **118**, 160 (2010).
- [42] A. K. Agarwal, R. P. Gude and R. D. Kalraiyia, *Bioch. Biophys. Res. Comm.* **449**, 332 (2014).
- [43] N. Takashima, H. Ishiguro, Y. Kuwabara, M. Kimura, et al., *Oncol. Rep.* **15**, 667 (2006).
- [44] K. M. Waters, R. L. Sontag and T. J. Weber, *Toxicol. Appl. Pharmacol.* **268**, 141 (2013).
- [45] R. Navakauskienė, G. Treigytė, V. V. Borutinskaitė, D. Matuzevičius, D. Navakauskas and K. E. Magnusson, *J. Proteomics* **75**, 3291 (2012).
- [46] Egr1 early growth response. NCBI Gene Database <http://www.ncbi.nlm.nih.gov/gene/1958>.
- [47] Y. Kanoh, N. Ohtani, T. Mashiko, S. Ohtani, et al., *Anticancer Res.* **21**, 551 (2001).
- [48] A. Artero-Castro, J. Castellvi, A. García, J. Hernández, et al., *Hum. Pathol.* **42**, 194 (2011).
- [49] Y. L. Ma, J. Y. Peng, P. Zhang, L. Huang, W. J. Liu, et al., *J. Proteome Res.* **8**, 4525 (2009).
- [50] C. D. Lopez, G. Martinovsky and L. Naumovski, *Cancer Lett.* **180**, 195 (2002).
- [51] J. I. Kim, P. Chakraborty, Z. Wang and Y. Daaka, *J. Urol.* **187**, 322 (2012).
- [52] Y. J. Jou, C. D. Lin, C. H. Lai, C. H. Tang, et al., *Clin. Chim. Acta* **412**, 1357 (2011).
- [53] R. Lovering and J. Trowsdale, *Nucleic Acids Res.* **19**, 2921 (1991).
- [54] Y. H. Wong, R. H. Chen and B. S. Chen, *J. Theor. Biol.* (2014).
- [55] E. Carlsson, G. Carlsson and V. De Silva, *Intl. J. Comput. Geom. Appl.* **16**, 291 (2006).
- [56] V. J. Hodge and J. Austin, *Artificial Intelligence Review* **22**, 85 (2004).
- [57] Y. Choi and C. Kendziorski, *Bioinformatics* **25**, 2780 (2009).
- [58] J. D. Moon and L. Moser, *Israel journal of Mathematics* **3**, 23 (1965).
- [59] J. Binchi, E. Merelli, M. Rucco, G. Petri, et al., *El. Notes in Theor. Comp. Sci.* **306**, 5 (2014).
- [60] T. K. Dey, A. N. Hirani and B. Krishnamoorthy, *SIAM Journal on Computing* **40**, 1026 (2011).
- [61] T. K. Dey, J. Sun and Y. Wang, *Proc. 26th Symp. Comp. Geom, SoCG '10*, 166 (2010).

DISTANT SUPERVISION FOR CANCER PATHWAY EXTRACTION FROM TEXT

HOIFUNG POON*, KRISTINA TOUTANOVA, CHRIS QUIRK

Microsoft Research, Redmond, WA, USA

*E-mail: hoifung@microsoft.com

Biological pathways are central to understanding complex diseases such as cancer. The majority of this knowledge is scattered in the vast and rapidly growing research literature. To automate knowledge extraction, machine learning approaches typically require annotated examples, which are expensive and time-consuming to acquire. Recently, there has been increasing interest in leveraging databases for distant supervision in knowledge extraction, but existing applications focus almost exclusively on newswire domains. In this paper, we present the first attempt to formulate the distant supervision problem for pathway extraction and apply a state-of-the-art method to extracting pathway interactions from PubMed abstracts. Experiments show that distant supervision can effectively compensate for the lack of annotation, attaining an accuracy approaching supervised results. From 22 million PubMed abstracts, we extracted 1.5 million pathway interactions at a precision of 25%. More than 10% of interactions are mentioned in the context of one or more cancer types, analysis of which yields interesting insights.

Keywords: Distant supervision, knowledge extraction, cancer pathways, Literome

1. Introduction

Cancer stems from the synergistic perturbation of multiple pathways by mutations.¹ Recent advances in sequencing technology offer a plethora of panomics data, holding the promise to make precision medicine and personalized treatment a reality. However, it remains a formidable challenge to identify cancer drivers, due to complex cross talks and feedback loops in cancer pathways.² Moreover, even when the drivers are identified, they may not be directly druggable, as in the case of RAS, where the promising targets lie in downstream signaling pathways.³ Pathways are thus essential to understanding cancer and developing targeted treatments. As a result, pathways have been increasingly applied to panomics analysis.⁴⁻⁸

The majority of pathway knowledge resides in free text such as journal articles, which has been undergoing its own exponential growth. For example, PubMed contains over 22 million papers and adds more than one million each year. It is hard for manual curation to keep pace with such a vast and rapidly growing literature, making it a priority to automate the curation process. Such automation was traditionally pursued via rule-based systems,⁹ but hand-coding extraction rules is expensive and time-consuming, and generally suffers low recall due to the varieties of ways for expressing the same meaning. Machine learning approaches offer a much more attractive alternative by effectively automating the rule engineering itself, but they in turn require annotated examples, which are still difficult to acquire in scale.

The lack of annotated examples can be compensated for by leveraging *distant supervision* from existing knowledge bases, as first proposed by Craven & Kumlien¹⁰ and recently pursued actively in the natural language processing (NLP) community.¹¹⁻¹³ However, no existing approach addresses pathway extraction, focusing instead almost exclusively on newswire.

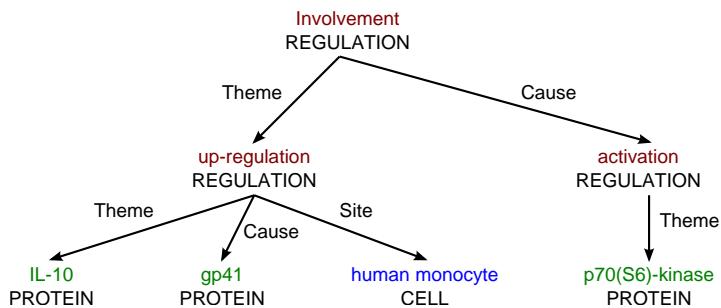


Fig. 1. Example pathway annotation of the sentence “Involvement of p70(S6)-kinase activation in IL-10 up-regulation in human monocytes by gp41 envelope protein”.

In this paper, we present the first attempt to apply distant supervision in pathway extraction. We formulate pathway extraction as a classification problem and apply a state-of-the-art distant-supervision method to it. To better evaluate the effectiveness of distant supervision in bridging the gap from supervised learning, we propose a novel evaluation methodology to create a controlled experimental setting using the GENIA event extraction dataset.¹⁴ Experimental results show that distant supervision outperforms baseline systems such as rule-based extraction by a wide margin, attaining an accuracy approaching supervised learning. Finally, we applied distant supervision to all PubMed abstracts, using prior pathway knowledge from the Pathway Interaction Database (PID).¹⁵ Our system extracted 1.5 million pathway interactions at a precision of 25%. More than 10% of these interactions are cancer-related, analysis of which yields a number of interesting observations.

2. Methods

2.1. Pathway Extraction from Text

Biological pathways capture the interactions among genes, gene products, and small molecules such as metabolites. Examples of interactions include transcriptional regulation (e.g., transcription factor binding for transcription initiation) and post-translational regulation (e.g., kinase phosphorylation for protein activity modulation). For simplicity, in this paper we focus on static pathways such as signaling transduction and gene regulation, rather than metabolic networks and dynamics. Figure 1 shows an example pathway fragment, with the corresponding text and annotation. In general, pathways form a hypergraph where each node represents a gene or gene product, and each hyperedge represents an interaction.

Decades of genetic studies have produced a wealth of pathway knowledge, which is scattered in the literature, with each paper or sentence covering only a few interactions. For example, the sentence “CTCF is a transcriptional repressor of the c-myc gene.” signifies a transcriptional regulation of c-myc by CTCF. In this paper, we focus on extracting such regulatory relations between two proteins, which identify the **Cause** argument such as CTCF, the **Theme** argument such as c-myc, and, if available, the regulation direction such as **negative_regulation**. Formally, the pathway extraction problem is to classify each ordered triple (P_T, S, P_C) into one of the following: {**positive_regulation**, **regulation**, **negative_regulation**, **NULL**},

Fig. 2. A simple distant supervision algorithm

Require: A set of sentences, with entity mentions identified

Require: A database of relation triples (entity, relation, entity)

- 1: For each relation triple, find all sentences containing the entity pair
- 2: Annotate those sentences with the corresponding relation
- 3: Sample unannotated sentences with co-occurring proteins as negative examples
- 4: Train a classifier using the annotated dataset
- 5: **return** the resulting classifier

where S is a sentence and P_T, P_C are protein mentions in S .*

2.2. Distant Supervision

The main idea of distant supervision is to use known relation instances in the database to automatically annotate training examples in unlabelled text,¹⁰ as shown in Figure 2. Suppose we know from the database that CTCF down-regulates c-myc, and in the text we find a sentence where CTCF and c-myc co-occur, such as “CTCF is a transcriptional repressor of the c-myc gene”. We thus have some reason to hypothesize that this sentence might be stating a **negative regulation** relation with **Theme** being c-myc and **Cause** being CTCF. A simple distant-supervision method would thus label such sentences as positive examples, and sample negative examples from random sentences where the co-occurring proteins do not have a known regulatory relation.

Of course, this simple approach would often introduce noise in the labels, since the sentence might not be about the given regulation, as in “In Bcl-deficient mice, expressions of both CTCF and c-myc showed marked decrease”. A more reasonable assumption is that *some* sentence in the text expresses this relation, though not necessarily every one with co-occurring CTCF, c-myc.[†] This assumption is adopted in state-of-the-art distant supervision methods.^{12,13}

In this paper, we use MULTIR¹³ because it is such a state-of-the-art method with a publicly available implementation. The key idea is to introduce a latent variable to signify whether a relation R holds between entities (E_1, E_2) for each sentence where E_1 and E_2 co-occur and are the **Theme** and **Cause**, respectively. Distant supervision is provided by enforcing during training that for each relational triple (E_1, R, E_2) , at least one latent assignment is true if the database contains the relation, and none otherwise. Each instance is represented by a linear model with features over the sentence and entities. The feature weights are learned using online learning with perceptron. In each iteration, for each protein pair, MULTIR first computes the best assignment to each instance according to the model. If the assignment is consistent with the database (each relation is expressed at least once, and no relations that don't appear in the database are expressed), no update is done on the protein pair. If not, it

*This formulation might sometimes lose information, such as co-factors required in a regulation, or experimental conditions. Lifting this limitation to handle n-ary, nested relations will be a key future direction.

[†]Note that this assumption still suffers a number of drawbacks. For example, it is possible that none of the available sentences mention the given relation. Moreover, the existing database is incomplete, so the absence of a relation might not necessarily signify its negation. Addressing these issues is an active research area.

uses a greedy algorithm to find the best assignment to each instance such that the assignment is consistent with the database, and does a perceptron update toward this assignment.

We used the following standard lexical and syntactic features¹¹ in our experiments, illustrated with sentence “Involvement of p70(S6)-kinase activation in IL-10 up-regulation in human monocytes by gp41 envelop protein”, and E_1, E_2 being “gp41” and “p70(S6)-kinase”, respectively.

Direction: 0 if the two protein spans overlap, +1 if E_2 follows E_1 in the sentence, and -1 if E_2 precedes E_1 . In the example, this feature is -1.

Distance: Four indicator features specifying whether E_1 is more than k tokens to the right of E_2 , with $k = 5, 10, 15, 20$. In the example, this feature is 1 for $k = 5$ and 0 otherwise.

Lexical: (1) the sequence of words between the two proteins, concatenated with direction, such as `Dir=-1 ∧ WordSeq="activation in IL-10 up-regulation in human monocytes by"`; (2) the sequence of lemmas of the words between the two proteins, concatenated with direction, such as `Dir=-1 ∧ LemmaSeq="activate in il-10 up-regulate in human monocyte by"`; (3) individual words between the two proteins, concatenated with direction, such as `Dir=-1 ∧ Word="activation"`, etc.; (4) individual lemmas between the two proteins, concatenated with direction, such as `Dir=-1 ∧ Lemma="activate"`, etc.

Dependency path: (1) unlexicalized dependency path between the two proteins (e.g. $\uparrow_{nn} \uparrow_{by} \uparrow_{in} \downarrow_{of} \downarrow_{nn}$); (2) lexicalized dependency path using the lemmas for lexicalization (e.g. $\uparrow_{nn} \text{protein} \uparrow_{by} \text{up-regulate} \uparrow_{in} \text{involve} \downarrow_{of} \text{activate} \downarrow_{nn}$); (3) upward (child to parent) and downward (parent to child) portions of the dependency path as separate features, concatenated with the lemma of the path root, e.g., `Upward=↑nn↑by↑in ∧ RootLemma="involve"` and `Downward=↓of↓nn ∧ RootLemma="involve"`.

MULTIR can be used in supervised learning by simply setting each relational assignment according to the sentence-level annotation (i.e., they are no longer latent).

2.3. Controlled Experiments using GENIA Event Extraction Dataset

Evaluating distant supervision is challenging as by definition there is no annotated dataset, so existing methods tend to resort to reporting sample precision and absolute recall (i.e., sample a small subset of system extractions, manually inspect them to determine the precision, and use it to estimate the number of correctly extracted instances). While this is useful for comparing distant supervision methods, the sampling process inevitably introduces bias and variance. Furthermore, it is difficult to assess the performance gap from supervised learning.

This motivates us to propose a new evaluation methodology by creating a simulated distant-supervision scenario from an annotated dataset, which enables us to assess the true precision and recall, and compare with supervised learning. Specifically, we used the GENIA event extraction dataset from BioNLP-09 Shared Task 1,¹⁴ where protein annotation is given as input, and pathway events are annotated as output (Figure 1).

We follow the formulation in Section 2.1 and reduce GENIA events to binary rela-

tions in $\{ \text{positive_regulation}, \text{regulation}, \text{negative_regulation}, \text{NULL} \}$.[‡] Specifically, for each protein pair E_1, E_2 in a training sentence, we compute all event paths between them from the annotation, and reduce each event path into a relation summarizing the path semantics as follows:

First, we identify the top event e in the path that has both proteins in its scope. E_1 should lie in the **Theme** branch from e , and E_2 in the **Cause** branch. If not, the relation is set to **NULL**.

Next, we check whether the **Theme** path contains any **Cause** argument, as between TP53 and MAPK1 in “TP53-induced BCL overexpression is inhibited by MARK1”. If so, the relation is also set to **NULL**. In general, as we can see from this example, we can not conclude a causal relation between these two proteins.

Otherwise, we assign a regulation relation with E_1 being the **Theme** argument and E_2 being the **Cause** argument. If any event in the path is **regulation**, we set the overall relation to **regulation** as well (i.e., we can not determine the direction). Otherwise, we set the relation to **positive_regulation** if there are an even number of **negative_regulation** events followed by a **Theme** argument, and to **negative_regulation** for an odd number.

For example, with sentence “Involvement of p70(S6)-kinase activation in IL-10 up-regulation in human monocytes by gp41 envelop protein”, and E_1, E_2 being “gp41” and “p70(S6)-kinase”, respectively, the non-**NULL** relations are **(positive_regulation, IL-10, gp41)**, **(regulation, IL-10, p70(S6)-kinase)**.

We thus form a database with the unique relation triples derived from the training set, which, together with the unlabelled text are served as input to the distant supervision method. The learned classifier is then applied to the test text to extract new events. The results can be evaluated using the test annotation and directly compared with the supervised learning approach, which is separately learned using sentence-level annotation in training text.

2.4. Rule-based Relation Extraction

In the past, rule-based extraction has been used extensively in relation extraction, such as the GeneWays system.⁹ This remains a preferred approach in the absence of annotated examples for applying machine learning approaches.¹⁶ To enable a head-to-head comparison with distant supervision, we followed standard practice as in GeneWays and other rule-based systems, and spent substantial effort to develop a high-performing rule-based system for pathway extraction.

Upon first consideration, relation extraction appears straightforward to automate. For example, in the sentence “CTCF is a transcriptional repressor of the c-myc gene.”, “transcriptional repressor” clearly indicates a **negative_regulation** relation. Unfortunately, the same information may be expressed in many variations, as shown in Table 1.

Some variations can be normalized in syntactic analysis. For example, lemmatization can normalize inflectional morphology (“inhibited” has the lemma “inhibit”), and derivational morphology (both “inhibitor” and “inhibition” are derived from “inhibit”). Syntactic parsing can normalize active/passive variations and identify related words even when they are far

[‡]GENIA also annotates simple unary events such as **expression** and **transcription**, which could be taken into account in future work.

Table 1. Linguistic variations describing the same pathway event.

Sentence	Variation
CTCF is an inhibitor of the c-myc gene	lexical
CTCF inhibits the c-myc gene	part of speech
The c-myc gene is inhibited by CTCF	active/passive voice
The ability of CTCF to inhibit c-myc	modality
CTCF has been shown to inhibit c-myc	background
Expression of the CTCF gene has been shown to inhibit c-myc	explanation
CTCF as well as other genes have been shown to inhibit c-myc	augmentation

apart in the sentence.

Other variations, however, are more difficult to handle in a domain-independent way, so past systems resorted to writing domain-specific rules to capture such variations.^{9,16} We followed this approach and developed our rules based on the event annotation in the training set of GENIA. The rules identify subtrees that trigger a particular relation, as well as child subtrees that identify the **Cause** and **Theme** of that relation. At the surface level, one might say that “Cause inhibits Theme” is a trigger for **negative regulation**. We represent these rules in terms of syntactic subtrees to better handle many of the aforementioned variations: “(inhibit nsubj: Cause dobj: Theme)” triggers **negative regulation**.

To identify relations for a candidate sentence and protein pair, we first parse the sentence using SPLAT¹⁷ and postprocess the parse into Stanford typed dependencies,¹⁸ which forms a tree where each node represents a word by its surface form, its lemma, and its part of speech, and each edge represents a typed dependency. For example, for “CTCF inhibits c-myc”, “inhibits” is the root, with a “subject” being “CTCF” and an “object” being “c-myc”. Next we attempt to match each trigger rule at each node in the tree. Here, matching means finding a correspondence between the nodes in the trigger subtree and the nodes in the candidate sentence. Each trigger subtree node must map to a unique sentence node, with matching relation types and node lemmas, except for **Cause** and **Theme**, which can match any nodes in the input tree. Finally, we check if the **Cause** and **Theme** are compatible with the protein arguments. In the simplest case, this amounts to matching **Cause** and **Theme** to the corresponding protein nodes. Additionally, we expand the criteria to account for variations in protein mentions (e.g., matching to “gene” in “the BCL2 gene”, or matching to “BCL1” in “BCL1 and BCL2”). If this is successful, we return the specified relation as the classification; otherwise, **NULL**.

This results in a set of 159 trigger rules and 63 protein expansion criteria, by hill-climbing on extraction accuracy over the GENIA training set.

2.5. PubMed Extraction with Distant Supervision from PID

Given the GENIA dataset, one might consider simply adopting supervised learning and applying the learned extractor to PubMed abstracts, as in EVEX.¹⁹ Such extractions would no doubt be very useful, but there remains a major concern over how representative the examples are, as with any supervised approach. The GENIA abstracts were chosen a decade ago and are rather dated by now. More importantly, they were sampled from a narrow subject area

(PubMed search with MeSH terms “human”, “blood cell” and “transcription factor”).

In contrast, a distant supervision approach can learn an extractor for any subject area, as long as there exist manually curated databases to leverage, which is generally the case. To demonstrate the feasibility of this direction, we used PID¹⁵ as the database for distant supervision, and applied the learned extractor to PubMed-scale pathway extraction.

PID represents each pathway as a hypergraph, where each node represents a gene (transcription), gene product (proteins and complexes), or process (e.g., apoptosis), and each hyper-edge represents an interaction, with multiple input and output nodes and their regulatory directions (induction or inhibition). We followed the formulation in Section 2.1 and reduced each interaction into binary regulations by creating relation triples between the component genes in each input/output pair, excluding the case when the two nodes represent identical molecules. This yields 15150 unique relation triples, such as (`negative_regulation`, APC, TP53). We filter out triples with conflicting regulation or causal directions between two proteins, such as (`positive_regulation`, CDKN1A, TP53), (`negative_regulation`, CDKN1A, TP53), and (`positive_regulation`, PLAU, PLG), (`positive_regulation`, PLG, PLAU). These are legitimate interactions representing feedback loops or contextual dependency, but their inclusion would confuse the distant-supervision learner given the lack of sentence-level annotation. There were 4,547 triples left after filtering, which were used for distant supervision.

In GENIA, gold protein annotation is given as input, which is not available for general PubMed abstracts. We thus used the protein extractor from Literome²⁰ to identify protein mentions in all PubMed abstracts. This extractor was built on various available resources with canonical mentions and synonyms for proteins, families, and complexes.

General PubMed abstracts are significantly more diverse and noisy compared to GENIA ones. At training time, we filtered out instances where the two proteins have overlapping spans or appear more than 15 words apart, which are unlikely to express a relation.

2.6. *Cancer Classification*

Given a pathway interaction, it would be useful to understand the context such as cell type, localization, experimental conditions, etc. As a first step toward this direction, we focus on identifying the cancer types mentioned in the same abstract. Specifically, we used the MeSH terms provided by Medline for the majority of PubMed abstracts, and extracted the ones that signify cancer types (i.e., ending in “Neoplasms”). An extraction instance is thus associated with the cancer types for the given abstract.

3. Results

3.1. *GENIA Experiments*

We used the GENIA dataset to create a controlled setting to evaluate distant supervision and compare it with rule-based and supervised approaches (Section 2.3). The GENIA dataset¹⁴ contains a set of annotated abstracts (800 training, 150 development). It also contains a test set, but its annotation is not made public so we can not use it to evaluate binary relation extraction. Therefore, we conducted training and development using the training data, and

Table 2. Test results on GENIA binary-relation classification comparing distant supervision with two baseline systems, supervised learning, and MSR11, a state-of-the-art system training on full event structures.

System	Precision	Recall	F1
Most-Frequent	3.4	69.7	6.5
Rule-Based	45.8	5.2	9.4
Distant Supervision	39.2	19.0	25.6
Supervised	37.5	29.9	33.2
MSR11	55.1	28.0	37.1

reserved the development set for test. We subsampled negative examples to avoid label imbalance (the ratio between positive and negative examples is about 1:3). For supervised learning, we also filtered out features with fewer than 3 counts in positive examples. Training and test both took less than a second.

We took all co-occurring protein pairs in test sentences as classification candidates and evaluated precision, recall and F1 of system extraction given the gold labels. We compared distant supervision with the rule-based system, a baseline that predicts the most frequent relation in training (`positive_regulation`) for co-occurring protein pairs, as well as the supervised system trained on sentence-level annotation, which provides an upper bound. To quantify a further upper bound when full event structures are taken into account, we also evaluated MSR11²¹ by converting its publicly available event extraction output to binary relations. MSR11 is an event extraction system trained on full event structures, with state-of-the-art event F1 score of 55.7 on GENIA development. (The best score of 58.7 is reported by Riedel et al.²²) Note that event F1 accounts for simple unary events such as `expression`, thus is not directly comparable with binary-relation F1. For example, MSR11 scores only 37.1 on binary-relation F1.

Table 2 shows the test results for all systems. Surprisingly, not only did distant supervision substantially outperform the two baseline systems, it also attained an accuracy that is rather close to the supervised upper bound. For example, compared to the rule-based system, distant supervision reduces the performance gap from the supervised upper bound by nearly 70%, while using much less information. (The rule-based system used sentence-level annotation during rule engineering.) Note that this supervised upper bound is very strong, comparable to state-of-the-art event extraction that leverages full event structures (33.2 vs. 37.1). Likewise, we spent substantial effort to engineer the rule-based system, attaining a strong performance on training (precision 80.3, recall 26.5, F1 39.8). Unfortunately, while its precision is relatively high, it suffers low recall and a big performance drop from training to test, which is typical of rule-based systems. Overall, these results clearly demonstrate the promise of distant supervision, which could attain competitive accuracy while requiring substantially less development effort compared to both rule-based and supervised approaches.

3.2. PubMed Experiments

We applied distant supervision using PID and PubMed abstracts to train an extractor (Section 2.5). We sampled negative examples following the positive/negative ratio used in GENIA, yielding 97,215 examples in total. We then ran the extractor to extract events from all PubMed

Table 3. Evaluation on 300 sample PubMed extractions, with annotation statistics and example instances. Mentions are annotated with Cause and Theme from automatic predictions.

Outcome	Count	Example
Correct	75	The polycomb protein <u>Bmi-1</u> ^{Cause} represses the INK4a locus , which encodes the tumor suppressors p16 and p14(ARF). ^{Theme}
Imperfect Sign	27	This regulated control of <u>STAM</u> ^{Theme} expression by <u>Hrs</u> ^{Cause} was independent of transcription.
Reversed Direction	46	This may possibly occur through inhibition of insulin receptor (IR) tyrosine kinase activity mediated by serine/threonine phosphorylation of the IR or <u>insulin receptor substrate 1 (IRS-1)</u> . ^{Theme} ^{Cause}
Protein Error	56	We found that the development of experimental autoimmune encephalomyelitis (EAE), the rodent model of multiple sclerosis, was significantly suppressed in <u>IL-17</u> (-/-)-mice. ^{Theme} ^{Cause}
Non-Regulation	96	Anti-dsDNA B cells, on the other hand, are functionally unresponsive to anti-IgM and <u>LPS</u> stimulation, and do not phosphorylate intracellular proteins, including <u>Syk</u> , upon mIg stimulation. ^{Theme} ^{Cause}

abstracts by classifying candidate sentences with co-occurring protein pairs. Note that we do not know the gold annotation for the positive examples used in training, nor could the learner memorize them since no protein-specific features are in use. Training took five minutes and extraction took 30 minutes using 900 cores. This PubMed-scale extraction yields 1,491,373 regulation instances, with 838,255 unique relation triples.

To assess the quality of the extraction, we sampled 300 extractions and manually annotated them. Table 3 summarizes our findings. Among the 300 sample extractions, 56 have wrong protein annotation, which is not surprising given that protein mentions are often highly ambiguous. Of the remaining 244 instances, 75 are correct, giving an end-to-end precision of 25% and a precision of 31% assuming gold protein annotation. Among the errors, 46 are actually correct regulation events, but in 21 of them the sign (`positive_regulation` or `negative_regulation`) is wrong or the sentence is ambiguous about it, and in the remaining 25 the causality direction is reversed. With this sample precision, we estimate that distant supervision from PID yields about 372,000 correct extractions, and 210,000 unique relation triples, which is an order of magnitude larger than PID.

These results are promising and testify to the feasibility of this direction. Of course, there is still much room for improvement. Protein errors occur in about one fifth of extractions. Currently, protein extraction does not benefit from distant supervision and is done separately from event extraction. Joint learning of protein and event extraction with distant supervision could potentially produce large improvement in both tasks. Relation errors often occur for two proteins in paths that are in conjunction, as in the example in Table 3, which might potentially be avoided with better filtering criteria. Finally, distant supervision can be used in

Table 4. Top ten most studied cancer types, along with the top ten genes for each type, both in the number of unique pathway extractions.

Cancer	Relations	Top ten most studied genes
Breast	27988	TP53, ESR1, MYBL2, BRCA1, ZFP36, EGFR, ZFP, ESR2, EGF, AKT
Prostate	10981	CBX8, SLC22A3, AR, KLK3, EPHB2, TP53, TDRD7, NPEPPS, AKT, SERPINB6
Lung	9423	EGFR, TP53, KRAS, VEGFA, CASP, MMP2, AKT, EPHB2, CDH1, CEACAM5
Liver	8438	TP53, EPHB6, HCCS, AFP, MYLIP, CCL2, VEGFA, RELA, NFKB1, NA
Colon	6092	TP53, APC, CTNNB1, AOM, TMED7, RELA, EGFG, NFKB1, SRC, PTGS2
Colorectal	5381	TP53, CTNNB1, APC, CEACAM5, KRAS, EGFR, MSI, CASP, VEGFA, PTGS2
Pancreatic	5178	INS, KRAS, EGFR, VEGFA, CEACAM5, TNF, AURKA, MIA, RELA, NFKB1
Ovarian	4331	LPA, TP53, EGFR, VEGFA, AKT, MMP, MUC16, BRCA1, BARX2, IFNG
Skin	4324	SERPINB3, TP53, KRT13, NFKB1, RELA, CD4, VIM, TNF, IL2, CD68
Brain	3988	RIPK1, CCDC88A, TP53, CSF2, NCAM1, AFP, MGMT, EGFR, ELAVL1, AKT

Table 5. Top ten most studied cancer types, along with top genes ranked by association score.

Cancer	Top ten most significantly associated genes
Breast	ESR1, BRCA1, ESR2, BRCA2, PGR, CYP19A1, ERBB2, EGF, IGF1, KRAS
Prostate	CBX8, KLK3, NPEPPS, AR, DYNLL1, SERPINB6, ERG, DPT, TMPRSS2, FOLH1
Lung	EGFR, KRAS, ALK, PCSK9, CBX8, ARCN1, KLK3, BRCA1, F7, TTF1
Liver	HCCS, EPHB6, AFP, TRIM26, HSPG2, ADAM17, LDLR, DNLZ, ALB, CCL15
Colon	AOM, DLD, CEACAM5, DDX53, APC, CTNNB1, GAST, PPARG, WNT16, SELE
Colorectal	KRAS, APC, MSI2, CTNNB1, MSI1, CEACAM5, MRC1, BRAF, FAP, MLH1
Pancreatic	INS, GCG, MIA, SST, CCK, KRAS, ZGLP1, PDX1, PRSS27, SMAD4
Ovarian	BRCA2, MUC16, BRCA1, LPA, DIRAS3, BRCA3, ARID1A, HEY1, GNRHR, ABCB1
Skin	SERPINB3, TNFRSF8, COL1A1, CMM, MLANA, EGFR, MC1R, CPD, CD8A, MCC
Brain	GFAP, MGMT, IDH1, MS, SMS, NEFH, KIAA1549, CSF2, GSC, NAA60

an integrative loop for eCuration:²³ distant supervision produces initial extractions, eCurators then verify them via an online interface such as Literome,²⁰ fixing errors by a click of buttons, which is much more efficient than annotation from scratch. The feedback could be fed back into distant supervision via online learning and used to continuously improve extraction quality. Active learning can also be incorporated by prioritizing the least confident extractions for annotator verification.

3.3. Cancer Pathway Analysis

With cancer contexts identified by MeSH terms (Section 2.6) and PubMed extractions produced from distant supervision (Section 3.2), we conducted an exploratory analysis to survey the research landscape and findings on cancer pathways.

Cancer Pathway Research Among the 1.5 million pathway extractions, 150,379 occur in the context of cancer, or about 10%; among the 838 thousand unique pathway relations, 108,373 occur in the context of cancer, or about 13%. Table 4 shows the top ten most studied cancer types, along with the top ten genes for each type, both in the number of unique pathway relations found in our extraction. Not surprisingly, many well-known cancer genes are in the list, such as TP53 and EGFR. Overall, the top ten most studied genes for cancers are: TP53, EGFR, VEGFA, CBX8, ESR1, SLC22A3, AKT, MYLIP, EGF, EPHB2. For non-cancer context: INS, TNF, CA2, TCF, CD4, TP53, IRF6, EPHB2, CALM3, CASP.

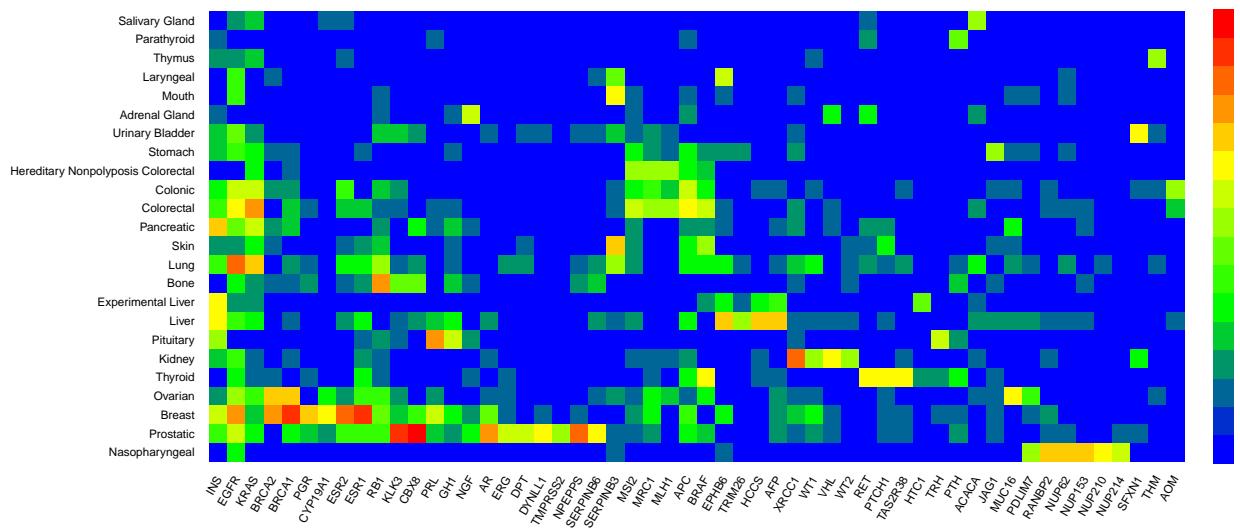


Fig. 3. Heatmap indicating the strength of association between cancer types and genes.

Cancer-Specific Genes Genes like TP53 are crucial for many cancer types and are well studied across the board. Additionally, we are interested in genes that act specifically to a cancer type. We thus searched for non-random association between genes and cancer types, using the log-likelihood ratio test where the null model assumes that the gene occurs independently from the cancer type in each extraction. Figure 3 shows a heatmap for the association scores between the top ten cancers and their most significantly associated genes. Table 5 shows the top ten most studied cancer types again, along with the top ten genes ranked by association score instead. Compared to top genes by relation counts, these lists are clearly type-specific, revealing many well-known associations, such as BRCA1, BRCA2, ESR1, ESR2 for Breast Cancer, EGFR, KRAS for Lung Cancer, BRCA1, BRCA2, BRCA3 for Ovarian Cancer, etc. Conversely, the highly ranked genes with lower extraction counts might reveal research opportunities for genes that are not yet well studied but potentially important for a cancer type. In general, our PubMed-scale extraction enables cancer pathway analysis encompassing the entire research landscape, revealing interesting insights as well as suggesting future research priorities.

4. Discussion

In this paper, we present the first attempt to apply distant supervision to pathway extraction. Evaluation on the GENIA event extraction dataset shows that distant supervision substantially outperforms rule-based extraction and other baselines, attaining an accuracy approaching supervised upper bounds. Application to all PubMed abstracts using the PID database yielded an order of magnitude more correct pathway interactions than the original database. Analysis of cancer-related interactions led to a number of interesting observations. The extracted pathways, sample annotation, and trigger rules used by the rule-based system will be made available at literome.azurewebsites.net/papers/psb15.

Overall, this demonstrates the great potential of distant supervision for cancer pathway

extraction and biological knowledge extraction in general: distant supervision could attain high accuracy while requiring substantially less development effort compared to both rule-based and supervised approaches.

This also opens up a number of interesting future research directions. Currently, pathway extraction is pursued in an isolated fashion, feeding on output from other tasks such as protein extraction. Joint learning with distant supervision is a promising direction for improving the accuracy in all pipeline tasks. Existing distant supervision methods are only applicable to binary relations; lifting this limitation to handle n-ary and nested relations is an important future direction, and is particularly important for identifying relevant contexts and reconciling seemingly conflicting relations. Our system currently ignores event modalities such as negation and hedging, which need to be incorporated in the future. Distant supervision can be seamlessly integrated in eCuration, combining with online learning from annotator feedback, and active learning for prioritizing verification requests. Finally, a particularly exciting prospect is to integrate the extracted pathways with high-throughput panomics data for automating discovery in genomic medicine.⁵

References

1. D. Hanahan and R. A. Weinberg, *Cell* **144**(5), 646 (2011).
2. B. Vogelstein, *et al.*, *Science* **339** (2013).
3. A. G. Stephen, *et al.*, *Cancer Cell* **25**(3), 272 (2014).
4. K. Wang, *et al.*, *Nature Reviews Genetics* **11**, 843 (2010).
5. C. J. Vaske, *et al.*, *Bioinformatics* **26**, 237 (2010).
6. T. Ideker, *et al.*, *Cell* **144**, 860 (2011).
7. S. Ng, *et al.*, *Bioinformatics* **28**, 640 (2012).
8. A. J. Sedgewick, *et al.*, *Bioinformatics* **29**, 62 (2013).
9. A. Rzhetsky, *et al.*, *J Biomed Inform.* **37**(1), 43 (2004).
10. M. Craven and J. Kumlien, Constructing biological knowledge bases by extracting information from text sources, in *Proc. of the 7th International Conference on Intelligent Systems for Molecular Biology*, 1999.
11. M. Mintz, *et al.*, Distant supervision for relation extraction without labeled data, in *Proc. of the Forty Seventh Annual Meeting of the Association for Computational Linguistics*, 2009.
12. S. Riedel, *et al.*, Modeling relations and their mentions without labeled text, in *Proc. of the Sixteen European Conference on Machine Learning*, 2010.
13. R. Hoffmann, *et al.*, Knowledge-based weak supervision for information extraction of overlapping relations, in *Proc. of the Forty Ninth Annual Meeting of the Association for Computational Linguistics*, 2011.
14. J.-D. Kim, *et al.*, Overview of BioNLP-09 Shared Task on event extraction, in *Proc. of the BioNLP Workshop*, 2009.
15. C. F. Schaefer, *et al.*, *Nucleic Acids Research* **37**, 674 (2009).
16. K. Ravikumar and H. Liu, Towards pathway curation through literature mining - a case study using PharmGKB, in *Proc. Pacific Symposium of Biology*, 2014.
17. C. Quirk, P. Choudhury, J. Gao, H. Suzuki, K. Toutanova, M. Gamon, W. Yih and L. Vanderwende, MSR SPLAT, a language analysis toolkit, in *Proc. of NAACL HLT Demonstration Session*, 2012.
18. M.-C. de Marneffe, *et al.*, Generating typed dependency parses from phrase structure parses, in *Proc. of the Fifth International Conference on Language Resources and Evaluation*, 2006.
19. S. V. Landeghem, *et al.*, *PLoS One* **8** (2013).
20. H. Poon, *et al.*, *Bioinformatics* (2014).
21. C. Quirk, *et al.*, MSR-NLP entry in BioNLP Shared Task 2011, in *Proc. BioNLP*, 2011.
22. S. Riedel and A. McCallum, Fast and robust joint models for biomedical event extraction, in *Proc. of the Conference on Empirical Methods in Natural Language Processing*, 2011.
23. C.-H. Wei, *et al.*, *Database* (2012).

UNSUPERVISED FEATURE CONSTRUCTION AND KNOWLEDGE EXTRACTION FROM GENOME-WIDE ASSAYS OF BREAST CANCER WITH DENOISING AUTOENCODERS

JIE TAN, MATTHEW UNG, CHAO CHENG, CASEY S GREENE*

Department of Genetics

Institute for Quantitative Biomedical Sciences

Norris Cotton Cancer Center

The Geisel School of Medicine at Dartmouth

Hanover, NH 03755, USA

**E-mail: Casey.S.Greene@dartmouth.edu*

Big data bring new opportunities for methods that efficiently summarize and automatically extract knowledge from such compendia. While both supervised learning algorithms and unsupervised clustering algorithms have been successfully applied to biological data, they are either dependent on known biology or limited to discerning the most significant signals in the data. Here we present denoising autoencoders (DAs), which employ a data-defined learning objective independent of known biology, as a method to identify and extract complex patterns from genomic data. We evaluate the performance of DAs by applying them to a large collection of breast cancer gene expression data. Results show that DAs successfully construct features that contain both clinical and molecular information. There are features that represent tumor or normal samples, estrogen receptor (ER) status, and molecular subtypes. Features constructed by the autoencoder generalize to an independent dataset collected using a distinct experimental platform. By integrating data from ENCODE for feature interpretation, we discover a feature representing ER status through association with key transcription factors in breast cancer. We also identify a feature highly predictive of patient survival and it is enriched by FOXM1 signaling pathway. The features constructed by DAs are often bimodally distributed with one peak near zero and another near one, which facilitates discretization. In summary, we demonstrate that DAs effectively extract key biological principles from gene expression data and summarize them into constructed features with convenient properties.

Keywords: Denoising Autoencoders; Breast Cancer; Feature Construction; Functional Genomics.

1. Introduction

Modern genomic technologies have dramatically reduced the cost of comprehensively interrogating biological systems, which has made genome-scale measurements a routine part of biology. This has produced vast amounts of data from which we can extract complex biological principles, potentially broadening our horizon from examining a single pathway to extracting and summarizing the global principles that underlie complex biological systems.

While “big data” introduce new opportunities, they also raise new computational challenges. Even though hardware advances will play a role, new algorithms that can extract, represent, and reason about the principles embedded in such data are also needed. Supervised machine learning algorithms have been developed to predict gene functions and interacting partners¹ or to identify new disease-related genes.² These methods identify new genes or interactions by building upon known examples and consequently are well suited to discovering key biological features, but are not well suited to discovering new processes. Unsupervised algorithms such as clustering can identify key relationships embedded within data, e.g. the

existence of tumor subtypes,³ but tend to identify only the strongest signals in the data.

To best utilize big data in reasoning systems, the feature extraction method should allow for the discovery of new pathways and principles, construct features with amenable distributions, and build features that generalize across datasets. Based on these key factors, we identified Denoising Autoencoders (DAs) as a promising approach. DAs are a variant of Artificial Neural Networks (ANNs), but unlike ANNs, which are frequently used for classification, the goal of DAs is to learn compact and efficient representations from input data.⁴ DAs improve upon the classic autoencoder by incorporating noise during training, a procedure which generates robust features.⁵ The training objective for DAs is to build features that reconstruct initial input data from corrupted data, i.e. input data with random noise added. DAs depart from supervised ANNs whose performance heavily relies on the quality of gold standards, and because the data define the objective function for the algorithms, DAs can directly use unlabeled data. DAs also serve as building blocks for deep networks⁶ which have gained popularity in fields such as image and audio processing for their high performance. Compared with commonly used feature extraction approaches such as PCA or ICA that linearly map input to features , DAs extract features in the non-linear space.

In this paper, we introduce an unsupervised feature construction approach based on DAs that summarizes available genomic data and extracts useful features. We apply this approach to a large compendium of breast cancer gene expression data. We demonstrate that the DA trained on one breast cancer dataset captures the same biological features in an independent dataset measured by a different technology. We use the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) cohort⁷ as a training set and the cohort from The Cancer Genome Atlas (TCGA)⁸ as an independent evaluation set. We demonstrate an approach that uses known characteristics to perform post-hoc interpretation of these features. The DA constructs features that distinguish tumor from normal samples, classify patients' estrogen receptor (ER) status, summarize intrinsic subtypes, and identify the activity of key transcription factors (TFs). We also observe a feature that is highly predictive of patient survival. This constructed feature is more predictive of survival than commonly used markers such as tumor grade or ER status. These results suggest that constructed features from denoising autoencoders shed light on unexplored knowledge in breast cancer and provide a fruitful ground for approaches that aim to reason from such data.

2. Methods

2.1. *Construction of Denoising Autoencoders from Genomic Data*

DAs provide a powerful means to analyze audio or image data where measurements are temporally or spatially linked.^{4,9} Here we describe a technique that allowed construction of DAs from genomic data, which do not have temporal or spatial linkage. These DAs summarized the gene expression characteristics of both breast cancers and normal tissues to automatically extract biologically and clinically relevant features. The constructed DA networks contained three layers: an input layer, a hidden layer and a reconstructed layer (Fig. 1A). The hidden layer represented the constructed features, with each node corresponding to one feature. The reconstructed layer and the input layer had the same dimensions, and the objective function

for the algorithm was to minimize the difference between the two layers. Noise was added to the input data during training to construct robust, high-quality features.⁶

Here we describe the detailed training process for DAs from genome-wide transcriptome measurements. We define the term “sample” to represent the complete gene expression vector for each collected tissue biopsy. To facilitate training, samples were randomly grouped into batches, and the number of samples contained in a batch was termed the batch size. For each sample in a batch, a set of genes matching a defined proportion of genes (termed the corruption level) were randomly chosen. The expression values for these genes in this sample were set to zero. Like other feed-forward ANNs, the hidden layer y was constructed by multiplying one sample x with a weight matrix W . A bias vector, b , was added before transformation by the sigmoid function (Formula 1). The value contained in the hidden vector y for each node was termed the activity value of that node. The reconstructed layer was generated from the hidden layer in a similar manner (Formula 2). We used tied weights, which meant that the transpose of W was used for W' . Cross-entropy (Formula 3) was used to measure the difference between the input layer (x) and the reconstructed layer (z). Thus, the problem became fitting appropriate weights and bias terms to minimize the cross-entropy. This optimization was achieved by stochastic gradient descent using the Theano¹⁰ library , with weight and bias being updated after each batch and the size of each update is controlled by learning rate. Training proceeded through epochs, and samples were rebatched at the beginning of each epoch. Training was stopped after a specified number of epochs (termed epoch size) was reached.

$$y = \text{sigmoid}(Wx + b) \quad (1)$$

$$z = \text{sigmoid}(W'y + b') \quad (2)$$

$$L_H(x, z) = - \sum_{k=1}^d [x_k \log z_k + (1 - x_k) \log (1 - z_k)] \quad (3)$$

To determine the appropriate parameter setting for the METABRIC dataset described in Section 2.3, we carried out a parameter sweep with 10-fold cross validation. We performed a full factorial design over all combinations of the following parameters: epoch size of 100, 200, 500; batch size of 1, 10, 20, 50; corruption level of 0.0, 0.1, 0.2; learning rate of 0.005, 0.01, 0.05. The dimension of hidden layer was manually set as 100 since it is a amenable size for interpretation and could be trained efficiently. With the selected parameters, all samples in the METABRIC dataset were used to train a DA network. Then we fixed the optimized weight matrix and bias vectors and calculated the activity values for each node in the hidden layer for each sample. We used the weights and bias terms derived from METABRIC to directly calculate the activity values for the same nodes for each sample in an independent set characterized by TCGA (Section 2.3). Specifically, the weight matrix, W , is derived during training the METABRIC dataset, and we calculate activity values for each node over each sample in both the METABRIC and TCGA datasets based on the same weight matrix W .

2.2. Interpretation of Constructed Features

A major weakness of traditional ANNs has been the difficulty of interpreting the constructed models. DAs have largely been used in image processing, where these algorithms construct

features that recognize key components of images, for example diagonal lines . Unlike pixels in image data, genes are not linked to their neighbors, and unlike audio data, they are not linked temporally. Instead they are linked by their transcription factors, their pathway membership, and other biological properties. To address this interpretation challenge, we developed strategies that allow constructed features to be linked to clinical and molecular features of the underlying samples.

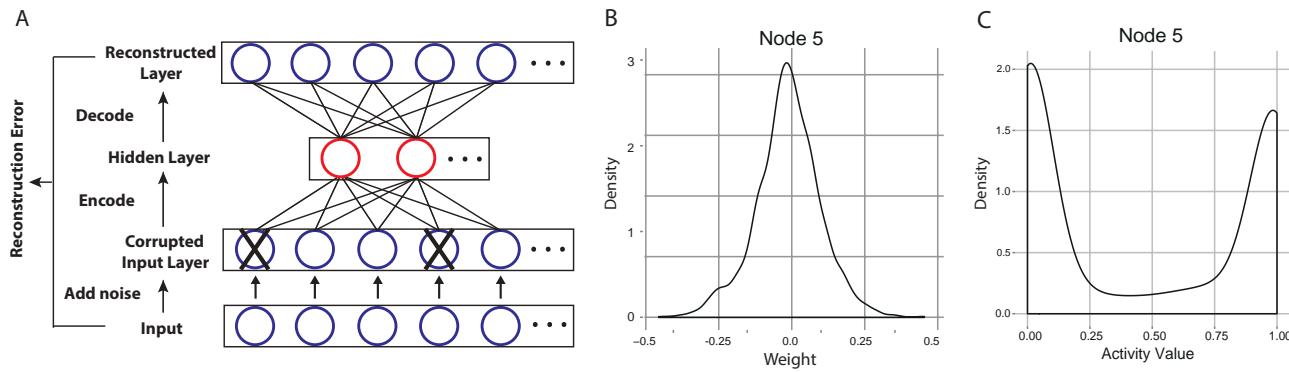


Fig. 1. A) The network structure of denoising autoencoders. B) The distribution of one node's weight vector. C) The distribution of activity values for a node are bimodally distributed. Here we use Node5 as an example.

2.2.1. Linking Constructed Features to Sample Characteristics

We first linked constructed features to specific sample characteristics. This included the identification of features that categorize tumor and normal samples, molecular subtypes, and ER status. We divided the METABRIC samples into two parts: two thirds of the sample set (1424 samples) was used for a discovery set, and one third of the sample set (712 samples) was reserved as a test set. For these tasks, we binarized each node activity by identifying each node's highest and lowest activity values among the samples in the discovery set and defined 10 equally spaced activation thresholds between these values. We evaluated the balanced accuracy for each node at each threshold to predict the desired sample characteristic. For each task, we chose the nodes with the highest balanced classification accuracies and recorded the corresponding thresholds. These high-accuracy nodes were tested on the test set using the activation threshold identified in the discovery set. To avoid sampling bias, the above procedure was repeated ten times with random partitioning of the discovery and testing sets. The final reported discovery/test accuracy for a node represents these accuracies averaged over the ten runs. We applied this procedure to identify nodes that could stratify tumor/normal samples, ER+/- samples, and categorize samples into molecular subtypes (i.e. Luminal A, Luminal B, Basal-like, HER2-enriched, and Normal-like). We verify that this procedure avoids overestimating the performance of constructed features by evaluating these features in the independent TCGA dataset without retraining. This separate dataset was never used at any prior stages, and in this dataset gene expression values were measured by an entirely different experimental platform. To evaluate each node from METABRIC in TCGA, we used the average of the thresholds identified over the ten METABRIC partitions as the activation threshold for TCGA samples. We termed this the “independent evaluation” performance of the node.

2.2.2. *Linking Constructed Features to Transcription Factors*

To interpret selected features in the context of transcriptional programs that they summarized, we developed an approach to link transcription factors to constructed features. The weights connecting the input and hidden layers determined how each gene in the input layer influenced the activity values of each node in the hidden layer. The distribution of weights for all genes to a single node approximately resembled a normal distribution centered at zero (Fig. 1B). Most genes gave zero or low weight to a hidden node; while a small number of genes gave high positive or high negative weights. Hence, we defined “high-weight genes” as those exhibiting weight values that lie outside two standard deviations from the mean of the weight distribution. We identified nodes whose high-weight genes were overrepresented by genes bound by one transcription factor and calculated the odds ratio. We employed the transcription factor ChIP-seq data from ENCODE via the UCSC genome browser at <http://genome.ucsc.edu/ENCODE/downloads.html>.^{11,12} ChIP-seq data were derived from experiments in T47d and MCF7, a pair of breast cancer cell lines that were expected to best match these breast cancer datasets. All transcription factor target genes were determined using a probabilistic model called Target Identification for Profiles (TIP).¹³

2.2.3. *Linking Constructed Features to Patient Survival*

We linked features constructed by the DAs to patient survival. We evaluated patient survival using the METABRIC dataset because it contained up to 15 years of follow-up on each patient. Because the distribution of activity values for each node is bimodal with one peak close to 0 and another close to 1 (Fig. 1C), we defined 0.5 as the activation cutoff to divide samples into two groups. We assessed the differences of these groups for each node by Kaplan-Meier curves and the non-parametric logrank test using the survival package in R.¹⁴ The node whose activities best separated two high and low survival groups was the one with the most prognostic power. To evaluate the importance of this constructed feature, we further compared this feature with frequently-used clinical markers of survival including tumor grade, molecular subtype and ER status.

2.2.4. *Linking Specific Features to Biological Pathways*

To interpret selected constructed features in the context of biological pathways, we developed an approach that leverages gene set enrichment analysis (GSEA)¹⁵ to associate known pathways to constructed features. GSEA was developed to identify enrichment within the distribution of differentially expressed genes from microarray experiments. We applied GSEA to a node’s weight vector, which identified pathways associated with genes that consistently gave high positive or high negative weight to a node. These genes also exhibited the most influence over the activity value of that node. We defined significantly associated pathways using a false discovery rate (FDR) q-value threshold of 0.1. We used cancer pathways available from the Pathway Interaction Database (PID) curated by the National Cancer Institute and Nature Publishing Group¹⁶ as gene sets for GSEA. In total, this set contained 196 pathways downloaded from the Molecular Signatures Database (MSigDB).¹⁵

2.3. The METABRIC and TCGA Breast Cancer Compendium

We constructed a compendium consisting of the two largest molecular characterizations of breast cancer to date. The first dataset, METABRIC, was collected from tumor banks in the UK and Canada and contained 1992 clinically annotated breast cancer specimens and 144 tumor-adjacent normal tissues.⁷ We used this dataset to train DAs and identify predictive features. The gene expression data from this study were downloaded from the European Genome-phenome Archive after approval by the specified Data Access Committee. The transcriptomes of tumor and normal tissue samples were profiled on the Illumina HT-12 v3 platform. We converted Illumina identifiers to gene symbols, and the expression values for identifiers that mapped to the same gene symbol were averaged. Missing values were imputed using KNNImputer from the Sleipnir library¹⁷ using 10 neighbors as recommended by Troyanskaya et al.¹⁸ We used median absolute deviation (MAD) to filter genes with invariant expression across samples and retained the top 3000 genes with the highest MAD values.

The second dataset was obtained from The Cancer Genome Atlas (TCGA). It consisted of 522 primary tumors, 3 metastatic tumors, and 22 tumor-adjacent normal samples.⁸ These data were obtained by measurement on three distinct platforms, none of which match the METABRIC platform. The TCGA consortium constructed a unified expression collection that summarized genes' measurements from three platforms into one mean-centered expression value per gene. In this unified expression dataset, genes were identified by their symbol. This dataset served as our independent evaluation dataset, and no training, discovery, or threshold selection was performed on this dataset.

In order to evaluate DAs constructed from the METABRIC dataset directly on the independent TCGA dataset, we removed genes that were not measured by TCGA. This results in a METABRIC set containing 2520 genes measured for 2136 samples and a TCGA set containing the same 2520 genes measured for 547 samples. DAs use values between zero and one in the input vector, so the range of expression values for each gene were linearly transformed to this range.

3. Results and Discussion

To summarize our breast cancer gene expression compendium and construct biologically meaningful features, DAs were trained using the METABRIC dataset and evaluated on the TCGA dataset. We interpreted the constructed features by sample characteristics classification, transcription factor enrichment, survival analysis, and pathway analysis. We identified features representing a variety of important clinical or molecular characters of breast cancer, including sample type (tumor or normal tissue), ER status, and intrinsic subtype. They were shown to be robust across datasets. In addition, breast cancer related transcription factors were found to be linked to these features. Finally, DAs constructed a novel feature that was highly predictive of patient survival, and from this we uncovered a variety of biological processes enriched in that feature.

3.1. Construction of a Denoising Autoencoder from Genomic Data

To apply DAs to genomic data for the first time, we performed a full factorial parameter sweep over the METABRIC dataset. We fixed the number of nodes in the hidden layer to 100 to fix the number of weights the algorithm is allowed to fit under each parameter setting. The remaining parameters were evaluated by the ability of autoencoders generated using each set to reconstruct held out test data. The parameters that we selected were: a batch size of 10, an epoch size of 500, a corruption level of 0.1, and a learning rate of 0.01. We note that by keeping other parameters constant while setting the corruption level to 0.0 (no noise added) results in the best performance. We chose a corruption level of 0.1 to avoid the risk of overfitting and to improve the quality and robustness of constructed features. Although METABRIC dataset is one of the largest available expression datasets of breast cancer, they have fewer examples than most datasets used for training neural networks. We observe that DAs still effectively summarize dataset of this size. After selecting these parameters, DAs were built by training on the METABRIC dataset. We named the constructed features “Node##” based on the order in which they appear in the hidden layer. The DAs trained on the METABRIC dataset were directly applied to the TCGA dataset to generate activity values for each already constructed feature in each sample. The results of our parameter sweep, as well as the activity scores for each feature in each sample, are available for download from the online supplement.

3.2. Features constructed by DAs represent clinical characteristics

In order to examine whether the features constructed by DAs exhibit clinical significance, we assessed the ability of hidden nodes to classify two important clinical features: sample type and ER status. We first identified hidden nodes that best classified whether samples were obtained from tumor or normal tissue using two thirds of the METABRIC samples (discovery set) and then tested the performance of these nodes in the remaining samples (test set). Tumor and normal tissues are very distinct from each other, and consequently there should be at least one feature that separates these with high accuracy. Table 1 shows the balanced classification accuracies calculated during discovery, testing, and an independent evaluation dataset. The top 5 hidden nodes are ordered by their performance in the discovery set. Nodes 64 and 99 achieved very high accuracy in distinguishing tumor from normal samples in the METABRIC dataset. More interestingly, Node64 and Node99 also classified TCGA samples with near-perfect accuracy. This indicates that training the weights matrix using all of the METABRIC samples does not appear to lead to test set contamination during the feature interpretation phase.

Next we sought to apply our framework to the more challenging task of constructing features with activities that reflect the ER status of breast cancer samples. The estrogen receptor plays an important role in the progression of breast cancer, as the majority of breast cancers start out as estrogen dependent and overexpress the estrogen receptor.¹⁹ A tumor’s expression status (ER+ vs. ER-) serves as an immunohistological biomarker that helps determine whether patients will benefit from endocrine therapy.²⁰ Table 2 shows that Node30 and Node58 achieve the highest accuracy for ER status classification. As expected, the accuracy associated with stratifying samples into ER categories is not as high as when stratifying tumor and

Table 1. Performance of hidden nodes in classifying tumor from normal samples.

Node	METABRIC		TCGA
	Discovery	Test	Evaluation
64	0.970	0.968	0.996
99	0.957	0.959	0.998
38	0.879	0.887	0.911
43	0.873	0.873	0.750
69	0.871	0.872	0.906

Table 2. Performance of hidden nodes in classifying ER + from ER - samples.

Node	METABRIC		TCGA
	Discovery	Test	Evaluation
89	0.848	0.833	0.749
30	0.824	0.822	0.856
58	0.808	0.801	0.828
6	0.798	0.799	0.771
69	0.784	0.779	0.820

normal samples. This is because ER signaling is a complex biological process involving many co-regulatory proteins and can be activated by both estrogen ligands and a variety of other pathways.²⁰ Thus, the expression pattern for ER positive samples and ER negative samples is more complicated, making classification more challenging. Again, the results for ER status stratification in METABRIC and TCGA datasets indicate that features constructed by DAs maintain robust performance across datasets. This is consistent with our observations from the tumor/normal separation and further indicates that the unsupervised training using all of METABRIC does not lead to contamination during the discovery/test phase.

Transcription factors FOXA1 and GATA3 are two key determinants of ER function and endocrine response. FOXA1 facilitates ER-chromatin interactions that are necessary for ER-mediated gene regulation,²¹ and GATA3 acts upstream of FOXA1 in modulating the accessibility of enhancers bound by ER.²² Therefore, to understand whether these nodes were capturing underlying biological principles, we evaluated whether these TFs were strongly associated with nodes categorizing ER status. We calculated odds ratios for each node/TF pair as described in section 2.2.2. We found that Node58 achieved the highest odds ratios for both FOXA1 and GATA3 among all nodes. Specifically, it was enriched of genes regulated by FOXA1 with an odds ratio of 4.02 and of genes regulated by GATA3 with an odds ratio of 3.16. These results suggest that Node58 was able to distinguish ER+ from ER- breast cancers with high accuracy because it contained genes that reflect the activity of key ER-associated TFs.

3.3. Features constructed by DAs recapitulate molecular subtypes

Table 3. Performance of hidden nodes in classifying each intrinsic subtype.

Subtype	Basal	Her2	LumA	LumB	Normal	LumA/B
Node	30	29	5	66	42	6
METABRIC Discovery	0.929	0.761	0.780	0.755	0.750	0.849
METABRIC Test	0.918	0.741	0.777	0.750	0.748	0.849
TCGA Evaluation	0.992	0.712	0.800	0.717	0.733	0.825

Breast cancer is a complex and heterogeneous disease caused by various molecular alterations in distinct signaling pathways. Breast cancer patients respond differently to treatment and show diverse clinical outcomes. Categorizing breast cancers into molecular subtypes that

exhibit similar characteristics opens the door to the development of subtype-specific prognostic markers and treatments. Parker et al. developed a 50-gene subtype predictor (PAM50) based on gene expression data.²³ PAM50 subtypes are available for both the METABRIC and TCGA cancers. Here, we evaluated whether features constructed by DAs can also predict these subtypes. We defined each subtype prediction as a binary classification problem and show the results in Table 3. The table contains the node that achieves the highest accuracy in the discovery set. In general, we observed that only one node was predictive of each subtype. There were two exceptions: for the Normal-like subtype Node38 performed similarly to Node42, and for the Luminal A subtype Node23 performed similarly to Node5. We observed that the Basal-like subtype reached the highest accuracy, which is consistent with clustering results showing Basal as the most distinct cluster.⁸ Luminal A (LumA) and Luminal B (LumB) subtypes were mixed together in the clustering analysis as well. Combining LumA and LumB into one subtype identified features that obtained higher accuracies. Both Node30 (Basal) and Node6 (LumA/B), identified here as subtype nodes, were also predictive of ER status. This is because ER positive patients usually fall into the luminal subtypes, and ER negative patients usually fall into the Basal subtype. In METABRIC, 77.6% of ER positive samples are LumA or LumB, while only 3.9% of ER positive samples are from the Basal subtype. Therefore the ER status signal is sufficient to identify these groups, but not differentiate between, for example, the two luminal subtypes. The specific LumA and LumB subtypes are best captured by different nodes. Methods that aim to reason based upon features constructed from the DA may be able to exploit these distinct but complementary features to obtain superior accuracy and better reflect the underlying complex biology of breast cancer.

3.4. Features constructed by DAs are associated with prognosis

Many factors have been correlated to prognosis, such as histologic grade,²⁴ ER status,²⁵ lymph node metastases,²⁶ and intrinsic molecular subtypes.²⁷ Here we assessed the correlation between features constructed by DAs and patient prognosis. As shown in Fig. 1C, the distribution of node activity was bimodal. We separated patients into two groups based on their node activity using a cutoff of 0.5 and then correlated activity values of each node to patient survival time. The node that best separated good and poor prognosis was Node5 (logrank p-value of $2.1e^{-20}$; Fig. 2A). To compare the performance of Node5 with other clinical or molecular features, we carried out survival analysis for ER status (Fig. 2B), each tumor grade, and each intrinsic molecular subtype. The LumA subtype had a significantly better prognosis than other subtypes and provided the strongest subtype-survival association in the METABRIC dataset (Fig. 2C). A tumor of grade 3 (clinically termed high-grade) was a sign of poor prognosis when compared to grades 1 and 2 (Fig. 2D). Comparing the ability of these survival-associated breast cancer features to Node5, Node5 was more strongly associated with survival. Interestingly, Node5 was also the node that best classifies the LumA subtype (Table 3). We investigated the associations of each node with tumor grade and found that Node5 was also most strongly associated with grade (online supplement). These results suggest that Node5 learned a combined expression pattern that captures features of both the LumA subtype and a tumor's grade, which contributed to its strong association with survival.

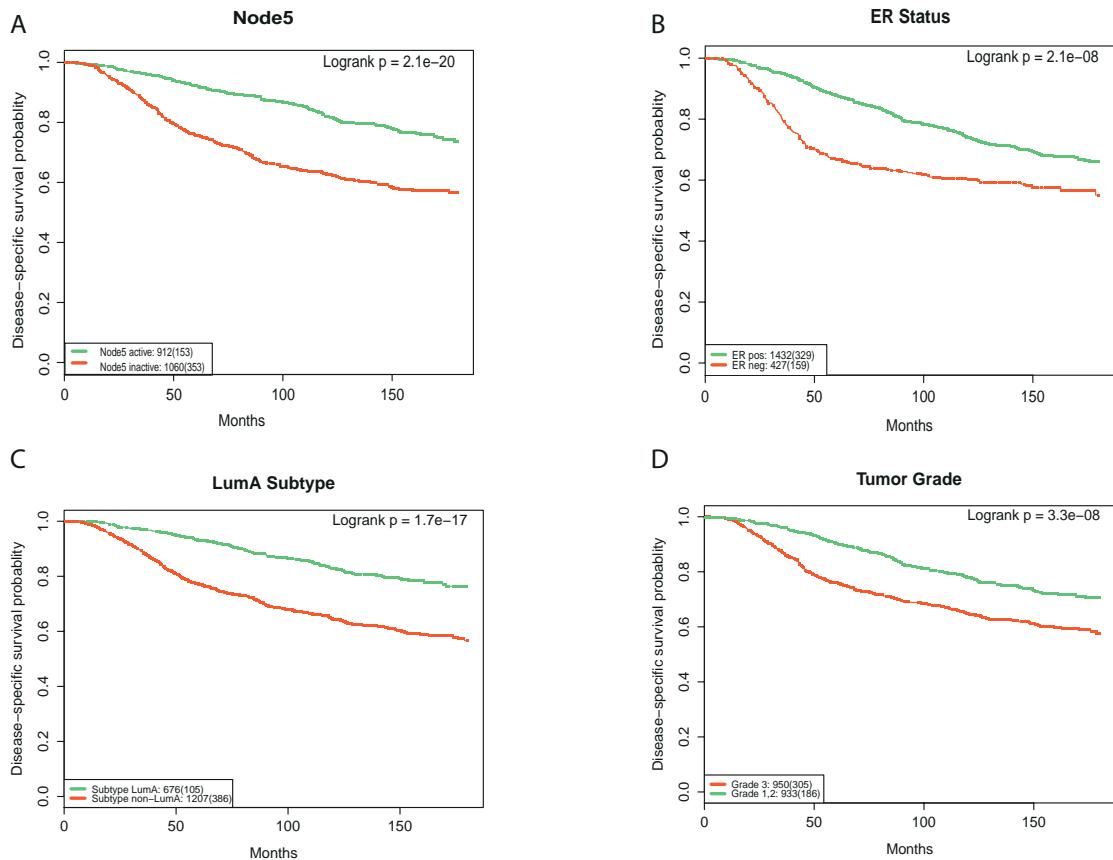


Fig. 2. Kaplan-Meier plots of disease-specific survival for Node5 (A), ER status (B), Luminal A subtype (C), and Tumor Grade (D) demonstrate that the constructed feature outperforms the other predictors.

To further investigate how Node5 learned this combined pattern, we performed pathway analysis using a modified GSEA. GSEA identified pathways that defined the activity values of Node5 (Table 4). The most significant pathway was the FOXM1 transcription factor network. FOXM1 is one of the most overexpressed genes in breast cancer²⁸ and its down-regulation has been shown to inhibit proliferation, migration and invasion of breast cancer cells.²⁹ The Aurora Kinase A, Aurora Kinase B, and PLK1 pathways affect cell cycle progression.^{30,31} Misregulation of the cell cycle is related to both tumor formation and growth and is consistent with FOXM1's role in modulating cell cycle progression. The number of cells undergoing division is used in determination of a tumor's grade, indicating a potential relation for Node5 to grade based on the number of mitotic events. The other two pathways have also been demonstrated to be key players in breast cancer,^{32,33} and c-Myb is a known marker of the luminal subtypes³⁴ indicating a potential connection between

Table 4. PID pathways enriched in Node5.

Pathway	FDR q-value
FOXM1 transcription factor network	< 1e ⁻⁴
Aurora B signaling	4.93e ⁻⁴
Aurora A signaling	0.001
PLK1 signaling	0.003
Integrin-linked kinase signaling	0.068
C-MYB transcription factor network	0.074

Aurora Kinase A, Aurora Kinase B, and PLK1 pathways affect cell cycle progression.^{30,31} Misregulation of the cell cycle is related to both tumor formation and growth and is consistent with FOXM1's role in modulating cell cycle progression. The number of cells undergoing division is used in determination of a tumor's grade, indicating a potential relation for Node5 to grade based on the number of mitotic events. The other two pathways have also been demonstrated to be key players in breast cancer,^{32,33} and c-Myb is a known marker of the luminal subtypes³⁴ indicating a potential connection between

Node5 and the Luminal A subtype.

4. Conclusion

While machine learning has made key contributions to biology, a gap still exists between data integration methods, which are largely supervised, and discovery-oriented approaches, which are unsupervised and don't condition on known biology. We have evaluated DAs as a means to fill this gap allowing us to develop unsupervised methods for data integration. We found that DAs effectively summarize key features in breast cancer data. We identified features that stratify tumor/normal samples, ER+/- samples, and molecular subtypes, in addition to identifying transcription factor activities. Moreover, DAs constructed a feature significantly associated with the FOXM1 pathway that was highly predictive of patient survival by combining information from the Luminal A subtype and tumor grade. A DA constructed from one dataset identifies the same features in an independent dataset, even though the strongest principle component of a combined dataset captures the underlying study highlighting the major methodological differences between these studies.³⁵

Future work will focus on developing new approaches to interpret features, especially features that cannot be mapped to existing knowledge but may represent new signals. We will also evaluate deep network architectures with multiple layers of stacked DAs. We anticipate that employing features generated by DAs in supervised learning will improve prediction accuracies, as we have shown that these features comprise clinical information and molecular patterns. Because the patterns observed are consistent across datasets, we also anticipate that DAs will provide a fruitful mechanism for data integration. Future work should explore the scope and limitations of this approach for large-scale data integration. Overall, we anticipate that DAs can provide a new mechanism to effectively summarize and integrate large compendia of genomic data in an unsupervised manner.

Acknowledgements: This work was supported in part by P20 GM103534 from the NIH and an American Cancer Society Research Grant, #IRG-82-003-27. JT is a Neukom Graduate Fellow supported by the William H. Neukom 1964 Institute for Computational Science.

Supplement: <http://discovery.dartmouth.edu/~cgreene/da-psb2015/>.

References

1. A. K. Wong, C. Y. Park, C. S. Greene, L. A. Bongo, Y. Guan and O. G. Troyanskaya, *Nucleic Acids Research* **40**, W484 (2012).
2. E. E. Schadt, J. Lamb, X. Yang, J. Zhu, S. Edwards, D. GuhaThakurta, S. K. Sieberts, S. Monks, M. Reitman, C. Zhang *et al.*, *Nature Genetics* **37**, 710 (2005).
3. T. Sørlie, R. Tibshirani, J. Parker, T. Hastie, J. Marron, A. Nobel, S. Deng, H. Johnsen, R. Pesich, S. Geisler *et al.*, *Proceedings of the National Academy of Sciences* **100**, 8418 (2003).
4. P. Vincent, H. Larochelle, Y. Bengio and P.-A. Manzagol, Extracting and composing robust features with denoising autoencoders, in *Proceedings of the 25th International Conference on Machine Learning*, (ACM, New York, NY, USA, 2008).
5. Y. Bengio, *Foundations and trends in Machine Learning* **2**, 1 (2009).
6. P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio and P.-A. Manzagol, *The Journal of Machine Learning Research* **11**, 3371 (2010).

7. C. Curtis, S. P. Shah, S.-F. Chin, G. Turashvili, O. M. Rueda, M. J. Dunning, D. Speed, A. G. Lynch, S. Samarajiwa, Y. Yuan *et al.*, *Nature* **486**, 346 (2012).
8. The Cancer Genome Atlas Network, *Nature* **490**, 61 (2012).
9. G. E. Dahl, D. Yu, L. Deng and A. Acero, *Audio, Speech, and Language Processing, IEEE Transactions on* **20**, 30 (2012).
10. J. Bergstra, O. Breuleux, F. Bastien, P. Lamblin, R. Pascanu, G. Desjardins, J. Turian, D. Warde-Farley and Y. Bengio, Theano: a CPU and GPU math expression compiler, in *Proceedings of the Python for Scientific Computing Conference (SciPy)*, June 2010. Oral Presentation.
11. M. B. Gerstein, A. Kundaje, M. Hariharan, S. G. Landt, K.-K. Yan, C. Cheng, X. J. Mu, E. Khurana, J. Rozowsky, R. Alexander *et al.*, *Nature* **489**, 91 (2012).
12. S. G. Landt, G. K. Marinov, A. Kundaje, P. Kheradpour, F. Pauli, S. Batzoglou, B. E. Bernstein, P. Bickel, J. B. Brown, P. Cayting *et al.*, *Genome Research* **22**, 1813 (2012).
13. C. Cheng, R. Min and M. Gerstein, *Bioinformatics* **27**, 3221 (2011).
14. T. M. Therneau, *A Package for Survival Analysis in S*, (2014). R package version 2.37-7.
15. A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander *et al.*, *Proceedings of the National Academy of Sciences* **102**, 15545 (2005).
16. C. F. Schaefer, K. Anthony, S. Krupa, J. Buchhoff, M. Day, T. Hannay and K. H. Buetow, *Nucleic Acids Research* **37**, D674 (2009).
17. C. Huttenhower, M. Schroeder, M. D. Chikina and O. G. Troyanskaya, *Bioinformatics* **24**, 1559 (2008).
18. O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein and R. B. Altman, *Bioinformatics* **17**, 520 (June 2001).
19. S. Saha Roy and R. K. Vadlamudi, *International Journal of Breast Cancer* **2012** (2011).
20. W. Shao, M. Brown *et al.*, *Breast Cancer Research* **6**, 39 (2004).
21. A. Hurtado, K. A. Holmes, C. S. Ross-Innes, D. Schmidt and J. S. Carroll, *Nature Genetics* **43**, 27 (2011).
22. V. Theodorou, R. Stark, S. Menon and J. S. Carroll, *Genome Research* **23**, 12 (2013).
23. J. S. Parker, M. Mullins, M. C. Cheang, S. Leung, D. Voduc, T. Vickery, S. Davies, C. Fauron, X. He, Z. Hu *et al.*, *Journal of Clinical Oncology* **27**, 1160 (2009).
24. C. Elston and I. Ellis, *Histopathology* **19**, 403 (1991).
25. E. E. Lower, E. L. Glass, D. A. Bradley, R. Blau and S. Heffelfinger, *Breast Cancer Research and Treatment* **90**, 65 (2005).
26. E. R. Fisher, J. Costantino, B. Fisher, A. S. Palekar and C. Redmond, *Cancer* **75**, 1310 (1995).
27. T. Sørlie, C. M. Perou, R. Tibshirani, T. Aas, S. Geisler, H. Johnsen, T. Hastie, M. B. Eisen, M. van de Rijn, S. S. Jeffrey *et al.*, *Proceedings of the National Academy of Sciences* **98**, 10869 (2001).
28. N. Bektas, A. Ten Haaf, J. Veeck, P. J. Wild, J. Lüscher-Firzlaff, A. Hartmann, R. Knüchel and E. Dahl, *BMC Cancer* **8**, p. 42 (2008).
29. A. Ahmad, Z. Wang, D. Kong, S. Ali, Y. Li, S. Banerjee, R. Ali and F. H. Sarkar, *Breast Cancer Research and Treatment* **122**, 337 (2010).
30. G. Vader and S. Lens, *Biochimica et Biophysica Acta (BBA)-Reviews on Cancer* **1786**, 60 (2008).
31. N. Takai, R. Hamanaka, J. Yoshimatsu and I. Miyakawa, *Oncogene* **24**, 287 (2005).
32. A. A. Troussard, P. C. McDonald, E. D. Wederell, N. M. Mawji, N. R. Filipenko, K. A. Gelmon, J. E. Kucab, S. E. Dunn, J. T. Emerman, M. B. Bally *et al.*, *Cancer Research* **66**, 393 (2006).
33. T. Gonda, P. Leo and R. Ramsay, *Expert Opinion on Biological Therapy* **8**, p. 713 (2008).
34. A. R. Thorner, J. S. Parker, K. A. Hoadley and C. M. Perou, *PLOS ONE* **5**, p. e13073 (2010).
35. C. S. Greene, J. Tan, M. Ung, J. H. Moore and C. Cheng, *Journal of Cellular Physiology* (2014).

AUTOMATED GENE EXPRESSION PATTERN ANNOTATION IN THE MOUSE BRAIN

TAO YANG^{1,2}, XINLIN ZHAO^{1,2}, BINBIN LIN², TAO ZENG³, SHUIWANG JI³, JIEPING YE^{1,2}

¹*Department of Computer Science and Engineering,*

²*Center for Evolutionary Medicine and Informatics, The Biodesign Institute,
Arizona State University, Tempe, AZ 85287, USA*

³*Department of Computer Science,*

Old Dominion University, Norfolk, VA 23529, USA

E-mail: ^{1,2}{T.Yang, Xinlin.Zhao, Binbin.Lin, Jieping.Ye}@asu.edu, ³{tzeng, sjj}@cs.odu.edu

Brain tumor is a fatal central nervous system disease that occurs in around 250,000 people each year globally and it is the second cause of cancer in children. It has been widely acknowledged that genetic factor is one of the significant risk factors for brain cancer. Thus, accurate descriptions of the locations of where the relative genes are active and how these genes express are critical for understanding the pathogenesis of brain tumor and for early detection. The Allen Developing Mouse Brain Atlas is a project on gene expression over the course of mouse brain development stages. Utilizing mouse models allows us to use a relatively homogeneous system to reveal the genetic risk factor of brain cancer. In the Allen atlas, about 435,000 high-resolution spatiotemporal *in situ* hybridization images have been generated for approximately 2,100 genes and currently the expression patterns over specific brain regions are manually annotated by experts, which does not scale with the continuously expanding collection of images. In this paper, we present an efficient computational approach to perform automated gene expression pattern annotation on brain images. First, the gene expression information in the brain images is captured by invariant features extracted from local image patches. Next, we adopt an augmented sparse coding method, called Stochastic Coordinate Coding, to construct high-level representations. Different pooling methods are then applied to generate gene-level features. To discriminate gene expression patterns at specific brain regions, we employ supervised learning methods to build accurate models for both binary-class and multi-class cases. Random undersampling and majority voting strategies are utilized to deal with the inherently imbalanced class distribution within each annotation task in order to further improve predictive performance. In addition, we propose a novel structure-based multi-label classification approach, which makes use of label hierarchy based on brain ontology during model learning. Extensive experiments have been conducted on the atlas and results show that the proposed approach produces higher annotation accuracy than several baseline methods. Our approach is shown to be robust on both binary-class and multi-class tasks and even with a relatively low training ratio. Our results also show that the use of label hierarchy can significantly improve the annotation accuracy at all brain ontology levels.

Keywords: Gene Expression Pattern, Image Annotation, Sparse Learning, Imbalanced Learning, Multi-label classification, Label Hierarchy

1. Introduction

Brain tumor is a fatal central nervous system disease and it is the second cause of cancer in children.¹ Previous studies indicate that preventing and detecting brain tumors at early stages are effective methods to reduce brain damage; these studies also show the potential benefit of utilizing the genetic determinants.² Accurate descriptions of the locations of where the relative genes are active and how these genes express are critical for understanding the pathogenesis of brain tumor and for early detection.

An accurate characterization of the gene expression and its role on brain tumor requires extensive experimental resources on brain. A recent study² uses mouse to reveal the genetic risk factor of brain cancer. However, such study was performed on a limited set of genes. The Allen Developing Mouse Brain Atlas (ADMBA) is an online public repository of extensive gene



(d) Part of Brain Ontology

* Detailed ontology of each node can be found in: http://developingmouse.brain-map.org/docs/Legend_2010_03.pdf

Fig. 1. Sample schemas of different gene expression metrics and brain ontology

expression and neuroanatomical data over different mouse brain developmental stages.^{3,4} The knowledge is documented as high-resolution spatiotemporal *in situ* hybridization (ISH) images for approximately 2,100 genes from embryonic through postnatal stages of brain development. In addition, a brain ontology has been designed to hierarchically organize brain structure for the developing form of mouse brain, which facilitates gene expression pattern annotation to specific brain areas. For a complete description of the status of gene expression revealed by *in situ* hybridization, three kinds of metrics, *i.e.*, pattern, density and intensity, are utilized at the Reference Atlas for ADMBA (R-ADMBA). These metrics were scored for each brain region according to a set of standard schemes; some examples are shown in Figure 1.

It is worthwhile to mention that such annotation tasks are very costly. The entire atlas contains around 435,000 ISH images and there are over 1,000 brain regions that need to be annotated in the designed brain ontology. To precisely assign gene expression metrics to specific brain areas, current reference atlas uses expert-guided manual annotation, which was performed by Dr. Martinez's team at Spain.^{4,5} However, it is labor-intensive since it requires expertise in neuroscience and image analysis, and it does not scale with the continuously expanding collection of images. Therefore, developing an effective and efficient automated gene expression pattern annotation method is of practical significance.

The gene expression pattern annotation problem can be formulated as an image annotation problem, which has been widely studied in computer vision and machine learning. Specifically, a key to solve the problem is to learn effective feature representations of images. The scale-invariant feature transform (SIFT) algorithm has been commonly applied to transform image content into local feature coordinates that are invariant to translation, rotation, scale, and other imaging parameters.⁶ SIFT has been shown to be a powerful tool to capture patch-level characteristics of images. Based on those local image descriptors, the next step is to construct

high-level feature representations of the ISH images. A common approach is to use the bag-of-words (BoW) model to represent high-level features, which has been used in a recent study.⁷ However, BoW is not efficient to learn a large number of keywords or deal with large scale data atlas. In this study, we employ sparse coding to construct high-level features, which has been demonstrated to be effective in many fields including image recognition.⁸ Sparse coding aims to use sparse linear combinations of basis vectors to reconstruct data vectors and learn a non-orthogonal and over-complete dictionary, which has more flexibility to represent the data.^{9–11} The previous study⁷ uses BoW instead of sparse coding mainly due to the high computational cost of solving the sparse coding problem especially for large-scale data in ADMBA. In this study, we adopt a novel implementation of sparse coding, called Stochastic Coordinate Coding (SCC),¹² which has been shown to be much more efficient than existing approaches.

Besides the image representation problem, many other difficulties are also inherent in the annotation tasks. First of all, for a specific set of ISH images, current reference atlas uses up to four categories [see Figure 1, (a)-(c)] to give an accurate description of the gene expression status for a specific metric. Thus, the annotation problem we are facing is indeed a multi-class classification problem. Secondly, the imbalanced class distribution is often involved in each annotation task, while traditional machine learning methods will often be biased and fail to provide reliable models.¹³ In addition, annotating gene expression pattern over the brain ontology is essentially a multi-label classification problem. However, if we simply treat each label separately, we do not make full use of the structural relationships among labels [as shown in Figure 1 (d)] in the learning procedure, resulting in suboptimal prediction performance.^{14,15}

In this paper, we propose an efficient computational approach to perform automated gene expression pattern annotation based on ADMBA ISH images. We first employ the SIFT method to construct local image descriptors. We next use sparse coding to efficiently learn the dictionary from SIFT descriptors of all ISH images and generate patch-level sparse feature representations of the images. Different pooling methods are utilized to combine patch-level representations to form image-level features, and further generate gene-level representations. To discriminate gene expression patterns over each brain area, we employ sparse logistic regression classifier and its multi-task extension to learn models for binary-class and multi-class classification. In addition, random undersampling and majority voting strategies are utilized to deal with imbalanced class distribution inherent within each annotation task. Furthermore, we make full use of the label hierarchy and dependency by developing a novel structure-based multi-label classification approach, which consists of two learning phases. In the first phase, a set of interested tasks (at the bottom of the label hierarchy) are learned individually, and in the second phase, knowledge learned from the first phase will be utilized to train models for the remaining tasks. We test our proposed approach on the four embryonic mouse developmental stages. Annotation results show that the adopted sparse coding approach outperforms the bag-of-words method. The proposed method provides favourable classification accuracy on both binary-class and multi-class tasks and even with a relatively low ratio of training. Experiment results also show that the structure-based multi-label classification approach can significantly improve the annotation accuracy at all brain ontology levels.

The remaining part of the paper is organized as follows: Section 2 details our feature extraction framework; Section 3 introduces several regularized learning methods, our strategies for learning from imbalanced data, and the proposed structure-based multi-label classification approach; Section 4 presents extensive empirical studies and Section 5 concludes the paper.

2. Proposed Feature Extraction Framework

2.1. *Image-level feature extraction*

Extracting and characterizing features from images is the key for image annotation. To capture as much gene expression details as possible over the entire brain ontology, ADMBA provides numerous spatiotemporal high-resolution ISH images. However, those raw images are not well aligned since they were taken from different samples and at different spatial slices. This makes it challenging to generate features from raw ISH images. A commonly used approach in such case is to employ the well-known scale-invariant feature transform method to construct local image descriptors. Specifically, the SIFT method first detects multiple localized keypoints (patches) from a raw image, and then transforms those image content into local feature coordinates that are invariant to translation, rotation, scale, and other imaging parameters. We use the SIFT detection in VLFeat¹⁶ and an average of 3,500 keypoints have been captured for each ISH image. In this study, each patch is represented by a 128-dimensional SIFT descriptor.

2.2. *High-level feature construction*

Based on the SIFT descriptors, we next apply sparse coding to construct high-level features. Sparse coding has been applied in many fields such as audio processing and image recognition. It refers to the process of using sparse linear combinations of basis vectors to reconstruct data and learning a non-orthogonal and over-complete dictionary. We can write the sparse coding problem as follows:

$$\begin{aligned} & \min_{\mathbf{D}, \mathbf{z}_1, \dots, \mathbf{z}_n} \sum_{i=1}^n \left(\frac{1}{2} \|\mathbf{D}\mathbf{z}_i - \mathbf{a}_i\|_2^2 + \lambda \|\mathbf{z}_i\|_1 \right) \\ & \text{s.t. } \|\mathbf{D}_{\cdot j}\|_2 \leq 1, 1 \leq j \leq p \end{aligned} \quad (1)$$

where $\mathbf{A} = [\mathbf{a}_1, \dots, \mathbf{a}_n] \in \mathbb{R}^{m \times n}$ is the set of SIFT descriptors constructed from image patches, each SIFT descriptor $\mathbf{a}_i \in \mathbb{R}^m$ is a m -dimension column vector with zero mean and unit norm, $\mathbf{D} \in \mathbb{R}^{m \times p}$ is the dictionary, λ is the regularization parameter, and $\mathbf{Z} = [\mathbf{z}_1, \dots, \mathbf{z}_n] \in \mathbb{R}^{p \times n}$ is the set of sparse feature representations of the original data. In addition, to prevent \mathbf{D} from taking arbitrarily large values, the constraint, $\mathbf{D}_{\cdot j}, 1 \leq j \leq p$, restricts each column of \mathbf{D} to be in a unit ball.

It has been known that solving the sparse coding problem is computationally expensive, especially when dealing with large-scale data and learning a large size of dictionary. The main computational cost comes from the updating of sparse codes and the dictionary. In our study, we adopt a new approach, called Stochastic Coordinate Coding (SCC), which has been shown to be much more efficient than existing methods.¹² The key idea of SCC is to alternately update the sparse codes via a few steps of coordinate descent and update the dictionary via second order stochastic gradient. In addition, by focusing on the non-zero components of the sparse codes and the corresponding dictionary columns during the updating procedure, the computational cost of sparse coding is further reduced.

In our study, the dictionary is learned from SIFT descriptors of all ISH images. The constraint, $\mathbf{z}_i \geq 0$, $1 \leq i \leq n$, is further added to ensure the non-negativity of sparse codes. To generate image-level features based on patch-level representations, we apply the max-pooling operation. Max-pooling takes the strongest signal among multiple patches to represent the image, which has been shown to be powerful in combining low-level sparse features.¹⁷

2.3. Gene-level feature pooling

Recall that a specific ISH image is obtained from particular brain spatial coordinates and it may not be able to present the gene expression pattern over the entire brain ontology. In order to describe expression pattern at all brain regions, we use a gene-level feature pooling. Since it remains unclear what kind of pooling methods will perform better on those high-level representations, both average-pooling and max-pooling are employed in our study.

3. Gene Expression Pattern Classification Methods

In this section, we introduce several regularized learning methods for gene expression pattern classification as well as our strategies for learning from imbalanced data. In addition, we present a structure-based multi-label classification approach for annotation.

3.1. Sparse logistic regression

We first consider the simple case: binary classification. Specifically, for a certain metric of gene expression, we convert the original annotation task into a binary classification problem by treating the category “undetected” as one class and all remaining categories as the other class. We employ the regularized supervised learning methods, which have been widely used in machine learning and bioinformatics. Let $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^n \in \mathbb{R}^{n \times p}$ denote a p dimensional data set with n observations, and $\mathbf{y} = \{y_i\}_{i=1}^n \in \mathbb{R}^{n \times 1}$, $y_i \in \{-1, 1\}$ be the corresponding labels. Then, we can write the sparse logistic regression problem as follows:

$$\min_{\mathbf{w}} \mathcal{L}(\mathbf{X}\mathbf{w}, \mathbf{y}) + \lambda \|\mathbf{w}\|_1, \quad (2)$$

where $\mathcal{L}(\cdot)$ denotes the logistic loss, $\mathbf{w} \in \mathbb{R}^{p \times 1}$ is the model weight vector and λ is the l_1 -norm regularization parameter. The solution of the above system will yield sparsity in \mathbf{w} , and the significant columns of \mathbf{X} are determined by the corresponding non-zero entries in \mathbf{w} . In our study, \mathbf{x}_i is a gene-level representation (after patch-level pooling and image-level pooling) and y_i encodes the annotation of gene expression status for a specific brain region.

3.2. Multi-task sparse logistic regression

We also propose to directly solve the multi-class annotation problem via multi-task learning. Suppose there are k classes ($k = 3$ or 4 in our study). We can represent the category of a sample by a k -tuple, where $y_{ik} = 1$ if sample i belongs to class k and $y_{ik} = -1$ otherwise. Then we can rewrite the response \mathbf{Y} as $\mathbf{Y} = \{\mathbf{y}_i\}_{i=1}^n \in \mathbb{R}^{n \times k}$. We employ the following multi-task sparse logistic regression formulation for the multi-class case:

$$\min_{\mathbf{W}} \mathcal{L}(\mathbf{X}\mathbf{W}, \mathbf{Y}) + \lambda \|\mathbf{W}\|_{2,1}, \quad (3)$$

where $\mathbf{W} \in \mathbb{R}^{p \times k}$, and the i -th column of \mathbf{W} is the model weight for the i -th task. The $l_{2,1}$ -norm penalty on \mathbf{W} results in grouped sparsity, which restricts all tasks to share a common set of features. In this paper, we employ this multi-task model to solve the multi-class annotation problem. The SLEP¹⁸ package is utilized to solve both Problem (2) and Problem (3).

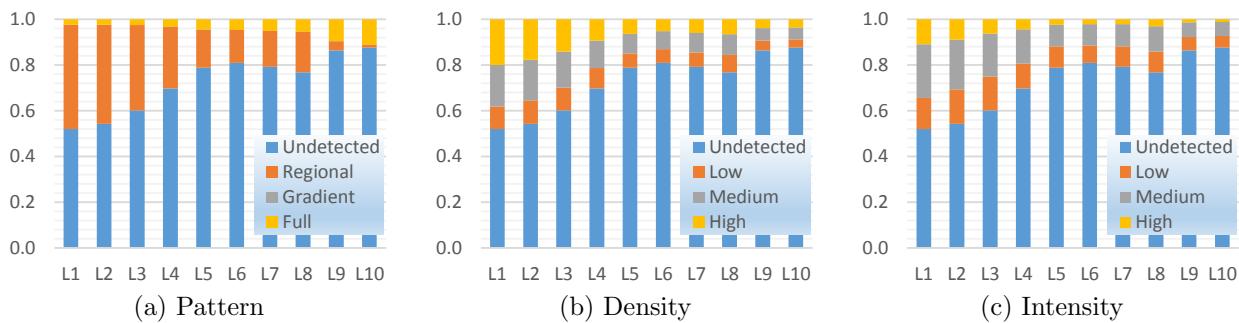


Fig. 2. Percentage of different categories of gene expression status at each brain level

3.3. Undersampling and majority voting

In this study, regions in the brain ontology are divided into 10 levels. Figure 2 shows the statistics of annotation distribution at each brain ontology level. It can be observed that even for the binary classification case, the data imbalance problem is particularly severe. A desired training set should contain approximately equal numbers of observations from each category. Traditional machine learning methods may be very sensitive to imbalance issue since the models will be biased toward the majority class of samples. To learn a better model from an imbalanced data set, a simple and intuitive idea is to balance the training set. Some existing studies suggest that random undersampling method is effective in dealing with data imbalance.¹³ Besides undersampling, model ensemble is also beneficial for learning from imbalanced data.¹⁹ Ensemble methods refer to the process of combining multiple models to improve predictive performance. The idea of classifier ensemble is to build a prediction model by combining a set of individual decisions from multiple classifiers.²⁰ In this study, we employ undersampling multiple times, combine a set of learning models, one for each undersampled data, and finally use majority voting to infer the predictions.

3.4. Structure-based multi-label annotation over brain ontology

Annotating gene expression patterns over the brain ontology is indeed a multi-label classification problem. In the reference atlas, the expression patterns of a single gene are recorded based on a hierarchically organized ontology of anatomical structures. In practice, it is possible to propagate annotation to parent or child structures under a set of systematic rules.⁴ Rather than simply treating each individual annotation task separately, if we build all prediction models together by utilizing the structure information among labels, the predictive performance can potentially be significantly improved.^{14,15}

In this study, we propose a novel structure-based multi-label classification approach. Suppose we are given n training data points $\{(\mathbf{x}^i, \mathbf{y}^i)\}_{i=1}^n$, where $\mathbf{x}^i \in \mathbb{R}^p$ is a data point of p features, and $\mathbf{y}^i \in \mathbb{R}^k$ is the corresponding label vector of k tasks. Let $j \in \{1, \dots, k\}$ denote the j -th learning task. We then divide the learning procedure into two phases. Assuming there are t tasks ($t < k$) at the bottom level of the hierarchy, in the first phase, each of those tasks is learned individually by:

$$\tilde{y}_j = \mathcal{F}_j(\tilde{\mathbf{x}}), \quad 1 \leq j \leq t < k, \quad (4)$$

where $\mathcal{F}_j(\cdot)$ denotes a learnt model by the j -th task, $\tilde{\mathbf{x}} \in \mathbb{R}^p$ is an arbitrary data point, and

$\tilde{y}_j \in \mathbb{R}$ is the prediction of $\tilde{\mathbf{x}}$ for the j -th task. The learned knowledge in (4) is then used to learn the remaining tasks (*i.e.*, $t + 1 \leq j \leq k$) in the second phase. Specifically, we augment the feature set by adding the prediction probabilities learnt in the previous phase, *i.e.*, we denote $\tilde{\mathbf{x}}' = [\tilde{\mathbf{x}}, (\tilde{y}_1, \dots, \tilde{y}_t)]$. Annotation tasks in the second phase will be performed based on this augmented feature set $\tilde{\mathbf{x}}'$.

The tasks in the first phase can be considered as the auxiliary tasks in the second phase.²¹ We apply the two-stage approach in our case since the tasks are not symmetric due to the hierarchical label structure. With the prediction probabilities from the previous learning phase, we make use of label dependency along with the original image representations. Intuitively, if a new learning task is related to some of the tasks learnt in the first phase, then such approach is expected to achieve better classification accuracy. In our study, since the tasks associated with the bottom of the label hierarchy are related to the remaining tasks in the hierarchy, the prediction performance is expected to be improved by the two-stage learning approach. This is confirmed in our experiments presented in the next section.

4. Experiments

We design a serial of experiments to evaluate the proposed approach for gene expression pattern annotation on the Allen Developing Mouse Brain Atlas. Specifically, we evaluate our approach in the following four aspects: (1) Comparison of sparse coding and bag-of-words, (2) Comparison of different training ratios, (3) Comparison of different multi-class annotation methods, and (4) Comparison of annotation with and without brain ontology.

4.1. Data description and experimental setup

The gene expression ISH images are obtained from the Allen Developing Mouse Brain Atlas. Specifically, to ensure the consistency of brain ontology over different mouse developmental stages, we focus our experiments on the four embryonic stages, namely, E11.5, E13.5, E15.5 and E18.5. The ADMBA provides approximately 2,100 genes within each stage and an average of 15~20 images are used for each gene to capture the expression information over the entire 3D brain. The total number of ISH images in these four stages are 142,425. We use the SIFT method to detect local gene expression and apply sparse coding to learn sparse feature representations for image patches. Considering the resolution of the ISH images and the number of areas of the mouse brain ontology, a dictionary size of 2,000 is chosen, *i.e.*, $\mathbf{D} \in \mathbb{R}^{128 \times 2000}$. To generate gene-level representations, both max-pooling and average-pooling are used.

To evaluate the effectiveness of the proposed methods, we compare our approach with the well-known bag-of-words (BoW) method. Specifically, the BoW is performed in two different settings: the first approach, called non-spatial BoW, concatenates three BoW representations of SIFT features, where each BoW is learned from the ISH images at a specific scale; the second approach, called spatial BoW, divides the brain sagitally into seven intervals according to the spatial coordinate of each image, and then 21 regional BoW representations are built (7 intervals \times 3 scales).⁷ At each scale, a fixed number of 500 clusters (keywords) are constructed from SIFT features and an extra dimension is used to count the number of zero descriptors.

R-ADMBa uses three different metrics including pattern, density and intensity, to evaluate the gene expression pattern on each brain ontology area. As discussed in the previous section,

we consider the annotation tasks as either binary-class or multi-class classification problem. For the simple binary-class case, the category “undetected” is treated as the negative class, which refers to the scenario that no gene expression pattern is detected at the specific brain area, and all remaining categories are treated as the positive class, which means some kind of expression pattern has been detected. It is worthwhile to note that, at such a binary-class situation, if the annotation metric “pattern” is marked as “undetected”, then metrics “density” and “intensity” must be “undetected”, and *vice versa*. That is, it is possible to use a single metric to evaluate the gene expression status at this case.

In addition, in order to balance the class distributions of training sets, random undersampling are performed for 11 times. To give a baseline performance of the traditional method, the experiment results of using Support Vector Machine (SVM) classifier²² is also reported. To better describe the classification performance under the circumstances of data imbalance, we use the area under the curve (AUC) of a receiver operating characteristic (ROC) curve as the performance measure for binary-class classification. The accuracy is used as the performance measure for the multi-class case.

4.2. Comparison of sparse coding and bag-of-words

We use the first serial of experiments to compare sparse coding with the bag-of-words method. Specifically, we generate the training data from raw gene expression ISH images using the following four methods: (1) SCC_Average, using SCC to learn image-level representations and average-pooling to generate gene-level features; (2) SCC_Max, similar to (1) but using max-pooling to generate gene-level features; (3) BoW_nonSpatial, generating single bag-of-words representation using all ISH images; (4) BoW_Spatial, generating multiple bag-of-words representations using ISH images from different spatial coordinates. Here we only consider the simple binary-class situation, and the entire data set is being randomly partitioned into training set and testing set for each annotation task using a ratio of 4:1. In addition, in comparison with the proposed majority voting strategy, the average classification performance of 11 times undersampling is also recorded. The overall classification performance for each brain ontology level at different developmental stages are summarized in Figure 3.

We can observe from Figure 3 that the proposed approach achieves the highest overall AUC of 0.9095, 0.8573, 0.8717 and 0.8903 at mouse brain developmental stages E11.5, E13.5, E15.5 and E18.5 respectively. For the comparison of different types of image representations, SCC_Average provides the best overall performance among all four stages. Although in some annotation tasks, BoW_Spatial provides competitive performance to SCC_Average, it is worthwhile to note that, the spatial BoW ensembles 21 single dictionaries and contains more than 10,000 features. Thus, spatial BoW is far more complex than SCC and involves higher computational costs. We can also observe that the use of undersampling and majority voting strategies improves the individual model by 1% ~ 3% in terms of AUC. Moreover, in comparison with SVM classifier, the sparse logistic regression classifier achieves better predictive performance. Those experimental results verify the superiority of our proposed methods.

4.3. Comparison of different training ratios

In this experiment, we compare the classification performance of using different training ratios. More specifically, we would like to verify the robustness of the presented approach when using a

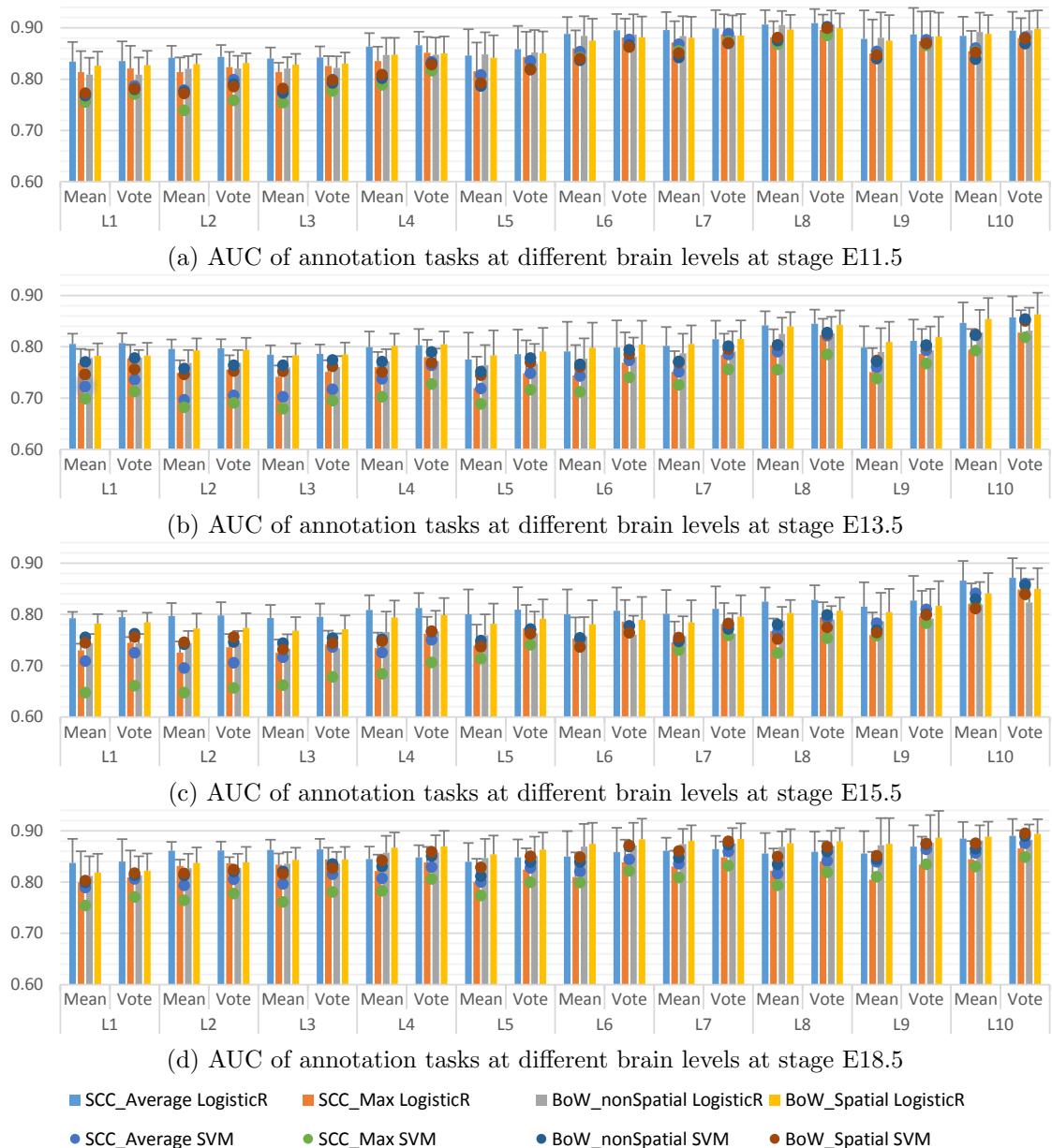


Fig. 3. Comparison of the proposed approach and bag-of-words method. Each column bar represents the performance of using sparse logistic regression classifier for a specific set of gene-level image representations. Each dot represents the performance of using SVM classifier for a specific set of gene-level image representations. The error bar of each column is the standard deviation of annotation performance within the corresponding brain level. “Mean” group records the average performance of 11 sub-models. “Vote” group records the performance of using majority voting.

relatively small number of samples for training. According to the first serial of experiments, we use the SCC_Average to construct features in this experiment. For each annotation task, we fix 10% of the samples as testing set and vary the ratio of training set in {50%, 60%, 70%, 80%, 90%}. The experimental results are summarized in Figure 4.

We can observe from the figure that, at all four mouse brain developmental stages and all brain levels, no significant difference is observed between different training ratios. We can

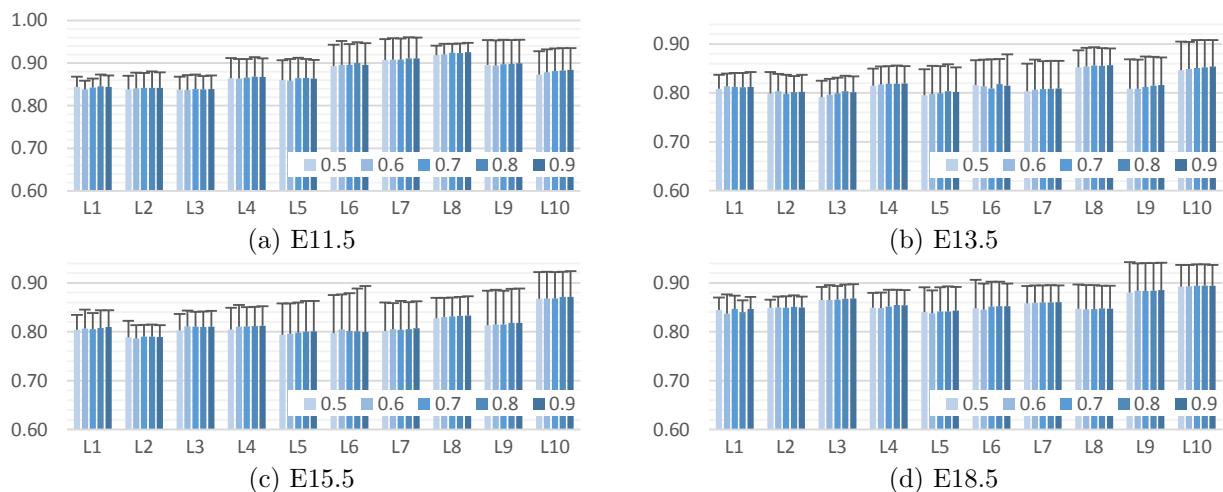


Fig. 4. Classification performance (AUC) of the proposed approach of using different training ratios. Shading of the bars from light to dark indicates training ratio from 0.5 ~ 0.9. The error bar of each column is the standard deviation of annotation performance within the corresponding brain level.

conclude from this experiment that our proposed approach is robust even with a low training ratio, thus accurate models for gene expression annotation can be learned based on a relative small number of manually annotated images.

4.4. Comparison of different multi-class annotation methods

In this experiment, we evaluate our multi-task sparse logistic regression (mcLR) approach in the multi-class annotation situation. Data set SCC_Average is employed and we use the multi-class SVM (mcSVM) as the baseline for performance comparison. In this experiment, 80% of the samples from each class are randomly chosen as the training set, and the remain 20% of the samples are used as the testing set. We only include annotation classes if there are more than 100 samples available for a specific class. The accuracy is used as the performance measure and the results are reported in Table 1.

We can observe that our proposed approach using sparse logistic regression with grouped sparsity constraint provides favourable predictive accuracy for this multi-class annotation task. Specifically, the classification accuracy of mcLR is significantly higher than mcSVM at all brain stages and levels. All detailed gene expression status measured by pattern, density and intensity can be well distinguished by our classifiers. These results imply that those multiple classes are inherently related and it is beneficial to learn four (or three) classification models simultaneously by restricting all models to share a common set of features. We plan to explore other multi-task learning models in our future work.²³

4.5. Comparison of annotation with and without brain ontology

Recall that the expert-guided manual annotations are based on a hierarchically organized ontology of anatomical structures. Rather than learning each task individually, it may be beneficial to utilize the hierarchy among the labels for a joint annotation. As we can observe from previous experiments, models learned in a lower level typically have better predictive performance. Thus it is natural to make use of the lower-level models and label structures to improve the prediction performance of high-level tasks.

Table 1. Classification accuracy (in percentage) of multi-class annotation.

	Pattern		Density		Intensity			Pattern		Density		Intensity			
	mcSVM	mcLR	mcSVM	mcLR	mcSVM	mcLR		mcSVM	mcLR	mcSVM	mcLR	mcSVM	mcLR		
(a) E11.5	L5	77.26	80.38	71.52	74.93	76.98	80.90	(b) E13.5	L5	73.53	80.10	67.91	73.87	70.51	76.62
	L6	79.29	81.78	80.68	82.97	79.93	83.57		L6	75.80	81.79	72.87	77.76	73.35	78.40
	L7	77.69	80.43	77.34	79.88	79.33	82.54		L7	72.07	77.56	72.09	77.05	73.71	78.85
	L8	81.61	84.35	83.40	85.46	83.98	85.80		L8	71.83	77.03	70.82	75.02	73.90	78.35
	L9	77.02	81.10	85.40	87.16	84.84	87.04		L9	78.54	82.26	81.12	84.66	80.57	84.34
	L10	— ^a	—	—	—	—	—		L10	—	—	85.36	87.48	83.22	86.00
(c) E15.5	L5	80.91	86.78	70.13	74.52	72.17	77.15	(d) E18.5	L5	75.41	78.93	72.73	76.18	75.22	78.96
	L6	79.66	83.09	76.42	80.03	76.28	80.83		L6	83.08	87.37	76.67	79.55	78.94	82.01
	L7	—	—	74.55	78.53	75.36	80.05		L7	77.34	79.65	75.78	78.77	77.67	80.79
	L8	74.97	79.79	70.93	75.36	72.83	78.23		L8	—	—	73.79	77.48	75.89	79.58
	L9	78.84	82.39	80.49	83.26	80.32	83.97		L9	79.49	81.42	79.17	81.56	80.59	83.01
	L10	—	—	86.16	87.61	87.03	88.26		L10	79.38	81.45	82.94	84.96	83.52	85.56

In this study, we compare our proposed structure-based multi-label learning (SMLL) method with the simple individual annotation, which builds models for different tasks independently. Again, we employ the SCC_Average method to construct the data. At each brain developmental stage, around 200 genes are randomly pre-selected as the testing set for the annotation tasks over the entire brain ontology and the remaining genes are included in the training set. For SMLL method, 432 tasks (regions) at level 10 (L10) are learned individually in the first phase. The prediction probabilities of L10 tasks will be used as the additional features in the data. In this experiment, we consider the binary-class situation and results are summarized in Table 2.

We can observe from Table 2 that the overall annotation performance achieved by SMLL is higher than the individual model. Improvements in terms of AUC can be observed at most of the brain ontology levels among all developmental stages. This verifies the effectiveness of the proposed structured-based multi-label learning approach.

5. Conclusion

In this paper, we propose an efficient computational approach to perform automated gene expression pattern annotation on mouse brain images. The key information in spatiotemporal *in situ* hybridization images is first captured by the SIFT method from local image patches. Image-level features are then constructed via sparse coding. To generate gene-level representations, different pooling method are adopted. Regularized learning methods are employed to build classification models for annotating gene expression pattern at different brain regions. To utilize hierarchy information among the brain ontology, a novel structure-based multi-label classification approach is proposed. Extensive experiments have been conducted on the atlas and results demonstrate the effectiveness of the proposed approach. One of our future directions is to explore deep learning models to learn feature representations from ISH images. In addition, we plan to explore other multi-task learning models to make more effective use of the label hierarchy in the annotation.

^a “—” means the experiment is not applicable for the specific brain ontology level.

Table 2. Classification performance (AUC) of structure-based multi-label annotation.

	E11.5				E13.5				E15.5				E18.5			
	LogisticR		SVM		LogisticR		SVM		LogisticR		SVM		LogisticR		SVM	
	Single	SMLL	Single	SMLL												
L1	0.837	0.811	0.806	0.837	0.793	0.781	0.737	0.778	0.744	0.749	0.657	0.699	0.890	0.878	0.845	0.879
L2	0.866	0.850	0.854	0.877	0.774	0.772	0.744	0.785	0.755	0.764	0.632	0.695	0.894	0.882	0.831	0.884
L3	0.898	0.884	0.884	0.903	0.799	0.797	0.766	0.808	0.781	0.788	0.634	0.710	0.893	0.885	0.833	0.885
L4	0.941	0.941	0.932	0.951	0.868	0.874	0.843	0.873	0.796	0.803	0.665	0.709	0.891	0.890	0.852	0.888
L5	0.905	0.908	0.904	0.922	0.843	0.855	0.822	0.848	0.838	0.844	0.710	0.746	0.871	0.876	0.837	0.850
L6	0.935	0.937	0.937	0.947	0.898	0.907	0.882	0.898	0.843	0.851	0.744	0.760	0.871	0.878	0.844	0.855
L7	0.951	0.950	0.950	0.959	0.860	0.866	0.842	0.863	0.846	0.858	0.743	0.777	0.894	0.896	0.874	0.890
L8	0.980	0.982	0.980	0.984	0.932	0.937	0.905	0.932	0.835	0.841	0.810	0.836	0.894	0.896	0.863	0.882
L9	0.966	0.969	0.971	0.972	0.890	0.896	0.877	0.884	0.865	0.873	0.811	0.816	0.871	0.872	0.852	0.843
L10	0.971	—	0.976	—	0.906	—	0.904	—	0.877	—	0.837	—	0.896	—	0.884	—

Acknowledgments

This work is supported in part by research grants from NIH (R01 LM010730) and NSF (IIS-0953662, IIS-1421057, IIS-1421100, DBI-1147134 and DBI-1350258).

References

- W. H. Organization, World cancer report 2014. (2014).
- K. M. Reilly, *Brain pathology* **19**, 121 (2009).
- ©2013 Allen Institute for Brain Science, Allen Developing Mouse Brain Atlas [Internet]. Available from: <http://developingmouse.brain-map.org>.
- ©2013 Allen Institute for Brain Science, *Allen Developing Mouse Brain Atlas*, tech. rep.
- C. L. Thompson, L. Ng, V. Menon, S. Martinez *et al.*, *Neuron* (2014).
- D. G. Lowe, Object recognition from local scale-invariant features, in *IEEE ICCV*, 1999.
- T. Zeng and S. Ji, *Automated Annotation of Gene Expression Patterns in the Developing Mouse Brain Atlas*, tech. rep. (2014).
- A. Szlam, K. Gregor and Y. LeCun, Fast approximations to structured sparse coding and applications to object classification, in *Computer Vision–ECCV 2012*, (Springer, 2012) pp. 200–213.
- B. A. Olshausen *et al.*, *Nature* **381**, 607 (1996).
- S. S. Chen, D. L. Donoho and M. A. Saunders, *SIAM J. Sci. Comput.* **20**, 33 (1998).
- D. L. Donoho and M. Elad, *Proceedings of the National Academy of Sciences* **100**, 2197 (2003).
- B. Lin, Q. Li, Q. Sun, M.-J. Lai, I. Davidson, W. Fan and J. Ye, *CoRR abs/1407.8147* (2014).
- H. He and E. A. Garcia, *Trans. Knowl. Data Eng., IEEE Trans. on* **21**, 1263 (2009).
- C. N. Silla Jr and A. A. Freitas, *Data Mining and Knowledge Discovery* **22**, 31 (2011).
- G. Tsoumakas, I. Katakis and I. Vlahavas, Mining multi-label data, in *Data mining and knowledge discovery handbook*, (Springer, 2010) pp. 667–685.
- A. Vedaldi and B. Fulkerson, VLFeat: An open and portable library of computer vision algorithms <http://www.vlfeat.org/>, (2008).
- Y.-L. Boureau, J. Ponce and Y. LeCun, A theoretical analysis of feature pooling in visual recognition, in *Proceedings of the 27th ICML*, 2010.
- J. Liu, S. Ji and J. Ye, *SLEP: Sparse Learning with Efficient Projections*. ASU, (2009).
- R. Dubey, J. Zhou, Y. Wang, P. M. Thompson and J. Ye, *NeuroImage* **87**, 220 (2014).
- R. Polikar, *Circuits and Systems Magazine, IEEE* **6**, 21 (2006).
- R. K. Ando and T. Zhang, *The Journal of Machine Learning Research* **6**, 1817 (2005).
- C.-C. Chang and C.-J. Lin, *ACM TIST* **2**, 27:1 (2011), www.csie.ntu.edu.tw/~cjlin/libsvm.
- J. Zhou, J. Chen and J. Ye, *MALSAR*. ASU, (2011).

**SESSION INTRODUCTION: CHARACTERIZING THE IMPORTANCE OF
ENVIRONMENTAL EXPOSURES, INTERACTIONS BETWEEN THE
ENVIRONMENT AND GENETIC ARCHITECTURE, AND GENETIC INTERACTIONS:
NEW METHODS FOR UNDERSTANDING THE ETIOLOGY OF COMPLEX TRAITS
AND DISEASE**

MOLLY A. HALL

Center for Systems Genomics
Department of Biochemistry and Molecular Biology
Pennsylvania State University
University Park, PA 16802, USA
E-mail: mah546@psu.edu

SHEFALI SETIA VERMA

Center for Systems Genomics
Department of Biochemistry and Molecular Biology
Pennsylvania State University
University Park, PA 16802, USA
E-mail: szs14@psu.edu

DENNIS P. WALL

Department of Pediatrics, Division of Systems Medicine
Department of Psychiatry
Program in Biomedical Informatics
Stanford University
Stanford, CA 94305, USA
E-mail: dpwall@stanford.edu

JASON H. MOORE

Institute for Quantitative Biomedical Sciences
Department of Genetics, Geisel School of Medicine
Dartmouth College
Hanover, NH 03755
E-mail: Jason.H.Moore@dartmouth.edu

BRENDAN KEATING

Center for Applied Genomics
Children's Hospital of Philadelphia
Philadelphia, PA 19104
E-mail: bkeating@mail.med.upenn.edu

DANIEL B. CAMPBELL

Department of Psychiatry and the Behavioral Sciences
Zilkha Neurogenetic Institute
Center for Genomic Psychiatry
Neuroscience Graduate Program
Keck School of Medicine
University of Southern California
Los Angeles, CA 90089
E-mail: dbcampbe@med.usc.edu

GREGORY GIBSON

Center for Integrative Genomics
Department of Biology
Georgia Institute of Technology

Atlanta, GA 30332
E-mail: greg.gibson@biology.gatech.edu

FOLKERT W. ASSELBERGS
Department of Cardiology, Division of Heart & Lungs
University Medical Center Utrecht
The Netherlands
E-mail: F.W.Asselbergs@umcutrecht.nl

SARAH PENDERGRASS
Center for Systems Genomics
Department of Biochemistry and Molecular Biology
Pennsylvania State University
University Park, PA 16802, USA
E-mail: sap29@psu.edu

While genome-wide association studies (GWAS) [1] have identified the genetic underpinning of a number of complex traits, large portions of the heritability of common, complex diseases are still unknown [2-6]. Beyond the association between genetic variation and outcomes, the impact of environment exposure, as well as gene-gene (GxG) and gene-environment (GxE) interactions, are undoubtedly fundamental mechanisms involved in the development of complex traits. Novel methods tailored to detect these predictors have the potential to (1) reveal the impact of multiple variations in biological pathways and (2) identify genes that are only associated with a particular disease in the presence of a given environmental exposure (e.g. smoking). Such knowledge could be used to assess personal risk and to choose suitable medical interventions, based on an individual's genotype and environmental exposures. Further, a more complete picture of the genetic and environmental aspects that impact complex disease can be used to inform environmental regulations to protect vulnerable populations.

Multiple challenges exist for undertaking GxG and GxE analyses. For instance, there can be computational burden and the need to adjust for Type-I error when GxG interactions are explored via an exhaustive combinatorial search. Comprehensive tests for interactions between genome-wide single nucleotide polymorphisms (SNPs) and multiple environmental variables can also be challenging and computationally intensive, especially when using millions of sequenced or imputed variants and more than a dozen exposure measures. Further challenges are faced when exploring GxG and GxE, and the goal of this paper session was to encourage advancements in research and tool development for greater understanding of the impact of environmental exposures, GxG, and GxE on outcomes and traits.

The papers selected for the session, “Characterizing the Importance of Environmental Exposures, Interactions between the Environment and Genetic Architecture, and Genetic Interactions: New Methods for Understanding the Etiology of Complex Traits and Disease”, address several current challenges faced in elucidating complex disease when taking into account environmental exposures, as well as GxG and GxE interactions.

Hu *et al.* explained their new approach for identifying GxG interactions using a human phenotype network method for GWAS data in *Genome-wide genetic interaction analysis of glaucoma using expert knowledge derived from human phenotype networks*. In this paper, the authors described their computational knowledge-based network as a filter to focus on the disease of interest. Subsequently, they employed statistical interaction tests to identify a

significant network of pairwise epistatic interactions among the prioritized SNPs to explain the complex nature of Glaucoma.

iPINBPA: An integrative network-based functional module discovery tool for genome-wide association studies by Wang *et al.* describes another novel method for handling genetic associations in a knowledge-driven manner. The authors used this method to identify and prioritize genetic associations by merging statistical data as well as biological evidence for protein interaction. The performance of this novel approach was compared to similar methods using two independent multiple sclerosis genome-wide association studies and one ImmunoChip study. iPINBPA demonstrated improvement over other methods in sub-network identification, and thus offers a novel approach for combining GWAS signal with knowledge of protein interactions.

In *Variable selection method for the identification of epistatic models*, Holzinger *et al.* described an approach for feature selection to detect epistatic interactions and applied the method to simulated data. This machine learning method integrates different selection parameters to identify the appropriate threshold between signal and noise and thus generate a list of variants with both interactions and main effects. These variables can further be assessed using powerful modeling tools for interpretation and prediction purposes to improve our understanding of complex human traits.

Patel *et al.* extended the environment-wide association study (EWAS) method by creating exposome correlation globes to account for inter-dependencies between exposure variables in *Development of exposome correlation globes to map out environment-wide associations*. The authors first mapped an “exposome globe”, displaying a comprehensive set of correlations that replicated across multiple datasets from the National Health and Nutrition Examination Survey (NHANES). Next, the “exposome globe” was employed to reveal interrelationships between EWAS results of two different phenotypes: type 2 diabetes and all-cause mortality. The results demonstrate the complex connections between exposures in the context of disease status, which can be applied to both EWAS and GxE interactions in future studies.

In *The global Exposome: a Bipartite Network Approach to Inferring Interactions between Environmental Exposures and Human Diseases*, Darabos *et al.* integrated environmental exposure data with human phenotypes and diseases to build a bird’s eye view of the connections between human diseases and chemical compounds. They used the CDC’s fourth national report on human exposure to environmental chemicals and obtained a total of 60 chemicals in 11 groups that are found in the environment and potentially harmful to human health. The authors also used the diseases of the NHGRI-GWAS catalog and a literature survey to compile a list of the diseases/traits that have been linked to their list of chemicals. The authors utilized this information to build a bipartite network of diseases and chemical substances and projected the network onto the disease/trait space as well as the substances space. There were many features identified in the resultant networks. The dyadicity and hetrophilicity of these networks were also explored, highlighting further characteristics of these environmental exposure-human phenotype networks.

A standard approach for GxE interactions is to fit one model per GxE model and then correct for multiple testing. However, two-stage methods can be used, in which potential GxE models are first filtered, testing only specific models passing the filtering step. In *A screening-testing approach for detecting gene-environment interactions using sequential penalized and unpenalized multiple logistic regression*, Frost *et al.* introduced a two-step method using filtering first: an elastic net-penalized multiple logistic regression model to estimate a marginal

association filter statistic, or a gene enrichment correlation statistic for all genetic markers. Next, a multiple logistic regression model was used to jointly assess marginal terms and GxE interactions for all markers passing the filtering step. A likelihood-ratio test was then used to determine if the interaction terms for the evaluated models were statistically significant. The authors evaluated their method with a bladder cancer dataset, showing the statistical benefits of using their novel method for filtering then evaluating GxE models.

The ways in which an exposure variable is measured (questionnaire versus direct lab measurements) may affect GxE analyses. In *Measures of exposure impact genetic association studies: An example in vitamin K levels and VKORC1*, Crawford *et al.* presented results from a study comparing types of exposure measurements for GxE analyses. SNPs in *VKORC1*, a rate-controlling enzyme in the vitamin cycle, were tested for interaction with 2 types of vitamin K measures: survey and serum levels in the third National Health and Nutrition Examination Studies. Results suggested that *VKORC1* associations with vitamin K levels vary by whether environmental measures were survey or serum levels. The authors cautioned that the results of this case study may be relevant for other GxE interaction analyses.

Both GxG and GxE were addressed by Jeff *et al.* in *Identification of genetic modifiers within the fibrogen gene cluster for fibrogen levels in three ethnically diverse populations*. This paper provides an application of GXG and GXE assessment to medically relevant complex traits in multiple ethnic groups using multivariate linear regression. The authors assessed interactions with and without main effects and described the importance of detecting epistasis to explain missing heritability.

Restrepo *et al.* provided an additional expansion of association testing beyond nuclear DNA SNP testing by exploring associations with mitochondrial DNA and complex disease in *Mitochondrial variation and the risk of age-related macular degeneration across diverse populations*. Mitochondria are known to play an important role in health and vision sustainability. There is some evidence that mitochondrial genetic variation may contribute to age-related macular degeneration (AMD). Studies of the contribution of mitochondrial genetic variation to AMD have been limited to populations of European decent. The authors of this paper explored associations between 50 mitochondrial SNPs and AMD in non-Hispanic whites, non-Hispanic blacks, and Mexican Americans, adjusting models for well-known environmental modifiers of BMI and smoking. The authors performed analyses stratified by ancestry using individual SNPs, as well as using three mitochondrial haplogroups. A total of five SNPs were found to be associated with AMD, including three located in the mitochondrial region responsible for the initiation of transcription of the MT genome. No SNPs were found associated in the other ancestry groups, although the sample size was limited within this study.

Knowledge of the nature of environment, genetic interactions, and GxE interactions will help to determine the impact of alterations in biological pathways and identification of genes that are only found to be associated with disease in the context of the environment. This valuable information can be used to assess personal risk and to choose the most appropriate medical interventions based on unique genotype and environmental variation. The papers selected for this section offer novel insights to address the challenges in elucidating the genetic and environmental underpinnings of complex human traits.

References:

1. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, et al. (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci U S A* 106: 9362-9367.
2. Zuk O, Hechter E, Sunyaev SR, Lander ES (2012) The mystery of missing heritability: Genetic interactions create phantom heritability. *Proc Natl Acad Sci U S A* 109: 1193-1198.
3. Frazer KA, Murray SS, Schork NJ, Topol EJ (2009) Human genetic variation and its contribution to complex traits. *Nat Rev Genet* 10: 241-251.
4. Eichler EE, Flint J, Gibson G, Kong A, Leal SM, et al. (2010) Missing heritability and strategies for finding the underlying causes of complex disease. *Nat Rev Genet* 11: 446-450.
5. Maher B (2008) Personal genomes: The case of the missing heritability. *Nature* 456: 18-21.
6. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, et al. (2009) Finding the missing heritability of complex diseases. *Nature* 461: 747-753.

MEASURES OF EXPOSURE IMPACT GENETIC ASSOCIATION STUDIES: AN EXAMPLE IN VITAMIN K LEVELS AND VKORC1

DANA C. CRAWFORD

Institute for Computational Biology, Department of Epidemiology and Biostatistics, Case Western Reserve University, Wolstein Research Building, 2103 Cornell Road, Suite 2527, Cleveland, OH 44106, USA

Email: dana.crawford@case.edu

KRISTIN BROWN-GENTRY

Center for Human Genetics Research, Vanderbilt University, 519 Light Hall, 2215 Garland Avenue, Nashville, TN 37232, USA

Email: Kristin.Gentry@healthspring.com

MARK J. RIEDER

Adaptive Biotechnologies Corporation, 1551 Eastlake Avenue East, Suite 200, Seattle, WA 98102, USA

Email: mrieder@adaptivebiotech.com

Studies assessing the impact of gene-environment interactions on common human diseases and traits have been relatively few for many reasons. One often acknowledged reason is that it is difficult to accurately measure the environment or exposure. Indeed, most large-scale epidemiologic studies use questionnaires to assess and measure past and current exposure levels. While questionnaires may be cost-effective, the data may or may not accurately represent the exposure compared with more direct measurements (e.g., self-reported current smoking status versus direct measurement for cotinine levels). Much like phenotyping, the choice in how an exposure is measured may impact downstream tests of genetic association and gene-environment interaction studies. As a case study, we performed tests of association between five common *VKORC1* SNPs and two different measurements of vitamin K levels, dietary ($n=5,725$) and serum ($n=348$), in the Third National Health and Nutrition Examination Studies (NHANES III). We did not replicate previously reported associations between *VKORC1* and vitamin K levels using either measure. Furthermore, the suggestive associations and estimated genetic effect sizes identified in this study differed depending on the vitamin K measurement. This case study of *VKORC1* and vitamin K levels serves as a cautionary example of the downstream consequences that the type of exposure measurement choices will have on genetic association and possibly gene-environment studies.

1. Introduction

Complex human diseases and traits are shaped both by genetics and the environment. The development of dense genotyping arrays in the past decade has enabled genome-wide association studies in large epidemiologic and clinic-based collections, and these studies have been successful in identifying thousands of common genetic variants associated with common disease [1]. The more recent advent of relatively

cost-effective sequencing technologies is now contributing towards knowledge of rare genetic variants contributing to human disease and traits [2].

While genomic technologies have made tremendous advancements in the number of variants that can be assayed or the number of reads that can be sequenced, the methods to analyze such complex data have not evolved as rapidly. Also, very few studies have attempted to incorporate environmental exposures or gene-environment interactions. Indeed, most studies that have examined the impact of gene-environment interaction on common diseases or traits have been limited to candidate gene studies [e.g., 3] as there is little consensus on how to best test for gene-environment interactions on the genome-wide scale [4].

Another challenge faced in examining the effects on environmental exposures on human health is how to measure these data. Most large-scale epidemiologic collections use questionnaires to assess past and present exposures, which can vary in accuracy and be associated with substantial biases depending on the exposure being measured [e.g., 5]. Clinic-based collections are further hampered by the fact that most clinics do not routinely collect exposure data in a standardized manner [6]. To further complicate the field, substantial across-study differences exist in how exposure data is collected, making data harmonization efforts difficult [7,8]. Oftentimes, the simplest or most frequently used measure of exposure found across studies is chosen for harmonization across studies despite the availability of more accurate exposure measures albeit at the cost of statistical power.

It is presently unclear what impact the common practices of exposure harmonization across studies have on genetic association study findings. To document potential genetic association study differences due to differences in exposures measures, we examined the association of vitamin K epoxide reductase complex subunit 1 (*VKORC1*) common genetic variation and two different measures of vitamin K levels in the Third National Health and Nutrition Examination Surveys (NHANES III).

Vitamin K is a fat-soluble vitamin essential for blood clotting and bone formation. Vitamin K is found in two forms, K₁ (phylloquinone) and K₂, the former of which is synthesized by green leafy plants such as kale, spinach, and collards. Vitamin K is a required co-enzyme for the γ -carboxylation of three glutamic acid (Glu) residues in osteocalcin, converting them to gamma-carboxyglutamic acid (Gla) as part of the vitamin cycle. *VKORC1* is a co-enzyme in this vitamin cycle required for the recycling of vitamin K and the eventual Glu to Gla conversion [9,10]. The anticoagulant drug Warfarin targets *VKORC1* and effectively blocks the recycling of vitamin K, which in turn blocks the blood clotting process.

It is well established that *VKORC1* genetic variants are associated with warfarin dosing [11,12]. Previous candidate gene studies have also suggested that *VKORC1* common variants are associated with vitamin K levels. *VKORC1* is a rate-controlling enzyme in the vitamin cycle and common genetic variants within *VKORC1* have previously been associated with vitamin K levels [13,14]. Data presented here suggest that *VKORC1* associations differ for vitamin levels measured from dietary questionnaires compared with serum and serve as a cautionary case study that may be applicable to further gene-environment studies.

2. Methods

2.1. Study population

The study population and DNA samples described here are from the Third National Health and Nutrition Examination Surveys (NHANES III) conducted by the National Center for Health Statistics (NCHS) at the Centers for Disease Control and Prevention (CDC). NHANES III was conducted between 1988 and 1994 in two phases, and biospecimens for DNA extraction were collected in phase 2 (1991-1994). NHANES is a series of cross-sectional surveys designed to document the health status of Americans at the time of ascertainment; as such, participants are ascertained regardless of health status. NHANES III was a complex survey design where minority (non-Hispanic blacks and Mexican Americans) participants as well as the elderly were oversampled. Health data on NHANES III participants is collected through questionnaires, a physical examination by health professionals, and laboratory measures. All physical examinations are performed in the Mobile Examination (MEC) unless the participant is physically unable, in which case the examinations are conducted in the participant's home. Design and operations of NHANES III has previously been described [15].

The present study was approved by the CDC Ethics Review Board. Because the study investigators did not have access to personal identifiers, this study was considered non-human subjects research by the Vanderbilt University Internal Review Board.

2.2. Genotyping

We selected five *VKORC1* tagSNPs as previously described [16,17] and allele frequencies for NHANES III have been previously published [12]. Briefly, tagSNPs were chosen to tag common *VKORC1* genetic variation for mostly European-descent populations. SNP rs2884737 was genotyped using Sequenom's iPLEX® Gold coupled with MassARRAY MALDI-TOF MS detection (San Diego, CA). SNPs rs9923231, rs9934438, rs8050894, and rs2359612 were genotyped using Applied Biosystem's (now Thermo Scientific) TaqMan® SNP Genotyping Assays (Foster City, CA). All SNPs were genotyped by the Vanderbilt University's Center for Human Genetics Research DNA Resources Core. All SNPs were in Hardy Weinberg Equilibrium. In addition to genotyping experimental NHANES III DNA samples, we genotyped 368 blinded duplicates required by CDC for additional quality control. All genotypes have been deposited into CDC's Genetic NHANES database and are available for secondary analysis.

2.3. Vitamin K

Vitamin K was collected on NHANES III participants in two ways. First, total nutrient intake of vitamin K (mcg) was collected from the 24-hour dietary recall performed in the MEC [15]. These data were collected in collaboration with the University of Minnesota's Nutrition Coordinating Center (NCC). Second, serum vitamin K (phylloquinone; ng/ml) was measured using reverse phase HPLC [18] in non-Hispanic white women ages 6-29 years in Phase 2 of NHANES III. According to NHANES documentation, the lower

limit of detection is 0.05 and the range of serum vitamin K levels in the subset of NHANES samples tested was 0.05ng/ml-6.799ng/ml.

2.4. Statistical methods

Single SNP tests of association were performed unadjusted and adjusted using linear regression assuming an additive genetic model stratified by self-identified race/ethnicity. Two dependent variables were tested for an association with each SNP: dietary vitamin K levels and serum vitamin K levels, both log transformed. Serum vitamin K adjusted models included age, body mass index, current smoking status, dietary calcium, phosphorous, magnesium, iron, zinc, copper, sodium, potassium, protein, carbohydrates, fiber, total vitamin A, total carotenes, total alpha-tocopherol equivalents, vitamin C, vitamin B₆, vitamin B₁₂, folic acid, and total calories as covariates. Body mass index (continuous variable) was calculated based on height and weight measured at the MEC. Current smoking status (binary variable) was defined by “do you smoke cigarettes now?” or cotinine levels > 15ng/ml. The remaining dietary covariates (continuous traits) were available in NHANES III from the 24-hour dietary recall performed in the MEC. Dietary vitamin K models included the same covariates as serum vitamin K models with the addition of sex as a covariate. All analyses were conducted remotely in SAS v9.2 (SAS Institute, Cary, NC) and SUDAAN (Research Triangle Institute, Research Triangle Park, NC) using the Analytic Data Research by Email (ANDRE) portal of the CDC Research Data Center in Hyattsville, MD. All analyses presented here were performed weighted to account for the complex survey design. Results of tests of association were visualized using Synthesis View [19].

3. Results

Study population characteristics are given in Table 1. On average, study participants with dietary vitamin K levels were more likely to be female among non-Hispanic whites and non-Hispanic blacks. The average ages for non-Hispanic blacks and Mexican Americans were younger than non-Hispanic whites and both populations on average overweight (body mass index >25 mg/k²) compared with non-Hispanic whites. Mean dietary vitamin K levels were similar across all three racial/ethnic groups.

Given that vitamin K levels from dietary recall data are more readily available than serum levels, we first performed single SNP tests of association with these data stratified by race/ethnicity and adjusted for demographic and relevant covariates (see Methods). From all tests of association, only rs8050894 was significantly associated with dietary vitamin K levels in non-Hispanic whites ($\beta = 0.06$; $p=0.01$; Figure 1). No test of association was significant all race/ethnicities for the same *VKORC1* SNP.

We then performed tests of association for all SNPs among non-Hispanic white females with serum vitamin K levels and compared these data with tests of association performed among non-Hispanic whites with dietary vitamin K levels (Figure 2). Similar to dietary vitamin K levels, only one SNP in non-Hispanic whites was associated with serum vitamin K levels ($\beta = 0.16$; $p = 0.03$; Figure 2). However, the SNP associated with serum vitamin K levels (rs2359612) is not the same SNP associated with dietary vitamin K levels (rs8050894). The genetic effect sizes (betas) estimated for dietary vitamin K levels were

lower than those estimated for serum vitamin K levels with the exception of the genetic effect size estimated for rs9923231 (Figure 2).

Based on the disparate results obtained in models where the dependent variable was either dietary or serum vitamin K levels, we tested for a correlation between the variables. A total of 229 non-Hispanic white females have both vitamin K measurements available in NHANES III. The Pearson Correlation Coefficient was 0.11 between the two vitamin K estimates, which was not statistically significant ($p=0.08$).

Table 1. NHANES III study population characteristics. Unweighted descriptive statistics are shown for basic demographic variables (sex, age, and body mass index) as well as the two measures of vitamin K levels.

Sample sizes shown are for participants with dietary vitamin K levels available. Serum vitamin K levels were only measured in non-Hispanic white women ages 6 – 29 years in Phase 2 of NHANES III (n=348).

Abbreviations: standard deviation (SD), natural log (ln).

	Non-Hispanic whites (n=2,344)	Non-Hispanic blacks (n=1,675)	Mexican Americans (n=1,706)
% female	61	58	50
Mean (\pm SD) age in years	53.46 (20.32)	40.79 (16.71)	41.15 (17.42)
Mean (\pm SD) body mass index (kg/m^2)	26.66 (5.6)	27.3 (369.6)	27.1 (422.6)
Mean (\pm SD) ln(dietary vitamin K) (mcg)	4.05 (0.97)	4.02 (1.29)	3.79 (0.99)
Mean (\pm SD) ln(serum vitamin K) (mg/dl)	-1.30 (0.79)	-	-

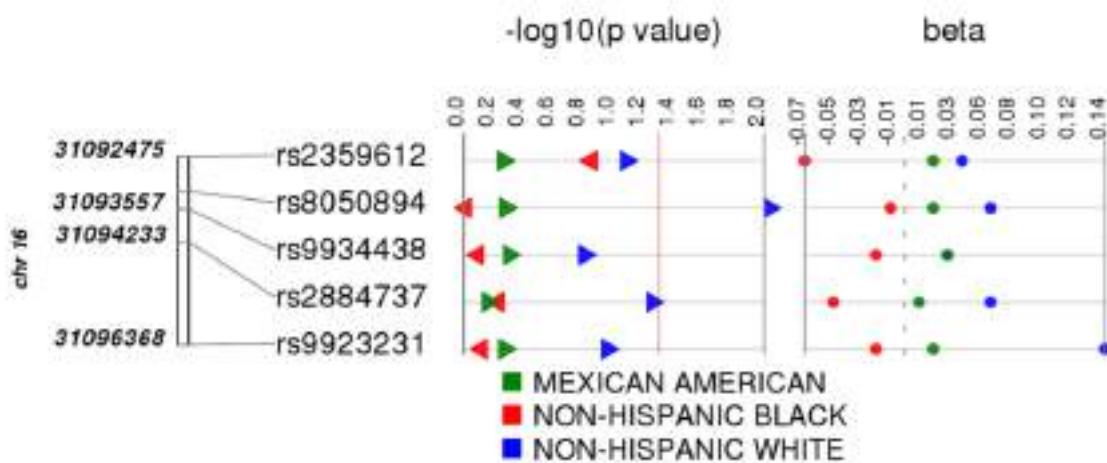
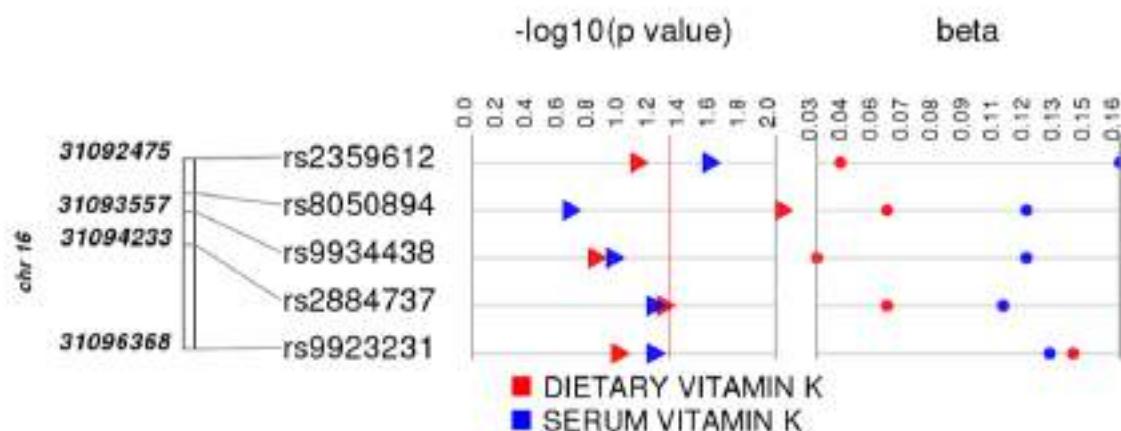


Figure 1. Results of tests of association between *VKORC1* common variation and dietary vitamin K levels stratified by race/ethnicity. Tests of association were performed using linear regression assuming an additive genetic model and adjusted for sex, age, body mass index, current smoking status, dietary calcium, phosphorous, magnesium, iron, zinc, copper, sodium, potassium, protein, carbohydrates, fiber, total vitamin A, total carotenes, total alpha-tocopherol equivalents, vitamin C, vitamin B₆, vitamin B₁₂,



folic acid, and total calories. Results (p-values and betas) are plotted by rs number and race/ethnicity (coded by color) using Synthesis View. The direction of the triangles represents the direction of the genetic effect. The red line represents a p-value threshold of 0.05.

Figure 2. Comparison of results of tests of association between *VKORC1* common variation and vitamin K levels by vitamin K measurement. Tests of association were performed using linear regression assuming an additive genetic model and adjusted for sex (for model with dietary vitamin K

levels), age, body mass index, current smoking status, dietary calcium, phosphorous, magnesium, iron, zinc, copper, sodium, potassium, protein, carbohydrates, fiber, total vitamin A, total carotenes, total alpha-tocopherol equivalents, vitamin C, vitamin B₆, vitamin B₁₂, folic acid, and total calories. Results (p-values and betas) are plotted by rs number and vitamin K measurement (coded by color) using Synthesis View. The direction of the triangles represents the direction of the genetic effect. The red line represents a p-value threshold of 0.05.

4. Discussion

To illustrate potential genetic association differences dependent on exposure measurement differences, we tested five *VKORC1* for an association with both dietary and serum-measured vitamin K levels. We identified suggestive associations for both dietary and serum-measured vitamin K levels; however, the results were not concordant between the two vitamin K measurements. Moreover, the genetic effect sizes estimated for each test of association differed between dietary and serum-measured vitamin K levels. Finally, we did not directly replicate previously reported associations between *VKORC1* common genetic variants and vitamin K levels in NHANES III.

Previously studies have suggested that vitamin K levels are associated with *VKORC1* genetic variants. More specifically, Nimptsch and colleagues [13] demonstrated in 548 German males and females that dietary vitamin K levels were inversely correlated with serum undercarboxylated osteocalcin/ total intact osteocalcin ratio levels dependent on rs2359612 genotypes. Crosier and colleagues [14] describe an association between plasma phylloquinone levels and rs8050894 in 416 older (60-80 years) men and women primarily of European descent. Interestingly, both rs2359612 and rs8050894 were suggestively associated with serum and dietary levels of vitamin K, respectively, at p<0.05 in this study. However, the tests of association and results described here are not a replicate of the tests performed by Nimptsch et al [13] and Crosier et al [14] given the different modeling assumptions measures of vitamin K associated with each SNP.

We have found that within NHANES III, serum levels of vitamin K are not significantly correlated with dietary measures of vitamin K levels. Unlike the present study, previous studies have suggested a correlation between dietary intake and serum levels of vitamin K levels [20, 21]. The lack of correlation observed in NHANES III may explain in part the differences in observed genetic associations. Indeed, the two measures of vitamin K levels examined here are likely measuring different traits with different underlying genetic architectures. Dietary vitamin K levels are likely measuring phylloquinone levels (K₁) [10]. For serum vitamin K levels, previous family studies have suggested that plasma phylloquinone levels have non-significant heritability [22].

This study has several limitation and strengths. A major limitation of the current study is sample size and power. Although we were properly powered to detect associations with dietary vitamin K levels with >5,000 participants, we have many fewer participants with serum vitamin K levels (n=348). Furthermore, serum vitamin K levels were only measured in a subset of non-Hispanic white females, which may impact the generalizability of these tests of association and comparison across studies. Also, NHANES III did not

collect data for serum undercarboxylated osteocalcin/ total intact osteocalcin ratio levels, another measure of vitamin K levels, on any of the participants. NHANES III also only measured vitamin K levels once per participant for both serum levels as well as dietary intake. It is possible that seasonal variations in vitamin K levels exist and may have impacted this study. Finally, neither ancestry informative markers nor GWAS-level data are available in NHANES III. It is possible that population stratification may have impacted the results observed here particularly for admixed populations such as Mexican Americans.

Despite these limitations, a major strength of NHANES III is the availability of both questionnaire-based and laboratory-based exposure data even for only a fraction of the study population. Most epidemiologic studies collect only questionnaire-based data given it is more cost-effective than the alternative. Self-reported data may be not be as accurate as a laboratory assay, the latter of which is more attractive for quantitative trait genetic association studies. Even within this limited dataset, our results suggest that choice of exposure measurement may have an impact on the results and interpretation of a genetic association study.

5. Acknowledgments

This work was supported, in part, by NIH grant NS053646 (MJR). We would like to thank Jody McLean and Dr. Geraldine McQuillan from the National Center for Health Statistics at the Centers for Disease Control and Prevention for their assistance with the NHANES III genetic data. The findings and conclusions in this report are those of the authors and do not necessarily represent the views of the National Institutes for Health or the Centers for Disease Control and Prevention. The Vanderbilt University Center for Human Genetics Research, Computational Genomics Core provided computational and/or analytical support for this work.

References

1. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA. *Proc Natl Acad Sci USA*. **106**(23):9362-7 (2009).
2. TG and HDL Working Group of the Exome Sequencing Project, National Heart, Lung, and Blood Institute, Crosby J, Peloso GM, Auer PL, Crosslin DR, Stitzel NO, Lange LA, Lu Y, Tang ZZ, Zhang H, Hindy G, Masca N, Stirrups K, Kanoni S, Do R, Jun G, Hu Y, Kang HM, Xue C, Goel A, Farrall M, Duga S, Merlini PA, Asselta R, Girelli D, Olivieri O, Martinelli N, Yin W, Reilly D, Speliotes E, Fox CS, Hveem K, Holmen OL, Nikpay M, Farlow DN, Assimes TL, Franceschini N, Robinson J, North KE, Martin LW, DePristo M, Gupta N, Escher SA, Jansson JH, Van Zuydam N, Palmer CN, Wareham N, Koch W, Meitinger T, Peters A, Lieb W, Erbel R, Konig IR, Kruppa J, Degenhardt F, Gottesman O, Bottinger EP, O'Donnell CJ, Psaty BM, Ballantyne CM, Abecasis G, Ordovas JM, Melander O, Watkins H, Orho-Melander M, Ardissono D, Loos RJ, McPherson R, Willer CJ, Erdmann J, Hall AS, Samani NJ, Deloukas P, Schunkert H, Wilson JG, Kooperberg C, Rich SS, Tracy RP, Lin DY, Altshuler D, Gabriel S, Nickerson DA,

Jarvik GP, Cupples LA, Reiner AP, Boerwinkle E, Kathiresan S. *N Engl J Med.* **371**(1):22-31 (2014).

3. Dumitrescu L, Carty CL, Franceschini N, Hindorff LA, Cole SA, Bůžková P, Schumacher FR, Eaton CB, Goodloe RJ, Duggan DJ, Haessler J, Cochran B, Henderson BE, Cheng I, Johnson KC, Carlson CS, Love SA, Brown-Gentry K, Nato AQ, Quibrera M, Shohet RV, Ambite JL, Wilkens LR, Le Marchand L, Haiman CA, Buyske S, Kooperberg C, North KE, Fornage M, Crawford DC. *Hum Genet.* **132**(12):1427-31 (2013).
4. Thomas D. *Nat Rev Genet.* **11**(4):259-72 (2010).
5. Mossavar-Rahmani Y, Tinker LF, Huang Y, Neuhouser ML, McCann SE, Seguin RA, Vitolins MZ, Curb JD, Prentice RL. *Nutr J.* **12**:63 (2013).
6. Wiley LK, Shah A, Xu H, Bush WS. *J Am Med Inform Assoc.* **20**(4):652-8 (2013).
7. Hamilton CM, Strader LC, Pratt JG, Maiiese D, Hendershot T, Kwok RK, Hammond JA, Huggins W, Jackman D, Pan H, Nettles DS, Beaty TH, Farrer LA, Kraft P, Marazita ML, Ordovas JM, Pato CN, Spitz MR, Wagener D, Williams M, Junkins HA, Harlan WR, Ramos EM, Haines J. *Am J Epidemiol.* **174**(3):253-60 (2011).
8. Pan H, Tryka KA, Vreeman DJ, Huggins W, Phillips MJ, Mehta JP, Phillips JH, McDonald CJ, Junkins HA, Ramos EM, Hamilton CM. *Hum Mutat.* **33**(5):849-57 (2012).
9. Stafford DW. *J Thromb Haemost.* **3**(8):1873-8 (2005).
10. Booth SL. *Food Nutr Res.* **56** (2012).
11. Rieder MJ, Reiner AP, Gage BF, Nickerson DA, Eby CS, McLeod HL, Blough DK, Thummel KE, Veenstra DL, Rettie AE. *N Engl J Med.* **352**(22):2285-93 (2005).
12. Limdi NA, Wadelius M, Cavallari L, Eriksson N, Crawford DC, Lee MT, Chen CH, Motsinger-Reif A, Sagreya H, Liu N, Wu AH, Gage BF, Jorgensen A, Pirmohamed M, Shin JG, Suarez-Kurtz G, Kimmel SE, Johnson JA, Klein TE, Wagner MJ; International Warfarin Pharmacogenetics Consortium. *Blood.* **115**(18):3827-34 (2010).
13. Nimptsch K, Nieters A, Hailer S, Wolfram G, Linseisen J. *Br J Nutr.* **101**(12):1812-20 (2009).
14. Crosier MD, Peter I, Booth SL, Bennett G, Dawson-Hughes B, Ordovas JM. *J Nutr Sci Vitaminol (Tokyo)* **55**(2):112-9 (2009).
15. <http://www.cdc.gov/nchs/nhanes/nh3data.htm>
16. Crawford DC, Ritchie MD, Rieder MJ. *Pharmacogenomics.* **8**(5):487-96 (2007).

17. Crawford DC, Brown-Gentry K, Rieder MJ. *PLoS One*. **5**(12):e15088 (2010).
18. Haroon Y, Bacon D, Sadowski J. Liquid-chromatographic determination of vitamin K 1 plasma with fluorometric detection. *Clin Chem* **32**:1925-9 (1986).
19. Pendergrass SA, Dudek SM, Crawford DC, Ritchie MD. *BioData Min*. **3**:10 (2010).
20. Thane CW, Bates CJ, Shearer MJ, Unadkat N, Harrington DJ, Paul AA, Prentice A, Bolton-Smith C. *Br J Nutr* **87**: 615-22 (2002).
21. Thane CW, Wang LY, Coward WA. *Br J Nutr* **96**:1116-11124 (2006).
22. Shea MK, Benjamin EJ, Dupuis J, Massaro JM, Jacques PF, D'Agostino RB, Sr, Ordovas JM, O'Donnell CJ, Dawson-Hughes B, Vasan RS, Booth SL. Genetic and non-genetic correlates of vitamins K and D. *Eur J Clin Nutr*. **63**(4):458-64. (2009).

A BIPARTITE NETWORK APPROACH TO INFERRING INTERACTIONS BETWEEN ENVIRONMENTAL EXPOSURES AND HUMAN DISEASES

CHRISTIAN DARABOS, EMILY D. GRUSSING, MARIA E. CRICCO,
KENZIE A. CLARK, JASON H. MOORE

*Institute for the Quantitative Biomedical Sciences, The Geisel School of Medicine at Dartmouth College,
Lebanon, NH 03756, U.S.A.
E-mail: Christian.Darabos@dartmouth.edu*

Environmental exposure is a key factor of understanding health and diseases. Beyond genetic propensities, many disorders are, in part, caused by human interaction with harmful substances in the water, the soil, or the air. Limited data is available on a disease or substance basis. However, we compile a global repository from literature surveys matching environmental chemical substances exposure with human disorders. We build a bipartite network linking 60 substances to over 150 disease phenotypes. We quantitatively and qualitatively analyze the network and its projections as simple networks. We identify mercury, lead and cadmium as associated with the largest number of disorders. Symmetrically, we show that breast cancer, harm to the fetus and non-Hodgkin's lymphoma are associated with the most environmental chemicals. We conduct statistical analysis of how vertices with similar characteristics form the network interactions. This dyadicity and heterophilicity measures the tendencies of vertices with similar properties to either connect to one-another. We study the dyadic distribution of the substance classes in the networks show that, for instance, tobacco smoke compounds, parabens and heavy metals tend to be connected, which hint at common disease causing factors, whereas fungicides and phytoestrogens do not. We build an exposure network at the systems level. The information gathered in this study is meant to be complementary to the genome and help us understand complex diseases, their commonalities, their causes, and how to prevent and treat them.

Keywords: Exposure; Complex Diseases; Substances; Bipartite Network; Dyadicity; Heterophilicity; Human Phenotype Network.

1. Introduction

The environment in which we live undeniably affects our health. Prolonged exposure to chemical substances present in water, soil or in the air directly impact our food sources, and are passed along to humans through ingestion or inhalation where they are the cause of many diseases and severe health issues.¹ Locally limited studies of specific chemical compounds are becoming common, linking tobacco smoke to cardiovascular and respiratory diseases, and asbestos dust to several types of cancer. However, in the same way these complex diseases are believed to be the result of multiple non-linear genetic interactions, one can speculate that they can also be caused by long-term exposure to multiple environmental factors.

Human phenotypes, including physical traits, diseases and behaviors, have been successfully linked through their shared biology and thoroughly studied using mathematical and statistical analyses of the networks they form.^{2,3} Indeed, networks offer a comprehensive array of solid analytical tools while at the same time offering an intuitive representation of interactions.⁴

The *exposome*⁵ encompasses all human environmental exposures and complements the genome for predicting disorders in “exposed” people. Starting with a systems biology approach,

combining exposome and network models,⁶ we propose to integrate the global interactions between environmental exposure and human phenotypes and diseases. This bird's eye view of the associations between human diseases and chemical compounds will help us establish relationships at the system's level – across disorder classes and our environment. While there are many resources available to map diseases to the human genome, such as the National Human Genome Research Institute GWAS Catalog⁷ or the National Center for Biotechnology Information's database of Genotypes and Phenotypes (dbGaP, <http://www.ncbi.nlm.nih.gov/gap>), there is no equivalent initiative to aggregate and freely offer known environmental exposure data. Using the Centers for Disease Control and Prevention's (CDC) National Report on Human Exposure to Environmental Chemicals, a subset of the whole exposome, we have established causal interaction data through a thorough survey of the specialized literature. We use the resulting data to build the human phenotype network (HPN) based on causal effects of environmental chemical exposure. Notable predecessors to this study were limited to a single disease,⁸ occupational exposure and diseases,⁹ infancy,¹⁰ or focused on health disparities in different populations.¹¹

We analyze the networks both in quantitative and qualitative terms; identifying most represented diseases, chemical substances, and most significant interactions among them and offering clinical and biomedical interpretation. Beyond the substances themselves, we statistically determine the chemical families or groups most responsible for diseases and how disorders and chemicals tend to cluster with those caused by some groups, but not others.

2. Methods

This section describes the steps necessary to compile the exposure data, starting from a list of environmental chemical substances and relating them to diseases and phenotypes. Then, we detail the method used to build the relationship network that will allow us to run a complete array of quantitative and qualitative analyses on the substance-to-disease relationship. Finally, we formally describe a method to study the global connection propensities of chemicals and disorders with respect to the associated substance classification group.

2.1. *Exposure Data*

Environmental exposure data, and information on the diseases that they cause have not, to the best of our knowledge, been aggregated in publicly accessible sources. To establish causal effects at a global level, we use the CDC's Fourth National Report on Human Exposure to Environmental Chemicals (<http://www.cdc.gov/exposurereport/>), including its subsequent updated tables, and the NHGRI GWAS Catalog, accessed on 05/06/2014. The former contains chemical substances, classified in families or groups that have been surveyed in the American population. We extract 60 chemicals in 11 groups, found in the environment, that form a plausible list of substances potentially harmful to our health. Table 1 recapitulates the groups and number of chemical substances in each.

For each chemical substance we perform a meticulous *PubMed* and *Google Scholar* manual literature survey and compile a list of the diseases and traits that it has been shown to (negatively) impact. Causal association between a chemical substance and a disease is based

Table 1. Substance Groups and the number of substances in each group.

Substance Classification Groups	Number of Substances
Disinfection By-Product	5
Environmental Phenols	3
Fungicides and Metalolites	1
Heavy Metals	13
Organochlorine Pesticides and Metabolites	12
Parabens	4
Perchlorate and Other Anions	3
Phytoestrogens and Metabolites	2
Polycyclic Aromatic Hydrocarbon Metabolites	1
Tobacco Smoke	2
Volatile Organic Compounds	12

on compelling evidence found in the literature and confirmed in multiple studies, limiting uncertain associations to a minimum. We subsequently use the phenotype list from the GWAS catalog and the International Classification of Diseases Ninth Revision (ICD-9) codes to classify all traits and identify redundancies. Our survey inventories 548 well-established causal effects between these 60 substances and 151 human phenotypic traits and disorders. We however note that the data collected might contain a bias towards phenotypes and exposures that are more heavily studied.

2.2. Building the Human Phenotype Network on Exposure Data

The expansion of systems biology has given rise to a trend toward studying disease from a global perspective, beyond the silos of traditional medicine. Graphs, or network, are commonly used to study the interactions between phenotype and genotype. In the Human Disease Network (HDN),² or its extension, the Human Phenotype Network,¹² nodes representing diseases and phenotypes are linked by edges that represent various connections between disorders. These connections can be established by identifying shared causal genes,² genetic variants (SNPs),¹³ linkage-disequilibrium SNP clusters,¹² biological pathways,³ or clinical symptoms.¹⁴ The underlying connections of these networks contribute to the understanding of the basis of disorders, which in turn lead to a better understanding of human disease.

Using the data collected during the substance-to-phenotype survey, we build a bipartite network.⁴ A bipartite network is a mathematical graph composed of two distinct sets of vertices – in our case, diseases and chemical substances. Vertices can only connect across sets (Figure 1b), never within. In other words, a phenotype can only connect to a substance and vice-versa. Bipartite networks can be projected onto the space of either vertex set (Figure 1a,c). In our study, we project the bipartite network onto the phenotype space, linking diseases via causal substances, and onto the substances space, which links chemicals causing common disorders. The actual networks resulting from our study and their statistical properties are presented in Section 3.

Furthermore, each node in the network is annotated with the substance classification group(s) to which it belongs. In the case of chemicals, the annotation is straight forward, as each substance belongs to exactly one class. For diseases, we identify all groups that contain

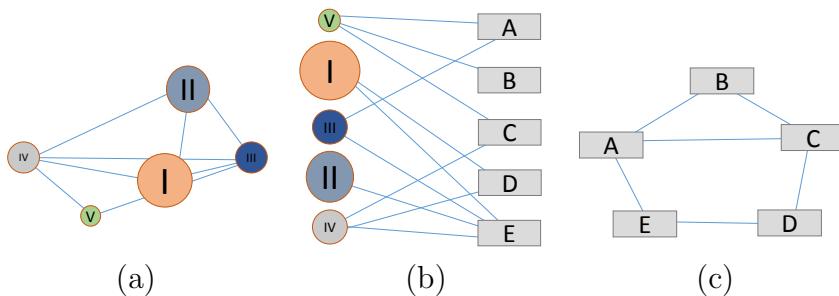


Fig. 1. Schematic representation of a Bipartite Network (b) and its projection in the space of either vertex set (a) and (c).

at least one causal substance. Additionally, we identify the “majority class” which represents the class most represented within the list of associated chemicals. The majority class is only used for coloring the network nodes in Section 3.

Assessing the Distribution of Vertex Characteristics within a Network

Beyond the standard properties, most network analyses focus on how the similar vertices are connected across the network. This type of study is very common in social sciences, where the detection of close-knit communities is a pivotal aspect of the analysis.¹⁵ However, modules or communities detection solely depends on the network structure. Alternatively, an important quantitative tool available to graph analysis is the distribution of the vertices’ characteristics across the network and how nodes with similar properties tend to link to one another. Park *et al.*¹⁶ formally define the tendency of vertices with similar characteristics or, on the contrary, vertices with dissimilar properties to connect as the *dyadicity D* and the *heterophilicity H* respectively.

The dyadicity and heterophilicity of a given vertex characteristic in a network relies on the *binary* nature of the property of interest. Either a vertex does have the studied property, in which case it is flagged accordingly (usually with a binary value 1), or a vertex does not have the given property (flagged with a 0). We define N as the total number of vertices in the network, n_1 as the number of nodes with the property and n_0 as the number of vertices without, therefore $N = n_1 + n_0$. Let M be the number of edges in the network. Each edge falls into one of three *dyads*: connecting two vertices with the given property (1–1), two vertices without (0–0), or one of each (1–0). We define m_{11} as the number of (1–1) edges, m_{10} as the number of (1–0) edges, and m_{00} the number of (0–0) edges. Therefore, $M = m_{11} + m_{10} + m_{00}$. Without losing information, we can use only m_{11} and m_{10} to analyze the dyadicity and heterophilicity of the network’s vertices properties. The mathematical formulation of D and H can be found in Fig. 2, where \bar{m}_{11} and \bar{m}_{10} are the expected values if the characteristic was distributed randomly among the vertices, and p is the average probability that two nodes are connected.

If $D > 1$, the property is called *dyadic*. It is called *anti-dyadic* otherwise. Intuitively, if a property is dyadic, nodes with that property tend to connect to one another. If anti-dyadic, then vertices without that property tend to connect. Similarly, if $H > 1$ the property is called *heterophilic*. Otherwise, it is *heterophobic*. Fig. 2(b) is a schematic representation of the $(D; H)$ coordinate space of properties. A property is heterophilic if nodes with and without the given

$$D \equiv \frac{m_{11}}{\bar{m}_{11}},$$

where $\bar{m}_{11} = \binom{n_1}{2} \times p = \frac{n_1(n_1-1)}{2} p$

and $p \equiv \frac{2M}{N(N-1)}$,

$$H \equiv \frac{m_{10}}{\bar{m}_{10}}$$

where $\bar{m}_{10} = \binom{n_1}{1} \binom{n_0}{1} \times p$

$$= n_1(N - n_1)p$$

(a)

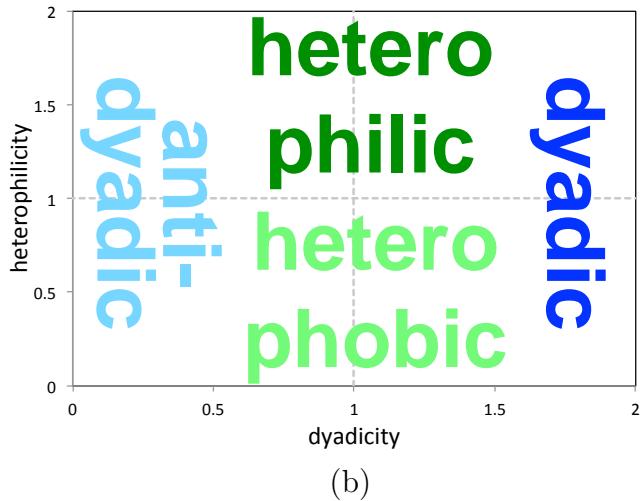


Fig. 2. Dyadicity and Heterophilicity. (a) mathematical definition and (b) schematic representation.

property tend to connect and heterophobic otherwise. The fact that a vertex characteristic can be dyadic and heterophilic (or anti-dyadic and heterophobic) at the same time is somewhat counter-intuitive. This is because D and H are defined as a statistically significant deviation of m_{11} (m_{10}) from its expected value \bar{m}_{11} (or \bar{m}_{10} respectively).

The binary properties can be virtually any attribute of the vertex to which the value is Boolean (either “yes” or “no”). In our study, we focus on the tendencies of nodes (phenotype or substance) associated with or within a certain substance class to connect to other members of that class. Results of this study are shown in the next section.

3. Results

In this section we present the bipartite network and both its projections, including a quantitative overview of the networks and degree distributions. Furthermore, we look into the most connected (hubs) in each network and into the strongest interactions within the projections to identify the highest risk factors and phenotype(s) at risk as well as the strongest connections between phenotypes.

3.1. Bipartite Network and Projections: Quantitative Study

The bipartite network is made of two distinct sets of vertices, the chemical substances and the diseases, resulting from the methods described in Section 2. This graph, represented in Fig. 3, is composed of 60 chemical substances (top row, red vertices) responsible for 151 human disorders (bottom row, light blue vertices), linked by 548 “causal-effect” edges. The node sizes are proportional to the vertex’s degree, i.e. the number of connections to the opposite set of vertices.

The “mono-partite” networks resulting from the projections in either vertex space are pictured in Fig. 4. Nodes are color coded according to their (majority) substance class. The phenotype network has 151 nodes and is very densely connected (average degree of 40+), where each edge signifies that the two endpoint diseases are associated with one or more common

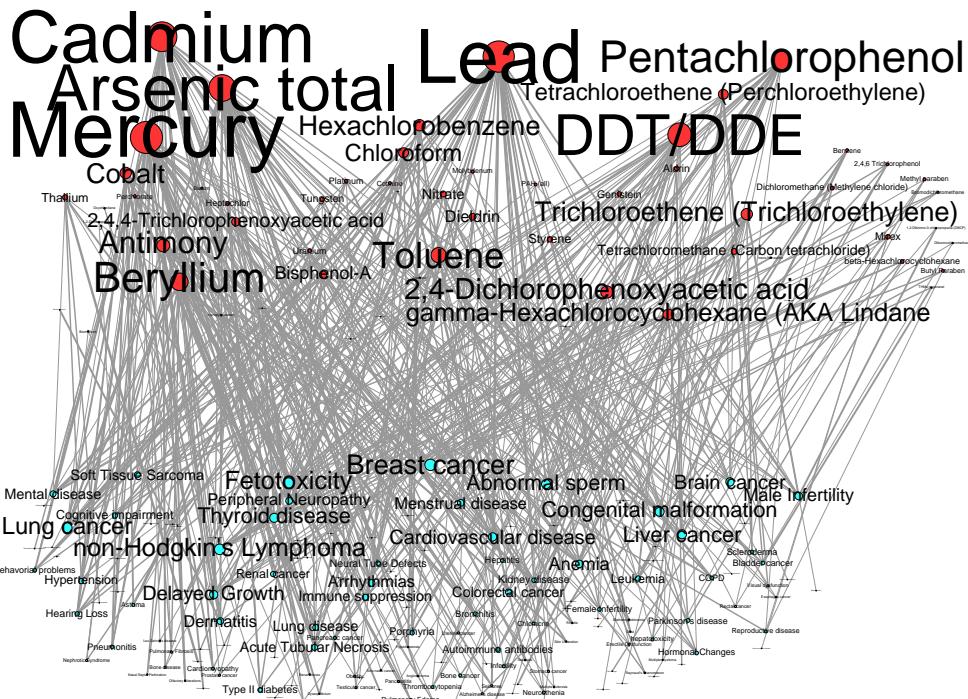


Fig. 3. Bipartite Phenotype-Substances Network. Top row, red vertices: environmental chemical substances. Bottom row, blue vertices: human phenotypes and diseases. Vertex size is proportional to the degree.

substances. The 60 substances represented in the chemicals network are each connected to about 20 other substances through shared disease(s) to which they have been associated.

3.2. Qualitative Observations and Biomedical Implications

In this section, we report qualitative observations and draw conclusions from detailed observation of the bipartite network and its projections. In the bipartite network in Figure 3, the nodes are ranked by degree or number of edges to the opposite set. For a phenotype (in blue), the edges represent the number of substances that are associated with the disease. For the substances (in red), vertices' sizes represent the number of phenotypes to which they are associated. Mercury is reported to be associated with the most phenotypes, followed by lead and cadmium. Therefore, we observe that heavy metals are the most prominent exposure class in our environment. Breast cancer is linked to the most substances, followed by lymphoma and lung cancer. Table 2 recapitulates these findings. On the right-hand side, we see the top 5 most connected nodes in each set of the bipartite network, in decreasing order of degree. The left-hand side shows the top 5 most connected vertices in either projection.

In the projection of substances, an edge represents a common phenotype associated with two different substances. DDT/DDE causes the most common diseases shared among environmental chemicals, closely followed by cadmium, lead, arsenic, and mercury. In the substance projection network, the highest edge weight is between lead and mercury, meaning that the two substances linked to many of the same diseases or share the most edges. Note that the

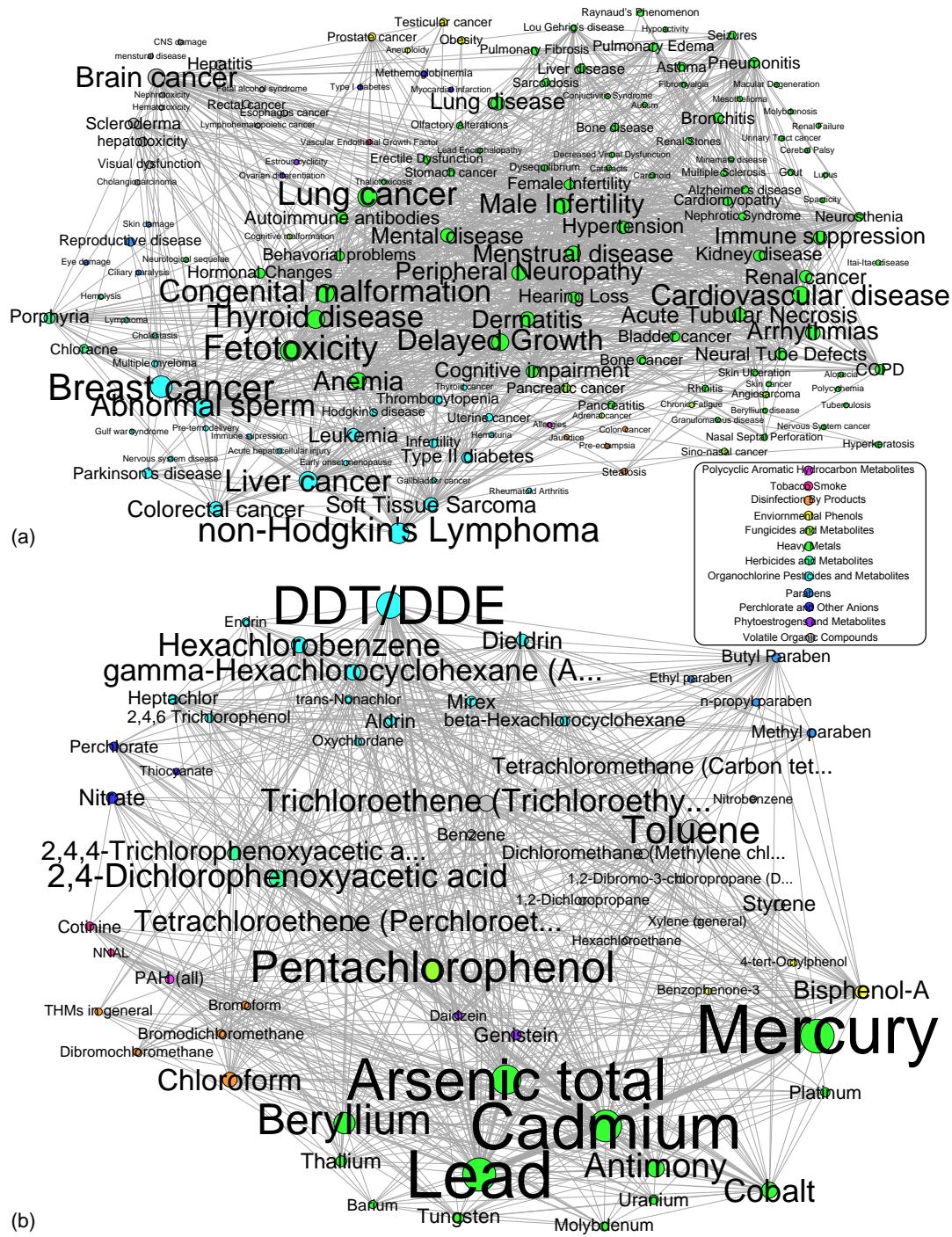


Fig. 4. Projections. Nodes are colored according to their (majority) substance group according to the legend. (a) projection of the bipartite network onto the disease/trait space. Node sizes are proportionate to the number of substances associated. Edges are weighted by the number of shared substances. (b) projection of the bipartite network on the substances space. Node sizes are proportionate to the number of diseases associated. Edges are weighted as the number of shared diseases.

Table 2. Top 5 Most Connected Vertices in both projections and in each set of the bipartite network. The number in parenthesis represents the degree of the vertex.

Projections		Bipartite	
Diseases	Substances	Diseases	Substances
fetotoxicity (112)	DDT/DDE (48)	breast cancer (15)	mercury (42)
congenital malformations (104)	cadmium (46)	fetotoxicity (14)	lead (41)
peripheral neuropathy (103)	lead (45)	non-Hodgkin's lymphoma (14)	cadmium (39)
immune suppression (100)	arsenic (45)	lung cancer (13)	arsenic (35)
lung cancer, anemia, delayed growth (99)	mercury (42)	abnormal sperm, thyroid disease, liver cancer, congenital malformations (12)	DDT/DDE (31)

two substances are among the largest nodes when ranked by degree. Other significant edges are those between lead and cadmium and cadmium and mercury. The edges with the highest weights link substances that are related beyond just the similar phenotypes they might cause. Lead and mercury residue from old mines are found together in the form of household dust.¹⁷ Mushrooms and vegetables can contain lead and cadmium, poisoning consumers.^{18,19} Cadmium and mercury are found in soil near mercury mines and also in utility batteries.²⁰ These substances have a tendency to be present together even beyond the above examples, like in automobile emissions and soil. The copresence of the above substances and heavy metals in general may contribute to their similar health effects and should be noted when considering their high edge weight. The node size can be ranked not just by degree, but also by the number of phenotypes a substance causes. When this ranking is done, the largest nodes are mercury, lead, cadmium, arsenic, and DDT/DDE, causing 42, 41, 39, 35, and 31 phenotypes, respectively. This seems logical, because if a substance is associated with a large number of diseases, there is a high probability other substances in the network share these diseases. Out of the substances, lead, cadmium, mercury, arsenic, and DDT/DDE seem to have the most significant health effects. The occurrence of some highly connected substances may be explained by their co-presence in the world.

The phenotypes in the bipartite network that are associated with the most substances are breast cancer, fetotoxicity, non-Hodgkin's lymphoma, lung cancer, abnormal sperm, thyroid disease, liver cancer, congenital malformations, cardiovascular disease, delayed growth, and brain cancer. Breast cancer is linked to a combination of 15 substances; fetotoxicity and non-Hodgkin's lymphoma by 14; lung cancer by 13; abnormal sperm, thyroid disease, liver cancer, congenital malformations by 12; and cardiovascular disease, delayed growth, and brain cancer by 11.

In Table 3, we present the strongest pairwise connection within each network projection. When the phenotype projection nodes are ranked by degree, the largest nodes are fetotoxicity, congenital malformations, peripheral neuropathy, immune suppression, lung cancer, anemia, and delayed growth. Ranking by degree indicates that these phenotypes are linked to the most shared substances with other phenotypes. Breast cancer, non-Hodgkin's lymphoma, abnormal

Table 3. Strongest connections among pairs of Environmental Chemical Substances causing the most common diseases, and among pairs of Diseases and the number of substances they have in common.

Substance pair	#shared diseases	Disease pair	#shared substances
lead – mercury	28	fetotoxicity – congenital malformations	9
lead – cadmium	23	fetotoxicity – delayed growth	8
mercury – cadmium	22	breast cancers – non-Hodgkin's lymphoma	8
arsenic – cadmium	17	lung cancer – cardiovascular diseases	7
arsenic – mercury	17	fetotoxicity – renal cancer	7

sperm, thyroid disease, liver cancer, cardiovascular disease and brain cancer are no longer among the largest nodes when degree ranking is utilized. These phenotypes, though common among the network, must be caused by substances that do not cause as many phenotypes as other substances. Phenotypes prevalent in degree ranking that are not among the most common phenotypes in the network are peripheral neuropathy, immune suppression, and anemia. These phenotypes are not related to the most substances but the substances that are responsible for them also cause many other phenotypes. It may be suspected that fetotoxicity, congenital malformations, lung cancer, and delayed growth are linked to the most prevalent substances since the phenotypes exhibit strong connections between many substances and phenotypes. Indeed, these phenotypes are caused by at least four out of the five most prevalent substances. Literature searches were done between each of the top edge weight phenotype pairings in an attempt to identify a genetic link. When the Klf4 gene was deleted, mice showed growth retardation and death before or just after birth.²¹ Thus, the relationship between Feto-toxicity and Delayed Growth via exposure to substances may supplement an existing genetic component. Liver cancer and non-Hodgkin's lymphoma also seemed to be associated with a shared gene: *p53*, a known cancer-causing gene. Though no specific genetic connection has been identified between the two phenotypes, both have been independently linked to *p53*.^{22,23} Again, this may partially explain the higher edge weight and indicate both genetic and environmental relationships between the two phenotypes. Thirdly, there is a documented interaction between cardiovascular disease and lung cancer outside of environmental exposure. A disruption in the SMAD proteins has been linked to both cardiovascular disease and lung cancer.²⁴ In addition to literature searches, we study overlap in the genome based HPN^{3,12} for phenotype connections. Unfortunately, there were no phenotypes observed in the original HPN similar or relating to fetotoxicity or congenital malformations, so the only pairings that could be searched were non-Hodgkin's lymphoma and liver cancer, breast cancer and non-Hodgkin's lymphoma, and lung cancer and cardiovascular disease. Only one shared edge, with a weight of thirty, was found between lung cancer and cardiovascular disease in the GWAS pathway analysis. The genetic and environmental relationships between fetotoxicity and delayed growth, liver cancer and non-Hodgkin's lymphoma, and lung cancer and cardiovascular disease may partially explain their higher edge weights. Other phenotype pairings with significant edges that yielded no genetic connections may be related only by environmental exposure or the genetics of the phenotype or interacting phenotypes have not been fully studied.

3.3. Distribution of Substance Classes with the Projection Networks

The dyadicity and heterophilicity analysis described in Section 2 is used to study the trends within both projection networks of substance class correlations. Each substance class (found in Table 1) is considered a binary attribute of the vertices. In the substances projection, a vertex is associated with a single chemical class. In the phenotype network, nodes may be associated with more than one substance class, the maximum being the number of substances themselves. We report the numerical and plotted results of the dyadicity study for the projection networks in Figure 5.*D* and *H* are considered significant when their respective p-value is < 0.05 (bold fonts in the table). We obtain significance measurements by performing 1,000 random permutation tests on the distribution of each characteristic within the network. On the right-hand side, for each projection separately, we plot all substances' coordinates in the (*D*; *H*) space to facilitate the interpretation of the results in a qualitative manner.

Substance Classification	projection:		substance	
	phenotype <i>D</i>	<i>H</i>	<i>D</i>	<i>H</i>
Disinfection By-Products	2.486	1.373	0.526	0.756
Environmental Phenols	3.037	1.377	0	0.815
Fungicides and Metabolites	3.398	1.234	0	1.649
Heavy Metals	1.789	0.431	2.225	1.024
Herbicides and Metabolites	3.022	1.161	2.630	1.406
Organochlorine Pesticides and Metabolites	2.23	0.955	2.152	0.991
Parabens	1.942	0.745	2.192	0.611
Perchlorate and Other Anions	2.574	1.263	0.877	0.661
Phytoestrogens and Metabolites	2.427	1.263	0	0.703
Polycyclic Aromatic Hydrocarbon Metabolites	3.398	1.277	0	0.892
Tobacco Smoke	3.398	1.474	2.630	0.589
Volatile Organic Compounds	2.011	1.088	0.956	0.822

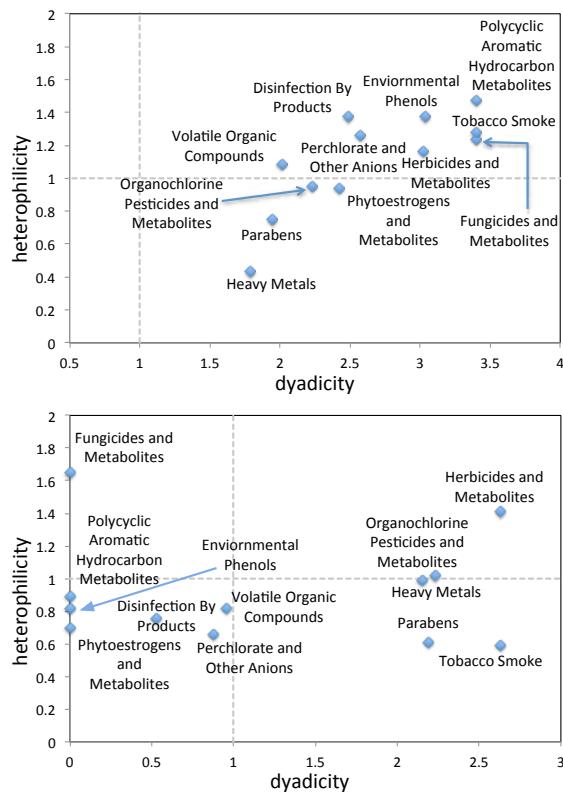


Fig. 5. Dyadicity and Heterophilicity Analysis of the Substances Classification Distribution in both Projection Networks. The statistically significant values ($p - \text{value} < 0.05$) are in bold and underlined. On the right-hand side, at the top: plotted values for the phenotype projection. Bottom: plotted values for the substances projection network.

Common to both networks, heavy metals, herbicides, organochlorine pesticides, and tobacco smoke are significantly dyadic. This means that diseases caused by chemicals in these classes and the chemicals themselves tend to connect to other members of their respective classes. In fact, in the phenotype network, all substance classes are significantly dyadic, ex-

cept for parabens. In the substances network, only parabens are additionally significantly dyadic.

Looking at the vertices that favor connecting to nodes that do not share their characteristic common to both networks, only volatile organic compounds have significant values. However, in the phenotype network, those nodes are heterophilic, whereas they are heterophobic in the substance network. In the phenotype network, disinfectants, phenols, fungicides, herbicides are all also significantly heterophobic. Heavy metals and pesticides are heterophobic.

From a clinical viewpoint, the dyadicity analysis tells us that diseases caused by dyadic classes, i.e. tobacco smoke and phenols tend to connect through their substances. In the phenotype network, there are no anti-dyadic classes. In the substances network, we see that tobacco smoke and herbicides are dyadic, and their group members (substances) tend to cause the same diseases. However, herbicides are also heterophilic, and are responsible for diseases belonging to different classes. Organochlorines and heavy metals are neither heterophilic nor heterophobic, but dyadic, causing the same subset of diseases. At the other end of the spectrum, fungicides are highly heterophilic, taking part in causing diseases in many different groups. Volatile organic compounds are the most neutral substances.

4. Conclusions & Future Work

Environmental exposure data are part of most recent GWAS. They are however limited to the disease of interest and centered around factors possibly impacting that particular disease. In this work, we take a global approach, conducting an in-depth literature search to identify chemical substances present in the environment and their possible adverse effects on our health. The result is the Human Phenotype Network, based on common causal substances. Breast cancer and injury to the fetus are the most connected phenotypes in the network, making them the most susceptible to environmental chemicals, namely the heavy metals: mercury, lead and cadmium. These are in turn the environmental substances associated with the most diseases. Moreover, the substance-class dyadicity analysis of both projected networks reveals that all substance classes in the phenotype network are dyadic, and tend to connect to similar classes. However, only about half of them are also heterophilic, also connecting to different substance families. The information gathered in this study is meant to be complementary to the genome in helping us understand complex diseases, their commonalities, their causes, and how to prevent and treat them. The current work is limited by the availability of reliable exposure data linked to human diseases.

We are planning on extending this work in several directions. First, we will add geographical information into the model, as most of the environmental chemical substances are limited in their physical locations. Secondly, it would be interesting, though challenging due to the lack of available data, to segregate the diseases by ethnic background. Finally, we will merge the chemical-substance based HPN to the genetic HPN,²⁵ analyzing the overlap and differences. Combined, this new global HPN has the potential to inform us on both genetic and environmental causes of a large array of common and complex disease.

Acknowledgments

Financial supported by NIH grants R01 EY022300, LM009012, LM010098, AI59694.

References

1. A. A. Rooney, A. L. Boyles, M. S. Wolfe, J. R. Bucher and K. A. Thayer, *Environ Health Perspect* **122**, 711 (Jul 2014).
2. K.-I. Goh, M. E. Cusick, D. Valle, B. Childs, M. Vidal and A.-L. Barabasi, *Proceedings of the National Academy of Sciences* **104**, 8685 (2007).
3. C. Darabos, M. J. White, B. E. Graham, D. N. Leung, S. Williams and J. H. Moore, *BioData Min* **7**, p. 1 (Jan 2014).
4. M. Newman, *Networks: An Introduction* (Oxford University Press, Inc., New York, NY, USA, 2010).
5. C. P. Wild, *Cancer Epidemiol Biomarkers Prev* **14**, 1847 (Aug 2005).
6. J. D. Pleil and L. S. Sheldon, *Biomarkers* **16**, 99 (2011), PMID: 21138393.
7. D. Welter, J. MacArthur, J. Morales, T. Burdett, P. Hall, H. Junkins, A. Klemm, P. Flicek, T. Manolio, L. Hindorff and H. Parkinson, *Nucleic Acids Res* **42**, D1001 (Jan 2014).
8. C. J. Patel, J. Bhattacharya and A. J. Butte, *PLoS ONE* **5**, p. e10746 (05 2010).
9. L. Faisandier, V. Bonneterre, R. De Gaudemaris and D. J. Bicout, *J Biomed Inform* **44**, 545 (Aug 2011).
10. M. Vrijheid, R. Slama, O. Robinson, L. Chatzi, M. Coen, P. van den Hazel, C. Thomsen, J. Wright, T. J. Athersuch, N. Avellana, X. Basagana, C. Brochot, L. Buccini, M. Bustamante, A. Carracedo, M. Casas, X. Estivill, L. Fairley, D. van Gent, J. R. Gonzalez, B. Granum, R. Grazuleviciene, K. B. Gutzkow, J. Julvez, H. C. Keun, M. Kogevinas, R. R. C. McEachan, H. M. Meltzer, E. Sabido, P. E. Schwarze, V. Siroux, J. Sunyer, E. J. Want, F. Zeman and M. J. Nieuwenhuijsen, *Environ Health Perspect* **122**, 535 (Jun 2014).
11. P. Juarez, *J Health Care Poor Underserved* **24**, 114 (Feb 2013).
12. C. Darabos, K. Desai, R. Cowper-Sallari, M. Giacobini, B. Graham, M. Lupien and J. Moore, Inferring human phenotype networks from genome-wide genetic associations, in *Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics*, eds. L. Vanneschi, W. Bush and M. Giacobini, Lecture Notes in Computer Science, Vol. 7833 (Springer Berlin Heidelberg, 2013) pp. 23–34.
13. H. Li, Y. Lee, J. L. Chen, E. Rebman, J. Li and Y. A. Lussier, *Journal of the American Medical Informatics Association : JAMIA* **19**, 295 (January 2012).
14. X. Zhou, J. Menche, A.-L. Barabasi and A. Sharma, *Nat Commun* **5** (06 2014).
15. D. J. Watts and S. H. Strogatz, *Nature* **393**, 440 (1998).
16. J. Park and A.-L. Barabási, *Proceedings of the National Academy of Sciences* **104**, 17916 (2007).
17. E. I. Hamilton, *Sci Total Environ* **249**, 171 (Apr 2000).
18. M. A. Rahman, M. M. Rahman, S. M. Reichman, R. P. Lim and R. Naidu, *Ecotoxicol Environ Saf* **100**, 53 (Feb 2014).
19. S. A. S. Petkovsek and B. Pokorny, *Sci Total Environ* **443**, 944 (Jan 2013).
20. M. Vahter, S. A. Counter, G. Laurell, L. H. Buchanan, F. Ortega, A. Schutz and S. Skerfving, *Int Arch Occup Environ Health* **70**, 282 (1997).
21. T. Yoshida, Q. Gan, A. S. Franke, R. Ho, J. Zhang, Y. E. Chen, M. Hayashi, M. W. Majesky, A. V. Somlyo and G. K. Owens, *J Biol Chem* **285**, 21175 (Jul 2010).
22. Y. Chen, Z. Xiang, H. Li, N. Yang and H. Zhang, *J Tongji Med Univ* **19**, 27 (1999).
23. R. Said, Y. Ye, D. S. Hong, F. Janku, S. Fu, A. Naing, J. J. Wheler, R. Kurzrock, C. Thomas, G. A. Palmer, K. R. Hess, K. Aldape and A. M. Tsimberidou, *Oncotarget* **5**, 3871 (Jun 2014).
24. M. Witkowska and P. Smolewski, *Postepy Hig Med Dosw (Online)* **68**, 301 (2014).
25. C. Darabos, M. White, B. Graham, D. Leung, S. Williams and J. Moore, *BioData Mining* **7**, p. 1 (2014).

A SCREENING-TESTING APPROACH FOR DETECTING GENE-ENVIRONMENT INTERACTIONS USING SEQUENTIAL PENALIZED AND UNPENALIZED MULTIPLE LOGISTIC REGRESSION

H. ROBERT FROST^{*1,2,3}, ANGELINE S. ANDREW^{1,2}, MARGARET R. KARAGAS^{1,2},
AND JASON H. MOORE^{1,2,3}

¹*Institute for Quantitative Biomedical Sciences,
Geisel School of Medicine, Dartmouth College,
Lebanon, NH 03756*

²*Section of Biostatistics and Epidemiology, Department of Community and Family Medicine,
Geisel School of Medicine, Dartmouth College,
Lebanon, NH 03756*

³*Department of Genetics,
Geisel School of Medicine, Dartmouth College,
Hanover, NH 03755*

Gene-environment ($G \times E$) interactions are biologically important for a wide range of environmental exposures and clinical outcomes. Because of the large number of potential interactions in genome-wide association data, the standard approach fits one model per $G \times E$ interaction with multiple hypothesis correction (MHC) used to control the type I error rate. Although sometimes effective, using one model per candidate $G \times E$ interaction test has two important limitations: low power due to MHC and omitted variable bias. To avoid the coefficient estimation bias associated with independent models, researchers have used penalized regression methods to jointly test all main effects and interactions in a single regression model. Although penalized regression supports joint analysis of all interactions, can be used with hierarchical constraints, and offers excellent predictive performance, it cannot assess the statistical significance of $G \times E$ interactions or compute meaningful estimates of effect size. To address the challenge of low power, researchers have separately explored screening-testing, or two-stage, methods in which the set of potential $G \times E$ interactions is first filtered and then tested for interactions with MHC only applied to the tests actually performed in the second stage. Although two-stage methods are statistically valid and effective at improving power, they still test multiple separate models and so are impacted by MHC and biased coefficient estimation. To remedy the challenges of both poor power and omitted variable bias encountered with traditional $G \times E$ interaction detection methods, we propose a novel approach that combines elements of screening-testing and hierarchical penalized regression. Specifically, our proposed method uses, in the first stage, an elastic net-penalized multiple logistic regression model to jointly estimate either the marginal association filter statistic or the gene-environment correlation filter statistic for all candidate genetic markers. In the second stage, a single multiple logistic regression model is used to jointly assess marginal terms and $G \times E$ interactions for all genetic markers that pass the first stage filter. A single likelihood-ratio test is used to determine whether any of the interactions are statistically significant. We demonstrate the efficacy of our method relative to alternative $G \times E$ detection methods on a bladder cancer data set.

1. Introduction

A significant body of recent research in the statistical genetics and genetic epidemiology communities has focused on the detection of statistical interactions between genetic markers and environmental variables ($G \times E$ interactions) using genome-wide association (GWA) data.¹

Such data sets are comprised by the measurements of thousands to over one million genetic markers, typically single nucleotide polymorphisms (SNPs), along with relevant clinical and environmental variables on a set of human subjects that number in the thousands to hundreds-of-thousands for large GWA studies. Since the number genetic markers, and therefore the number of potential $G \times E$ interactions for a single environmental variable, is usually larger than the number of subjects, statistical testing of $G \times E$ interactions has typically been accomplished by fitting separate models for each genetic marker and applying multiple hypothesis correction (MHC) to the generated p-values to control the type I error rate. Although a $G \times E$ interaction can be defined as a departure from additivity on either a log odds or absolute risk scale, we focus on the former type of interaction in this paper. Statistically, such an interaction is commonly tested using a logistic regression model of the form:

$$\text{logit}(P(D = 1|G, E)) = \beta_0 + \beta_E E + \beta_G G + \beta_{GE} GE \quad (1)$$

where D is a binary outcome variable, E is the environmental variable and G is one of the genetic markers. In this paper, we assume that both D and E are binary, e.g., disease case/control status and exposed/non-exposed indicator, and that G represents a SNP specified using additive coding, i.e., 0, 1 or 2 based on the number of copies of the minor allele. Using this model, the null hypothesis of no $G \times E$ interaction on a log odds scale can be specified as $H_0 : \beta_{GE} = 0$ with significance tested via either a Wald test associated with $\hat{\beta}_3$ or a likelihood ratio test. Variations on this basic approach that also use one model per potential $G \times E$ interaction include the case-only gene-environment association test, the test of marginal association and the combined test of marginal gene association and $G \times E$ interaction.²

Although methods that test a separate model for each potential $G \times E$ interaction (so-called one-step methods) are easy to understand, simple to implement and can, in many instances, identify biologically plausible interactions, they have two serious drawbacks. First, the power to detect $G \times E$ interactions, already much lower than the power to detect main effects at a given sample size,³ is severely degraded for even a moderate number of genetic markers due to the requirement for MHC to control the type I error rate across the separate models. Second, because each interaction is assessed independently, the estimated interaction coefficients will be biased if associations exist between the genetic markers.

To address the poor statistical power of standard one-step methods, researchers have recently explored two-stage, or screening-testing, methods.^{4–9} Screening-testing methods first filter the set of candidate genetic markers and, for the markers that pass the first stage filter, test $G \times E$ interactions. As long as the statistic used to filter the genetic markers in the first stage is statistically independent of the second stage test statistic under the null hypothesis, type I error rates will be correctly controlled with MHC applied to just the smaller number of hypotheses that pass the first stage filter.^{7,10} Two popular independent filters for $G \times E$ interaction detection are the marginal association filter⁴ and the gene-environment correlation filter.⁵ The marginal association filter measures the statistical association between the outcome variable and the genetic marker using a logistic regression model of the form:

$$\text{logit}(P(D = 1|G)) = \beta_0 + \beta_G G \quad (2)$$

where the filter statistic is the p-value associated with the $\hat{\beta}_1$ coefficient estimate. The gene-

environment correlation filter measures the statistical association between the environmental variable and each genetic marker using a logistic regression model of the form:

$$\text{logit}(P(E = 1|G)) = \beta_0 + \beta_G G \quad (3)$$

This is the same model used with the case-only gene-environment association test for $G \times E$ interaction, but fit, for the correlation filter, using pooled cases and controls instead of just cases. Similar to the marginal association filter, the correlation filter uses the p-value associated with the $\hat{\beta}_G$ coefficient estimate as a filter statistic. To be effective at improving power, a filter statistic must not only be independent of the second stage test statistic under the null hypothesis, but must also be associated with the test statistic under the alternative hypothesis of $G \times E$ interaction. While the first requirement has been proven for both the marginal association and correlation filter statistics in the context of $G \times E$ interaction detection using logistic regression models of the form in equation 1,⁷ there is no guarantee that the second requirement will hold for the data set under analysis. For some data sets, the marginal association filter will be optimal, for others the correlation filter will perform best and success has been reported using an ensemble of both filter types, e.g., the cocktail method.⁸ Although current screening-testing methods can be effective at improving $G \times E$ detection power, the fact that these methods use separate models for each candidate genetic marker during both the filter and testing stages means that both MHC correction and omitted variable bias remain issues. For data sets with small sample-to-marker or signal-to-noise ratios, even the reduced MHC penalty after filtering is sufficient to negate $G \times E$ detection power.

Another recently developed approach to $G \times E$ interaction detection involves the use of penalized regression to jointly estimate all possible $G \times E$ interactions as well as main effects in a single model. Such penalization approaches typically enforce a hierarchical constraint that will only consider interaction terms for significant main effects. Many variations of the joint penalized model approach exist, including the hierarchical LASSO by Bien et al.,¹¹ the penalized hierarchical approach of Liu et al.,¹² the progressive hierarchical penalization approach of Zhu et al.,¹³ the multi-stage LASSO method of Wu et al.¹⁴ and approaches that fit a single LASSO-penalized model with all possible marginal and interaction terms (termed all-pairs LASSO (APL) by Bien et al.)¹¹ The approach of Wu et al.¹⁴ is especially relevant since it employs a LASSO penalized multiple logistic regression model in the first stage to filter genetic markers based on marginal association and then tests for interactions in a second stage model. The fact that Wu et al. use LASSO penalization in the second stage model, however, means that their approach cannot generate valid measures of interaction statistical significance and is therefore not a valid screening-testing method. Also, Wu et al. focus on gene-gene as opposed to gene-environment interactions. Methods that fit a single penalized model have the significant benefit of jointly estimating all potential $G \times E$ interactions along with marginal gene and environmental effects and can therefore be very effective for prediction; however, the shrunken coefficients may be severely biased with unclear statistical significance. Although some authors, e.g. Wu et al.,¹⁴ advocate refitting a non-penalized model for just the interaction terms with non-zero coefficients in the penalized model to generate more meaningful coefficients and measures of statistical significance, this approach fails to account for the prior penalized selection process and thus cannot correctly compute statistical

significance or interaction effect size.

To address the limitations of inadequate power and biased coefficient estimation associated with one-step approaches, we have developed a novel G×E interaction detection method that combines aspects of screening-testing with hierarchical penalized regression. In the first stage, our approach uses a single elastic net-penalized multiple logistic regression model to jointly estimate either the marginal association filter statistic or the gene-environment correlation filter statistic for all candidate genetic markers. In the second stage, a single multiple logistic regression model is used to jointly assess marginal effects and G×E interactions for all genetic markers that pass the first stage filter. An important feature of our approach is that a single omnibus test can be used to detect the presence of statistically significant G×E interactions. As we demonstrate using a bladder cancer genotype data set with smoking status as the environmental variable, our method provides the statistical benefits of joint estimation along with significantly improved G×E detection power relative to competing approaches.

2. Methods

2.1. Proposed screening-testing method for G×E interaction detection

2.1.1. Screening stage

Our approach filters the set of measured genetic markers using a penalized multiple logistic regression model that jointly computes a filter statistic, either the marginal association filter⁴ or the gene-environment correlation filter,⁵ for all measured genetic markers. For the marginal association filter, a penalized multiple logistic regression model of the following form is used:

$$\text{logit}(P(D = 1|G)) = \beta_0 + \beta_{G_1}G_1 + \dots + \beta_{G_p}G_p \quad (4)$$

This model is fit using an elastic net¹⁵ penalty via the *glmnet* R package implementation.¹⁶ This procedure computes coefficient estimates to maximize an objective function with both L1, i.e., LASSO, and L2, i.e., ridge, penalties:

$$-\frac{\log(L(\beta_1, \dots, \beta_p|G))}{n} + \lambda\left(\frac{1-\alpha}{2}\sum_{i=1}^p \beta_{G_i}^2 + \alpha\sum_{i=1}^p |\beta_{G_i}|\right) \quad (5)$$

where the α coefficient is the elastic net mixing parameter ($\alpha = 1$ corresponds to just LASSO penalization and $\alpha = 0$ corresponds to just ridge penalization). The elastic net penalty parameter λ can be selected according to cross-validation or to achieve a specific number of non-zero coefficients. For the gene-environment correlation filter, a penalized multiple logistic regression model of the following form is used:

$$\text{logit}(P(E = 1|G)) = \beta_0 + \beta_{G_1}G_1 + \dots + \beta_{G_p}G_p \quad (6)$$

Modeling fitting in this case follows the same approach used for the model in equation 4.

2.1.2. Testing stage

To test for G×E interactions, a single multiple logistic regression model is fit using marginal and interaction terms for all genetic markers selected during the screening stage:

$$\text{logit}(P(D = 1|G, E)) = \beta_0 + \beta_E E + \beta_{G_1} G_1 + \dots + \beta_{G_p} G_p + \beta_{G_1 E} G_1 E + \dots + \beta_{G_p E} G_p E \quad (7)$$

If desired, covariates can also be included in this model. To determine if any of the $G \times E$ interaction coefficients are statistically significant, the null hypothesis $H_0 : \beta_{G_1 E} = \dots = \beta_{G_p E} = 0$ is tested using a LR test between a version of the model in equation 7 without the interaction terms and the model with interaction terms. If p-value from this LR test is significant, this indicates that at least one of the $G \times E$ interaction coefficients is significantly non-zero. The interactions can then be prioritized for further investigation based on the estimated interaction coefficient size and the associated Wald test p-values, perhaps after MHC. If the p-value from this LR test is not significant, no further investigation is performed.

If only one of the two supported filter statistics is used during the first stage, then the model in equation 7 is fit just a single time and only one LR test is performed to detect potential $G \times E$ interactions. If both filters are applied, the model in equation 7 is fit separately using the output from each filter, LR tests are performed on both models and the generated p-values are adjusted via the Bonferroni method, i.e., 2^*p-value . If neither model has a significant LR p-value after MHC, no further investigation is performed, otherwise, the model with the most significant LR test result is used.

Although an ensemble approach, similar to the cocktail method,⁸ could be adopted that combines the results from the marginal association and gene-environment correlation filters to build a single stage two model, such an approach would eliminate a key benefit of filtering based on a single penalized regression model, namely the reduction of multi-collinearities. Because the marginal association and gene-environment correlation models would be estimated separately, each model could identify genetic markers highly correlated with the predictors output by the other model, resulting in estimation instability for the second stage multiple logistic regression model.

2.1.3. Interpretation

Because the coefficients in a multiple logistic regression model with interaction terms represent conditional effects on the log-odds of the outcome variable, they do not have a straight-forward interpretation. This complex, conditional interpretation can be seen as a disadvantage of the proposed approach relative to separate logistic regression models for each interaction. Specifically, the effect size of each interaction term must be evaluated by considering the estimated coefficients for both the G_i predictor and the $G_i E$ predictor. To be precise, the change in the log odds of the outcome per change in the number of minor allele copies (assuming additive coding) when there is no environmental exposure and all other predictors are held constant is represented by the estimated coefficient for the G_i predictor and the change in the log odds of the outcome per change in the number of minor allele copies when there is environmental exposure is represented by the sum of the estimated coefficients for the G_i and $G_i E$ predictors. In more simplified terms, the estimated coefficient for the $G_i E$ predictor reflects increased risk of disease for environmentally exposed individuals who carry the risk allele compared to unexposed risk allele carriers.

2.2. Bladder cancer data

We analyzed genetic variation in hypothesized cancer susceptibility genes and cigarette smoking in a population-based case-control study of bladder cancer. Detailed methods have been described previously in Karagas et al.¹⁷ and Andrew et al.¹⁸ Briefly, the cases were New Hampshire residents of ages 25 to 74 years, diagnosed with bladder cancer from July 1, 1994 to June 30, 2001 identified via the New Hampshire State Cancer Registry. Controls less than 65 years of age were selected using population lists obtained from the New Hampshire Department of Transportation, while controls aged 65 and older were chosen from data provided by the Centers for Medicare & Medicaid Services (CMS) of New Hampshire. The overwhelming majority (~ 98%) of the subjects were of Caucasian origin. Given the large proportion of Caucasian subjects, population structure should not be an issue for this data set, as confirmed in Andrew et al.¹⁹ We interviewed a total of 857 patients with bladder cancer, which was 85% of the cases confirmed to be eligible for the study, and 1191 controls without cancer. Informed consent was obtained from each participant and all procedures and study materials were approved by the Committee for the Protection of Human Subjects at Dartmouth College. DNA was isolated from peripheral circulating blood lymphocyte or buccal specimens using Qiagen genomic DNA extraction kits (QIAGEN Inc., Valencia, CA). Genotyping was performed on all DNA samples of sufficient concentration using the GoldenGate Assay system (Illumina, Inc., San Diego, CA). Out of the submitted samples, 99.5% were successfully genotyped, and samples repeated on multiple plates yielded the same call for 99.9% of the SNPs.²⁰ Excluding subjects who did not have genotype calls for more than 50% of the SNPs, and one additional case due to missing data on smoking status, resulted in 610 cases and 865 controls included in our analysis. After removing SNPs with missing genotype values for more than 10% of the 1475 samples, we analyzed genotype data for a total of 1488 SNPs. Remaining missing genotype values in this dataset were imputed using a simple frequency-based approach in which the missing value was set to the most common genotype in the study population. Genotyped SNPs were mostly those included on the Illumina Cancer Panel, representing ≈ 400 hypothesized cancer-related genes. SNPs were selected within coding, intronic and flanking regions hypothesized to be potentially functional for the genes of interest, including a median of three SNPs per gene.

2.3. $G \times E$ interaction detection for bladder cancer data

To support comparative evaluation of our proposed method using the bladder cancer data set described above, potential $G \times E$ interactions between smoking status (recoded as never (0) or ever (1)) and SNPs relative to bladder cancer case/control status were computed using the proposed $G \times E$ interaction detection method and standard one-step and two-stage approaches. Analysis details for each method are outlined in Sections 2.3.1-2.3.3 below. For all methods, age and gender were included as covariates in the logistic regression models used to test for $G \times E$ interactions, i.e., the models specified by equations 1 and 7.

2.3.1. One-step $G \times E$ interaction test for bladder cancer data

For each of the 1488 analyzed SNPs, a SNP-smoking interaction was tested using a separate logistic regression model of the form specified in equation 1 above with MHC performed using the false discovery rate (FDR) method of Benjamini and Hochberg.²¹

2.3.2. Standard two-stage $G \times E$ interaction test for bladder cancer data

The 1488 analyzed SNPs were first filtered using either the marginal association filter statistic, as computed via the logistic regression model in equation 2, or the gene-environment correlation filter statistic, as computed via the logistic regression model in equation 3. So that the results from the two-stage method would be comparable to those generated by our proposed $G \times E$ detection method, the number of SNPs retained after the first stage filtering was fixed at the same number used for the proposed method (103 SNPs, see Section 2.3.3 below). For the SNPs that passed the first stage filter, the presence of a SNP-smoking interaction was tested using the same logistic regression model and MHC method employed for the one-step $G \times E$ test.

2.3.3. Proposed $G \times E$ interaction detection method for bladder cancer data

The screening-testing approach outlined in Section 2.1 above was executed on the bladder cancer data using both the marginal association filter and the gene-environment correlation filter. For each filter, the screening stage penalized logistic regression model was fit with the elastic net mixing parameter α set to .999 to provide estimation stability via a small L2 penalty¹⁶ and the λ penalty parameter was set to achieve a ratio of observations-to-predictors in the unpenalized stage 2 model of 7 (the middle of the 5-to-9 recommended by Vittinghoff and McCulloch for multiple logistic regression²²). For the 1475 analyzed subjects, the observations-to-predictors ratio of 7 allowed approximately 103 SNPs to be kept after screening with corresponding λ values of 0.0112 and 0.0118 for the gene-environment correlation and marginal association filters, respectively.

3. Results

3.1. One-step $G \times E$ interaction test results

Table 1 shows the ten most significant smoking-SNP interactions computed via the one-step method detailed in Section 2.3.1. The $\hat{\beta}_{GE}$ values in the table represent the estimated interaction term coefficients from the logistic regression model specified in equation 1 with age and gender as covariates. The p-values were generated via a LR test comparing a model without the $G \times E$ term to the model with the $G \times E$ term and the false discovery rate (FDR) values were generated for all p-values using the method of Benjamini and Hochberg.²¹ Although some of the interaction LR p-values appear significant, after MHC to control the FDR, all findings appear consistent with H_0 . In addition to the poor power after MHC, half of the top ten interactions returned by the one-step method involve highly correlated SNPs from the same gene, MASP1; a direct result of testing each interaction in a separate regression model.

Table 1. Ten most significant smoking-SNP interactions computed using the standard one-step method.

dbSNP ID	Gene name	$\hat{\beta}_{GE}$	LR p-val	FDR
rs12635264	MASP1	-0.604	0.00142	0.848
rs13089330	MASP1	-0.596	0.00165	0.848
rs13094773	MASP1	-0.556	0.00332	0.848
rs2972418	GHR	-0.519	0.00383	0.848
rs9282553	ABCA6	-1.07	0.00419	0.848
rs3864099	MASP1	-0.541	0.00437	0.848
rs4376034	MASP1	-0.536	0.00447	0.848
rs3213216	IGF2	0.585	0.00457	0.848
rs3217773	CCNA2	0.6	0.0056	0.848
rs2229765	IGF1R	0.497	0.00589	0.848

3.2. Standard two-stage $G \times E$ interaction test results

Table 2. Ten most significant smoking-SNP interactions computed via the standard two-stage $G \times E$ detection method using either a marginal association filter or a gene-environment correlation filter.

Marginal association filter					G-E correlation filter				
dbSNP ID	Gene name	$\hat{\beta}_{GE}$	LR p-val	FDR	dbSNP ID	Gene name	$\hat{\beta}_{GE}$	LR p-val	FDR
rs2233679	PIN1	-0.381	0.0463	0.87	rs6347	SLC6A3	0.512	0.0219	0.983
rs2266690	CCNH	-0.376	0.0585	0.87	rs4696480	TLR2	-0.418	0.0255	0.983
rs5923	LCAT	-0.774	0.0711	0.87	rs1126667	ALOX12	-0.356	0.0492	0.983
rs3755557	GSK3B	-0.375	0.0776	0.87	rs1127717	FTHFD	0.442	0.0644	0.983
rs1799802	ERCC4	-0.652	0.101	0.87	rs2855262	SOD3	-0.354	0.0656	0.983
rs3937387	BZRP	0.27	0.144	0.87	rs998074	IGF2R	0.309	0.0947	0.983
rs1051740	EPHX1	0.298	0.153	0.87	rs7921327	AKR1C3	0.327	0.0983	0.983
rs5742926	PMS1	-0.362	0.154	0.87	rs676387	HSD17B1	-0.34	0.1	0.983
rs12801239	KIRREL3	0.272	0.167	0.87	rs998075	IGF2R	0.298	0.107	0.983
rs3448	GPX1	-0.277	0.187	0.87	rs4817027	BIC	-0.341	0.111	0.983

Table 2 displays the ten most significant smoking-SNP interactions computed via the standard two-stage method using either a marginal association filter or a gene-environment correlation filter, as detailed in Section 2.3.2. In this case, the top ten interactions returned by the two filters are completely disjoint. Although the first stage filter reduced the number of smoking-SNP interactions tested via logistic regression models from the 1488 examined by the one-step method to only 103, none of the results appear significant after MHC. In fact, the top uncorrected p-values after filtering, although independently significant, are substantially higher than the top uncorrected p-values found when all SNPs are tested via the one-step approach, indicating that there is only a weak correlation between the two filter statistics and $G \times E$ interaction (as tested using logistic regression models of the form in equation 1) for this data set under the alternate hypothesis of $G \times E$ interaction.

3.3. Proposed $G \times E$ interaction detection method results

Table 3 shows the significant smoking-SNP interactions computed using the proposed method, as detailed in Section 2.3.3. Specifically, the table includes interactions whose Wald test p-

Table 3. Smoking-SNP interactions computed via the proposed G×E detection method with Wald test p-values for the estimated interaction coefficients in the test stage multiple logistic regression model below 0.05. LR test p-values are the Bonferroni-corrected p-values from a likelihood ratio test comparing a test stage model without interaction terms to a model with interaction terms.

Marginal association filter Corrected LR p-value: 0.174					G-E correlation filter Corrected LR p-value: 0.022				
dbSNP ID	Gene name	$\hat{\beta}_{GE}$	Wald p-val	FDR	dbSNP ID	Gene name	$\hat{\beta}_{GE}$	Wald p-val	FDR
rs2233679	PIN1	-1.69	0.00102	0.108	rs3213223	IGF2	-1.21	0.00103	0.0585
rs26279	MSH3	1.49	0.0021	0.111	rs2266690	CCNH	-1.06	0.00118	0.0585
rs1584415	<i>na</i>	-0.978	0.00703	0.18	rs861539	XRCC3	0.906	0.00169	0.0585
rs9642880	CASC11	-1.08	0.0095	0.18	rs1381841	GSK3B	-1.26	0.00238	0.062
rs698090	MASP1	0.969	0.0111	0.18	rs4696480	TLR2	-0.596	0.018	0.294
rs8173	STK6	-1.12	0.0112	0.18	rs2877796	RGS6	-0.598	0.0194	0.294
rs4986765	BRIP1	-1.01	0.0119	0.18	rs113515	TSPO	0.592	0.0198	0.294
rs2676530	HSD17B1	-0.883	0.0176	0.233	rs7921327	AKR1C3	0.616	0.0234	0.304
rs760589	OPRD1	-0.872	0.0224	0.253	rs4619	IGFBP1	-0.553	0.03	0.337
rs4988340	BRIP1	-0.846	0.0278	0.253	rs869975	GPX3	-1.28	0.0324	0.337
rs3740066	ABCC2	0.744	0.0284	0.253	rs1059519	GDF15	0.636	0.0378	0.343
rs13167280	TERT	-1.12	0.0287	0.253	rs2233679	PIN1	-0.543	0.0403	0.343
rs8037	KRT23	0.763	0.0351	0.286	rs1126667	ALOX12	-0.498	0.0458	0.343
rs3847862	CELA1	-0.699	0.0465	0.324	rs6347	SLC6A3	0.625	0.0465	0.343
rs1650697	MSH3	-0.944	0.0499	0.324	rs872072	TEP1	0.5	0.0495	0.343

values for the estimated interaction coefficients in the test stage model specified in equation 7 are below 0.05. The corresponding FDR value was computed for the family of Wald tests on all interaction coefficients. The second stage model was fit for the genetic markers generated using both the marginal association filter (equation 4) and the gene-environment correlation filter (equation 6) screening stage models. Similar to the results from the standard two-stage method, the marginal association and gene-environment correlation filters generate largely independent sets of smoking-SNP interactions. In both cases, the presence of smoking-SNP interactions in the test stage model was assessed using a LR test comparing the likelihood of the model without interaction terms to the likelihood of the model with interaction terms. Because LR tests were performed for both test stage models, a Bonferroni correction was applied to each LR p-value. After MHC, only the LR test for the model fit using the 103 SNPs output by the gene-environment correlation filter was significant (adjusted p-value=0.022). To measure the quality of the SNPs in this significant second stage model, a Hardy Weinberg test of equilibrium was performed among the controls, resulting in an average test p-value of 0.48.

Further investigation of the most significant smoking-SNP interactions from the test stage model for the gene-environment correlation filter revealed several SNPs with prior evidence of association with bladder cancer and/or smoking in independent populations, most notably, a confirmed interaction between smoking and cyclin H (CCNH).²³ SNPs in Toll-like receptor 2 (TLR2) increased overall bladder cancer risk, however, the smoking interaction was not statistically significant in this smaller study.²⁴ Variation in the regulator of G-protein signaling 6 (RGS6) reduced bladder cancer risk, with suggestion of an interaction with smoking.²⁵ The AKR1C3 association is consistent across several studies^{26,27} with a potential relationship with smoking.²⁸ Likewise, our TEP1 bladder cancer association,²⁷ was independently confirmed.²⁹

An interaction with smoking may explain some of the heterogeneity observed among prior studies of the X-ray repair complementing defective in Chinese hamster 3 (XRCC3) SNP.³⁰ SLC6A3 variations lead to stress-induced cigarette craving.³¹ While data on SNP associations are lacking, growth differentiation factor 15 (GDF15) is being promoted as a biomarker of urothelial cell cancer.³² Insulin growth factor 2 (IGF2) is over-expressed in bladder tumors.³³

4. Discussion

In this paper, we have detailed a novel approach for G×E interaction detection that combines elements of screening-testing methods with hierarchical penalized regression. Similar to existing screening-testing techniques, our approach first filters all measured genetic markers according to a filter statistic that is independent from the G×E test statistic under H_0 , and, for the markers that pass the filter, performs a G×E interaction test. The key difference between our approach and existing two-stage methods lies in the structure of the screening and test stage models and the associated statistical G×E interaction tests. Whereas standard two-stage methods fit a separate logistic regression model for each potential G×E interaction in both the screening and testing stages, our method jointly evaluates all markers in a single multiple logistic regression model during both the screening and test stages. Because the number of measured markers is typically much larger than the number of subjects, the screening stage model must be fit using penalization and our approach employs an elastic net penalty that combines L1 and L2 penalty terms.¹⁵ The use of penalized multiple logistic regression enables either the marginal association filter statistic⁴ or the gene-environmental correlation filter statistic⁵ to be jointly computed for all markers, and, because LASSO-penalization tends to retain only one predictor from a set of correlated predictors,³⁴ the set of terms with non-zero coefficients will contain few significant collinearities. Generating a fairly small set of high-quality candidate markers in the screening stage that is free from collinearities is critical when attempting to fit a single unpenalized multiple logistic regression for these markers in the test stage. Assessing G×E interactions in the test stage using a single multiple logistic regression model, as opposed to separate models for each interaction, has two major benefits. First, estimating coefficients jointly decreases the bias associated with omitted predictors in regression. Second, and most importantly, fitting a single model for all markers that pass the screening stage enables the use of a single omnibus test to assess whether any statistically significant G×E interactions exist. Use of just one statistical test completely eliminates the penalty of MHC on power for basic G×E interaction detection. If the number of markers kept after screening is relatively small and the filter statistic correctly retains those markers with high likelihood of being in a G×E interaction, it is quite reasonable to limit inference to a single omnibus test. Wald test p-values and effect size estimates are then used to prioritize the interactions for further investigation and experimental validation. In situations with sample size constraints or poor data quality, a single omnibus test on a filtered set of markers may in fact be the only adequately powered test of G×E interactions.

The benefits of our proposed method relative to standard approaches are clearly demonstrated by the analysis of the bladder cancer data set for smoking-SNP interactions. Neither the one-step nor the standard two-stage methods were able to find any statistically significant

smoking-SNP interactions after MHC. The inability of the one-step and two-stage methods to identify significant interactions mirrors the results from other investigations into smoking-SNP interactions relative to bladder cancer, such as the recent study by Figueroa et al.³⁵ that failed to find significant additive or multiplicative interactions after MHC using a one-step analysis. Our proposed method, on the other hand, successfully found evidence of statistically significant interactions when using the gene-environment correlation filter, as evidenced by the corrected LR test p-value of 0.022 and multiple interactions coefficients with Wald test FDR values below 0.1. The significant interactions identified in this model have not been previously discovered via statistical G×E interaction tests using this data set. A subsequent investigation of this significant test stage model found biological support in the research literature for many of the most significant smoking-SNP interactions.

Although our approach has important methodological and statistical benefits relative to existing G×E interaction detection methods, there are some key limitations to note. First, interpretation of interaction coefficients may be more difficult using a joint model than when using separate models per interaction. Second, the use of an omnibus test just indicates that at least one of the G×E interactions is significant, it does not specify which interaction; unless MHC is applied to the Wald test p-values, these can be only be used for qualitative prioritization and not as strict measures of statistical significance. Finally, the evaluation detailed in this paper was for a data set with a small number of markers; it will be important to assess how well the method scales to genomic data sets measuring upwards of one million markers. For such large data sets, the computational complexity of the elastic net implementation may be a key constraint. In future work, it will be important to test our approach on a diverse collection of GWAS data sets for a range of different environmental exposures and outcome variables.

Acknowledgement

Funding: National Institutes of Health R01 grants LM010098, LM011360, LM009012, EY022300, GM103506 and GM103534. Conflict of Interest: None declared.

References

1. D. J. Hunter, *Nat Rev Genet* **6**, 287 (Apr 2005).
2. A. Ziegler and I. R. König, *A statistical approach to genetic epidemiology*, 2nd edn. (Wiley-VCH, Weinheim, 2010).
3. D. Wahlsten, *Behavioral and Brain Sciences* **13**, 109 (Mar 1990).
4. C. Kooperberg and M. Leblanc, *Genet Epidemiol* **32**, 255 (Apr 2008).
5. C. E. Murcray, J. P. Lewinger and W. J. Gauderman, *Am J Epidemiol* **169**, 219 (Jan 2009).
6. C. E. Murcray, J. P. Lewinger, D. V. Conti, D. C. Thomas and W. J. Gauderman, *Genet Epidemiol* **35**, 201 (Apr 2011).
7. J. Y. Dai, C. Kooperberg, M. Leblanc and R. L. Prentice, *Biometrika* **99**, 929 (Dec 2012).
8. L. Hsu, S. Jiao, J. Y. Dai, C. Hutter, U. Peters and C. Kooperberg, *Genet Epidemiol* **36**, 183 (Apr 2012).
9. J. Millstein, *Front Genet* **4**, p. 306 (2013).
10. R. Bourgon, R. Gentleman and W. Huber, *Proc Natl Acad Sci U S A* **107**, 9546 (May 2010).
11. J. Bien, J. Taylor and R. Tibshirani, *The Annals of Statistics* **41**, 1111 (2013).

12. J. Liu, J. Huang, Y. Zhang, Q. Lan, N. Rothman, T. Zheng and S. Ma, *Genomics* **102**, 189 (Oct 2013).
13. R. Zhu, H. Zhao and S. Ma, *Genet Epidemiol* **38**, 353 (May 2014).
14. T. T. Wu, Y. F. Chen, T. Hastie, E. Sobel and K. Lange, *Bioinformatics* **25**, 714 (Mar 2009).
15. H. Zou and T. Hastie, *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **67**, 301 (2005).
16. J. H. Friedman, T. Hastie and R. Tibshirani, *Journal of Statistical Software* **33**, 1 (Feb 2010).
17. M. R. Karagas, T. D. Tosteson, J. Blum, J. S. Morris, J. A. Baron and B. Klaue, *Environ Health Perspect* **106 Suppl 4**, 1047 (Aug 1998).
18. A. S. Andrew, J. Gui, T. Hu, A. Wyszynski, C. J. Marsit, K. T. Kelsey, A. R. Schned, S. A. Tanyos, E. M. Pendleton, R. M. Ekstrom, Z. Li, M. S. Zens, M. Borsuk, J. H. Moore and M. R. Karagas, *BJU International* , n/a (2014).
19. A. S. Andrew, T. Hu, J. Gu, J. Gui, Y. Ye, C. J. Marsit, K. T. Kelsey, A. R. Schned, S. A. Tanyos, E. M. Pendleton, R. A. Mason, E. V. Morlock, M. S. Zens, Z. Li, J. H. Moore, X. Wu and M. R. Karagas, *PLoS One* **7**, p. e51301 (2012).
20. A. S. Andrew, H. H. Nelson, K. T. Kelsey, J. H. Moore, A. C. Meng, D. P. Casella, T. D. Tosteson, A. R. Schned and M. R. Karagas, *Carcinogenesis* **27**, 1030 (May 2006).
21. Y. Benjamini and Y. Hochberg, *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* , 289 (1995).
22. E. Vittinghoff and C. E. McCulloch, *Am J Epidemiol* **165**, 710 (Mar 2007).
23. M. Chen, A. M. Kamat, M. Huang, H. B. Grossman, C. P. Dinney, S. P. Lerner, X. Wu and J. Gu, *Carcinogenesis* **28**, 2160 (Oct 2007).
24. V. Singh, N. Srivastava, R. Kapoor and R. D. Mittal, *Arch Med Res* **44**, 54 (Jan 2013).
25. D. M. Berman, Y. Wang, Z. Liu, Q. Dong, L.-A. Burke, L. A. Liotta, R. Fisher and X. Wu, *Cancer Res* **64**, 6820 (Sep 2004).
26. J. D. Figueroa, N. Malats, M. García-Closas, F. X. Real, D. Silverman, M. Kogevinas, S. Chanock, R. Welch, M. Dosemeci, Q. Lan, A. Tardón, C. Serra, A. Carrato, R. García-Closas, G. Castaño-Vinyals and N. Rothman, *Carcinogenesis* **29**, 1955 (Oct 2008).
27. A. S. Andrew, J. Gui, A. C. Sanderson, R. A. Mason, E. V. Morlock, A. R. Schned, K. T. Kelsey, C. J. Marsit, J. H. Moore and M. R. Karagas, *Hum Genet* **125**, 527 (Jun 2009).
28. T. Hu, Q. Pan, A. S. Andrew, J. M. Langer, M. D. Cole, C. R. Tomlinson, M. R. Karagas and J. H. Moore, *BioData Min* **7**, p. 5 (2014).
29. J. Chang, C. P. Dinney, M. Huang, X. Wu and J. Gu, *PLoS One* **7**, p. e30665 (2012).
30. Q. Ma, Y. Zhao, S. Wang, X. Zhang, J. Zhang, M. Du, L. Li and Y. Zhang, *Tumour Biol* **35**, 1473 (Feb 2014).
31. J. Erblich, C. Lerman, D. W. Self, G. A. Diaz and D. H. Bovbjerg, *Pharmacogenomics J* **4**, 102 (2004).
32. V. L. Costa, R. Henrique, S. A. Danielsen, S. Duarte-Pereira, M. Eknaes, R. I. Skotheim, A. Rodrigues, J. S. Magalhães, J. Oliveira, R. A. Lotte, M. R. Teixeira, C. Jerónimo and G. E. Lind, *Clin Cancer Res* **16**, 5842 (Dec 2010).
33. G. Pignot, A. Vieillefond, S. Vacher, M. Zerbib, B. Debre, R. Lidereau, D. Amsellem-Ouazana and I. Bieche, *Br J Cancer* **106**, 1177 (Mar 2012).
34. R. Tibshirani, *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **73**, 273 (2011).
35. J. D. Figueroa *et al.*, *Carcinogenesis* **35**, 1737 (Aug 2014).

VARIABLE SELECTION METHOD FOR THE IDENTIFICATION OF EPISTATIC MODELS

EMILY ROSE HOLZINGER[†]

*Computational and Statistical Genomics Branch (NHGRI, NIH)
Baltimore, MD 21224, USA
Email: emily.holzinger@nih.gov*

SILKE SZYMCZAK

*Computational and Statistical Genomics Branch (NHGRI, NIH)
Baltimore, MD 21224, USA
Email: s.szymczak@ikmb.uni-kiel.de*

ABHIJIT DASGUPTA

*Clinical Trials and Outcomes Branch (NIAMS, NIH)
Bethesda, MD 20892-1468, USA
Email: dasgupab@mail.nih.gov*

JAMES MALLEY

*Center for Information Technology (NIH)
Bethesda, MD 20892-1468, USA
Email: jmalley@mail.nih.gov*

QING LI

*Computational and Statistical Genomics Branch (NHGRI, NIH)
Baltimore, MD 21224, USA
Email: liq4@mail.nih.gov*

JOAN E. BAILEY-WILSON

*Computational and Statistical Genomics Branch (NHGRI, NIH)
Baltimore, MD 21224, USA
Email: jebw@mail.nih.gov*

[†] This work was supported in part by the Intramural Research Programs of the National Human Genome Research Institute, the National Institute for Arthritis, Musculoskeletal and Skin Disorders, and the Center for Information Technology, all part of the National Institutes of Health.

Standard analysis methods for genome wide association studies (GWAS) are not robust to complex disease models, such as interactions between variables with small main effects. These types of effects likely contribute to the heritability of complex human traits. Machine learning methods that are capable of identifying interactions, such as Random Forests (RF), are an alternative analysis approach. One caveat to RF is that there is no standardized method of selecting variables so that false positives are reduced while retaining adequate power. To this end, we have developed a novel variable selection method called *relative recurrency variable importance metric* (r2VIM). This method incorporates recurrency and variance estimation to assist in optimal threshold selection. For this study, we specifically address how this method performs in data with almost completely epistatic effects (i.e. no marginal effects). Our results show that with appropriate parameter settings, r2VIM can identify interaction effects when the marginal effects are virtually nonexistent. It also outperforms logistic regression, which has essentially no power under this type of model when the number of potential features (genetic variants) is large. (All Supplementary Data can be found here: http://research.nhgri.nih.gov/manuscripts/Bailey-Wilson/r2VIM_epi/).

1. Introduction

1.1. Variable selection that allows for interactions

Thousands of variants have been identified that are associated with complex human traits [1]. However, a large portion of the estimated heritability remains unexplained for many traits [2]. Additionally, these variants often do not improve prediction of complex traits in independent data sets over metrics that are relatively easier to collect (e.g. age, sex, body mass index, family history)[3]. This is likely due, in part, to overly simplistic study designs and modeling methods. The complex nature of biological pathways makes it unlikely that additive main effects explain all of the heritability. Empirical observations in animal model studies show that complex effects are actually pervasive in nature [4]. The identification of these effects would require variant discovery and modeling methods that are robust to interactions, even when main effects are very small or non-existent.

The first step in solving this problem is to separate true signal from noise. Machine learning methods are promising candidates for this task and are currently used in other scientific fields, including drug design [5]. One type of machine learning method is Random Forests (RF) [6]. One limitation of RF is that no standard method exists for selecting a set of “associated” variants with low levels of false positives and adequate power. Parametric analyses produce statistics with generally accepted values for error rates, assuming the parametric model has been exactly and correctly specified. One way to obtain equivalent values is to generate empirical distributions by running thousands of permutation analyses. This is computationally impractical for studies that use high-throughput genomic data, which usually consists of thousands to millions of variables. We propose a more efficient method called r2VIM, which integrates different selection parameters to identify the appropriate threshold between signal and noise [7]. The ultimate goal of this

method is to generate variant sets that include main and interaction effects. These sets can then be assessed using modeling tools for interpretation and prediction purposes to further our understanding of complex human traits.

2. Methods

2.1. r2VIM

The method r2VIM uses a novel variable selection algorithm based on RF results. RF generates a collection of regression (quantitative outcome) or classification (categorical outcome) trees. In RF, bootstrap samples are drawn to train tree models, and the performance of the trained model is evaluated by testing the tree on the “out-of-bag” (OOB) sample, i.e., the observations not included in the bootstrap sample used for training. This process is repeated over many bootstrap samples and the optimal RF is based on evaluating performance across all the OOB samples. This process reduces the likelihood of overfitting as the model is optimized based on OOB data and not the training data [8]. The VIM is calculated as the difference of an error metric before and after random variable permutation. Variables that result in greater error due to permutation have higher VIMs and are considered more important for prediction purposes. While methods do exist for interpreting the VIM, there is no gold standard method for determining the threshold that best differentiates between noise and functional variables. The random nature of the algorithm can result in variables with high VIMs in one run and low VIMs in another with only a different random seed. To address this, we combine recurrency and a threshold optimization procedure, as described below and illustrated in Supp. Figure 1:

1. *Permutation-based importance score*: Unscaled permutation is used based on previous studies that found this VIM estimation method to be the most reliable [9].
2. *Estimate of null variance*: Assuming that the smallest VIM is negative, the absolute value of the difference between the smallest VIM and zero should be a reliable estimate of VIM variance in null data, as variables with no effect should be symmetrically and randomly distributed around zero [10]. Variables with VIMs greater than the estimated null variance are less likely to be noise variables. In preliminary analyses, we observed that this estimate may be too conservative or liberal for data with different effect types. Therefore, we use the distribution of VIMs to guide threshold selection for this analysis [7].
3. *Recurrency*: Due to the randomness of RF, variables that are deemed important in one run may be declared not important in a second run having only a different random number seed. Variables with relatively high VIMs across many runs are more likely to be true signals. The reasoning here is that stronger predictors will have a higher probability

of being in a top VIM list, and this rate of *recurrency* for a predictor is an estimate of this probability. For this analysis, we run RF five times for each of the 100 simulated datasets to assess false positive and true positive selection at various thresholds. We calculate relative importance score (RIS), which is the VIM divided by the variance estimate. This allows us to compare VIMs across the five runs, and it allows us to select a more appropriate threshold based on the RIS distribution that results from the five runs. We use the median or minimum of the RIS values from the five runs as a “recurrency-corrected” metric. For summarization purposes, we report the median value of this metric for all 100 datasets from each simulation model, unless stated otherwise

2.2. Data Simulation

A previous study has assessed r2VIM using simulated data with only main effects [7]. r2VIM performed comparably to linear regression in terms of power and false positive rate. Our study specifically addresses the performance of r2VIM in the presence of interactions with no main effects.

We simulated data sets using genomeSIMLA [11], [12]. We simulated four different types of models, which had either 100 or 1000 total SNPs, with and without correlation between the SNPs (i.e. linkage disequilibrium, or LD). For each model, 100 data set replicates were generated. The genetic effect for the four models consists of two SNPs with an interaction effect and no marginal effects. This genetic effect was used to generate data sets with a penetrance table. This table provides the probability of being a case for each of the nine genotype combinations. The model was generated using the simpen algorithm in genomeSIMLA. The minor allele frequency for both SNPs is 0.4. The target heritability and odds ratio of the effect model are 0.10 and 2.0, respectively. The marginal effects for the genotypes at each locus are all very close to 0. The penetrance values for the genetic models are shown in Supp. Table 1. The outcome is binary (case/control status) with 250 cases/250 controls in the 100 SNP data and 500 cases/500 controls in the 1000 SNP data. Datasets that included LD were generated using forward time population

Table 1. RF parameters for each of the simulated data analyses

Model	RF Parameter Values
100 SNPs (no LD and LD)	mtry = 20, 40 ntree = 200, 600 nodesize = 10
1000 SNPs (no LD and LD)	mtry = 300, 400 ntree = 2000, 6000 nodesize = 100

simulation, as previously described [11]. LD models were selected that had moderate correlation overall, but virtually no correlation between the two functional SNPs. LD plots showing these correlation patterns are shown in Supp. Figure 2.

3. Results

3.1. Simulated Data

We ran r2VIM on the four simulation models specifying five runs of RF per simulated dataset and obtaining the RIS scores for each run and the median and minimum RIS for all five runs for each dataset. We report the median value of the median or minimum RIS for the 100 datasets. We also calculated the detection power (or rate), which is defined as the number of times in the 100 datasets that the median RIS or minimum RIS exceeded a set threshold. We ran RF with different combinations of variable subset sizes (*mtry*) and number of trees (*ntree*), as these have the largest effect on performance. The minimum number of samples allowed in a terminal node, called terminal node size (*nodesize*), is also important, and

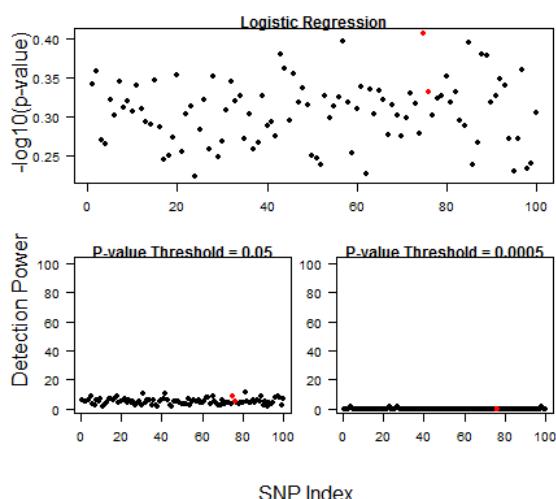


Fig. 1. Results for univariate logistic regression analysis of 100 SNP data with no LD. The top plot shows the median $-\log_{10}$ (p-value) for each SNP across 100 dataset replicates. The bottom two plots show the number of times out of 100 datasets that the p-value was smaller than the specified significance threshold for each SNP. Functional SNPs 75 and 76 are shown in red.

varied across models according to sample size. Table 1 shows the RF parameter values applied. Other parameter settings were consistent across runs. We ran univariate logistic regression for comparison.

Figure 1 shows the results for the logistic regression analysis on the 100 datasets for the 100 SNP data with no LD. We report the median p-value for each SNP across all datasets. As expected, the lack of marginal effects in the simulated model results in virtually no power, even at very liberal selection thresholds. Even if an exhaustive search for all possible two-way interactions was performed, the multiple testing correction would hinder the identification of most true models. Moreover, an ideal analysis would recode the SNPs genotypically, which doubles the number of variables and makes the correction even more stringent.

The results for the 100 SNP data with no LD for the r2VIM analysis are shown in Figure 2. The r2VIM analyses took ~13 seconds per dataset to complete. The median RIS for the analysis with $mtry=40$ and $ntree=600$ is shown. Results for all parameter settings can be found in Supp. Figures 3-8. The functional SNPs (75 and 76) are highlighted in red. Note that the positions here are not relevant as all of the SNPs were simulated to be independent. Again, RIS is calculated as raw VIM / variance estimate for that dataset. This allows us to compare importance scores across datasets.

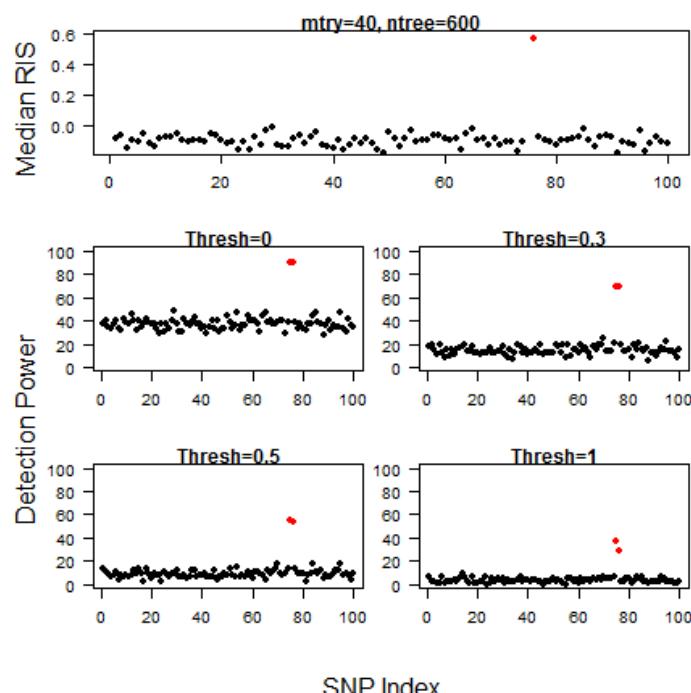


Fig. 2. Results for r2VIM analysis of datasets with 100 SNPs and no LD for $mtry=40$ and $ntree=600$. The top plot shows the median RIS for each SNP across 100 dataset replicates. The bottom plots show the number of times out of 100 replicates that the median RIS was smaller than the specified significance threshold for each SNP. Functional SNPs 75 and 76 are shown in red.

We show the number of times each variable was selected at four different selection threshold levels. The thresholds (0, 0.3, 0.5, and 1) were chosen based on the distribution of median RIS values. The results suggest that the median of RIS values produces less false positives than the minimum RIS in this data (Supp. Figures 3, 4, 6, and 7). For this genetic model, a factor threshold of 0.3 optimizes both detection power and false positive rate when applied to the median RIS.

Next, we assessed the effect of LD on r2VIM performance. LD is an important characteristic of genetic data that can have a large effect on power and false positive rate [13]. Figure 3 shows the distribution of median RIS values for all 100 datasets with moderate LD. The functional SNP positions for these simulations (4 and 26) are relevant, as they were selected to be nearly uncorrelated. The detection power thresholds are higher here, as the RIS scores were much higher than those for the data with no LD.

Interestingly, LD increases detection power and the RIS scale for the functional and non-functional SNPs. VIM inflation with variable correlation has been observed before in RF [9]. This feature should be considered when performing SNP filtering based on pairwise LD measures. Of note, LD results in higher detection rates for functional *and* non-functional SNPs due to inflated RIS values and could result in more false positives. However, for this small number of SNPs, many of the nearby non-functional SNPs are in high LD with at least one of the functional SNPs and are not true “false positives” but instead are the result of RF detecting association of a “chromosomal region” with the trait.

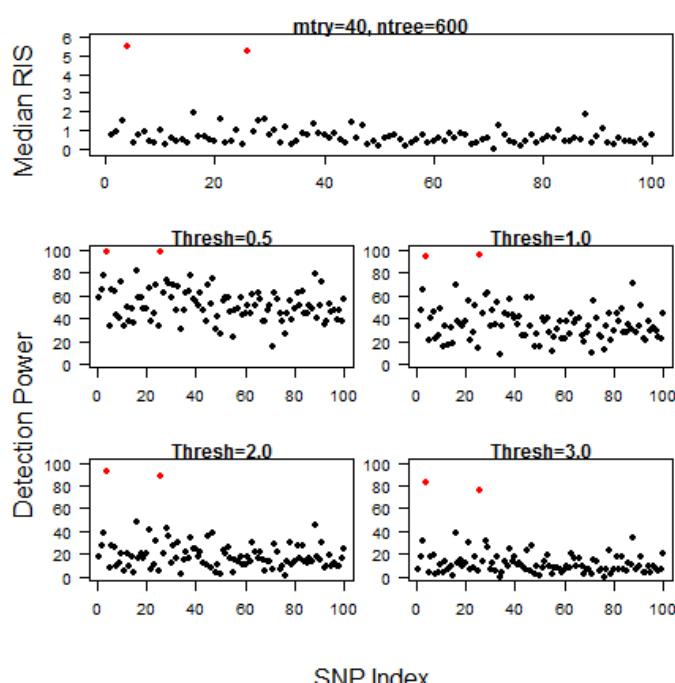


Fig. 3. Results for r2VIM analysis of datasets with 100 SNPs and LD for $mtry=40$ and $ntree=600$. The top plot shows the median RIS for each SNP across 100 dataset replicates. The bottom plots show the number of times out of 100 replicates that the median RIS was smaller than the specified significance threshold for each SNP. Functional SNPs 4 and 26 are shown in red.

To determine the effect of added noise, we simulated data sets with 1000 SNPs (998 non-functional and 2 functional). Each analysis took ~40 seconds per dataset to complete. This is still not near the scale of a typical GWAS analysis; however, it is impractical to run r2VIM on 100 GWAS-sized datasets many times. Figures 4 and 5 show the results for the 1000 SNP analyses without and with LD, respectively. When no LD was present, detection power was lower than the 100 SNP data for the RIS thresholds shown; however, the median RIS values still differentiate between the non-functional and functional SNPs. With LD, we observe the same increase in RIS values and detection power for all SNPs. This is even more pronounced in the 1000 SNP data.

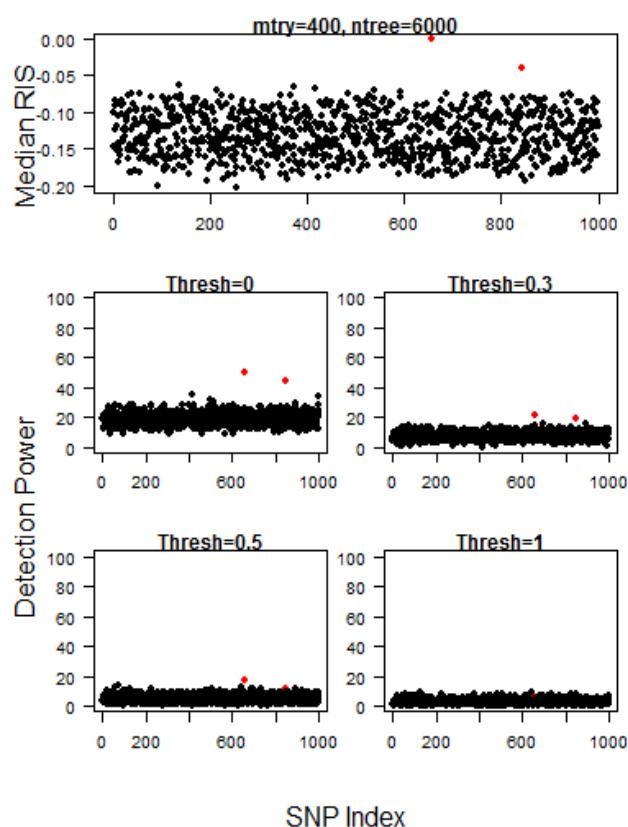


Fig. 4. Results for r2VIM analysis of datasets with 1000 SNPs and no LD for $mtry=400$ and $ntree=6,000$. The top plot shows the median RIS for each SNP across 100 dataset replicates. The bottom plots show the number of times out of 100 replicates that the median RIS was smaller than the specified significance threshold for each SNP. Functional SNPs 657 and 844 are shown in red.

Also, the detection power was higher with a larger $ntree$ (Supp. Figures 9-12). This emphasizes the importance of using the correct parameter settings and selection criteria for data with a high noise to signal ratio.

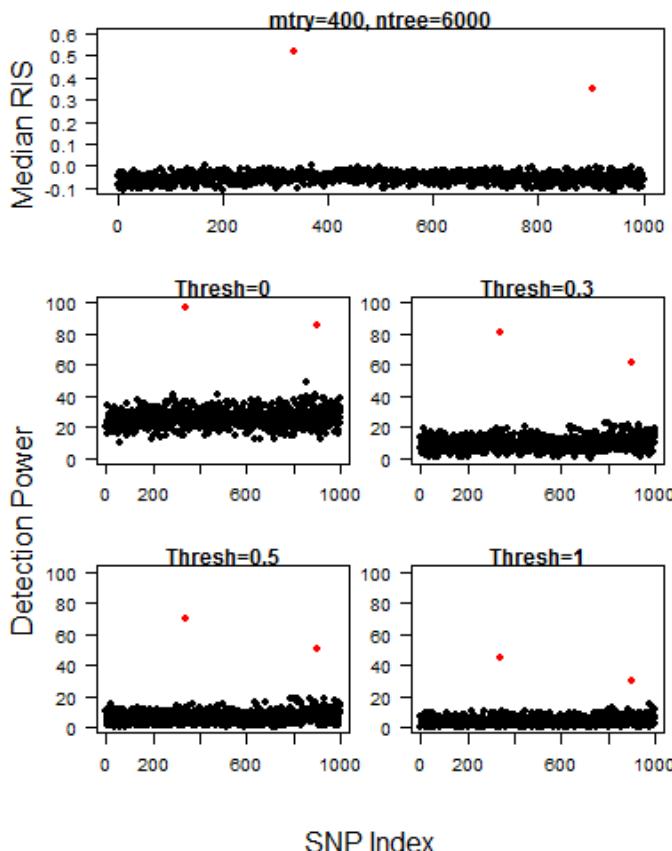


Fig. 5. Results for r2VIM analysis of datasets with 1000 SNPs and LD for $mtry=400$ and $ntree=6,000$. The top plot shows the median RIS for each SNP across 100 dataset replicates. The bottom plots show the number of times out of 100 replicates that the median RIS was smaller than the specified significance threshold for each SNP. The functional SNPs 336 and 903 are shown in red.

Recurrency requires more computational resources than a single-run RF analysis, so we assessed the level of performance gain due to recurrency by comparing false positive detection at the RIS threshold of 0.3 for all models (Figure 6). To compare 100 SNP data with 1000 SNP data we divided the number of FPs selected by the total variable count. For the five single runs and the recurrency-corrected runs (median and minimum), both functional SNPs were identified at this threshold. For a single data set, the minimum RIS value reduces false positive selection over all other RIS values. This is in contrast to the summary data analyses where the median RIS value appeared to be optimal (Supp. Figures 2-11). This could be a factor of the summarization itself, as this essentially adds another layer of recurrency to the results. In applied analyses, it will be important to plot both the minimum and median RIS values to assess the distributions of each.

4. Discussion

One of the biggest hurdles in performing a successful analysis of high-throughput data is selecting variables least likely to be noise. The most commonly used methods thus far often

take the results from a univariate analysis to identify predictor variables with some level of marginal effects. This would not be appropriate if interaction effects exist with little or no marginal effects. Our method addresses this by performing relatively fast variable selection that can identify interaction effects when marginal effects are virtually nonexistent.

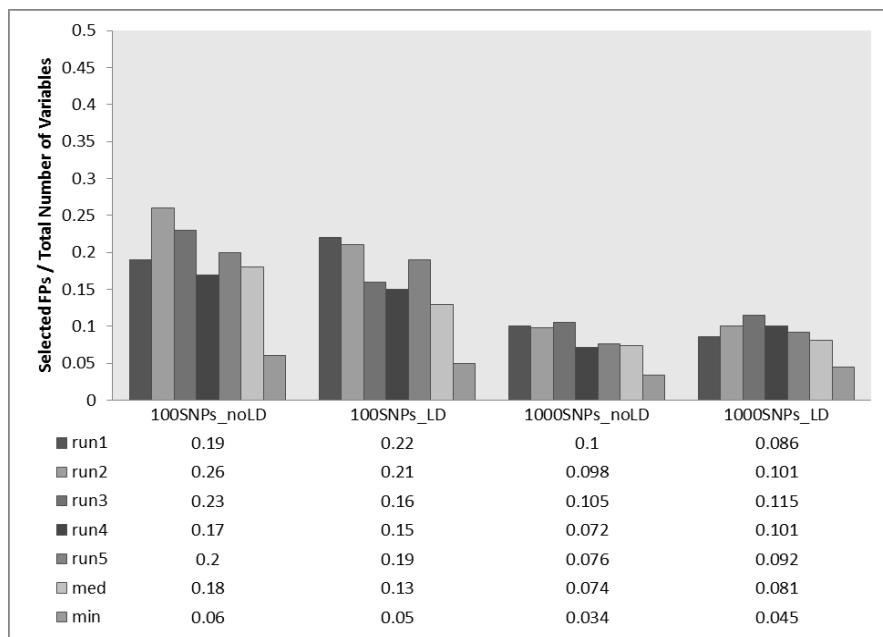


Fig. 6. False positives selected at an RIS threshold of 0.3 for a single dataset for each of the simulation models. The number shown is the number of FPs selected divided by the total number of variables in dataset for comparison purposes.

Importantly, there are limitations to this method. For example, RF does not allow for any missing data. Missingness is very common in high-throughput data; therefore procedures such as imputation or complete removal of variables with missing data are required to run RF. Additionally, this paper only tests data sets with a binary (case/control) outcome using classification trees. A quantitative outcome would require regression trees, and more testing needs to be done to determine how this affects performance. Although this method supplies a more regimented procedure for RF threshold selection, it is still highly dependent on many factors. For example, the minimum RIS was optimal for certain analyses, while the median was preferable for others. Currently, RF variable selection requires exploring various aspects of the results to determine the best selection method. Finally, sequence data is becoming increasingly available for research. To this end, it will be necessary to optimize r2VIM for rare variant detection, as well as quantitative high-throughput data, such as RNA sequence levels.

For a fair comparison to other methods, tools designed for interaction identification should be assessed on the same data. However, this is not a trivial task. An exhaustive interaction analysis

using a regression model is also underpowered to identify highly epistatic models if the SNPs are coded in an allelic, or additive, manner as they were for the RF analysis. To quickly illustrate this, we tested the correct SNP pair and nine random SNP pairs using logistic regression ($y=snp1 +.snp2 +.snp1*snp2$) in a 100 SNP / no LD simulated dataset. The true model interaction term had a p-value of 0.24, which is not borderline significant even before multiple testing correction. Additionally, three of the nine random SNP pair analyses had interaction p-values lower than the correct model. The logistic regression model would have more power to find the interactions if they are coded in a genotypic manner due to the simulation design. However, this encoding would require doubling the number of variables that need to be exhaustively tested. RF, on the other hand, often identifies the model without re-coding the SNPs. Further tests will involve recoding the SNPs genetically and testing in RF, logistic regression, and other methods.

The impact of adding variables with main effects to the interaction model must also be assessed, as biological data is likely to include many different types of effects. Future work will involve simulating different effect types to assess the impact on RIS distribution and detection power.

After variable selection, the next step is to model the subset for interpretation and prediction purposes. This requires a method robust to interaction and marginal effects. Machine learning methods are also an attractive candidate for this step [14], [15]. It will be important to recode the data so that genotypic effects can be seen, especially for possible interactions. This could also be done at the selection step; however, as it doubles the number of variables in the dataset, it is not an ideal procedure in data with already high levels of noise. Fortunately, r2VIM appears to be able to identify non-additive interaction effects even with the standard additive encoding. After selection, however, it is more computationally feasible to recode the subset of candidate SNPs to generate more informative prediction models. It is also useful to note that the two methods proposed here (noise detection by removal of SNPs with negative variable importance measures and recurrency), could be used with many other machine learning schemes, such as neural nets, boosting and support vector machines.

The ultimate goal of r2VIM is to provide a tool that can perform powerful selection while taking into account main and interaction effects. Non-linear interactions are especially difficult to identify unless specific tools robust to these effects are used. Our results suggest that using proper threshold selection procedures, RF can identify these types of effects even in the extreme situation of virtually no marginal effects.

References

- [1] D. Welter, J. MacArthur, J. Morales, T. Burdett, P. Hall, H. Junkins, A. Klemm, P. Flicek, T. Manolio, L. Hindorff, and H. Parkinson, "The NHGRI GWAS Catalog, a curated resource of SNP-trait associations," *Nucleic Acids Res.*, vol. 42, no. D1, pp. D1001–D1006, Dec. 2013.
- [2] T. A. Manolio, F. S. Collins, N. J. Cox, D. B. Goldstein, L. A. Hindorff, D. J. Hunter, M. I. McCarthy, E. M. Ramos, L. R. Cardon, A. Chakravarti, J. H. Cho, A. E. Guttmacher, A. Kong, L. Kruglyak, E. Mardis,

- C. N. Rotimi, M. Slatkin, D. Valle, A. S. Whittemore, M. Boehnke, A. G. Clark, E. E. Eichler, G. Gibson, J. L. Haines, T. F. C. Mackay, S. A. McCarroll, and P. M. Visscher, "Finding the missing heritability of complex diseases," *Nature*, vol. 461, no. 7265, pp. 747–753, Oct. 2009.
- [3] B. Godman, A. E. Finlayson, P. K. Cheema, E. Zebedin-Brandl, I. Gutiérrez-Ibarluzea, J. Jones, R. E. Malmström, E. Asola, C. Baumgärtel, M. Bennie, I. Bishop, A. Bucsics, S. Campbell, E. Diogene, A. Ferrario, J. Fürst, K. Garuoliene, M. Gomes, K. Harris, A. Haycox, H. Herholz, K. Hviding, S. Jan, M. Kalaba, C. Kvalheim, O. Laius, S.-A. Lööv, K. Malinowska, A. Martin, L. McCullagh, F. Nilsson, K. Paterson, U. Schwabe, G. Selke, C. Sermet, S. Simoens, D. Tomek, V. Vlahovic-Palcevski, L. Voncina, M. Wladysiuk, M. van Woerkom, D. Wong-Rieger, C. Zara, R. Ali, and L. L. Gustafsson, "Personalizing health care: feasibility and future implications," *BMC Med.*, vol. 11, p. 179, 2013.
 - [4] W. Huang, S. Richards, M. A. Carbone, D. Zhu, R. R. H. Anholt, J. F. Ayroles, L. Duncan, K. W. Jordan, F. Lawrence, M. M. Magwire, C. B. Warner, K. Blankenburg, Y. Han, M. Javaid, J. Jayaseelan, S. N. Jhangiani, D. Muzny, F. Ongeri, L. Perales, Y.-Q. Wu, Y. Zhang, X. Zou, E. A. Stone, R. A. Gibbs, and T. F. C. Mackay, "Epistasis dominates the genetic architecture of *Drosophila* quantitative traits," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 109, no. 39, pp. 15553–15559, Sep. 2012.
 - [5] J. C. Gertrudes, V. G. Maltarollo, R. A. Silva, P. R. Oliveira, K. M. Honório, and A. B. F. da Silva, "Machine learning techniques and drug design," *Curr. Med. Chem.*, vol. 19, no. 25, pp. 4289–4297, 2012.
 - [6] A. Cutler, D. R. Cutler, and J. R. Stevens, "Random Forests," in *Ensemble Machine Learning*, C. Zhang and Y. Ma, Eds. Boston, MA: Springer US, 2012, pp. 157–175.
 - [7] S. Szymczak, E. Holzinger, A. Dasgupta, J. Malley, and J. Bailey-Wilson, "A new variable selection method for random forests in genome-wide association studies," *Bioinformatics*, In Preparation.
 - [8] L. Breiman, "Random Forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.
 - [9] K. K. Nicodemus, J. D. Malley, C. Strobl, and A. Ziegler, "The behaviour of random forest permutation-based variable importance measures under predictor correlation," *BMC Bioinformatics*, vol. 11, no. 1, p. 110, 2010.
 - [10] C. Strobl, J. Malley, and G. Tutz, "An introduction to recursive partitioning: Rationale, application, and characteristics of classification and regression trees, bagging, and random forests," *Psychol. Methods*, vol. 14, no. 4, pp. 323–348, 2009.
 - [11] T. L. Edwards, W. S. Bush, S. D. Turner, S. M. Dudek, E. S. Torstenson, M. Schmidt, E. Martin, and M. D. Ritchie, "Generating Linkage Disequilibrium Patterns in Data Simulations Using genomeSIMLA," *Lect. Notes Comput. Sci.*, vol. 4793, pp. 24–35, 2008.
 - [12] S. M. Dudek, A. A. Motsinger, D. R. Velez, S. M. Williams, and M. D. Ritchie, "Data simulation software for whole-genome association and other studies in human genetics," *Pac.Symp.Biocomput.*, vol. 11, pp. 499–510, 2006.
 - [13] S. A. Tishkoff and B. C. Verrelli, "Role of evolutionary history on haplotype block structure in the human genome: implications for disease mapping," *Curr.Opin.Genet.Dev.*, vol. 13, no. 6, pp. 569–575, Dec. 2003.
 - [14] E. Holzinger, S. M. Dudek, A. T. Frase, R. M. Krauss, M. W. Medina, and M. D. Ritchie, "ATHENA: A Tool for Meta-Dimensional Analysis Applied to Genotypes and Gene Expression Data to Predict HDL Cholesterol Levels," presented at the Pacific Symposium on Biocomputing, 2013, vol. 18, pp. 385–396.
 - [15] A. Dasgupta, S. Szymczak, J. H. Moore, J. E. Bailey-Wilson, and J. D. Malley, "Risk estimation using probability machines," *BioData Min.*, vol. 7, no. 1, p. 2, 2014.

GENOME-WIDE GENETIC INTERACTION ANALYSIS OF GLAUCOMA USING EXPERT KNOWLEDGE DERIVED FROM HUMAN PHENOTYPE NETWORKS

TING HU[†], CHRISTIAN DARABOS[†], MARIA E. CRICCO, EMILY KONG, JASON H. MOORE

*Institute for the Quantitative Biomedical Sciences, Geisel School of Medicine, Dartmouth College
Hanover, NH 03755, U.S.A.*

E-mail: jason.h.moore@dartmouth.edu

The large volume of GWAS data poses great computational challenges for analyzing genetic interactions associated with common human diseases. We propose a computational framework for characterizing epistatic interactions among large sets of genetic attributes in GWAS data. We build the human phenotype network (HPN) and focus around a disease of interest. In this study, we use the GLAUGEN glaucoma GWAS dataset and apply the HPN as a biological knowledge-based filter to prioritize genetic variants. Then, we use the statistical epistasis network (SEN) to identify a significant connected network of pairwise epistatic interactions among the prioritized SNPs. These clearly highlight the complex genetic basis of glaucoma. Furthermore, we identify key SNPs by quantifying structural network characteristics. Through functional annotation of these key SNPs using Biofilter, a software accessing multiple publicly available human genetic data sources, we find supporting biomedical evidences linking glaucoma to an array of genetic diseases, proving our concept. We conclude by suggesting hypotheses for a better understanding of the disease.

Keywords: GWAS; Epistasis; Gene-gene interaction; Human phenotype network; Statistical epistasis network; Biofilter; Eye diseases; Glaucoma; SNPs; Pathways.

1. Introduction

With the rapid development of genotyping technologies and exponential increase in computational power, we are able to leverage the wealth of genome-wide association studies (GWAS) data to test millions of genetic variations (single nucleotide polymorphisms, SNPs) for their associations with common human diseases.^{1,2} However, common disorders are usually the result of the non-linear combined effect of many variations. The study of these complex, epistatic interactions among multiple genetic attributes is crucial in explaining portions of the genotype-to-phenotype associations.^{3,4}

Detecting and analyzing complex genetic interactions in GWAS data pose great statistical and computational challenges. Epistatic effects potentially involve any number of SNPs, from a couple to hundreds or even thousands. In turn, the uncertainty of size of the interacting SNP set would require interaction studies to comprehensively explore all possible combinations of SNPs. A typical GWAS dataset encompasses hundreds of thousands of variations, making the enumeration of all pairwise SNP interactions computationally infeasible, and the computational cost increases exponentially with the number of features at hand. This inevitably becomes a computational bottleneck of processing and analyzing GWAS data.

Therefore, genetic interaction studies meant to process modern GWAS data require the development of advanced informatics methodologies. These novel algorithms must be able to

[†]Co-first authors – TH and CD have contributed equally to this work.

simultaneously analyze large sets of interactions and identify genetic attributes potentially involved in higher-order interactions. Network modeling has emerged as a suitable framework for such purposes,^{5–7} thanks to its ability to represent a large number of entities, as vertices, and their relationships, as edges.

In this study, we propose an informatics framework of detecting and analyzing genetic interactions for GWAS data. We use the example of the GLAUGEN study, a GWAS on glaucoma consisting of over a million attributes. We pre-screen the dataset using a knowledge-based filter by building human phenotype network (HPN) to prioritize SNPs and reduce the search space according to their known relationship to the disease/phenotype in question. This reduced SNP list is then used to quantitatively evaluate all pairwise epistatic interactions and to build statistical epistasis network (SEN) to characterize their global interaction structure. The key SNPs identified by the SEN are functionally annotated and evaluated for their interaction with other disorders. This new framework has the potential to elucidate large parts of the complex genetic architecture of common human diseases, such as glaucoma, and identify key SNPs for further biological validations.

2. Dataset and Methods

2.1. *Glaucoma and the GLAUGEN Study*

Glaucoma, a neurodegenerative disease, is the primary cause of irreversible blindness, affecting over 60 million people worldwide. The most common kind of glaucoma, in all populations, is primary open-angle glaucoma (POAG). Currently the only modifiable risk factor of POAG is intraocular pressure (IOP), but even lowering that will only slow the process, not stop it.⁸ However, a substantial amount of POAG has been shown to have a genetic basis. Familial aggregation of POAG has been long recognized and studied to find multiple loci linked to them, causing the discovery of glaucoma-causing genes myocilin (MYOC), optineurin (OPTN) and WD^a-repeat domain 36.⁹ About 5% of POAG is presently attributed to a single-gene or Mendelian forms of glaucoma. More cases of POAG are caused by the combined epistatic effects of many genetics risk factors.¹⁰

The Glaucoma Gene Environment (GLAUGEN) seeks to illuminate the origin of the disease, to discover genetic loci associated with POAG, and to identify gene-gene and gene-environment associations.⁸ GLAUGEN is a GWAS, case-control study, with about 2,000 unrelated cases and over 2,300 controls. To be involved in the study, both the cases and the controls had to be at least 40 years of age and European derived or Hispanic Caucasian. Subjects were genotyped with 1,048,965 SNPs examined.⁸

2.2. *Human Phenotype Network (HPN)*

The human disease network,¹¹ or the more general human phenotype networks (HPNs),^{12,13} are mathematical graph models where nodes represent human genetic disorders and edges link those nodes with shared biology.¹⁴ The underlying connections of the HPN contribute to

^atryptophan-aspartic acid

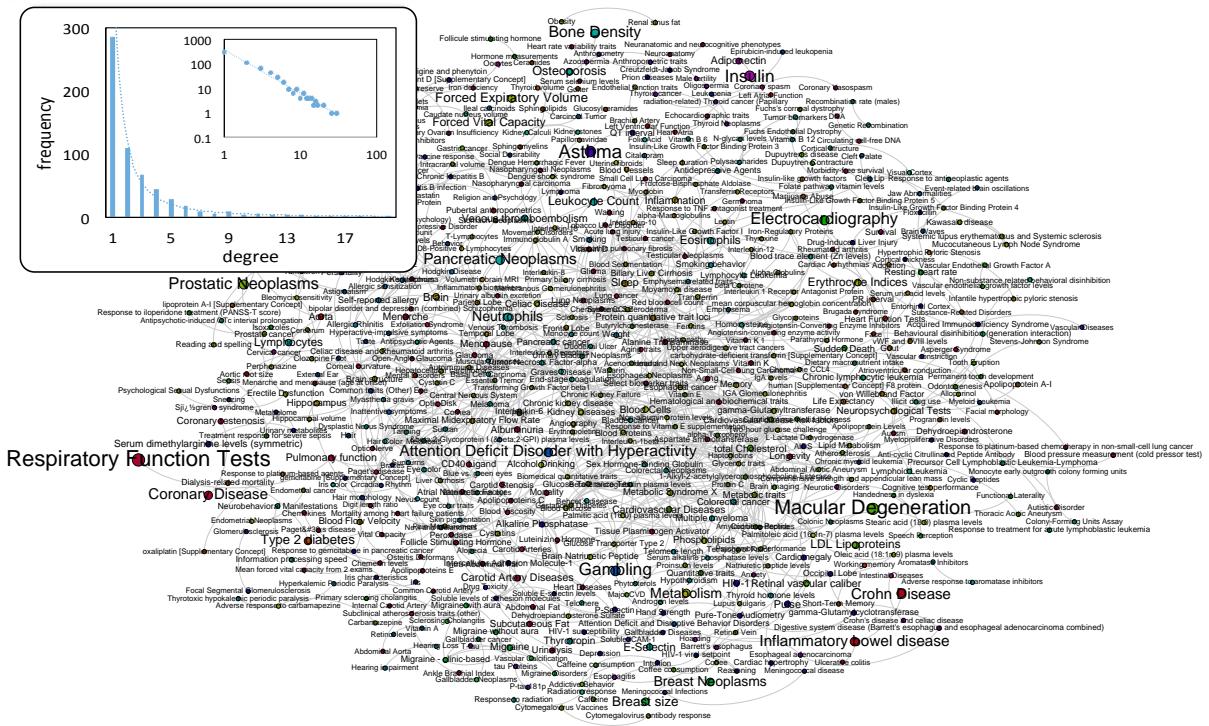


Fig. 1. The SNP-based human phenotype network, filtered (edge weight cutoff = 0.002). Vertices are colored by “modules”¹³ to increase readability. The vertex sizes are proportional to the number of associated SNPs. The degree distribution is given on both linear and logarithmic scales.

the understanding of the basis of disorders, which in turn leads to a better understanding of human diseases.

In the present work, we rely on the SNP-based HPN, where connected diseases share common SNPs. We use two sources for the phenotype-to-SNPs mapping: the National Human Genome Research Institute GWAS catalog,¹⁵ a manually curated list of all the National Institute of Health (NIH) funded GWAS studies, and the NIH’s database of Genotypes and Phenotypes (dbGaP, <http://www.ncbi.nlm.nih.gov/gap>). To build the HPN, all traits listed in the GWAS catalog and dbGaP are annotated with their risk-associated SNPs. All the traits become nodes in our network. Then, traits that have one or more SNPs in common are linked in the HPN by an edge, the weight of which is proportional to the size of the SNPs’ overlap. The GWAS catalog and dbGaP report 1,252 traits combined, annotated with 37,681 SNPs in 16,411 loci. The resulting bipartite network is projected in the space of phenotype vertices to obtain the HPN shown in Fig. 1.

The HPN contains 985 vertices and over 26,000 edges, and encompasses all phenotypes listed in the GWAS catalog and dbGaP, provided that they are connected to at least one other trait. The resulting network is extremely dense, with an average degree greater than 500.

2.3. Statistical Epistasis Network (SEN)

Statistical epistasis network (SEN) is designed to study a global aggregation of pairwise epistatic interactions among a large number of factors.^{16,17} First, pairwise epistatic interaction

is quantified using the information-theoretic measure called *information gain*¹⁸ for all possible pairs of genetic attributes. Specifically, the genotypes of two SNPs A and B , and the discrete phenotypic class C are regarded as variables. *Mutual information* $I(A; C)$ or $I(B; C)$ measures the shared information between (A or B) and C calculated as $I(A; C) = H(A) + H(C) - H(A, C)$ or $I(B; C) = H(B) + H(C) - H(B, C)$, where *entropy* H measures the uncertainty of a random variable or joint multiple random variables. Therefore, mutual information $I(A; C)$ or $I(B; C)$ describes the reduction of the uncertainty of the phenotypic class C given the knowledge of the genotype of A or B , known as the main effect of A or B on C . Similarly, the joint mutual information $I(A, B; C)$ measures the total effect of combining A and B on explaining the phenotype C . Subtracting the individual mutual information $I(A; C)$ and $I(B; C)$ from $I(A, B; C)$ captures the gained information on C by considering A and B together rather than individually. This *information gain* $IG(A; B; C)$ is a practical measure for the epistatic interaction effect of SNPs A and B on phenotype C , and has been used as an efficient non-parametric and model-free statistical quantification of pairwise epistasis in genetic association studies.^{19–23}

Second, all pairwise SNP-SNP interactions are quantified and ranked. We build statistical epistasis networks where vertices are SNPs and edges are weighted pairwise interactions. We only include pairs of SNPs if their interaction strengths are stronger than a preset threshold. We analyze the network topological properties at each cutoff value of the threshold, such as the size of the network (the number of its vertices and the number of its edges), the connectivity of the network (the size of its largest connected component), and its vertex degree distribution.

Then, a threshold of the pairwise interaction strength is determined systematically by finding the cutoff where the topological properties of the real-data network differentiate the most from the null distribution of permutation testing. Such a SEN provides a significant global structure of clustered strong pairwise epistatic interactions associated with a particular phenotype. It serves as a map for further network properties investigation and key SNPs prioritization as discussed in the following section.

2.4. Network Property Analysis of SEN

The *assortativity* of a network measures the propensities of vertices with similar characteristics to connect to one another.^{24,25} In the context of SENs, we are interested in looking into the main effect assortativity, i.e., whether there exists a correlation of main effects between pairs of interacting SNPs. Such main effect assortativity is calculated as the Pearson correlation coefficient r of the main effects at either ends of an edge in a SEN,

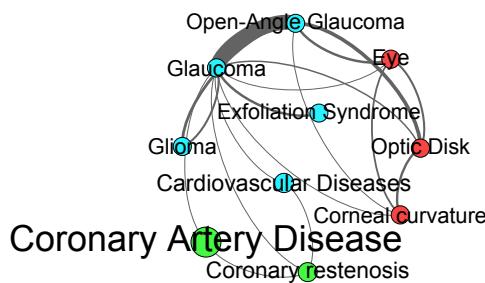
$$r = \frac{M^{-1} \sum_i j_i k_i - [M^{-1} \sum_i \frac{j_i + k_i}{2}]^2}{M^{-1} \sum_i \frac{j_i^2 + k_i^2}{2} - [M^{-1} \sum_i \frac{j_i + k_i}{2}]^2}, \quad (1)$$

where M is the total number of edges, and j_i and k_i are the main effects, calculated as the mutual information of a SNP and the phenotypic class of the vertices (SNPs) at the ends of the i -th edge, with $i = 1, 2, \dots, M$. The coefficient r lies between -1 and 1, with $r = 1$ indicating perfectly assortative, $r = 0$ for non-assortative, and $r = -1$ for complete disassortative networks.

At the vertex level, *centrality* measures the importance of individual vertices in a network. The most commonly used centrality measure is the node's *degree* $C_D(v)$, which is the total

Phenotype	#SNPs	Degree
Exfoliation Syndrome	1	1
Coronary Artery Disease	639	2
Cardiovascular Diseases	70	2
Corneal curvature	13	4
Optic Disk	17	4
Open-Angle Glaucoma	6	5
Glioma	12	2
Eye	13	4
Glaucoma	18	9
Coronary restenosis	56	3

(a)



(b)

Fig. 2. Glaucoma network neighborhood. (a) Phenotypes, including the number of SNPs associated with each phenotype and the “strength” of the interaction (degree) with the rest of the sub-network. (b) a depicted representation of the Glaucoma-centered sub-network, in which the node sizes are proportional to the number of associated SNPs.

number of edges connected to vertex v . The degree of a SNP in the statistical epistasis network shows the number of other SNPs with which it is interacting. *Betweenness* centrality is a more sophisticated metric that quantifies the number of times a vertex is part of the shortest path between any pair of vertices,²⁶ represented as $C_B(v) = \sum_{s \neq v \neq t \in V} \frac{\sigma_{st}(v)}{\sigma_{st}}$, where σ_{st} is the total number of shortest paths from vertex s to vertex t and $\sigma_{st}(v)$ is the number of those paths that pass through vertex v . *Closeness* centrality, defined as $C_C(v) = \frac{1}{\sum_{s \neq v} d_{vs}}$, where d_{vs} is the distance between vertices v and s .^{27,28} This metric describes how easily a given vertex can reach all other vertices in a network. In the context of SENs, centrality measures are used to identify key SNPs that play an essential role in the global interaction structure.

3. Results

3.1. Pre-screening of Glaucoma Data Using HPN

To reduce the computational cost (in time and resources) of analyzing a GWAS dataset, the data must be filtered, selected and/or prioritized prior to processing. Algorithmic filtering methods fall into two main families: knowledge-driven and data-driven. Knowledge-driven filters relies on the actual known biology behind the data, whereas data-driven filtering solely relies on the statistical pre-processing of the data, regardless of their type. ReliefF, SURF, and TuRF^{29,30} are examples of data-driven statistical methods of preprocessing GWAS data. For a complete review of the knowledge-based vs. data-driven filtering, refer to Sun *et al.*³¹

The method we propose to use here starts with the full HPN described in Section 2.2. We identify the glaucoma vertex and establish a list of the disorders and phenotypes in its immediate neighborhood. The 10 members of the “glaucoma neighborhood” are presented in table in Fig. 2.

Using the sub-HPN, we compile a list of 948 SNPs that have been associated with any of the phenotypes in the sub-network. We subsequently annotate each of the risk-associated SNPs with its linkage disequilibrium³² SNPs, indirectly associated with an increased risk. The full list of 4,388 features containing both direct and indirect risk-associated SNPs is used to filter the full one million SNPs of the GLAUGEN GWAS data. The subset made of the overlap between our priority-list and the GLAUGEN dataset, composed of 2,380 SNPs, is

used to build the SNP interaction network presented in the following section.

3.2. Statistical Epistasis Network of Glaucoma

Using the list of 2,380 SNPs, we evaluate all the pairwise epistatic interactions (>2.8 million) using the information gain measure. The strongest interacting pair is SNPs rs13123790 and rs6572380 with a strength of 1.163% association of the disease class. We set the pairwise interaction strength threshold decreasing from the strongest observed value with a step of 0.0001. At each cutoff, a SEN is built by including pairs of SNPs as edges and two end vertices if their interaction strength is above the cutoff. Then for each network we evaluate its topological properties including the number of non-isolated vertices, the number of edges, the sum of weighted edges, and the size of its largest connected component. A 100-fold permutation test assesses the significance of these observed network properties. We permute the data 100 times by randomly shuffling the disease class column, and for each permuted dataset all the pairwise interaction strengths are calculated and networks are built using the same cutoffs as the read-data networks.

We observe a network at cutoff 0.693% that has a largest connected component including significantly more vertices ($p = 0.05$) than those permuted-data networks at the same cutoff. This largest connected component has 713 vertices and 789 edges (Fig. 3).

The main effect assortativity of this network is -0.053 with a significance of $p = 0.04$ using a 100-fold edge swapping permutation test where we randomly pick $10 \times |E|$ ($|E|$ is the total number of edges) pairs of edges and swap their end vertices. This significant negative assortativity indicates that the main effects of interacting SNPs are negatively correlated, i.e., high main-effect vertices tend to interact more with low main-effect vertices.

The degree, betweenness, and closeness centralities of all the vertices in the network are evaluated, and their distributions are shown in Fig. 4. Most of vertices have degrees equal to or less than 3. However there are a minority vertices with a degree higher than 5, meaning that they interact with larger number of other SNPs. In the distribution of betweenness centrality, the majority of vertices have low betweenness. However, some vertices have a much higher betweenness than the rest. The closeness centrality follows a normal distribution indicating that most vertices have similar access to all other vertices. We sort the vertices by descending centrality. At the top, we find the key SNPs that are potentially involved the genetic processes responsible for glaucoma. We further investigate the clinical and biomedical implications of these SNPs through Biofilter³³ functional annotations. Biofilter provides a single interface to multiple publicly available online human genetic datasources (<https://ritchielab.psu.edu/software/biofilter-download>).

4. Discussion

In this study, we proposed a genetic interaction analysis framework of using human phenotype network (HPN) as knowledge-based data filter and subsequently statistical epistasis network (SEN) for quantitative epistatic interaction detection and visualization. Our method was successfully applied to GLAUGEN GWAS data and we were able to identify a connected network of interacting SNPs that included significantly more SNPs than expected randomly. Such a

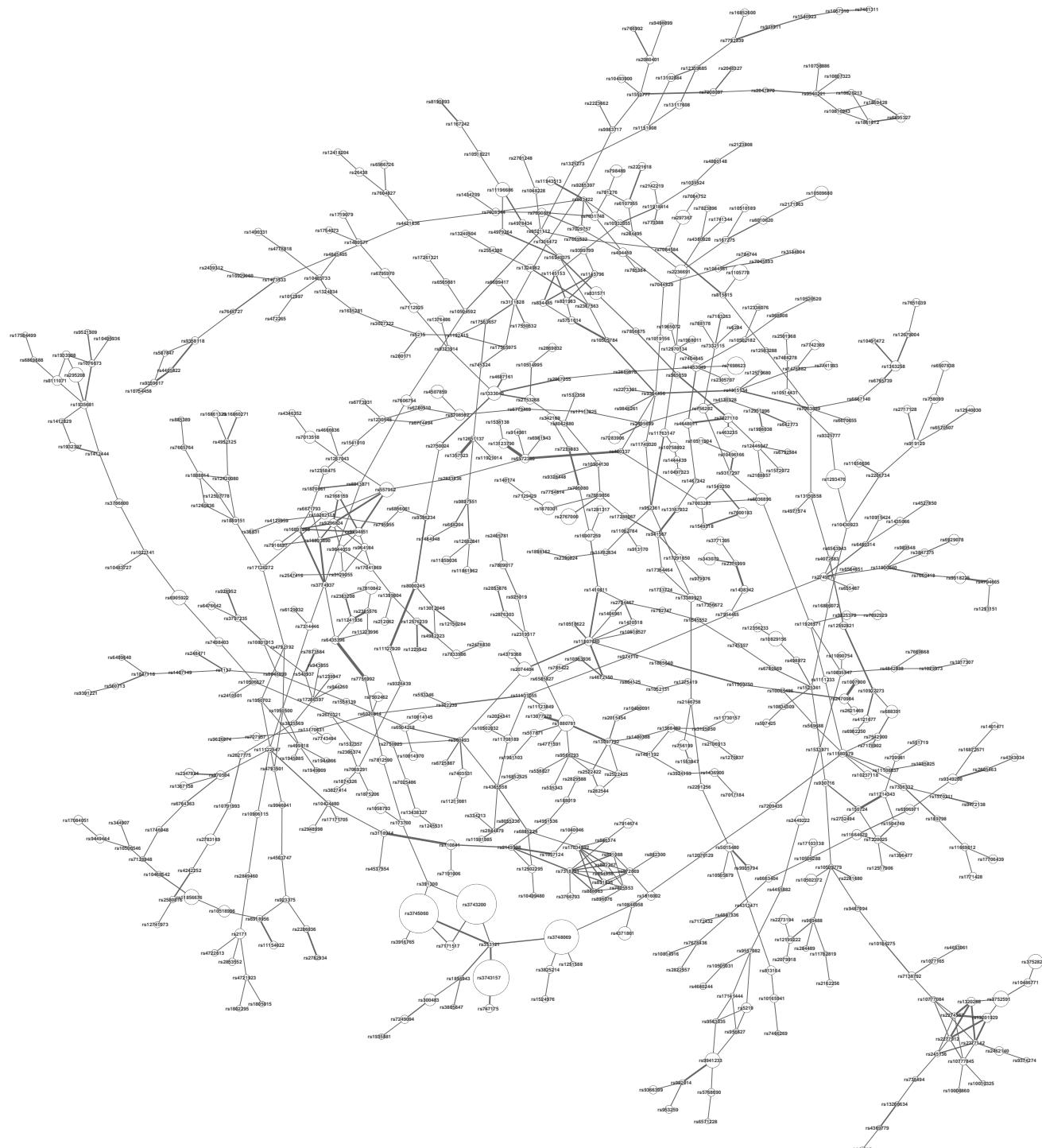


Fig. 3. The SEN of Glaucoma. The network includes 713 SNPs (vertices) and 789 pairwise interactions (edges). The size of a vertex represents the strength of the main effect of its corresponding SNP, with the disease association ranging from $\ll 0.001\%$ to 1.124% . The width of an edge indicates the strength of its corresponding interaction, with the disease association ranging from 0.693% to 1.163% .

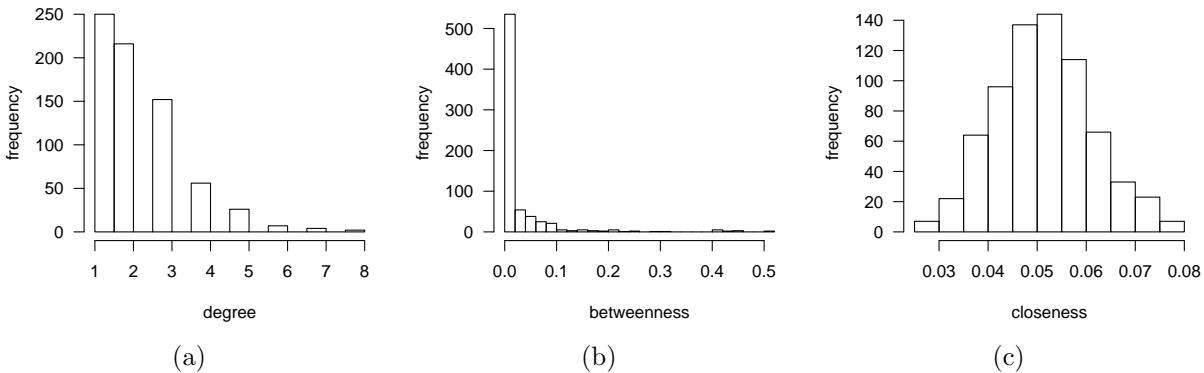


Fig. 4. The distributions of (a) degree, (b) betweenness, and (c) closeness centralities of vertices in the SEN of Glaucoma.

large connected SNP network indicated a complex genetic basis of glaucoma.

HPN is an computational model that systematically organizes and clusters various phenotypes and diseases based on their shared genetic association factors or related basic functional units. It also serves a powerful knowledge-based filter for GWAS data pre-screening and prioritization. Indeed, when compared to established data-driven filters, including the family of Relief algorithms, the HPN filtering performed equally well. This was evaluated by quantifying all the pairwise interactions of filtered sets of SNPs using different data-driven filters (data not shown).

SEN is able to detect and analyze genetic interactions among a large set of attributes in genetic association studies. The derived SNP interaction network of glaucoma in this study had a significant connected structure and may serve as a map for identifying key genetic factors associated with the disease with further biological validations. The significant negative main effect assortativity of such a network indicated that epistatic interactions were likely to happen between SNPs with negatively correlated main effects. This finding suggests that most existing main-effect-based GWAS data filtering algorithms may overlook potential important interactions since some SNPs with low main effects would not be pre-selected using such filters.

The set of SNPs with highest ranked centralities were annotated using Biofilter³³ to their associated genes and, in turn, the genes to pathways and functional groups. We used the annotation to identify other complex diseases possibly sharing biology with glaucoma and suggest novel hypotheses for undocumented shared genetic interactions.

It is known that glaucoma and type II diabetes (T2D) are associated, but direct genetic relationship remains unknown. We found SNPs connecting T2D to glaucoma: rs1333040 ($C_D = 5, C_B = 0.043$) and rs1053049 ($C_D = 5$). The genes CDKN2B-AS1 and PPARD, which these two SNPs are respectively located in, are known to be associated with diabetes. Peng *et al.* observed that CDKN2A/B gene had a 1.17-fold (95% CI: 1.1-1.23) risk estimate for T2D.³⁴ PPARD has been found to be associated with higher fasting glucose, glycated hemoglobin (HbA1c), and increased risk of T2D and combined impaired fasting glucose/T2D in a population-based Chinese Han sample.³⁵ SNP rs9540221 ($C_D = 5$) is also intergenic between NFYAP1 and

LGMNP1, of which the former has also been associated with diabetes and obesity. Indeed, population-based studies suggest that persons with T2D have an increased risk of developing open-angle glaucoma (OAG): the Beaver Dam Eye Study,³⁶ the Framingham Eye Study,³⁷ and the Los Angeles Latino Eye Study.³⁸ Chopra *et al.* has noted that there is a positive association between T2D and OAG in a sample Latino population study in Los Angeles, USA.³⁸ The Rotterdam Study³⁹ has also previously reported a positive association between DM and OAG, with a relative risk of 3.11 (95% CI, 1.12-8.66). Laboratory evidence for T2D and glaucoma linkages also exists. Some have observed that there is an inner retinal cell death, determined by the presence of hyaline bodies in the ganglion cell and retinal nerve fiber layer of the eyes of post-mortem eyes of patients with T2D.⁴⁰ Possible hypotheses of T2D to glaucoma direct linkage include how altered biochemical pathways and vascular changes in T2D reduce blood flow and impaired oxygen diffusion, thus increasing oxidative stress and endothelial cell injury.⁴¹ Excessive glial cell activation may contribute to chronic inflammation, making the retinal ganglion cells more susceptible to glaucomatous damage.⁴² Connective tissue remodeling might also promote increased intraocular pressure (IOP) and greater optic nerve head mechanical stress, which can result in glaucomatous optic neuropathy.⁴³ Diabetes is known to cause microvascular damage and may affect the vascular autoregulation of the optic nerve and retina. Thus a longer duration of T2D would be associated with a higher risk of OAG.

Obesity, like T2D, is a trait often associated with glaucoma, but whose connection remains elusive. We found associations between obesity and glaucoma in SNPs rs12970134 ($C_D = 5, C_C = 0.062$), rs11807640 ($C_D = 8, C_B = 0.516, C_C = 0.071$), rs9540221 ($C_D = 5$), rs10777845 ($C_D = 5$), and rs4365558 ($C_D = 5, C_C = 0.063$), all of which showed significance in our glaucoma study. SNP rs12970134 is located upstream of gene MC4R, the defects of which have been identified as a cause for autosomal dominant obesity.⁴⁴ SNP rs11807640 is located downstream to gene CDK4PS, which has been associated with obesity traits among some postmenopausal women.⁴⁵ Both rs9540221, which is intergenic between NFYAP1 and LGMNP1 and rs4365558, which is intergenic between A4GALT and RPL5P34, are SNPs situated between genes that have been associated with obesity as well. Finally, rs10777845 is upstream to gene RMST, which has been observed to be associated with severe early-onset obesity.⁴⁶ In the investigation to determine direct linkage between glaucoma and obesity, several hypotheses exist. Newman-Casey *et al.* observed that not only was the frequency of OAG was higher among obese individuals (3.1%) than among non-obese individuals (2.5%), but that the relationship between OAG and obesity had significant correlations with gender.⁴⁷ Obese women had a 6% increased hazard of developing OAG when compared to non-obese women, while there was no significant effect in men. Similar findings were made by Zang and Wynder, in which they observed OAG diagnosis as twice as likely in women with a body mass index (BMI) above 27.5, but no impact of BMI in men.⁴⁸ Some hypothesize that increased orbital pressure as a result of excess fat tissue may cause a rise in venous pressure and a consequent increase in intraocular pressure (IOP).⁴⁷

Our study also show links between Alzheimer's disease (AD) and glaucoma. The SNP rs803422 ($C_D = 5, C_C = 0.060$), which we found to affect glaucoma is also associated with

AD. In addition, the two genes SORCS3 and MTHFD1L which contain two of the SNPs with the strongest association are also known to be associated with AD. Patients with AD have a significantly increased rate of glaucoma occurrence. In a study in four nursing homes in Germany, 112 Alzheimer's patients were taken as a case group. 29 of those 112 were found to have Glaucoma, a rate of 25.9% as opposed to a 5.2% rate in the control group.⁴⁹ However, in spite of significant research on this topic, there have still been no clear clinical or genetic relationships found between the two diseases.⁵⁰

Perhaps unsurprisingly, our research has also led to us connecting glaucoma with other phenotypes relating to the eye, particularly cataracts and myopia. The association between cataracts and glaucoma can be seen in the similar genes CDK4PS, NFYAP1, and LGMNP1, all of which contain SNPs found to be highly associated with glaucoma. Myopia and glaucoma are both directly related to the SNP rs569688 ($C_B = 0.415, C_C = 0.074$). There have been many recent studies about the connection between glaucoma and myopia, particularly primary open-angle glaucoma (POAG) and high myopia. Though the connection has still not been found, it is believed the high myopia is another risk factor of POAG, though it is not as important as intraocular pressure. However, it is recommended that those with high myopia be screen for glaucoma more frequently.⁵¹

Of the SNPs and genes associated to glaucoma in our study, the majority that are not directly eye-related, are heart related. Glaucoma is connected therefore probably connected to cardiovascular disease, coronary artery disease, and coronary restenosis. SNPs rs17034592 ($C_D = 7, C_C = 0.060$) and loci LDB2, CDKN2B and TRNAQ46P are all associated with coronary artery disease. SNPs rs499818 ($C_D = 5, C_C = 0.052$), rs1333040 ($C_D = 5, C_C = 0.052$), and loci A4GALT, RPL5P34, and CST7 are all associated with myocardial infarction, and A4GALT and RPL5P34 are associated with cardiovascular disease. Studies have shown that cardiovascular disease may play an important role in the development and progression of glaucoma. However, it is also hypothesized that this relationship may be indirect, due to the variable of age.⁵² Additional studies about the relationship between intraocular pressure, the main risk factor for glaucoma, and cardiovascular disease exist. However, any genetic relationship between these two diseases is yet to be found.

An undocumented interaction found in this study is the link between glaucoma and colorectal carcinoma. Relevant SNPs that were significant were rs972869 ($C_D = 7, C_C = 0.052$), located near the EPHB1 gene, and rs300493 ($C_D = 6, C_B = 0.311, C_C = 0.066$), located in the gene NAV3, both of which are linked to colorectal carcinoma. Studies have shown that obesity and T2D are associated risks of colorectal cancer, which could lead to the hypothesis that colorectal carcinoma and glaucoma may be co-morbidities and are affected by similar pathways.

In summary, the successful application of our methodology on GLAUGEN GWAS data proves the effectiveness of our bioinformatics framework, and suggests its great potential in detecting and characterizing gene-gene interactions in GWAS data. In future studies, we expect to extend the interaction analysis beyond the order of pairs. The increased computational demand would, however, require a stricter data pre-filtering. The more stringent SNP filtering can be achieved using a HPN based on SNP overlap to restrict the interactions to the most

significant ones. Other avenues of research would include using eye-specific tissue genotype data to further refine the analysis and the drug response of the nodes of interest.

Acknowledgements

This work was supported by the National Institute of Health (NIH) grants R01-EY022300, R01-LM010098, R01-LM009012, R01-AI59694, P20-GM103506, P20-GM103534.

References

1. R. Sachidanandam, D. Weissman, S. C. Schmidt, J. M. Kakol, L. D. Stein and et al., *Nature* **409**, 928 (2001).
2. W. Y. S. Wang, B. J. Barratt, D. G. Clayton and J. A. Todd, *Nature Review Genetics* **6**, 109 (2005).
3. H. J. Cordell, *Human Molecular Genetics* **11**, 2463 (2002).
4. J. H. Moore, *Nature Genetics* **37**, 13 (2005).
5. N. A. Davis, J. E. Crowe Jr., N. M. Pajewski and B. A. McKinney, *Genes and Immunity* **11**, 630 (2010).
6. T. Hu and J. H. Moore, Network modeling of statistical epistasis, in *Biological Knowledge Discovery Handbook: Preprocessing, Mining, and Postprocessing of Biological Data*, eds. M. Elloumi and A. Y. Zomaya (Wiley, 2013) pp. 175–190.
7. A. Pandey, N. A. Davis, B. C. White, N. M. Pajewski, J. Savitz, W. C. Drevets and B. A. McKinney, *Translational Psychiatry* **2**, p. e154 (2012).
8. J. L. Wiggs, M. A. Hauser, W. Abdrabou, R. R. Allingham, D. L. Budenz, E. Delbono, D. S. Friedman, J. H. Kang, D. Gaasterland, T. Gaasterland, R. K. Lee, P. R. Lichter, S. Loomis, Y. Liu, C. McCarty, F. A. Medeiros, S. E. Moroi, L. M. Olson, A. Realini, J. E. Richards, F. W. Rozsa, J. S. Schuman, K. Singh, J. D. Stein, D. Vollrath, R. N. Weinreb, G. Wollstein, B. L. Yaspan, S. Yoneyama, D. Zack, K. Zhang, M. Pericak-Vance, L. R. Pasquale and J. L. Haines, *J Glaucoma* **22**, 517 (Sep 2013).
9. M. Takamoto and M. Araie, *Ophthalmology Journal* **58**, 1 (2014).
10. J. H. Fingert, *Eye (Lond)* **25**, 587 (May 2011).
11. K.-I. Goh, M. E. Cusick, D. Valle, B. Childs, M. Vidal and A.-L. Barabasi, *Proceedings of the National Academy of Sciences* **104**, 8685 (2007).
12. C. Darabos, K. Desai, R. Cowper-Sallari, M. Giacobini, B. Graham, M. Lupien and J. Moore, *Lecture Notes in Computer Science* **7833**, 23 (2013).
13. C. Darabos, M. J. White, B. E. Graham, D. N. Leung, S. Williams and J. H. Moore, *BioData Mining* **7**, p. 1 (Jan 2014).
14. H. Li, Y. Lee, J. L. Chen, E. Rebman, J. Li and Y. A. Lussier, *Journal of the American Medical Informatics Association : JAMIA* **19**, 295 (January 2012).
15. D. Welter, J. MacArthur, J. Morales, T. Burdett, P. Hall, H. Junkins, A. Klemm, P. Flieck, T. Manolio, L. Hindorff and H. Parkinson, *Nucleic Acids Res* **42**, D1001 (Jan 2014).
16. T. Hu, N. A. Sinnott-Armstrong, J. W. Kiralis, A. S. Andrew, M. R. Karagas and J. H. Moore, *BMC Bioinformatics* **12**, p. 364 (2011).
17. T. Hu, Y. Chen, J. W. Kiralis and J. H. Moore, *Genetic Epidemiology* **37**, 283 (2013).
18. T. M. Cover and J. A. Thomas, *Elements of Information Theory: Second Edition* (Wiley, 2006).
19. A. Jakulin and I. Bratko, Analyzing attribute dependencies, in *Proceedings of the 7th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD 2003)*, , Lecture Notes in Artificial Intelligence Vol. 2838 (Springer-Verlag, 2003).

20. J. H. Moore, J. C. Gilbert, C.-T. Tsai, F.-T. Chiang, T. Holden, N. Barney and B. C. White, *Journal of Theoretical Biology* **241**, 252 (2006).
21. D. Anastassiou, *Molecular Systems Biology* **3**, p. 83 (2007).
22. J. H. Moore, N. Barney, C.-T. Tsai, F.-T. Chiang, J. Gui and B. C. White, *Human Heredity* **63**, 120 (2007).
23. B. A. McKinney, J. E. Crowe, J. Guo and D. Tian, *PLoS Genetics* **5**, p. e1000432 (2009).
24. M. E. J. Newman, *Networks: An Introduction* (Oxford University Press, 2010).
25. M. E. J. Newman, *Physical Review Letters* **89**, p. 208701 (2002).
26. L. C. Freeman, *Sociometry* **40**, 35 (1977).
27. A. Bavelas, *Journal of the Acoustical Society of America* **22**, 725 (1950).
28. G. Sabidussi, *Psychometrika* **31**, 581 (1966).
29. J. H. Moore and B. C. White, *Lecture Notes in Computer Science* **4447**, 166 (2007).
30. C. S. Greene, N. M. Penrod, J. Kiralis and J. H. Moore, *BioData Min* **2**, p. 5 (2009).
31. X. Sun, Q. Lu, S. Mukherjee, P. K. Crane, R. Elston and M. D. Ritchie, *Front Genet* **5**, p. 106 (2014).
32. C. International HapMap, *Nature* **437**, 1299 (Oct 2005).
33. S. A. Pendergrass, A. Frase, J. Wallace, D. Wolfe, N. Katiyar, C. Moore and M. D. Ritchie, *BioData Min* **6**, p. 25 (2013).
34. F. Peng, D. Hu, C. Gu, X. Li, Y. Li, N. Jia, S. Chu, J. Lin and W. Niu, *Gene* **531**, 435 (Dec 2013).
35. L. Lu, Y. Wu, Q. Qi, C. Liu, W. Gan, J. Zhu, H. Li and X. Lin, *PLoS One* **7**, p. e34895 (2012).
36. B. E. Klein, R. Klein and S. C. Jensen, *Ophthalmology* **101**, 1173 (Jul 1994).
37. P. Mitchell, W. Smith, T. Chey and P. R. Healey, *Ophthalmology* **104**, 712 (Apr 1997).
38. V. Chopra, R. Varma, B. A. Francis, J. Wu, M. Torres and S. P. Azen, *Ophthalmology* **115**, 227 (Feb 2008).
39. S. de Voogd, M. K. Ikram, R. C. W. Wolfs, N. M. Jansonius, J. C. M. Witteman, A. Hofman and P. T. V. M. de Jong, *Ophthalmology* **113**, 1827 (Oct 2006).
40. J. R. Wolter, *Am J Ophthalmol* **51**, 1123 (May 1961).
41. V. H. Y. Wong, B. V. Bui and A. J. Vingrys, *Clin Exp Optom* **94**, 4 (Jan 2011).
42. M. Nakamura, A. Kanamori and A. Negi, *Ophthalmologica* **219**, 1 (Jan-Feb 2005).
43. V. C. Lima, T. S. Prata, C. G. V. De Moraes, J. Kim, W. Seiple, R. B. Rosen, J. M. Liebmann and R. Ritch, *Br J Ophthalmol* **94**, 64 (Jan 2010).
44. K. M. Ling Tan, S. Q. D. Ooi, S. G. Ong, C. S. Kwan, R. M. E. Chan, L. K. Seng Poh, J. Mendoza, C. K. Heng, K. Y. Loke and Y. S. Lee, *J Clin Endocrinol Metab* **99**, E931 (May 2014).
45. P. Coveney and R. Highfield, *Frontiers of Complexity: The Search for Order in a Chaotic World* (Faber and Faber, London, 1995).
46. E. Wheeler, N. Huang, E. G. Bochukova, J. M. Keogh, S. Lindsay, S. Garg, E. Henning, H. Blackburn, R. J. F. Loos, N. J. Wareham, S. O'Rahilly, M. E. Hurles, I. Barroso and I. S. Farooqi, *Nat Genet* **45**, 513 (May 2013).
47. P. A. Newman-Casey, N. Talwar, B. Nan, D. C. Musch and J. D. Stein, *Ophthalmology* **118**, 1318 (Jul 2011).
48. E. A. Zang and E. L. Wynder, *Nutr Cancer* **21**, 247 (1994).
49. A. U. Bayer, F. Ferrari and C. Erb, *Eur Neurol* **47**, 165 (2002).
50. P. Wostyn, K. Audenaert and P. P. De Deyn, *Br J Ophthalmol* **93**, 1557 (Dec 2009).
51. S.-J. Chen, P. Lu, W.-F. Zhang and J.-H. Lu, *Int J Ophthalmol* **5**, 750 (2012).
52. S. S. Hayreh, *Survey of Ophthalmology* **43**, Supplement 1, S27 (1999).

**IDENTIFICATION OF GENE-GENE AND GENE-ENVIRONMENT
INTERACTIONS WITHIN THE FIBRINOGEN GENE CLUSTER FOR
FIBRINOGEN LEVELS IN THREE ETHNICALLY DIVERSE POPULATIONS ***

JANINA M. JEFF[†]

*Charles Bronfman Institute for Personalized Medicine , Icahn School of Medicine at Mount Sinai,
1468 Madison Ave.
New York, NY 10128, United States of America
Email: janina.jeff@mssm.edu*

KRISTIN BROWN-GENTRY

*Cigna-Health Spring, 500 Great Circle Drive
Nashville, TN 37228, United States of America
Email: kristin.gentry@healthspring.com*

DANA C. CRAWFORD

*Institute for Computational Biology, Department of Epidemiology and Biostatistics, Case
Western Reserve University, Wolstein Research Building, 2103 Cornell Road, Suite 2527,
Cleveland, OH 44106, USA
Email: dana.crawford@case.edu*

Elevated levels of plasma fibrinogen are associated with clot formation in the absence of inflammation or injury and is a biomarker for arterial clotting, the leading cause of cardiovascular disease. Fibrinogen levels are heritable with >50% attributed to genetic factors, however little is known about possible genetic modifiers that might explain the missing heritability. The fibrinogen gene cluster is comprised of three genes (*FGA*, *FGB*, and *FGG*) that make up the fibrinogen polypeptide essential for fibrinogen production in the blood. Given the known interaction with these genes, we tested 25 variants in the fibrinogen gene cluster for gene x gene and gene x environment interactions in 620 non-Hispanic blacks, 1,385 non-Hispanic whites, and 664 Mexican Americans from a cross-sectional dataset enriched with environmental data, the Third National Health and Nutrition Examination Survey (NHANES III). Using a multiplicative approach, we added cross product terms (gene x gene or gene x environment) to a linear regression model and declared significance at $p < 0.05$. We identified 19 unique gene x gene and 13 unique gene x environment interactions that impact fibrinogen levels in at least one population at $p < 0.05$. Over 90% of the gene x gene interactions identified include a variant in the rate-limiting gene, *FGB* that is essential for the formation of the fibrinogen polypeptide. We also detected gene x environment interactions with fibrinogen variants and sex, smoking, and body mass index. These findings highlight the potential for the discovery of genetic modifiers for complex phenotypes in multiple populations and give a better understanding of the interaction between genes and/or the environment for fibrinogen levels. The need for more powerful and robust methods to identify genetic modifiers is still warranted.

[†] This work was supported in part by NIH U01 HG004798 and its ARRA supplements.

1. Introduction

Plasma fibrinogen is a key player in the pathogenesis of cardiovascular disease, specifically arterial thrombosis. In the coagulation pathway, fibrinogen is converted to fibrin, the main protein component of an arterial clot.^{1; 2} Fibrinogen levels are highly heritable with reports of up to 50% of the trait variability attributable to genetic factors.² These estimates include genes within the fibrinogen gene cluster (*FGA*, *FGB*, and *FGG*) as well as other genes that regulate the production of fibrinogen.³ However, collectively these loci do not account for the expected genetic component for fibrinogen levels. For complex traits such as fibrinogen, the search for highly penetrant disease loci that accurately predict fibrinogen levels has not been as successful compared to Mendelian diseases. Like many traits associated with cardiovascular disease, the role of the environment, several loci in the genome, or the interaction of amongst genes and/or with the environment can account for the heritability yet to be described for fibrinogen levels.

Fibrinogen synthesis is highly dependent on interactions between fibrinogen polypeptide pairs. Fibrinogen has three pairs of polypeptides A α , B β , γ that are encoded by three genes: *FGA*, *FGB*, and *FGG*, respectively.⁴ These polypeptides form disulfide bonds at the near their N-termini and once connected make up fibrinogen.^{5; 6} Therefore it is highly plausible that interactions between genes within the fibrinogen gene cluster may affect fibrinogen levels. There have been reports in the literature on possible gene-gene interactions associated with fibrinogen levels from variants in the fibrinogen gene cluster and Factor XIII.⁷ Missing from the literature is an association study of the possible gene x gene interactions between fibrinogen genes.

In addition to possible gene-gene interactions, the interplay between genes and environmental risk factors can explain the inter-individual variability of fibrinogen levels. Risk factors such as smoking, age, lack of the exercise, sex, use of contraceptives, and body mass index (BMI) significantly contribute to the variability of fibrinogen levels amongst individuals.⁸ However, after accounting for these risk factors, a significant amount of the variance in fibrinogen levels remains unknown.

One of the primary goals for the first phase of the Population Architecture using Genomics and Epidemiology (PAGE I) consortium was to identify genetic modifiers of common disease.⁹ Likewise, the Epidemiologic Architecture of Genes Linked to Environment (EAGLE), a member of the PAGE I consortium, has access to a diverse, cross-sectional survey, the Third National Health and Nutrition Examination Survey (NHANES III). NHANES III is rich in environmental data to enable the identification of genetic modifiers. Here, we examine the effect of genetic modifiers (gene-gene and gene-environment) with plasma fibrinogen levels in three populations: non-Hispanic blacks, non-Hispanic whites, and Mexican Americans from NHANES III. Using 25 SNPs in the fibrinogen gene cluster, we tested for and identified gene-gene interactions within this cluster, as well as gene-environment interactions with these SNPs and known environmental risk factors across all three NHANES populations.

2. Methods

2.1. Study population & genotyping

The National Health and Nutritional Examination Surveys are cross-sectional surveys conducted across the United States by the National Center for Health Statistics (NCHS) at the Centers for Disease Control and Prevention (CDC). NHANES III was conducted between 1988-1990 (phase 1) and 1991-1994 (phase 2)^{10, 11} as a complex survey that over-sampled minorities, the young, and the elderly. All NHANES have interviews that collect demographic, socioeconomic, dietary, and health-related data. Participant interviews, physical examinations, and laboratory measures were used to collect environmental factors for this study. Current smoking status was determined by the question “Do you smoke cigarettes now?” and by cotinine levels >15 ng/ml. Participants that answered “yes” and have cotinine levels >15ng/ml were classified as a current smoker. Environmental variables are described in Table 1.

All NHANES study participants undergo a detailed medical examination at a central location, the Mobile Examination Center (MEC). Body Mass Index (BMI) was measured directly in the MEC as height and weight. Biomarkers were also collected in the MEC including plasma fibrinogen on participants > 40 years using the Clauss clotting method.¹² Participants with fibrinogen levels greater than > 4.0 g/L were excluded from the analysis since extreme measurements can indicate acute inflammation or recent trauma. Plasma fibrinogen levels were normally distributed in all three study populations (Table 1).

Beginning with phase 2 of NHANES III, DNA samples were collected from study participants aged 12 years and older. DNA was extracted from crude cell lysates from lymphoblastoid cell lines in NHANES III as part of genetic NHANES.¹³ A total of 25 SNPs were genotyped using the Illumina GoldenGate assay (as part of a custom 384-oligonucleotide pool assay (OPA)) by the Center for Inherited Disease Research (CIDR) through the National Heart Lung and Blood Institute’s Resequencing and Genotyping Service. A total of 7,159 samples were genotyped, including 2,631 non-Hispanic whites, 2,108 non-Hispanic blacks and 2,073 Mexican Americans. TagSNPs were selected using LDSelect¹⁴ for multiple populations at $r^2 > 0.80$ for common variants (minor allele frequency (MAF) >5%) in three candidate genes (*FGA*, *FGB*, and *FGG*) based on data available for European Americans and African Americans in SeattleSNPs¹⁵. Quality control measurements were calculated locally using the Platform for the Analysis, Translation, and Organization of large-scale data.¹⁶ We flagged SNPs that deviated from Hardy Weinberg Equilibrium expectations (p-value <0.001), MAF<0.05, and SNP call rates <95% for each subpopulation. In addition to these quality control metrics, we genotyped blinded duplicates as required by CDC, and all SNPs reported passed quality control metrics required by CDC. All genotype data reported here were deposited into the NHANES III Genetic database and are available for secondary analysis through the CDC. After all inclusion/exclusion criteria were applied our final study population was comprised of 620 Non-Hispanic Blacks, 1,385 Non-Hispanic Whites, and 664 Mexican Americans.

Table 1. Study population characteristics and hematological trait descriptive statistics for NHANES III participants. Un-weighted means (\pm standard deviations) or percentages are given for demographics and plasma fibrinogen by subpopulation for adults >40 years in age in phase 2 of NHANES III. Abbreviations: non-Hispanic black (NHB), non-Hispanic whites (NHW), Mexican Americans (MA)

Variable	Mean or %			Standard Deviation		
	NHB	NHW	MA	NHB	NHW	MA
Age (yrs)	40.7	53.4	41	± 16	± 20	± 17
Female (%)	58	60	50	-	-	-
Current Smokers (%)	37	26	24	-	-	-
Body Mass Index (kg/m ²)	28.2	26.6	27.7	± 6.67	± 5.56	± 5.42
Plasma Fibrinogen (g/L)	2.95	2.93	2.97	± 0.51	± 0.51	± 0.50

2.2 Statistical Methods

All analyses were performed using the Statistical Analysis Software (SAS v.9.2; SAS Institute, Cary, NC) either locally or via the Analytic Data Research by Email (ANDRE) portal of the CDC Research Data Center (RDC) in Hyattsville, MD. Using multivariate linear regression, cross product terms (all pair-wise gene-gene or gene-environment terms for 25 fibrinogen variants) were added to the regression models for plasma fibrinogen. All SNPs were coded additively and all models were adjusted for the main effect of the SNP (both the GxG and GxE models) and environmental factor (only the GxE models). In addition to adjusting for main effects, all analyses were adjusted for known covariates age, BMI, sex, and smoking status. Age and BMI were coded as continuous variables whereas sex and smoking status were coded as binary variables in all models. Any results or variables with counts less than 5 were ‘suppressed’ and unavailable for reporting by the CDC. All p-values are presented uncorrected for multiple testing, and we considered an interaction significant at $p<0.05$. We chose a liberal significance threshold for several reasons: 1.) we only tested 25 variants in a candidate gene setting, 2.) several variants are correlated yielding redundant results, and 3.) we have limited power (due to sample size and limited loci) to detect moderate (or small) effects.

3. Results

3.1 Gene-gene interactions

We identified a total of 18 unique significant gene-gene (GxG) interactions out of 828 tests performed that effect fibrinogen levels in at least one NHANES III population across all three NHANES III populations (Table 2). We define a significant GxG interaction here as having significant interaction term ($p<0.05$) from two different genes in the fibrinogen gene cluster.

Of the GxG interactions we detected, 50% were identified in non-Hispanic whites, all of which included a *FGB* SNP in the interaction term (Table 2). In fact, all of the significant interaction terms included either *FGB* rs2227395 or *FGB* rs4220 in non-Hispanic whites (Table 2). In our previous work, we demonstrated that these SNPs are in strong linkage disequilibrium ($r^2 = 0.86$) and that both SNPs had significant main effects.¹⁷ After accounting for this correlation, only seven unique gene-gene interactions reached the significance threshold in non-Hispanic whites. Of the seven unique interaction terms, three were associated with increased fibrinogen levels and four were associated with decreased levels of fibrinogen.

In non-Hispanic blacks a total of six unique GxG interactions were associated with fibrinogen levels. Three GxG interactions were associated with decreased fibrinogen levels and three were associated with increased fibrinogen levels (Table 2). Missense variant rs6050 (Thr312Ala) in the *FGA* gene is a known association with fibrinogen levels in Europeans. In previous work, we were unable to detect this association with non-Hispanic blacks¹⁷; however, the interaction term rs6050 x rs2227395 (*FGG*) was significantly associated with decreased fibrinogen levels in non-Hispanic blacks ($\beta = 0.15$, p-value = 0.017, Table 2). Significant interactions terms with *FGB* variant rs2227395 were consistently observed with SNPs in the *FGA* gene (Table 2, last column). Based on our previous single SNP analysis, none of the SNPs identified in these interaction terms had a significant main effect (Table 2, column 3).¹⁷ Overall there were four associations we observed across non-Hispanic whites and non-Hispanic blacks, all of which include rate the rate-limiting gene, *FGB* (Table 2).

In Mexican Americans, we identified six unique interaction terms at the $p < 0.05$ level. Based on the genetic similarity between non-Hispanic whites and Mexican Americans in this population, the associations including rs2227395 and rs4220, and may be correlated in this Mexican American population. After accounting for the correlation between SNPs, three interactions are associated with increased fibrinogen levels and two are associated with decreased levels (Table 2).

Table 2. Significant fibrinogen cluster gene-gene interactions in NHANES III for plasma fibrinogen. Using multivariate linear regression, we added cross product terms (all pair-wise SNP-SNP) to the regression models for plasma fibrinogen. Below are the association results for each SNP individually as well as the association results for the interaction term. Abbreviations: SE = standard error.

Gene-Gene Interactions	SNP-SNP Interactions	SNP 1 Effect		SNP 2 Effect		Interaction Term	
		Beta (SE)	P	Beta (SE)	P	Beta (SE)	P
Non-Hispanic Blacks							
<i>FGB</i> x <i>FGG</i>	rs1800791 x rs1800792	-0.08 (0.04)	0.07	-0.01 (0.04)	0.71	0.20 (0.08)	0.02
<i>FGA</i> x <i>FGB</i>	rs2070006 x rs2227395	0.01 (0.03)	0.85	1.00 (0.03)	0.94	0.14 (0.07)	0.04
<i>FGA</i> x <i>FGB</i>	rs2070009 x rs2227395	<0.001 (0.03)	0.94	1.00 (0.03)	0.94	-0.16 (0.07)	0.02
<i>FGA</i> x <i>FGG</i>	rs2070022 x rs1049636	-0.08 (0.04)	0.06	0.06 (0.03)	0.07	-0.18 (0.08)	0.03
<i>FGB</i> x <i>FGG</i>	rs2227395 x rs2066861	1.00 (0.03)	0.94	0.05 (0.03)	0.12	0.17 (0.07)	0.01
<i>FGA</i> x <i>FGB</i>	rs6050 x rs2227395	0.02 (0.03)	0.52	1.00 (0.03)	0.94	-0.15 (0.06)	0.02
Non-Hispanic Whites							
<i>FGB</i> x <i>FGB</i>	rs1800791 x rs2227395	<0.001 (0.03)	0.89	-0.08 (0.02)	6.7E-4	0.15 (0.06)	0.02
<i>FGA</i> x <i>FGB</i>	rs2070006 x rs2227395	0.02 (0.02)	0.20	-0.08 (0.02)	6.7E-4	-0.09 (0.03)	0.005
<i>FGA</i> x <i>FGB</i>	rs2070006 x rs4220	0.02 (0.02)	0.20	-0.09 (0.02)	2.1E-3	-0.09 (0.04)	0.01
<i>FGA</i> x <i>FGB</i>	rs2070009 x rs2227395	-0.03 (0.02)	0.16	-0.08 (0.02)	6.7E-4	0.09 (0.04)	0.005
<i>FGA</i> x <i>FGB</i>	rs2070009 x rs4220	-0.03 (0.02)	0.16	-0.09 (0.02)	2.1E-3	0.09 (0.04)	0.01
<i>FGA</i> x <i>FGB</i>	rs2070011 x rs2227395	0.03 (0.02)	0.14	-0.08 (0.02)	6.7E-4	-0.08 (0.04)	0.03
<i>FGA</i> x <i>FGB</i>	rs2070011 x rs4220	0.03 (0.02)	0.14	-0.09 (0.02)	2.1E-3	-0.09 (0.04)	0.01
<i>FGA</i> x <i>FGB</i>	rs2070022 x rs2227395	-0.03 (0.02)	0.23	-0.08 (0.02)	6.7E-4	0.18 (0.06)	0.002

<i>FGA</i> x <i>FGB</i>	rs2070022 x rs4220	-0.03 (0.02)	0.23	-0.09 (0.02)	2.1E-3	0.14 (0.06)	0.02
<i>FGB</i> x <i>FGG</i>	rs2227395 x rs2066861	-0.08 (0.02)	6.7 E-4	0.05 (0.02)	0.01	-0.13 (0.04)	0.002
<i>FGB</i> x <i>FGG</i>	rs4220 x rs2066861	-0.09 (0.02)	2.1 E-3	-0.03 (0.02)	0.23	-0.08 (0.04)	0.05
<i>FGA</i> x <i>FGB</i>	rs6050 x rs2227395	-0.04 (0.02)	0.04	-0.08 (0.02)	6.7E-4	0.13 (0.04)	0.001
<i>FGA</i> x <i>FGB</i>	rs6050 x rs4220	-0.04 (0.02)	0.04	-0.09 (0.02)	2.1E-3	0.09 (0.04)	0.02
Mexican Americans							
<i>FGA</i> x <i>FGB</i>	rs2070008 x rs1800791	-0.02 (0.04)	0.57	-0.03 (0.03)	0.26	0.18 (0.08)	0.022
<i>FGA</i> x <i>FGG</i>	rs2070033 x rs1049636	-0.16 (0.15)	0.26	0.04 (0.03)	0.08	0.59 (0.30)	0.051
<i>FGA</i> x <i>FGB</i>	rs2070033 x rs1800791	-0.16 (0.15)	0.26	-0.03 (0.03)	0.26	-0.66 (0.33)	0.048
<i>FGB</i> x <i>FGG</i>	rs2227395 x rs1049636	-0.02 (0.04)	0.58	0.04 (0.03)	0.08	0.17 (0.06)	0.006
<i>FGB</i> x <i>FGG</i>	rs2227395 x rs1800792	-0.02 (0.04)	0.58	-0.03 (0.03)	0.36	-0.13 (0.06)	0.025
<i>FGB</i> x <i>FGG</i>	rs4220 x rs1049636	-0.04 (0.04)	0.36	0.04 (0.03)	0.08	0.14 (0.07)	0.035
<i>FGB</i> x <i>FGG</i>	rs4220 x rs1800792	-0.04 (0.04)	0.36	-0.03 (0.03)	0.36	-0.12 (0.06)	0.032

3.2 Gene-environment Interactions

We tested for gene-environment (GxE) interactions with fibrinogen cluster variants and known environmental factors: sex, age, smoking status, and BMI, which were all associated with elevated fibrinogen levels.⁸ We tested each environmental variable separately for an association with plasma fibrinogen. We detected 13 significant SNP environment interactions at $p < 0.05$ out of 299 totals test performed.

Over 60% of significant GxE interaction terms were observed in non-Hispanic blacks. Interaction terms with rs2227434 (*FGB*, Pro337Ser) were consistently associated with fibrinogen levels and represented more than 37% (3/8) of the significant associations in non-Hispanic blacks (Table 3). However, this SNP is rare (MAF < 0.05) in all NHANES III populations and did not have significant main effects; thus, these associations are likely false positives (Table 3). We observed four significant GxE interactions in non-Hispanic whites and one in Mexican Americans.

We identified six interaction terms with any given fibrinogen SNP and sex, after excluding rare SNPs rs2227434 and rs6063 (Table 3). All interaction terms identified in non-Hispanic whites were with sex and three were associated with increased levels of fibrinogen (Table 3). There was only one SNP x sex interaction significantly associated with increased fibrinogen levels in non-

Hispanic blacks, rs6058 x sex (Table 3). There were no significant interaction terms with age after excluding rare SNPs that did not have significant main effects, as previously mentioned.

Three SNP x BMI interactions were associated with fibrinogen levels in non-Hispanic blacks or Mexican Americans. The interaction term rs6050 x BMI was significantly associated with increased fibrinogen levels (Table 3). Non-synonymous SNP rs6050 (Thr331Ala) is located in the *FGA* gene and was associated with decreased levels of serum fibrinogen in non-Hispanic whites but not non-Hispanic blacks based on previous single SNP analyses.¹⁷ Another interaction term with BMI, rs2070006 x BMI, was significantly associated with decreased fibrinogen levels (Table 3). Rs2070006 is also located in the *FGA* gene; however this SNP was not significant in any NHANES III population for any single SNP model.¹⁷ There was one GxE interaction term significant in Mexican Americans, rs2006879 x BMI, which was significantly associated with decreased fibrinogen levels, Table 3.

There were two SNP x smoking interaction terms significantly associated with decreased serum fibrinogen, rs2070033 and rs2070008, in non-Hispanic blacks (Table 3). Both SNPs are located in the *FGA* gene and were not associated with fibrinogen levels in previous the single SNP test of association.¹⁷

Table 3. Gene-environment interactions with fibrinogen variants and age, sex, BMI, and smoking status. Using linear regression we added the multiplicative terms (SNP x environmental factor) to the regression model while simultaneously adjusting for covariates: age, sex, BMI, and smoking status. Below are the association results for the single SNP or environmental factor tests of association as well as the association results for the interaction term. Suppressed = results not released by CDC because counts were < 5 (see ‘Statistical Methods’).

Interaction	Gene	SNP Effect		Environment Effect		Interaction Effect	
		Beta (SE)	P	Beta (SE)	P	Beta (SE)	P
Non-Hispanic Blacks							
rs2227434 x age	<i>FGB</i>	0.26 (0.49)	0.59	0.008 (0.002)	3.76E-07	0.004 (<0.01)	3.76E-07
rs2227434 x BMI	<i>FGB</i>	0.26 (0.49)	0.59	0.02 (0.003)	4.32E-06	0.01 (<0.01)	4.32E-06
rs6050 x BMI	<i>FGA</i>	0.01 (0.03)	0.52	0.01 (0.003)	5.87E-06	0.01 (0.05)	8.10E-03
rs2227434 x sex	<i>FGB</i>	0.26 (0.49)	0.59	0.09 (0.04)	0.02	0.05 (0.02)	0.02
rs6058 x sex	<i>FGB</i>	0.06 (0.05)	0.18	0.10 (0.04)	0.01	0.22 (0.10)	0.03
rs2070033 x smoking	<i>FGA</i>	0.07 (0.05)	0.16	-0.06 (0.04)	0.14	-0.22 (0.10)	0.03
rs2070006 x BMI	<i>FGA</i>	0.005 (0.03)	0.85	0.02 (0.003)	6.94E-06	-0.01 (<0.01)	0.03
rs2070008 x smoking	<i>FGA</i>	-0.01 (0.04)	0.75	-0.05 (0.04)	-0.16	-0.19 (0.01)	0.04
Non-Hispanic Whites							
rs2070017 x sex	<i>FGA</i>	0.24 (0.22)	0.28	0.07 (0.03)	0.01	0.03	0.01
rs2066861 x sex	<i>FGG</i>	0.05 (0.02)	0.01	0.06 (0.03)	0.02	-0.08	0.05
rs2070009 x sex	<i>FGA</i>	-0.03 (0.02)	0.16	0.06 (0.03)	0.02	0.07	0.05
rs6050 x sex	<i>FGA</i>	-0.04 (0.02)	0.04	0.06 (0.03)	0.04	0.08	0.05
Mexican Americans							
rs2066879 x BMI	<i>FGG</i>	-0.37 (0.24)	0.12	0.02 (0.004)	3.1E-5	-0.28	0.03

4. Discussion

In the present study, we performed an exhaustive analysis to detect gene-gene and gene-environment interactions that may affect fibrinogen levels in NHANES III. Identifying genetic modifiers of fibrinogen levels is important since the genetic contribution of variable fibrinogen

levels at the population level is poorly understood. Additionally, current heritability estimates do not include possible gene-environment interactions. Here using a multiplicative approach, we tested for statistical epistasis or the deviation from linearity in the presence and absence of main effects. We tested for interactions amongst 25 tagSNPs in the fibrinogen gene cluster (gene-gene) as well as tested each SNP with environmental risk factors: age, sex, BMI and smoking status (gene-environment). We observed 19 gene-gene interactions and 13 gene-environment interactions with an empirical p-value < 0.05.

One criticism of exhaustively testing for interactions is the biological interpretation of the results. The fibrinogen gene cluster is an ideal candidate for identifying gene-gene interactions since each gene is dependent on the other to produce active fibrin in the blood.^{5, 6} In this gene-gene analysis, we detected several interactions between different genes in the fibrinogen cluster (Table 2, last column). More importantly, we identified several interactions with rs2227395, which is in the *FGB* gene, also known as the rate-limiting gene in fibrinogen production. It is important to note that this SNP did not have a significant main effect in the single SNP test of association (Table 2, first column).¹⁷ Interestingly, all gene-gene interactions we identified were observed in the absence of main effects (Table 2, first column). It is conceivable that SNPs such as rs2227395, where we did not observe a significant main effect, are interacting with other fibrinogen SNPs and ultimately affecting fibrinogen levels, an effect that would not be detected at the single SNP level.

Environmental variables are well-known risk factors of elevated fibrinogen levels.⁸ We identified several gene-environment interactions that were associated with fibrinogen levels across multiple populations (Table 3). We observed consistent associations with interaction terms containing rs2227434 or rs6063 and environmental factors age, sex and BMI in non-Hispanic blacks (Table 3). More often than not, consistent associations among different SNPs are indicative of high correlation or linkage disequilibrium between SNPs. We attempted to calculate linkage disequilibrium between these SNPs in our study population but failed to obtain accurate results due to low genotype counts for both SNPs. These SNPs were selected based on frequency and biological relevance for two populations as previously mentioned. However, due to poor genotyping efficiency, we have low genotype counts for these SNPs. We have added this statement to the discussion section. We then attempted to use reference HapMap population YRI to calculate linkage disequilibrium between these SNPs but like our study population, genotype frequencies were too low for an accurate measurement.¹⁸ As previously mentioned, both SNPs are rare in non-Hispanic blacks with minor allele frequencies less 0.001 and we only had 5% power to detect these interactions in our study population. After evaluating the main effect of the environmental factors we noticed that the association results for these factors and the interaction term with rs2227434 or rs6063 were identical as well (Table 3). Given the low minor allele frequencies, it is likely these interaction terms are false positives and actually represent the environmental factors alone. For these reasons we caution the interpretation of these results. In contrast, we identified three SNP x BMI interactions both of which had significant environmental main effects as well as two SNP x smoking interactions and five SNP x sex interactions.

While NHANES III contains over 33,000 study participants, our study population sample size was limited for several reasons. The Centers for Disease Control and Prevention (CDC) only collected DNA samples from a subset of the entire dataset, which includes 7,159 participants from phase 2 of NHANES III. DNA was not collected from participants who reported having hemophilia or chemotherapy within four weeks of collection.¹⁹ Our outcome variable for this study was limited to plasma fibrinogen, which was only collected on a subset of NHANES participants. Plasma fibrinogen levels in NHANES III was only measured for participants > 40 years of age¹⁹, which drastically reduced our sample size. Due to sample size, on average we were only 70% powered to detect effects that explain less < 0.5% of the trait variability at $\alpha = 0.05$ and minor allele frequency of at least 0.05. One caveat of exhaustively testing all pair-wise interactions is the increase in type I error or false positives. It is important to note for this study that we did not correct for multiple testing, which increases the likelihood of reporting false positives, thereby reducing the confidence of our reported findings.

Despite the small sample sizes, we identified several gene-gene and gene-environment associations at a liberal significance threshold of $p < 0.05$. To date this is the first study that examines the role genetic modifiers has on plasma fibrinogen levels in admixed populations such as non-Hispanic blacks and Mexican Americans. Despite the statistical evidence of epistasis, we cannot confirm the effect these interactions have on fibrinogen levels from a biological standpoint and future analyses are warranted. This work serves as first step in uncovering genetic epistasis for complex diseases such as cardiovascular disease that is mediated to an extent by fibrinogen levels. Furthermore, since the majority of our findings are in admixed populations this works highlights the importance and need for conducting genetic association studies in these high-risk populations for discovery.

REFERENCES

1. Collet, J.P., Soria, J., Mirshahi, M., Hirsch, M., Dagonnet, F.B., Caen, J., and Soria, C. (1993/10/15). Dusart syndrome: a new concept of the relationship between fibrin clot architecture and fibrin clot degradability: hypofibrinolysis related to an abnormal clot structure. *Blood* 82, 2462-2469.
2. Voetsch, B., and Loscalzo.J. (2003). Genetic Determinants of Arterial Thrombosis. *Arterioscler Thromb Vasc Biol* 24, 216-229.
3. de Maat, M.P. (2005/3). Genetic variation in the fibrinogen gene cluster. *ThrombHaemost* 93, 401-402.
4. Mosesson, M.W. (1998). Fibrinogen structure and fibrin clot assembly. *SeminThrombHemost* 24, 169-174.
5. Zhang, J.Z., and Redman, C.M. (1994/1/7). Role of interchain disulfide bonds on the assembly and secretion of human fibrinogen. *JBiolChem* 269, 652-658.
6. Zhang, J.Z., and Redman, C. (1996/11/22). Fibrinogen assembly and secretion. Role of intrachain disulfide loops. *JBiolChem* 271, 30083-30088.
7. Lim, B.C., Ariens, R.A., Carter, A.M., Weisel, J.W., and Grant, P.J. (2003/4/26). Genetic regulation of fibrin structure and function: complex gene-environment interactions may modulate vascular risk. *Lancet* 361, 1424-1431.

8. de Maat, M.P. (2001). Effects of diet, drugs, and genes on plasma fibrinogen levels. *AnnNY AcadSci* 936, 509-521.
9. Matise, T.C., Ambite, J.L., Buyske, S., Carlson, C.S., Cole, S.A., Crawford, D.C., Haiman, C.A., Heiss, G., Kooperberg, C., Marchand, L.L., et al. (2011/10/1). The Next PAGE in understanding complex traits: design for the analysis of Population Architecture Using Genetics and Epidemiology (PAGE) Study. *AmJ Epidemiol* 174, 849-859.
10. CDC. (1996). Centers for Disease Control and Prevention Third National Health and Nutrition Examination Survey, 1988-94, Plan and Operations Procedures Manual. In., pp -.
11. CDC. (2004). Centers for Disease Control and Prevention Plan and Operation of the Third National Health and Nutrition Examination Survey, 1988-94. In. (Bethesda, MD), pp -.
12. Daum, P., Estergreen, J., and Wener.M. (2000). Laboratory Procedure Manual:Rate of Clot Formation on the STA-Compact. In., pp 1-12.
13. Steinberg, K.K., Sanderlin, K.C., Ou, C.Y., Hannon, W.H., McQuillan, G.M., and Sampson, E.J. (1997). DNA banking in epidemiologic studies. *EpidemiolRev* 19, 156-162.
14. Carlson, C.S., Eberle, M.A., Rieder, M.J., Yi, Q., Kruglyak, L., and Nickerson, D.A. (2004/1). Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. *AmJ HumGenet* 74, 106-120.
15. Crawford, D.C., Akey, D.T., and Nickerson, D.A. (2005). The patterns of natural variation in human genes. *AnnuRevGenomics HumGenet* 6, 287-312.
16. Grady, B.J., Torstenson, E., Dudek, S.M., Giles, J., Sexton, D., and Ritchie, M.D. (2010). Finding unique filter sets in plato: a precursor to efficient interaction analysis in gwas data. *PacSympBiocomput*, 315-326.
17. Jeff, J.M., Brown-Gentry, K., and Crawford, D.C. (2012). Replication and characterisation of genetic variants in the fibrinogen gene cluster with plasma fibrinogen levels and haematological traits in the Third National Health and Nutrition Examination Survey. *Thrombosis and haemostasis* 107, 458-467.
18. Project, I.H. (2003/12/18). The International HapMap Project. *Nature* 426, 789-796.
19. W, E., Lewis, B., and Koncikowski, S. (1996). Laboratory Procedures Used for the Third National Health and Nutrition Examination Survey (NHANES III), 1988-1994. In., pp 537-546.

DEVELOPMENT OF EXPOSOME CORRELATION GLOBES TO MAP OUT ENVIRONMENT-WIDE ASSOCIATIONS

CHIRAG J PATEL *

*Center for Biomedical Informatics, Harvard Medical School, 10 Shattuck Street
Boston, MA. 02215, USA
Email: chirag_patel@hms.harvard.edu*

ARJUN K MANRAI

*Center for Biomedical Informatics, Harvard Medical School, 10 Shattuck Street
Boston, MA. 02215, USA
Harvard-MIT Division of Health Sciences and Technology
Cambridge, MA. 02138, USA
Email: manrai@post.harvard.edu*

The environment plays a major role in influencing diseases and health. The phenomenon of environmental exposure is complex and humans are not exposed to one or a handful factors but potentially hundreds factors throughout their lives. The *exposome*, the totality of exposures encountered from birth, is hypothesized to consist of multiple inter-dependencies, or correlations, between individual exposures. These correlations may reflect how individuals are exposed. Currently, we lack methods to comprehensively identify robust and replicated correlations between environmental exposures of the *exposome*. Further, we have not mapped how exposures associated with disease identified by environment-wide association studies (EWAS) are correlated with other exposures. To this end, we implement methods to describe a first “exposome globe”, a comprehensive display of replicated correlations between individual exposures of the exposome. First, we describe overall characteristics of the dense correlations between exposures, showing that we are able to replicate 2,656 correlations between individual exposures of 81,937 total considered (3%). We document the correlation within and between broad a priori defined categories of exposures (e.g., pollutants and nutrient exposures). We also demonstrate utility of the exposome globe to contextualize exposures found through two EWASs in type 2 diabetes and all-cause mortality, such as exposure clusters putatively related to smoking behaviors and persistent pollutant exposure. The exposome globe construct is a useful tool for the display and communication of the complex relationships between exposure factors and between exposure factors related to disease status.

* Corresponding author

1. Introduction

1.1. A need to identify correlations between exposures

The environment is hypothesized to play a significant role in health and disease, but we lack methods to elucidate how multiple environmental exposures are associated together with disease. Along this line, Wild and Rappaport and Smith have documented a new way to conceptualize the environment called the “exposome” [1, 2], the environmental analog of the genome that consists of the totality of exposures from birth to death. Recently, others and we have proposed a new method to search for environmental factors associated with disease called the environment-wide association study (EWAS) (e.g.,[3-6]). EWAS is analytically analogous to the genome-wide association study (GWAS), a comprehensive way to search for genetic variants associated with disease.

While EWAS and GWAS are operationally similar, genotypes and exposures are very different data types and correlation structures. Genotypes are static and often assume a fixed number of discrete values (e.g., homozygous/heterozygous for single nucleotide polymorphisms). Correlation between genetic variants is a function of chromosomal location due to the phenomenon of “linkage disequilibrium” (LD). The closer variants are located along the genome, the greater the chance they will be inherited together and correlated.

On the other hand, environmental exposures are heterogeneous (in measurement modality and data type) and are dependent on geographic location, human behavior, and time. Their correlation is known to be “denser” than that of genetic variants [3, 7] as many exposures are correlated with many others [8, 9]. Importantly, given an exposure identified from an EWAS, it is very difficult to infer if the exposure is independently associated with the disease, the direction of association (“what causes what”), or if the exposure is simply a correlate [7, 9].

Given these challenges, it is a priority to develop methods to identify robust correlations between exposures. Correlation between exposures may allow investigators to describe how exposures can lead to other exposures (as identified in an EWAS). For example, many nutrients are consumed together. A non-optimal diet (however it may be defined) may lead to a deficiency in a whole group of vitamins and nutrients. As another example, individuals who are exposed to air pollution may have high levels of several products of combustion, including hydrocarbons, volatile compounds, and heavy metal levels. In the environmental health sciences it is hypothesized that prevalent “mixtures”, or combinations of exposures, may dictate health [10]; understanding how exposures are correlated is one step toward defining what mixtures are relevant to human health.

Many methods have been proposed to describe the correlation between multiple variables, often under the analytical category of “unsupervised learning”, and have been used successfully in the genomics field (e.g, [11-13]). We have yet to apply these methods to describe relationships between exposures. In this report, we describe correlations between exposure variables to construct an “exposome globe”, extending methods developed for unsupervised learning with genomic data called “relevance networks” [13]. We utilize the exposome globe to identify clusters of exposures correlated with exposures identified in EWAS (“EWAS-identified exposure”). We

hypothesize it is possible to attain a broader and more interpretable view of EWAS-identified exposures with an exposome globe.

1.2. Methods

1.2.1. About the National Health and Examination Survey (NHANES) data

As documented earlier (e.g., [4]), we attained four NHANES surveys data each representing independent samplings from the US population in years 1999-2000, 2001-2002, 2003-2004, and 2005-2006. Each NHANES survey dataset ascertains an array of environmental factors, sociodemographic factors (e.g., income), and clinical indicators (e.g., serum glucose, time to death). NHANES is a representative sampling of the US population and therefore covers the entire age, sex, and demographic distribution of the US.

We constructed a correlation globe with factors of the exposome. These factors include direct and quantitative measurement of environmental exposures representing chemicals, nutrients, or infectious agents (assayed directly in human tissue, such as blood serum, urine and hair). For example, quantitative measurements of nutrient (e.g., vitamins, carotenes) and pollutant (e.g., heavy metals, polychlorinated biphenyls) levels in human tissue are ascertained via mass spectrometry (MS), such as gas chromatography and inductively coupled plasma MS. Infectious agents (e.g., bacteria) were measured via immunological assays. Second, the CDC ascertained other indicators of environmental exposure including participant self-reported nutrient consumption (derived from a food questionnaire on foods consumed prior to the interview), physical activity, and prescribed pharmaceutical drugs.

1.2.2. Construction of an exposome globe of replicated environmental correlations

Our method is similar to that of the “relevance network” framework to find correlations of expressed genes [13]. We computed the non-parametric correlation coefficient between each pair of environmental factors (e.g., biomarkers of exposures and self-reported information) for each independent survey (e.g., 1999-2000, 2001-2002, 2003-2004, and 2005-2006). These coefficients are bi-serial coefficients between pair of binary factors and Spearman correlations for continuous factors. There are many ways to compute correlations between variables. We chose a non-parametric metric as to not make any distributional assumptions regarding the environmental factors.

We computed 37,207 correlations in the 1999-2000 survey (we denote the set of all correlations by $\rho^{1999-2000}$), 59,412 in 2001-2002 ($\rho^{2001-2002}$), 128,715 in 2003-2004 ($\rho^{2003-2004}$) and 51,340 in 2005-2006 ($\rho^{2005-2006}$). We filtered out correlations that were present in only one survey and therefore could not be replicated and those that had sample sizes less than 10. After filtering, we were left with 35,835, 56,557, 80,401, 47,203 correlations in each of the four surveys respectively.

These correlations represented interdependencies between 289 unique environmental factors in 1999-2000, 357 in 2001-2002, 456 in 2003-2004, and 313 in 2005-2006 surveys. A total of 575 unique factors were observed across all surveys. The sample sizes for computing correlations

ranged from 11 to 9965 (median 1883) in 1999-2000, 11 to 11,039 (median 2237) in 2001-2002, 11 to 10122 (median 2271) in 2003-2004, and 33 to 10348 (median 3267) in 2005-2006.

There were a different number of correlations measured in 2, 3, or 4 surveys. This is because the CDC had different sample sizes for different exposures. Specifically, 41,158 correlations were ascertained in 2 surveys (e.g., 1999-2000 and 2003-2004), 25,436 in 3 surveys (e.g., 1999-2000, 2003-2004, and 2005-2006), and 15,343 in all 4 surveys, resulting in a total of 81,937 correlations considered.

We used a permutation-based approach to estimate the two-sided p-value of significance for each pair of correlations within each independent survey. Specifically, each environmental factor was randomly permuted (sampled without replacement) and the correlations were re-computed to create a set of correlations that reflected the null distribution of no correlation. Briefly, given an exposure X and Y in one dataset of NHANES, we shuffled values of X to a new array \tilde{X} and computed the correlation between \tilde{X} and Y . We repeated this procedure for all pairs of correlations for each survey. We denote distribution of correlations derived from the randomly permuted datasets as $\tilde{\rho}^{1999-2000}, \tilde{\rho}^{2001-2002}, \tilde{\rho}^{2003-2004}, \tilde{\rho}^{2005-2006}$ for each of the 4 surveys respectively. The p-value for an individual correlation from ρ was the fraction of correlations from the permuted dataset $\tilde{\rho}$ with greater absolute value. For example, for a correlation ρ_x from $\rho^{1999-2000}$, the p-value equals $\sum_{i=1}^{35835} I(|\rho_x| < |\tilde{\rho}_i^{1999-2000}|) / 35835$ where I is the indicator function.

We then estimated the false discovery rate (FDR) q-value for each correlation in each of the surveys using the Benjamini-Hochberg step-down approach [14], resulting in a vector of q-values for each survey, denoted as $q^{1999-2000}, q^{2001-2002}, q^{2003-2004}, q^{2005-2006}$. We deemed a correlation to be replicated if its q-value was less than 5% in at least 2 surveys.

A replicated correlation can exist in 2, 3, or 4 independent dataset surveys. We computed a single “overall” correlation that summarized the correlation from multiple surveys with the inverse variance weighting method as used in fixed-effect meta-analyses[15]. In summary, we computed the overall coefficient as weighted average of the coefficients from each of the survey where weights are the standard errors of coefficient. The *exposome globe* consisted of overall summarized correlations in a set of tuples called \mathbf{P} . Each tuple contains the relationship between exposures A and B and their correlation coefficient (ρ). Specifically, if a correlation between exposure A and B was replicated, its overall correlation is inserted in \mathbf{P} as the tuple $[(A, B), \rho]$, where (A, B) links A to B and ρ is the summarized correlation coefficient.

We visualized replicated overall correlations with the Circos visualization toolkit version 0.67 [16]. Each individual environmental factor is grouped and arranged in a circle. Lines between factors on the inside of the circle depict replicated correlations between factors and the thicknesses of the lines depict the absolute values of the correlations. Red and blue lines represent positive and negative correlations respectively.

1.2.3. Environment-wide association findings in type 2 diabetes and all-cause mortality

Previously, we have conducted EWASs for type 2 diabetes (T2D [4]) and all-cause mortality [5]. In T2D, we searched for association between 252 serum and urine biomarkers of exposure with

serum fasting glucose and validated 10 factors. These 10 factors included nutrients such as trans/cis- β -carotene and vitamin C/D and pollutants such as PCB170 and heptachlor epoxide. In all-cause mortality, we searched for association between 249 environmental exposures and self-reported consumption behaviors and validated 7 factors, including urine-measured and serum-measured cadmium, smoking behaviors (e.g., number of cigarettes smoked per day), and physical activity behaviors (e.g., metabolic equivalents).

We visualized the EWAS findings from these studies in the exposome globe. First, we plotted the $-\log_{10}(p\text{-value})$ of association between the environmental factor and outcome (e.g., T2D, all-cause mortality) as a scatter plot in the Circos plot (referred to as an “EWAS track” below). Next, given a set **E** of validated factors (e.g. the 10 factors validated in T2D), we filtered and visualized correlations of pairs (A, B) from **P** that contained any factor in **E** (e.g., all pairs (E, B) or (A, E) where E is a validated exposure in **E**). In other words, we visualized all “first-degree neighbors” of the validated EWAS findings **E** from **P**.

2. Results

2.1. Distribution of correlations of the exposome globe

We considered 81,937 total correlations of the exposome. Correlations among factors of the exposome were modest; specifically, the median of the absolute value of all correlations was 0.025 (interquartile range of 0.010 to 0.06, Figure 1A [red line]).

Of the 81,937 correlations, 12,385 (15%) had a q-value less than 5% in at least 1 survey dataset. Of these, the median absolute value of correlation was 0.122 (interquartile range of 0.071 to 0.282, Figure 1A [green line]).

We define the “exposome globe” as the network of correlations that were replicated (q-value less than 5% in at least two independent surveys). Of the 81,937 correlations, 2,656 (3%) were replicated and made up the exposome globe. The median absolute value of correlations of the exposome globe was markedly higher than the median of all correlations at 0.5 (interquartile range of 0.385 to 0.635, Figure 1A [blue line]). Most of the replicated correlations (2,513 of 2,656) had positive sign (Figure 1B). The median of positive and negative replicated correlations was 0.508 and -0.282 respectively (Figure 1B).

2.1.1. Concordance of replicated correlations

We observed that correlations were concordant between surveys. The concordance of the exposure correlations between the different surveys was greater than 0.8 (assessed via Pearson ρ , Table 1). For example, the concordance between all correlations in the 2001-2002 and the 2003-2004 survey was 0.82 (Table 1). All correlations were highly significant ($p < 10^{-10}$). As expected, when only considering replicated correlations (relationships of the exposome globe), the concordance was greater (e.g., concordance between 2001-2002 and 2003-2004 survey was 0.90). Therefore, while correlations were modest/small (Figures 1AB) they were reproducible across cohorts.

2.1.2. *The exposome globe reflects correlations within and between categories of factors*

While the globe was dense (2,656 of all possible 81,937 correlations were replicated) we observed interpretable broad patterns in the exposome globe (Figure 2). First, we observed positive correlations within each exposure category (“intra-category” correlations), such as between serum nutrients, nutrients ascertained from food recall questionnaires, volatile organic compounds, hydrocarbons, polychlorinated biphenyls (PCBs), dioxins, phthalates, bacteria (co-infection), and pesticides. We observed positive correlations between categories of exposures, such as between phthalates and hydrocarbons, PCBs and dioxins, dioxins and furans, furans and PCBs, pesticides and PCBs. Of note, there were positive correlations between some nutrients and pollutants, such as PCBs, dioxins, and furans. Briefly, PCBs, dioxins, hydrocarbons, and furans are “persistent pollutants”. Persistent pollutants are lipophilic (accrue in fatty tissue) and accumulate in the food chain. PCBs had been used for manufacturing materials whose use has been banned during the 1970s. Dioxins, furans and hydrocarbons are by-products of industrial processes such as pesticide manufacturing and combustion. Demographic factors, including age, sex, and race/ethnicity were also correlated with multiple groups of exposures.

2.1.3. *Describing EWAS-identified factors with the exposome globe*

We used the exposome globe to describe the first-degree correlations of factors validated in previous EWAS investigations of T2D and all-cause mortality. We only selected correlation links in the exposome globe that were between validated EWAS exposures and other exposures. We observed qualitatively different globes for EWAS factors found in T2D and mortality (Figure 3).

In all-cause mortality, we observed clusters of correlated exposures putatively related to smoking but little related to healthy behaviors, such as physical activity or diet (Figure 3A). Specifically, we observed that the self-reported variables of current and past smoking, which had been identified via EWAS as risk factors for death (red points in the EWAS track, Figure 3A), were correlated with hydrocarbons (e.g., naphthols) and volatile organic compounds (e.g., toluene). Further still, urine and serum cadmium, both also positively associated with death (red points on the EWAS track), were also correlated with smoking status and a biomarker of nicotine (cotinine). There were relatively fewer correlates for factors that were associated with protection from death, such as trans lycopene and physical activity.

In T2D, we observed that serum measures of PCB170 and heptachlor epoxide, two types of banned and polychlorinated compounds used in materials manufacturing and pesticides respectively, and positively associated with T2D (red point in EWAS track [Figure 3B]), correlated with other exposures of the same category, such as other polychlorinated biphenyls and pesticides. Therefore, PCB170 and heptachlor epoxide could be a marker of correlated chlorinated exposures, all which may play a role in T2D. Cumulative role of groups of persistent pollutants is indeed one hypothesis for T2D [17]. Serum levels of Vitamin A (e.g., retinol, retinyl stearate, and retinyl palmitate), were positively correlated with heptachlor epoxide. Further, serum-measures of γ -tocopherol, a type of vitamin E (positive association with T2D, red point in EWAS track, Figure 3B), was negatively correlated with serum-measured folate (blue correlation line, Figure 3B); individuals with high levels of γ -tocopherol had lower levels of folate.

Table 1. Pearson ρ of exposure correlations between each independent NHANES dataset. Number of correlations compared are in parentheses.

	1999-2000	2001-2002	2003-2004	2005-2006
1999-2000	1.00	0.84 (33191)	0.84 (34337)	0.92 (16955)
2001-2002		1.00	0.82 (55025)	0.93 (22931)
2003-2004			1.00	0.94 (47070)
2005-2006				1.00

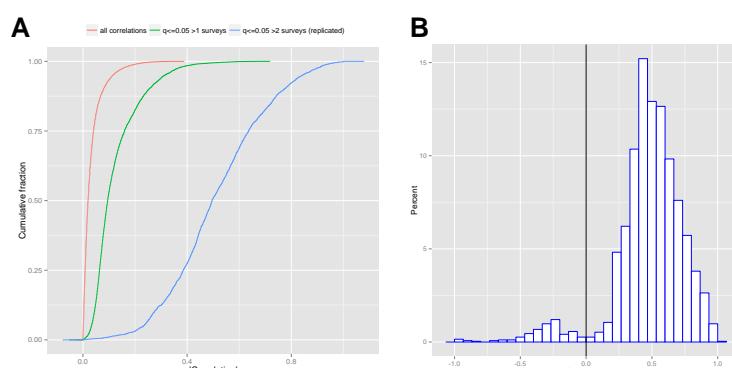


Fig 1. A.) Cumulative distribution of absolute value of correlations. The red line denotes the summarized correlation coefficients for all pairs of exposures possible. The green line denotes correlations that achieved q-value less than 5%. The blue line denotes correlations that were replicated (and part of the exposome globe), or had q-value less than 5% in at least 2 surveys. **B.) Histogram of all replicated correlations of the exposome globe.** Vertical black line denotes 0 correlation.

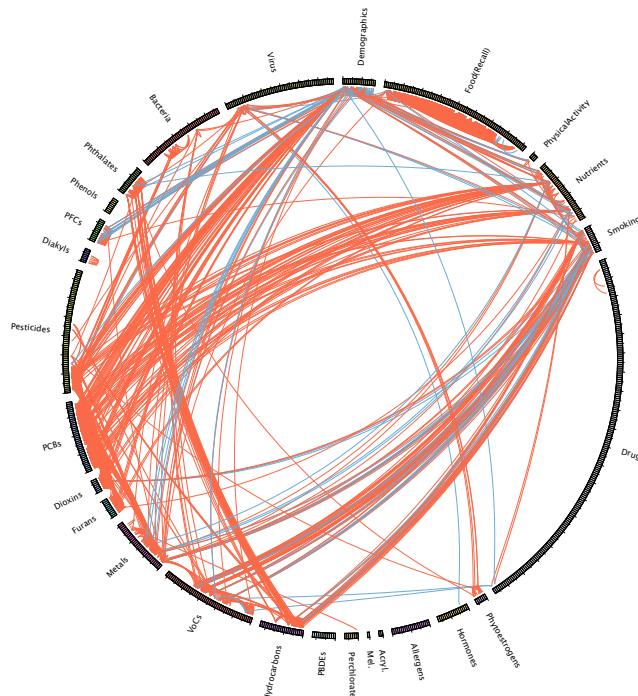


Fig 2. Overall Exposome Correlation Globe. 575 exposures are grouped by a priori defined environmental health categories and displayed in different colors in the globe. Replicated correlations are shown in red (positive correlation) or blue (negative correlation) lines between exposures. Line thickness is proportional to size of the absolute value of correlation coefficient. Only replicated correlation links are displayed (distribution shown in Figure 1B).

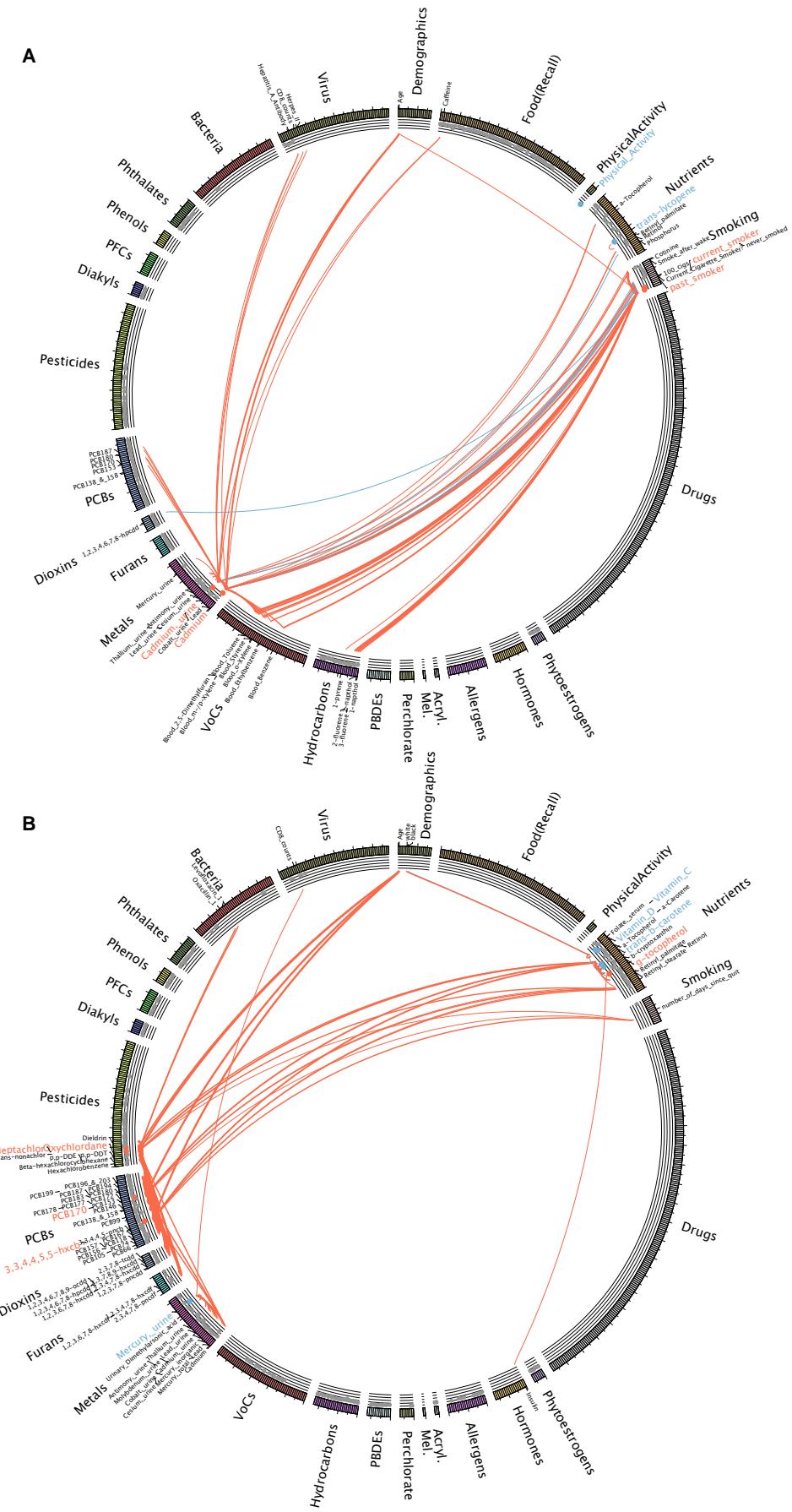


Fig 3 (above). **A.) Exposome Correlation Globes for EWAS in All-Cause Mortality and B.) T2D.** Association p-values from EWAS are shown as a separate track (“EWAS track”) above each exposure (red points denote EWAS validated associations with positive effect size [indicating risk] blue points indicate an EWAS validated negative effect size [indicating protective]). Validated EWAS associations for T2D and all-cause mortality are offset in labeled in red or blue text. Only “first-degree” correlations (correlations for validated EWAS findings) are displayed in the globes and displayed in black text. Acryl.=acrylamide; Mel=Melamine; VoC=volatile organic compounds; PCBs=polychlorinated biphenyls; PFCs=polyfluorinated compounds

3. Discussion

3.1.1. *Summary of findings*

By relating all possible exposures with one another by comprehensively computing correlations and replicating these correlations across multiple independent survey datasets, we were able to produce a first exposome correlation globe. We observed that this globe contains many reproducible correlations between exposures of the same environmental health category or group, but also between these groupings. The correlations of these exposures may be indicative of ways human populations are exposed (“routes of exposure”), such as behaviors and/or shared metabolic fate of biomarkers of exposure. Relatedly and importantly, by selecting correlations that are related to a disease outcome and identified by EWAS (via the EWAS track), we can create hypotheses regarding disease-related exposures, such as smoking correlates in mortality and persistent pollutants in T2D.

3.1.2. *Strengths of exposome globes*

There are several advantages of the proposed exposome globes. First, exposome globes allow the presentation and visualization of the clusters of co-existing exposures, or mixtures, in humans. These mixtures may be a result of common routes of exposure or behaviors (e.g., foods are mixtures of nutrients or smoking behavior can result in a mixture of hydrocarbons and heavy metals). These systematic correlations may also help identify shared characteristics of exposures; for example, chlorinated persistent pollutants were all densely correlated with one another perhaps due to shared routes of exposure, but also because they happen to be lipophilic and have similar metabolic fates.

Secondly, knowing how exposures are correlated with one another may aid inference in disease association studies, such as EWAS or gene-environment (GxE) interaction studies. For example, displaying EWAS identified factors with correlation globes may enable investigators to pin down behaviors that underlie the correlations. For example, we observed that many of the exposures found in an EWAS in all-cause mortality, such as smoking, were strongly correlated with hydrocarbons and volatile organic compounds. These compounds may be indicative of the complex chemical matrix of cigarette smoke (e.g., metals, hydrocarbons, and volatile compounds may be found in cigarette smoke). Such a visualization is analogous to a “manhattan plot” in GWAS, where the correlation (LD) between genetic variants and their p-value of association is visualized jointly to enable assessment of independence of associations between genotype and disease [18].

Relatedly, in GxE investigations, exposome globes may present alternative scenarios for interaction between the environment and genetic variants. Because of power and sample size constraints, GxE investigations test a few environmental factors at a time [19]. For example, we recently documented an interaction between serum levels of trans- β -carotene and a GWAS-identified SNP, rs13266634 in the *SLC30A8* gene in T2D [20]. However, evidence of statistical interaction is not evidence of biological interaction. But, other exposures correlated with trans- β -carotene (e.g., Figure 3B) may provide clues to other possible alternative molecular pathways that are centered on the *SLC30A8* gene.

Third, correlated exposures may enable investigators to identify biases, such as confounders in association studies, including EWAS or GxE interaction studies. Confounded exposures are those that are not causal, but associated with the disease of interest (e.g., diabetes or mortality) and the causal exposure (similar to genetic loci in linkage as discussed above). Once correlated exposures are identified through the globe, investigators can attempt to “control” and condition for them in their statistical models to observe how they influence the strength of association between exposures and the disease. Conditioning by correlated exposures also enables investigators to assess independence of associations between exposure and disease or even find other exposures associated with the disease [21], such as in GWAS [22]. Further, as we have claimed before, exposome globes may also enable investigators to compare effect sizes for disease associations among different categories of correlated exposures appropriately [9].

Fourth, exposome globes enable coordination, collaboration, and communication between individual investigators. For example, because of heterogeneous nature of exposures (such as measurement modality), single investigators may have expertise on but a few of these exposures (e.g., phenols, heavy metals, infectious agents). The exposome globe presents a way to relate exposures to another and across domains of expertise (e.g., between chemical exposure to infectious agents). Exposure globes may also help organize broad follow-up efforts, across exposures of different categories and correlated exposures.

3.1.3. Future directions

With the exposome globe in place, other analyses can follow. First, one could quantitatively identify highly correlated subsets of the exposome, analogous to “haplotypes” in the genome using methods such as weighted network analyses [11]. Haplotype blocks are contiguous regions of the genome that contain genetic variants that are correlated because they are inherited together, a phenomenon known as linkage disequilibrium. There are several benefits of explicitly identifying clusters of the exposome, including assessing only a subset of exposures that are correlated with one another in future EWAS. This is likely to be a more cost effective measurement of the exposome. By analogy, in GWAS, a comprehensive view of common frequency genetic associations is achieved by measuring only a subset (“tagging” variants) of all possible common genetic variants. Tag variants are in linkage disequilibrium and correlated with unmeasured variants. While providing “tag” exposures that are proxies of others, exposome haplotypes themselves will allow derivation of new categories of exposure that reflect the mixtures present in

humans. Further, we may begin to hypothesize how interventions on few exposures may modulate many others and even how seemingly distinct pathologies may share a common etiology.

We emphasize that the exposome globe is descriptive does not capture independent relationships, causal, and/or time-dependent relationships between exposures. Extending globes to partial correlation networks (e.g., [23, 24]) may be informative regarding independent relationships, but an outstanding challenge is adapting these methods to missing exposure information and assessing exposures over time (both issues with NHANES, a cross-sectional survey). Understanding the directionality of relationships between exposures will require longitudinal exposure data on individuals coupled with multivariate computational methods to model time-dependent changes of entire correlation globes. Exposures are highly time-dependent, and it would be worthwhile to test whether and how exposome globes differ between an individual from child to adulthood.

Our method could be expanded to incorporate geospatial and/or clinical data. Exposures reflect where individuals live and work; for example, the correlation globe for individuals in urban settings will likely be very different than those living in a rural place. Last, we plan to move beyond just T2D and mortality and consider relationships between the exposome globe and other clinical and physiological variables. In doing so, we hope to get a broader glimpse of the complex role of the exposome in disease.

4. Acknowledgments

This work is supported by a NIH National Institute of Environmental Health Sciences (NIEHS) K99/R00 Pathway to Independence Award (1K99ES023504-01) and a fellowship award from the Pharmaceutical Research and Manufacturers Association of America (PhRMA) to CJP.

References

- 1.Wild, C.P., *Complementing the genome with an "exposome": the outstanding challenge of environmental exposure measurement in molecular epidemiology*. Cancer Epidemiol Biomarkers Prev, 2005. **14**(8): p. 1847-50.
- 2.Rappaport, S.M. and M.T. Smith, *Environment and Disease Risks*. Science, 2010. **330**(6003): p. 460-461.
- 3.Patel, C.J. and J.P. Ioannidis, *Studying the elusive environment in large scale*. J Am Med Assoc, 2014. **311**(21): p. 2173-4.
- 4.Patel, C.J., J. Bhattacharya, and A.J. Butte, *An Environment-Wide Association Study (EWAS) on type 2 diabetes mellitus*. PLoS ONE, 2010. **5**(5): p. e10746.
- 5.Patel, C.J., et al., *Systematic evaluation of environmental and behavioural factors associated with all-cause mortality in the United States National Health and Nutrition Examination Survey*. Int J Epidemiol, 2013. **42**(6): p. 1795-810.
- 6.Hall, M.A., et al., *Environment-wide association study (EWAS) for type 2 diabetes in the Marshfield Personalized Medicine Research Project Biobank*. Pac Symp Biocomput, 2014: p. 200-11.
- 7.Ioannidis, J.P.A., et al., *Researching Genetic Versus Nongenetic Determinants of Disease: A Comparison and Proposed Unification*. Sci Transl Med, 2009. **1**(7): p. 8.
- 8.Smith, G.D., et al., *Clustered environments and randomized genes: a fundamental distinction between conventional and genetic epidemiology*. PLoS Med, 2007. **4**(12): p. e352.
- 9.Patel, C.J. and J.P. Ioannidis, *Placing epidemiological results in the context of multiplicity and typical correlations of exposures*. J Epidemiol Community Health, 2014.
- 10.Carlin, D., et al., *Unraveling the Health Effects of Environmental Mixtures: An NIEHS Priority*. Environ Health Perspect, 2012. **121**(1): p. a6-a8.

- 11.Horvath, S., *Weighted Network Analysis: Applications in Genomics and Systems Biology*2011, New York: Springer.
- 12.Eisen, M.B., et al., *Cluster analysis and display of genome-wide expression patterns*. Proc Natl Acad Sci U S A, 1998. **95**(25): p. 14863-8.
- 13.Butte, A.J. and I.S. Kohane, *Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements*. Pac Symp Biocomput, 2000: p. 418-29.
- 14.Benjamini, Y. and Y. Hochberg, *Controlling the false discovery rate: a practical and powerful approach to multiple testing*. J R Stat Soc B, 1995. **57**(1): p. 289-300.
- 15.Borenstein, M., et al., *Introduction to Meta-Analysis*2009, Chichester, UK: John Wiley and Sons.
- 16.Krzywinski, M., et al., *Circos: an information aesthetic for comparative genomics*. Genome Res, 2009. **19**(9): p. 1639-45.
- 17.Porta, M., *Persistent organic pollutants and the burden of diabetes*. Lancet, 2006. **368**(9535): p. 558-9.
- 18.Pearson, T.A. and T.A. Manolio, *How to interpret a genome-wide association study*. J Am Med Assoc, 2008. **299**(11): p. 1335-44.
- 19.Patel, C.J., R. Chen, and A.J. Butte, *Data-driven integration of epidemiological and toxicological data to select candidate interacting genes and environmental factors in association with disease*. Bioinformatics, 2012. **28**(12): p. i121-6.
- 20.Patel, C.J., et al., *Systematic identification of interaction effects between genome- and environment-wide associations in type 2 diabetes mellitus*. Hum Genet, 2013. **132**(5): p. 495-508.
- 21.Park, S.K., et al., *Environmental risk score as a new tool to examine multi-pollutants in epidemiologic research: an example from the NHANES study using serum lipid levels*. PLoS One, 2014. **9**(6): p. e98632.
- 22.Yang, J., et al., *Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits*. Nat Genet, 2012. **44**(4): p. 369-75, S1-3.
- 23.Magwene, P.M. and J. Kim, *Estimating genomic coexpression networks using first-order conditional independence*. Genome Biol, 2004. **5**(12): p. R100.
- 24.Friedman, J., T. Hastie, and R. Tibshirani, *Sparse inverse covariance estimation with the graphical lasso*. Biostatistics, 2008. **9**(3): p. 432-41.

MITOCHONDRIAL VARIATION AND THE RISK OF AGE-RELATED MACULAR DEGENERATION ACROSS DIVERSE POPULATIONS

NICOLE A. RESTREPO, SABRINA L. MITCHELL, ROBERT J. GOODLOE, DEBORAH G. MURDOCK

*Center for Human Genetics Research, Vanderbilt University, 2215 Garland Avenue, 519 Light Hall
Nashville, TN 37232, USA*

*Email: n.restrepo@vanderbilt.edu, sabrina.l.mitchell@vanderbilt.edu, robert.goodloe@gmail.com,
Deborah.murdock@vanderbilt.edu*

JONANTHAN L. HAINES, DANA C. CRAWFORD

*Institute for Computational Biology, Department of Epidemiology and Biostatistics, Case Western Reserve University,
Wolstein Research Building, 2103 Cornell Road, Cleveland, OH 44106, USA
Email: jonathan.haines@case.edu, dana.crawford@case.edu*

Substantial progress has been made in identifying susceptibility variants for age-related macular degeneration (AMD). The majority of research to identify genetic variants associated with AMD has focused on nuclear genetic variation. While there is some evidence that mitochondrial genetic variation contributes to AMD susceptibility, to date, these studies have been limited to populations of European descent resulting in a lack of data in diverse populations. A major goal of the Epidemiologic Architecture for Genes Linked to Environment (EAGLE) study is to describe the underlying genetic architecture of common, complex diseases across diverse populations. This present study sought to determine if mitochondrial genetic variation influences risk of AMD across diverse populations. We performed a genetic association study to investigate the contribution of mitochondrial DNA variation to AMD risk. We accessed samples from the National Health and Nutrition Examination Surveys, a U.S population-based, cross-sectional survey collected without regard to health status. AMD cases and controls were selected from the Third NHANES and NHANES 2007-2008 datasets which include non-Hispanic whites, non-Hispanic blacks, and Mexican Americans. AMD cases were defined as those > 60 years of age with early/late AMD, as determined by fundus photography. Targeted genotyping was performed for 63 mitochondrial SNPs and participants were then classified into mitochondrial haplogroups. We used logistic regression assuming a dominant genetic model adjusting for age, sex, body mass index, and smoking status (ever vs. never). Regressions and meta-analyses were performed for individual SNPs and mitochondrial haplogroups J, T, and U. We identified five SNPs associated with AMD in Mexican Americans at $p < 0.05$, including three located in the control region (mt16111, mt16362, and mt16319), one in *MT-RNR2* (mt1736), and one in *MT-ND4* (mt12007). No mitochondrial variant or haplogroup was significantly associated in non-Hispanic blacks or non-Hispanic whites in the final meta-analysis. This study provides further evidence that mitochondrial variation plays a role in susceptibility to AMD and contributes to the knowledge of the genetic architecture of AMD in Mexican Americans.

1. Introduction

Vision loss and blindness are powerful detriments to the economic and social well-being of individuals. Economically it costs the United States upwards of \$51 billion a year in medical expenses

and lost worker productivity.¹ More importantly, loss of vision has been ranked as one condition with the greatest impact on an individual's day-to-day life.² Age-related macular degeneration (AMD), a disease of the retina, afflicts more seniors in developed countries than any other form of blindness. The number of incident AMD cases is expected to grow from 11 million today to approximately 22 million by the year 2050.³

AMD is a late-onset, multifactorial disease that eventually results in loss of central vision. The retina is a photosensitive tissue that lines the back of the human eye. It collects visual data in the form of photons and through a series of chemical and electrical events, sends a neurological impulse to regions in the brain which can then interpret them as images.⁴ Central to these images is a region of the retina called the macula, responsible for center vision, which contains a high concentration of photoreceptor cells responsible for color vision (i.e. cones).

Clinically AMD is defined as the degeneration of the macula in which central vision becomes impaired. It occurs in two subtypes termed early and late AMD. Early AMD is typically defined as eyes that present with mild to moderate drusen deposits and pigment abnormalities. Late stage AMD is subdivided into dry (atrophic) and wet (exudative) AMD. Dry AMD is the result of geographic atrophy of the retinal pigment epithelial (RPE) layer which lies directly under the retina and provides metabolic and cellular support to the retina and photoreceptors. Exudative AMD leads to vision loss due to abnormal blood vessel growth. These abnormal blood vessels can break and cause blood to leak below the surface of the retina leading to irreversible damage to the macula and subsequent blindness.

Although AMD has a complex etiology, many environmental and genetic variables have been identified that alter the risk of this disease. Modifiable factors include smoking,⁵⁻⁹ body-mass-index (BMI), and blood lipid¹⁰⁻¹² levels such as high density lipoprotein cholesterol (HDL-C). Non-modifiable factors include race/ethnicity, age,¹³ gender, and genetics. Individuals of European-descent are at highest risk with prevalence rates in individuals over the age of 40 years reaching 7.3% compared to African Americans (2.4%), Asian Americans (6.8%),¹⁴ and Mexican Americans (5.1%).¹⁵ In European Americans over the age of 80, nearly 1 out of 10 is likely to be diagnosed with some form of AMD.¹⁶ Major genetic risk loci include *CFH*,¹⁷⁻¹⁹ *ARMS2*,^{20,21} and *C2/C3*²² which account for most of the known heritable risk of AMD. In all, 20 risk loci have been associated with risk of AMD, however, these only account for upwards of 60% of the heritable risk.²³ Nearly all of these loci are located within the nuclear genome as there have been few studies investigating the potential role that mitochondrial genetic variation plays in the development of AMD.

In vitro studies have found that mitochondrial DNA (mtDNA) variants can affect the replication rate of the mitochondrial genome and thus mtDNA copy number.²⁴ Mitochondria are both particularly sensitive to and a major contributor of cellular reactive oxygen species (ROS), which are a byproduct of oxidative phosphorylation. These free radicals play a large role in chronic inflammation, the complement system pathways, and cardiovascular disease.²⁵ Exposure to excessive oxidative stress can lead to mitochondrial dysfunction in the RPE layer,^{26,27} a decrease in cellular bioenergetics

imperative to photoreceptor initiation/maintenance,^{24,28,29} and susceptibility to apoptosis. Additionally, mitochondrial genetic variation has been associated with AMD risk in European Americans. MtDNA variants on mitochondrial haplogroup H, the most common European haplogroup, have been associated with decreased risk of AMD,^{30–32} while mitochondrial haplogroups J³³ and T³⁴ are associated with increased risk. Collectively, these data suggest that the health of ocular mitochondria may play a role in AMD pathology.

Determining the role that mitochondrial genetic variation plays in AMD risk across populations may provide new insights into the underlying disease pathology. More research is necessary to understand the genetic architecture of disease states in diverse populations. This study explores the contribution of mitochondrial genetic variation to AMD risk not only for European Americans, but also in African Americans and Mexican Americans, populations for which studies are few or missing.

2. Materials and Methods

2.1. Study population and genotyping

The National Health and Nutrition Examination Surveys (NHANES) are conducted by the National Center for Health Statistics (NCHS) at the Centers for Disease Control and Prevention (CDC). NHANES are U.S. population-based, cross-sectional surveys designed to ascertain non-institutionalized Americans without regard to health status. All NHANES include detailed demographic, health, lifestyle, laboratory, clinical, and physical examination data from study participants. Genetic NHANES consists of DNA samples collected from the Third NHANES (III) and subsets of Continuous NHANES (1999-2002 and 2007-2008), which began collection in 1999 and occurs annually. NHANES III was conducted in two phases between 1988-1994 and 1991-1994, with DNA samples having been collected in the second phase.^{35,36} The method of collection for Genetic NHANES has been previously described.^{37,38} We, as the Epidemiologic Architecture for Genes Linked to Environment (EAGLE) study, targeted a total of 63 mitochondrial SNPs for genotyping using Sequenom iPLEX® Gold MassArray as previously described.³⁹

We accessed study participant data exclusively from NHANES III and NHANES 2007-2008, as NHANES 1999-2002 did not collect ophthalmologic data. NHANES III, of which 3,131 participants had available fundus photographs and laboratory measurements of serum cotinine (ng/mL), oversampled non-Hispanic blacks and Mexican-Americans. NHANES 2007-2008 oversampled Mexican Americans and other-Hispanic blacks and had a total of 3,172 participants who completed the fundus exam. Current smokers were defined as those responding “yes” to the question “do you smoke cigarettes now?” or those with serum cotinine levels > 15ng/ml.

All procedures were approved by the CDC Ethics Review Board and written informed consent was obtained from all participants. Because no identifying information is available to the investigators, Vanderbilt University’s Institutional Review Board determined that this study met the criteria of “non-human subjects.”

2.2 AMD phenotype definition

Participants over the age of 40 were selected to have a non-stereoscopic, 45° color fundus photograph taken of one randomly selected eye in NHANES III and a 45° non-mydriatic digital photo taken of both eyes in NHANES 2007-2008. Fundus photographs were graded according to a modified version of the Wisconsin Age-related Maculopathy Scale.⁴⁰ Early AMD cases were at least 60 years of age and included participants with: 1) soft drusen, 2) depigmentation of the retinal pigment epithelium in the presence of soft and/or hard drusen, or 3) hyperpigmentation in the presence of soft and/or hard drusen. Late AMD included participants 60 years of age and older with 1) geographic atrophy, 2) sub-retinal hemorrhage, 3) sub-retinal fibrous scarring, or 4) sensory serous sub-retinal detachments. Controls were at least 60 years old with gradable retinal photographs showing an absence of hallmark AMD features. Controls were not excluded if they presented with another retinal disease.

2.3 Statistical analysis

We tested for an association between each individual mtSNP and mitochondrial haplogroups J, T, and U with AMD. Given that AMD largely occurs on a disease spectrum, data for early and late AMD was pooled for analyses so as to increase power to detect an association. Each mitochondrial variant was tested for an association with AMD using logistic regression assuming a dominant genetic model stratified by self-described race/ethnicity (e.g. non-Hispanic white, non-Hispanic black, and Mexican American). Of the SNPs that passed quality control (QC) standards (call rate >95%), a total of 55 SNPs were included in analyses for NHANES III and 60 SNPs were analyzed in NHANES 2007-2008. A total of 50 SNPs were available for meta-analysis. Haplogroups were assigned to each NHANES participant as previously described.³⁹ Haplogroup analyses were conducted in the same manner as the individual mtSNPs but with participants identified as having either haplogroup J, T, and U each being compared to participants in all other haplogroups. All models were adjusted for age, sex, BMI, and smoking status (current versus ever/never). Analyses were conducted using SAS v9.2 via the Analytic Data Research by Email (ANDRE) portal of the CDC Research Data Center in Hyattsville, MD. All p-values presented are uncorrected for multiple testing.

3. Results

3.1 Population

The study population consisted of a total of 416 AMD cases (312 non-Hispanic whites, 37 non-Hispanic blacks, and 67 Mexican Americans) and 2,200 controls (1,349 non-Hispanic whites, 430 non-Hispanic blacks, and 421 Mexican Americans) 60 years or older at the time of examination (Table 1). In the combined NHANES III/2007-2008 dataset, cases were generally female, except in Mexican Americans (49% female), and overweight defined as $BMI > 25 \text{ kg/m}^2$. Non-Hispanic black cases were nearly twice as likely to be smokers (62% smokers) compared to non-Hispanic white (29%

smokers) and Mexican American (36%) cases. On average, controls were younger compared to cases across all race/ethnicities and were nearly as likely to be smokers compared to cases with the exception of non-Hispanic black cases (62% smokers) versus controls (50% smokers).

3.2 Meta-analysis of mitochondrial variants

A total of 50 mitochondrial SNPs passed QC and were tested across the three race/ethnicities in NHANES III and NHANES 2007-2008. Not all SNPs were available across each population as some SNPs were monomorphic in one population or did not pass QC. Of these 50 SNPs, 41 were available for analysis in non-Hispanic whites, 44 in non-Hispanic blacks, and 42 Mexican Americans (Figure 1).

Table 1: Combined NHANES III and 2007-2008 study population demographics by case/control status and race/ethnicity

	non-Hispanic whites		non-Hispanic blacks		Mexican Americans	
	Case	Control	Case	Control	Case	Control
N	312	1349	37	430	67	421
Age (years)	76.0	71.0	69.2	67.8	69.0	67.1
% Female	60	51	57	49	49	48
% Smoker	29	30	62	50	36	31
BMI (kg/m^2)	27.2	27.8	31.1	29.0	29.2	28.9

Means are presented unless otherwise noted.

In Mexican Americans, five mtSNPs were associated with AMD at $p < 0.05$ (Table 2). Of these, three are located in the mitochondrial control region, the non-coding region responsible for the initiation of transcription of the MT-genome: mt16111 ($p = 0.005$; OR = 2.90; 95% CI 1.38 – 6.11), mt16362 ($p = 0.007$; OR = 2.80; 95% CI 1.32 – 5.95), and mt16319 ($p = 0.01$; OR = 2.59; 95% CI 1.24 – 5.42). A synonymous variant, mt12007, located within the NADH dehydrogenase subunit 4 (MT-ND4) gene, was also found to increase risk of AMD in Mexican Americans ($p = 0.018$; OR = 2.47; CI 1.18 – 5.18). Lastly, mt1736 located in the mitochondrial 16S ribosomal RNA (MT-RNR2) gene, was found to be protective ($p = 0.01$; OR = 0.40; CI 0.19 – 0.83) in this population.

In non-Hispanic blacks and non-Hispanic whites, no test of association was significant at $p < 0.05$ in the adjusted model following inclusion of individual study results into the meta-analysis.

Table 2: Mitochondrial genetic variants associated with AMD risk in Mexican Americans

SNPID	Gene	OR	lower CI	upper CI	p-value	CA	CAF (%)	Race/Ethnicity
mt16111	control region	2.90	1.38	6.11	0.005	A	0.31	Mexican American
		—	—	—	0.98	A	0.01	non-Hispanic white
		—	—	—	0.98	A	0.02	non-Hispanic black

mt16362	control region	2.80	1.32	5.95	0.01	C	0.42	Mexican American
		1.70	0.93	3.10	0.08	C	0.08	non-Hispanic white
		2.29	0.84	6.27	0.10	C	0.13	non-Hispanic black
mt16319	control region	2.59	1.24	5.42	0.01	A	0.34	Mexican American
		0.42	0.05	3.52	0.43	A	0.01	non-Hispanic white
		3.23	0.33	31.60	0.31	A	0.02	non-Hispanic black
mt1736	<i>MT-RNR2</i>	0.40	0.19	0.83	0.01	A	0.65	Mexican American
		—	—	—	0.98	A	0.99	non-Hispanic white
		—	—	—	0.99	A	—	non-Hispanic black
mt12007	<i>MT-ND4</i>	2.47	1.18	5.18	0.01	A	0.34	Mexican American
		1.18	0.31	4.50	0.81	A	0.02	non-Hispanic white
		0.91	0.11	7.50	0.93	A	0.06	non-Hispanic black

Most significant meta-analysis results for the model adjusted by age, sex, body mass index, and smoking status. Results are listed for tests with the smallest p-value

Abbreviations: Odds Ratio (OR), confidence interval (CI), coded allele (CA), coded allele frequency (CAF)

“—“ denotes genetic association tests with uninterpretable results due to very few case counts or monomorphic allele

3.3 Mitochondrial Haplogroups

Previous studies have suggested that mitochondrial haplogroups J, T, and U are associated with AMD.³⁰⁻³⁴ In NHANES III, none of the three haplogroups were associated with AMD at $p < 0.05$ although haplogroup J was associated in non-Hispanic whites at $p = 0.057$ (OR = 2.03; 95% CI 0.98 – 4.20). In NHANES 2007-2008, only haplogroup T was significantly associated at $p < 0.05$ in non-Hispanic whites (OR = 2.50; 95% CI 1.17 – 5.33). No haplogroup was found to be associated with AMD at $p < 0.05$ in any of the racial/ethnic groups in the NHANES III/2007-2008 meta-analysis.

Table 3: Haplogroup frequencies in the combined NHANES III, NHANES 1999-2002, and NHANES 2007-2008 populations as previously published³⁹

Haplogroup	non-Hispanic whites	non-Hispanic blacks	Mexican Americans
J	9.2 %	0.4%	1.4%
T	9.6%	0.4%	0.9%
U	13.6%	1.4%	1.6%

4. Discussion

In this study, haplogroup analyses did not replicate previous associations of the European haplogroups J, T, and U with risk of AMD in non-Hispanic whites.^{31,33,34} Although a little surprising, other studies have not always replicated these associations,^{30,41} which may be due in part to

heterogeneity across these studies or else suggesting a weak role of mitochondrial variation in the risk of AMD. However, we did observe that individual variants on the T haplogroup in the NHANES 2007-2008 non-Hispanic whites were associated with AMD risk in this population. Unsurprisingly, neither individual variants nor haplogroups that were previously associated with AMD in European-descent populations generalized to non-Hispanic blacks. African-descent populations suffer from a lesser burden of AMD, and previous studies have suggested that African-descent populations may have a different genetic architecture contributing to AMD etiology⁴²⁻⁴⁵ compared with other populations. We observed that mitochondrial variants mt16111, mt16362, mt16319, mt1736, and mt12007 were associated with AMD risk within the meta-analyzed Mexican American population after adjusting for well-known non-modifiable factors and environmental modifiers. The direction of the genetic effect was the same across the individual NHANES analyses for these SNPs in the Mexican American populations as follows: mt16111 OR = 2.90 and 2.33; mt16362 OR = 2.49 and 3.35; mt16319 OR = 2.17 and 3.72; mt1736 OR = 0.55 and 0.22; mt12007 OR = 1.83 and 4.26 in NHANES III and NHANES 2007-2008 respectively.

Limited studies have been performed to assess the genetic factors of AMD in Mexican or Latino populations. A handful of studies have examined whether *CFH*, *ARMS2*, and *C2/C3*, strong risk loci in European populations, contribute to risk of AMD in Mexican-descent populations.^{43,46,47} These studies, although limited in case size, did find a correlation between these European-derived variants and risk of advanced AMD in Mexicans and Latinos, suggesting that risk of AMD is being driven in part by European risk variants in these admixed populations. All five of the mtSNPs associated with AMD in Mexican Americans in this study are located on the A-A2 haplogroup background. Haplogroup A developed in Asia over 30,000 years ago and occurs most frequently in the Indigenous peoples of the Americas, with its subgroup A2 found to be the most common haplogroup in many of the indigenous ethnic groups of Central and North America.⁴⁸ In the combined NHANES III, 1999-2002, and 2007-2008 populations, haplogroup A is the most prevalent among Mexican Americans³⁹ with a frequency of 34.2% while composing less than 1% in non-Hispanic whites and non-Hispanic blacks. This observation is interesting given that Mexican Americans, who experience similar rates of AMD as that observed in European-descent groups (5.1% vs 7.3%),¹⁵ may contain a set of genetic risk factors on this haplogroup that are driving AMD risk in addition to or in combination with the already known European-derived variants.

Three of the significant mtSNPs (mt1611, mt16362, and mt16319) are located within the control region of the mitochondrial genome containing the origin of replication and the origin of transcription. These SNPs have not previously been associated with AMD but have been identified as contributors to various forms of cancer. A high load of somatic mtSNPs in the mitochondrial control region (i.e. mt16111) was found in patients suffering from prostate cancer.⁴⁹ In a study examining the effect of mtSNPs located more specifically within the D-loop of the control region on risk for renal cell carcinoma, mt16319 was specifically found to be reduced in cases of clear cell renal cell carcinoma.⁵⁰ Lastly, mt16362 was found to be a risk factor associated with familial breast cancer.⁵¹

Strengths of this study include the systematic fashion in which all participants over the age of 40 years were included in ophthalmologic exams to ascertain eye health and AMD status. This ensures a strong degree of homogeneity in case and control status across the various NHANES cohorts and minimizes between study heterogeneity. Over sampling of minority groups also likely increased the number of cases available for study in these underrepresented groups. Limitations include differences in data collection between NHANES III and NHANES 2007-2008, as NHANES III only performed fundus photography on one randomly selected eye whereas NHANES 2007-2008 performed fundus photography on both eyes. Other limitations include low statistical power and lack of correction for multiple hypothesis testing. When considering a significance threshold for mitochondrial variation analyses it should be noted that independence of each test is questionable as all variation is inherited as a whole. Lastly, we relied on self-reported race/ethnicity as opposed to genetic ancestry of the mitochondrial genome which may lead to false positive associations that are in actuality identifying differences in haplogroup ancestry. Future studies are needed to validate the results of the present study.

Despite limited sample sizes available for AMD analyses, NHANES is one of only a few surveys to include ophthalmologic exams of minority populations. As large scale epidemiology surveys become more cost prohibitive, a stronger emphasis on the utilization of electronic medical records (EMR) to identify cases and controls for inclusion in future studies will become more pronounced. Given that many of these EMR systems are still predominately composed of European-descent patients, a concerted effort must be made to increase the number of minorities with access to routine, continuous health care.

5. Conclusion

In conclusion, we identified potential novel associations between mitochondrial variants and risk of AMD in the NHANES Mexican Americans. All of the associated mitochondrial variants are found on the A-A2 haplogroup which is common among this racial/ethnic group and varies from the haplogroups previously reported to influence risk in European-descent populations. Future studies in larger Mexican-descent datasets may clarify some of these findings. These new findings may offer insight into the cellular mechanics underlying disease etiology.

6. Acknowledgments

This work was supported by NIH U01 HG004798 and its ARRA supplements. The Vanderbilt University Center for Human Genetics Research, Computational Genomics Core provided computational and/or analytical support for this work. The findings and conclusions in this report are those of the authors and do not necessarily represent the views of the Centers for Disease Control and Prevention.

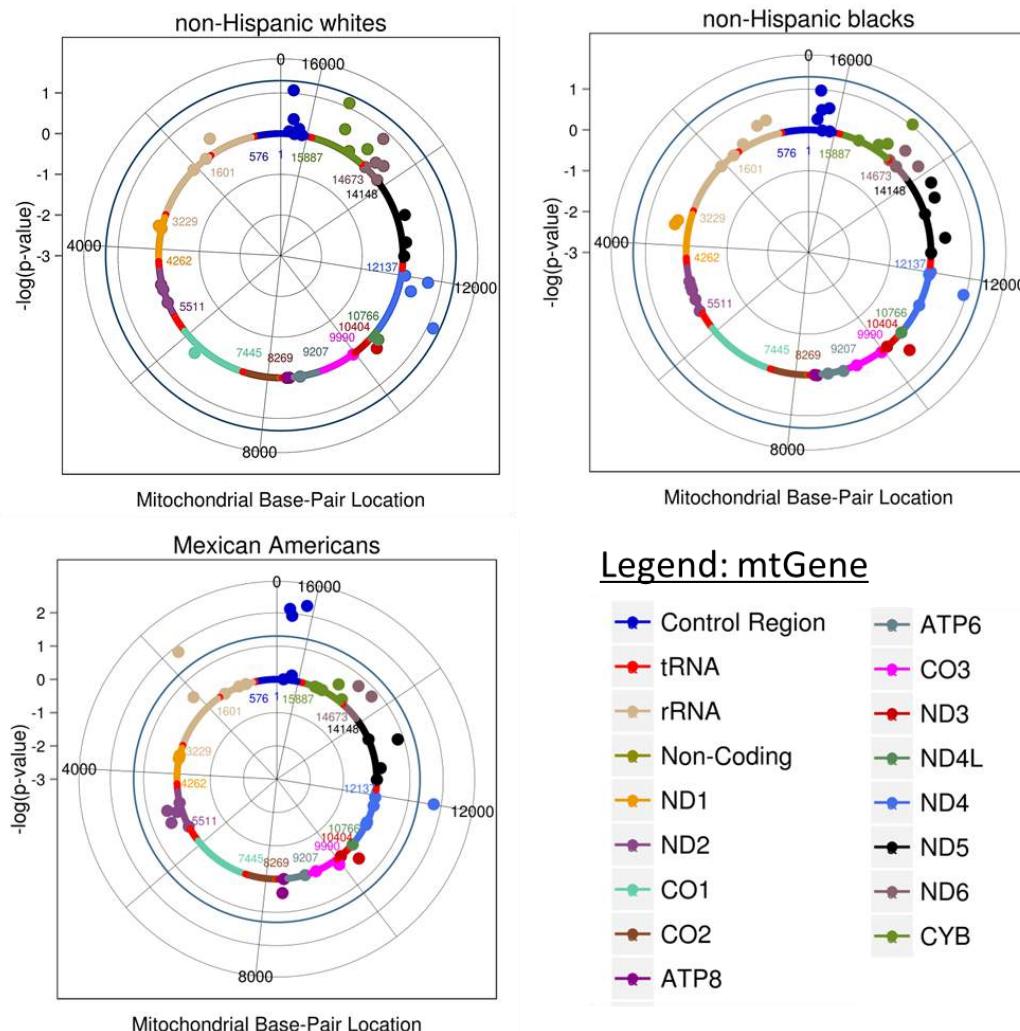


Figure 1: Meta-analysis single SNP association results by race/ethnicity. Log(p) values were plotted using R. The outer blue circle represents a significance threshold of $p = 0.05$. SNPs are color coded by mitochondrial gene/regions as denoted in the legend. Cases/controls for the three populations are as follows: non-Hispanic whites (case = 312, control = 1,349), non-Hispanic blacks (case = 37, control = 430), and Mexican Americans (case = 67, control = 421).

7. References

- National Alliance for Eye and Vision Research. *The Silver Book: Vision Loss: Chronic Disease and Medical Innovationn in an Aging nation.* (2006). at <<http://www.eyeresearch.org/pdf/VisionLossSilverbook.pdf>>
- Lions Clubs International Foundation, N. E. I. 2005 Survey of public knowledge, attitudes, and practices related to eye health and disease. (2007). at <<http://www.nei.nih.gov/nehep/kap>>

3. David S. Friedman, O'Colmain, B. and Ilona Mestril. 2012 Fifth Edition of Vision Problems in the U.S. (2012). at <<http://www.visionproblemsus.org/introduction/acknowledgments.html>>
4. Luo, D.-G., Xue, T. and Yau, K.-W. How vision begins: an odyssey. *Proc. Natl. Acad. Sci. U. S. A.* **105**, 9855–9862 (2008).
5. Risk factors associated with age-related macular degeneration. A case-control study in the age-related eye disease study: Age-Related Eye Disease Study Report Number 3. *Ophthalmology* **107**, 2224–2232 (2000).
6. Chakravarthy, U. *et al.* Clinical risk factors for age-related macular degeneration: a systematic review and meta-analysis. *BMC Ophthalmol.* **10**, 31 (2010).
7. Schmidt, S. *et al.* Cigarette smoking strongly modifies the association of LOC387715 and age-related macular degeneration. *Am. J. Hum. Genet.* **78**, 852–864 (2006).
8. Schaumberg, D. A., Hankinson, S. E., Guo, Q., Rimm, E. and Hunter, D. J. A prospective study of 2 major age-related macular degeneration susceptibility alleles and interactions with modifiable risk factors. *Arch. Ophthalmol.* **125**, 55–62 (2007).
9. Neuner, B. *et al.* LOC387715, smoking and their prognostic impact on visual functional status in age-related macular degeneration-The Muenster Aging and Retina Study (MARS) cohort. *Ophthalmic Epidemiol.* **15**, 148–154 (2008).
10. Reynolds, R., Rosner, B. and Seddon, J. M. Serum lipid biomarkers and hepatic lipase gene associations with age-related macular degeneration. *Ophthalmology* **117**, 1989–1995 (2010).
11. Van Leeuwen, R. *et al.* Cholesterol and age-related macular degeneration: is there a link? *Am. J. Ophthalmol.* **137**, 750–752 (2004).
12. Klein, R., Klein, B. E. K., Tomany, S. C. and Cruickshanks, K. J. The association of cardiovascular disease with the long-term incidence of age-related maculopathy: the Beaver Dam Eye Study. *Ophthalmology* **110**, 1273–1280 (2003).
13. Klein, R. *et al.* The prevalence of age-related macular degeneration and associated risk factors. *Arch. Ophthalmol.* **128**, 750–758 (2010).
14. Kawasaki, R. *et al.* The Prevalence of Age-Related Macular Degeneration in Asians: A Systematic Review and Meta-Analysis. *Ophthalmology* **117**, 921–927 (2010).
15. Klein, R. *et al.* Prevalence of age-related macular degeneration in the US population. *Arch. Ophthalmol.* **129**, 75–80 (2011).
16. Friedman, D. S. *et al.* Prevalence of age-related macular degeneration in the United States. *Arch. Ophthalmol.* **122**, 564–572 (2004).
17. Haines, J. L. *et al.* Complement factor H variant increases the risk of age-related macular degeneration. *Science* **308**, 419–421 (2005).
18. Klein, R. J. *et al.* Complement factor H polymorphism in age-related macular degeneration. *Science* **308**, 385–389 (2005).
19. Edwards, A. O. *et al.* Complement factor H polymorphism and age-related macular degeneration. *Science* **308**, 421–424 (2005).
20. Jakobsdottir, J. *et al.* Susceptibility genes for age-related maculopathy on chromosome 10q26. *Am. J. Hum. Genet.* **77**, 389–407 (2005).

- 21.Rivera, A. *et al.* Hypothetical LOC387715 is a second major susceptibility gene for age-related macular degeneration, contributing independently of complement factor H to disease risk. *Hum. Mol. Genet.* **14**, 3227–3236 (2005).
- 22.Gold, B. *et al.* Variation in factor B (BF) and complement component 2 (C2) genes is associated with age-related macular degeneration. *Nat. Genet.* **38**, 458–462 (2006).
- 23.Fritsche, L. G. *et al.* Age-Related Macular Degeneration: Genetics and Biology Coming Together. *Annu. Rev. Genomics Hum. Genet.* **15**, 151–171 (2014).
- 24.Kenney, M. C. *et al.* Molecular and bioenergetic differences between cells with African versus European inherited mitochondrial DNA haplogroups: implications for population susceptibility to diseases. *Biochim. Biophys. Acta* **1842**, 208–219 (2014).
- 25.Cristina Kenney, M. *et al.* Inherited mitochondrial DNA variants can affect complement, inflammation and apoptosis pathways: insights into mitochondrial-nuclear interactions. *Hum. Mol. Genet.* **23**, 3537–3551 (2014).
- 26.Liang, F.-Q. and Godley, B. F. Oxidative stress-induced mitochondrial DNA damage in human retinal pigment epithelial cells: a possible mechanism for RPE aging and age-related macular degeneration. *Exp. Eye Res.* **76**, 397–403 (2003).
- 27.Cano, M. *et al.* Oxidative stress induces mitochondrial dysfunction and a protective unfolded protein response in RPE cells. *Free Radic. Biol. Med.* **69**, 1–14 (2014).
- 28.Sheu, S.-J. *et al.* Resveratrol stimulates mitochondrial bioenergetics to protect retinal pigment epithelial cells from oxidative damage. *Invest. Ophthalmol. Vis. Sci.* **54**, 6426–6438 (2013).
- 29.Kenney, M. C. *et al.* Mitochondrial DNA variants mediate energy production and expression levels for CFH, C3 and EFEMP1 genes: implications for age-related macular degeneration. *PloS One* **8**, e54339 (2013).
- 30.Jones, M. M. *et al.* Mitochondrial DNA haplogroups and age-related maculopathy. *Arch. Ophthalmol.* **125**, 1235–1240 (2007).
- 31.Udar, N. *et al.* Mitochondrial DNA Haplogroups Associated with Age-Related Macular Degeneration. *Invest. Ophthalmol. Vis. Sci.* **50**, 2966–2974 (2009).
- 32.SanGiovanni, J. P. *et al.* Mitochondrial DNA variants of respiratory complex I that uniquely characterize haplogroup T2 are associated with increased risk of age-related macular degeneration. *PloS One* **4**, e5508 (2009).
- 33.Mueller, E. E. *et al.* Mitochondrial Haplogroups and Control Region Polymorphisms in Age-Related Macular Degeneration: A Case-Control Study. *PLoS ONE* **7**, e30874 (2012).
- 34.Canter, J. A. *et al.* Mitochondrial DNA Polymorphism A4917G Is Independently Associated with Age-Related Macular Degeneration. *PLoS ONE* **3**, e2091 (2008).
- 35.US Department of Health and Human Services (DHHS). Centers for Disease Control and Prevention Third National Health and Nutrition Examination Survey, 1988-94, Plan and Operations Procedures Manuals[CD-ROM]National Center for Health Statistics, Centers for Disease Control and Prevention, Hyattsville, MD. (1996).
- 36.Centers for Disease Control and Prevention Plan and Operation of the Third National Health and Nutrition Examination Survey, 1988-94 Bethesda, MD. (2004).
- 37.Steinberg, K. K. *et al.* DNA banking in epidemiologic studies. *Epidemiol. Rev.* **19**, 156–162 (1997).

- 38.Chang, M.-H. *et al.* Prevalence in the United States of selected candidate gene variants: Third National Health and Nutrition Examination Survey, 1991-1994. *Am. J. Epidemiol.* **169**, 54–66 (2009).
- 39.Mitchell, S. L. *et al.* Characterization of mitochondrial haplogroups in a large population-based sample from the United States. *Hum. Genet.* **133**, 861–868 (2014).
- 40.Klein, R. *et al.* The Wisconsin age-related maculopathy grading system. *Ophthalmology* **98**, 1128–1134 (1991).
- 41.Tilleul, J. *et al.* Genetic association study of mitochondrial polymorphisms in neovascular age-related macular degeneration. *Mol. Vis.* **19**, 1132–1140 (2013).
- 42.Sadigh, S. *et al.* Drusen and Photoreceptor Abnormalities in African-Americans with Intermediate Non-neovascular Age-related Macular Degeneration. *Curr. Eye Res.* 1–9 (2014).
- 43.Spencer, K. L., Glenn, K., Brown-Gentry, K., Haines, J. L. and Crawford, D. C. Population differences in genetic risk for age-related macular degeneration and implications for genetic testing. *Arch. Ophthalmol.* **130**, 116–117 (2012).
- 44.Friedman, D. S., Katz, J., Bressler, N. M., Rahmani, B. and Tielsch, J. M. Racial differences in the prevalence of age-related macular degeneration: the Baltimore Eye Survey. *Ophthalmology* **106**, 1049–1055 (1999).
- 45.Klein, R. *et al.* Subclinical atherosclerotic cardiovascular disease and early age-related macular degeneration in a multiracial cohort: the Multiethnic Study of Atherosclerosis. *Arch. Ophthalmol.* **125**, 534–543 (2007).
- 46.Buentello-Volante, B. *et al.* Susceptibility to advanced age-related macular degeneration and alleles of complement factor H, complement factor B, complement component 2, complement component 3, and age-related maculopathy susceptibility 2 genes in a Mexican population. *Mol. Vis.* **18**, 2518–2525 (2012).
- 47.Contreras, A. V. *et al.* CFH haplotypes and ARMS2, C2, C3, and CFB alleles show association with susceptibility to age-related macular degeneration in Mexicans. *Mol. Vis.* **20**, 105–116 (2014).
- 48.Fagundes, N. J. R. *et al.* Mitochondrial population genomics supports a single pre-Clovis origin with a coastal route for the peopling of the Americas. *Am. J. Hum. Genet.* **82**, 583–592 (2008).
- 49.Chen, J. Z., Gokden, N., Greene, G. F., Mukunyadzi, P. & Kadlubar, F. F. Extensive somatic mitochondrial mutations in primary prostate cancer using laser capture microdissection. *Cancer Res.* **62**, 6470–6474 (2002).
- 50.Zhang, J. *et al.* Identification of sequence polymorphisms in the displacement loop region of mitochondrial DNA as a risk factor for renal cell carcinoma. *Biomed. Rep.* **1**, 563–566 (2013).
- 51.Cheng, M. *et al.* Identification of sequence polymorphisms in the mitochondrial displacement loop as risk factors for sporadic and familial breast cancer. *Tumour Biol. J. Int. Soc. Oncodevelopmental Biol. Med.* **35**, 4773–4777 (2014).

**IPINBPA: AN INTEGRATIVE NETWORK-BASED FUNCTIONAL MODULE
DISCOVERY TOOL FOR GENOME-WIDE ASSOCIATION STUDIES**

LILI WANG

*School of Computing, Queen's University
25 Union Street, Goodwin Hall, Kingston, Ontario, K7L 3N6, Canada
Email: lili@cs.queensu.ca*

PARVIN MOUSAVI

*School of Computing, Queen's University
25 Union Street, Goodwin Hall, Kingston, Ontario, K7L 3N6, Canada
Email: pmousavi@cs.queensu.ca*

SERGIO E. BARANZINI

*Department of Neurology, University of California San Francisco
675 Nelson Rising Lane, Room 215, San Francisco, CA 94158, USA
Email: sebaran@cgl.ucsf.edu*

We introduce the integrative protein-interaction-network-based pathway analysis (iPINBPA) for genome-wide association studies (GWAS), a method to identify and prioritize genetic associations by merging statistical evidence of association with physical evidence of interaction at the protein level. First, the strongest associations are used to weight all nodes in the PPI network using a *guilt-by-association* approach. Second, the gene-wise converted p-values from a GWAS are integrated with node weights using the Liptak-Stouffer method. Finally, a greedy search is performed to find enriched modules, *i.e.*, sub-networks with nodes that have low p-values and high weights. The performance of iPINBPA and other state-of-the-art methods is assessed by computing the concentrated receiver operating characteristic (CROC) curves using two independent multiple sclerosis (MS) GWAS studies and one recent ImmunoChip study. Our results showed that iPINBPA identified sub-networks with smaller sizes and higher enrichments than other methods. iPINBPA offers a novel strategy to integrate topological connectivity and association signals from GWAS, making this an attractive tool to use in other large GWAS datasets.

1. Introduction

In the last decade, Genome-wide association studies (GWAS) have been a powerful tool to identify statistically significant differences in allelic frequencies between cases and controls at each tested single nucleotide polymorphism (SNP) for hundreds of phenotypes.¹ In order to consider a signal of genome-wide significance, a Bonferroni correction is usually applied ($p\text{-value} < 5 \times 10^{-8}$ for 1 million markers) under the assumption of independence among SNPs.² While this method ensues a low ratio of false positives, it inevitably increases the ratio of false negatives, thus neglecting a sizable proportion of risk SNPs and limiting the overall utility of this approach. Furthermore, the results of GWAS do not directly provide any functional information of the variants. Recent advances in our understanding of biological networks, especially the large-scale human protein interaction network (PIN), have enabled its use to investigate statistically modest associations in GWAS in the context of functional modules (referred to as pathways) to elucidate the underlying molecular mechanisms of several human diseases.³⁻⁵

Pathway analysis approaches can be divided into three broad classes. The first class of methods attempts to compute the over-representation of a given list of genes in gene ontology (GO) or pre-computed pathway databases (*e.g.*, KEGG, Biocarta, *etc.*). Examples include DAVID⁶ and INRICH.⁷ The second class of methods involve functional class scoring (FCS) approaches, such as GenGen,⁸ SSEA⁹ and PARIS.¹⁰ The input data to these tools is SNP-based statistics, such as *p*-values. FCS methods aggregate SNP-wise statistics into a single score for each pre-defined pathway. While potentially revealing, both the first and second classes of methods ignore the functional connections between genes and assume independence between pathways. A third class is composed of network-based analyses, and largely overcomes the assumption of pathway independence. These approaches commonly use a scaffold of protein interactions to build connections between gene products, where nodes represent proteins and edges represent physical or functional interactions between pairs of proteins. Rather than focusing on individual markers, network-based analysis methods take into account multiple loci in the context of molecular networks. Due to this critical feature, these methods can afford to use sub-genome-wide statistical significance and yet increase the power to detect new associations and functional relationships between genes in complex traits.⁵

To date, several network-based methods have been proposed to identify functional modules in the form of sub-networks of a given larger ensemble. For example, protein interaction network-based pathway analysis (PINBPA) of GWAS data was developed to identify over-represented modules in a large multiple sclerosis (MS) GWAS.⁵ This approach, adapted from a similar method for gene expression analysis, uses a greedy algorithm¹¹ to identify sub-networks based on aggregated gene-wise statistics. Dense module searching of GWAS (dmGWAS) also extensively searches for sub-networks enriched with low p-value genes in GWAS datasets.¹² Aside from using the human PIN as a scaffold, neither PINBPA nor dmGWAS exploit topological properties of this network. Another tool, called network interface miner for multigenic interactions (NIMMI),¹³ combines topological connectivity with association signals from GWAS. In this tool, sub-networks are generated for each node by adding their neighbors up to the second-order. Nodes are weighted using a modified Google PageRank algorithm, and then the pre-generated sub-networks are scored.¹³ More recently, the disease association protein-protein link evaluator (DAPPLE)¹⁴ was reported to prioritize novel associations in Crohn's disease and rheumatoid arthritis datasets. Using a fixed PIN,

DAPPLE builds direct and indirect networks among a list of genes or SNPs and computes the probabilities that those connections may have arisen by chance. However, DAPPLE does not take into account gene-wise or SNP-wise GWAS statistics.

In this paper, we introduce the integrative protein-interaction-network based pathway analysis (iPINBPA), a novel network-based pathway analysis strategy. This approach, based on the same principles of PINBPA, integrates topological connectivity among genes in PIN space and the association signals from GWAS to extensively search for sub-networks enriched in significant GWAS signals. We tested the performance of iPINBPA against PINBPA and dmGWAS using two independent well-powered datasets in multiple sclerosis (MS). Our results show that iPINBPA can identify sub-networks involving MS genes with much higher precision than the other tested methods.

2. Data and Methods

2.1. Data

The data used in this work include two independent GWAS data sets in MS, human PIN and benchmark MS genes for evaluation of our proposed method.

2.1.1. GWAS data sets

Two large-scale GWAS data sets have been used to evaluate the proposed method. The first data set is a meta-analysis (denoted as *Meta2.5*) of seven independent moderately powered GWAS and one meta-analysis and includes 137,432 SNPs mapped to 17,425 unique genes for 5,545 cases and 12,153 controls.¹⁵ The second data set is the largest GWAS in MS to date (denoted as *WTCCC2*), and is composed of 137,457 SNPs mapped to 17,401 unique genes in 9,772 cases and 17,376 controls.¹⁶ For both data sets, the SNP-wise statistical significance (SNP-wise p-value), was transformed into gene-wise significance (gene-wise p-value), using the versatile gene-based association study (VEGAS) tool.¹⁷ If the gene-wise $p\text{-value} \leq 0.05$ in GWAS data set, the corresponding gene is defined as nominally significant. There are 1,982 unique nominally significant genes in *WTCCC2* and 1,690 in the *Meta2.5* data set.

2.1.2. Human protein interaction network

We used a high-confidence, manually curated human protein interaction network previously reported,⁵ which is composed of 8,960 proteins and 27,724 interactions. The PIN is represented as an undirected graph.

2.1.3. Benchmarking

We assess the performance of our proposed method for two sets of benchmark genes: 1) the 45 genes emerging from GWAS as of 2011 (denoted as *WTCCC2* genes), and 2) the 135 genes from the most recent MS study, the ImmunoChip custom genotyping array¹⁸ (denoted as *iChip* genes).

2.2. Method

Given a human protein interaction network, and gene-wise p-values from a GWAS data set, iPINBPA detects enriched sub-networks in three steps as shown in Fig. 1. First, for a given GWAS data set, the nominally significant genes (gene-wise $p\text{-value} \leq 0.05$) are selected as seed genes, and then nodes in the network are weighted (*i.e.* network smoothing) via the random walk with restart algorithm according to their connectivity to seed genes based on *guilt-by-association*. Second, a network score is defined by the combination of the gene-wise p-values with node weights using the Liptak-Stouffer method.¹⁹ The background distribution for the network score is calculated using random sampling for various network sizes. Finally, a heuristic algorithm extensively searches for modules enriched in genes with low p-values and high weights, *i.e.*, high network score. We will explain each step in detail in the following subsections.

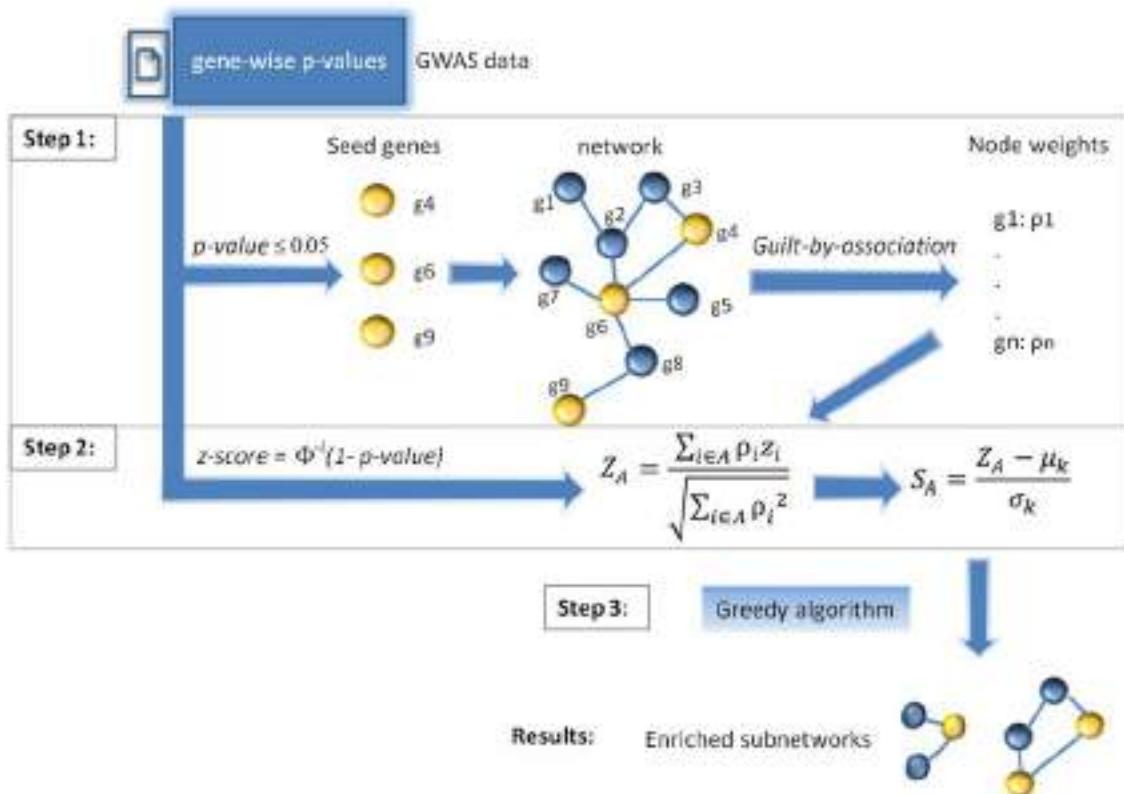


Fig. 1 Work flow of iPINBPA

2.2.1. Random walk with restart

Based on the assumption of *guilt-by-association*, Köhler *et al.*²⁰ developed a random walk with restart method to prioritize disease-associated genes. In this method, a walker starts moving from a seed node to connecting neighbors randomly. Nodes in the network are scored according to the probabilities of the walker reaching them at the end of the process. iPINBPA extends Köhler's approach by weighting the edge e_{ij} connecting n_i and n_j using corresponding gene-wise p-values as: $W_{ij} = ((1 - p_i) + (1 - p_j))/2$, where p_i and p_j are gene-wise p-values of n_i and n_j , and normalizing the adjacency matrix W by its columns. A score vector is calculated after each step of the walker as follows:

$$P(t) = (1 - r)W \cdot P(t - 1) + rP(0) \quad (1)$$

where $P(t)$ is the score after t steps of walking, and r is the restart ratio. The initial score vector $P(0)$ represents *a-priori* knowledge of genes (in our case, nominal significance), where 1 is assigned for seed genes and 0 for the rest. Finally, all nodes are scored according to their values in the vector $P(T)$, which quantitatively measures the topological connection to seed genes.

As mentioned above, iPINBPA requires a group of seed genes to start random walks. In this study, we used the nominally significant genes (gene-wise $p\text{-value} \leq 0.05$) in the GWAS data set as seed genes. This network smoothing step refines the searching for enriched sub-networks, as nominally significant genes will be assigned higher scores than the non-significant ones.

2.2.2. Network scoring

In the second step of our approach, each p-value p_i is transformed into its standard normal deviate z_i using the inverse normal CDF: $z_i = \Phi^{-1}(1 - p_i)$, and then a score for a network A containing k nodes is defined using a weighted Z transform test¹⁹ (also called Liptak-Stouffer formula), as shown in Equation (2). By using this formula, the gene-wise significance from GWAS is combined with node connectivity to known disease genes. According to this algorithm, nodes with low $p\text{-value}$ and close to known associated genes will score higher.

$$Z_A = \frac{\sum_{i \in A} P(T)_i z_i}{\sqrt{\sum_{i \in A} P(T)_i^2}} \quad (2)$$

To determine the significance of the network score calculated above, we performed a random sampling¹¹ of gene sets of size $k \in [1, 500]$ for 1000 times. For gene sets at size k , we computed their scores Z_A , then calculated the mean of network score μ_k and the standard deviation σ_k . The adjusted network score is defined as:

$$S_A = \frac{Z_A - \mu_k}{\sigma_k} \quad (3)$$

2.2.3. Greedy algorithm

The last step of iPINBPA is to find locally optimal sub-networks according to the adjusted network scores. A greedy algorithm starts searching for the optimal sub-network G for each node v_{start} in the network. It searches all neighbors of G as long as their shortest path to v_{start} is less than or equal to 2, if adding a neighbor increases the network score S_G , then add the neighbor with the largest increase. It stops adding until there is no increasing of S_G . Then it starts searching any node inside G as long as this node is not v_{start} and removable, which means G is still a connected sub-network after removing this node. If removing a node will increase S_G , then remove the one with the largest increase. The algorithm stops searching until there is no increasing of S_G . The pseudo code of this algorithm is as follows:

(1) $G \leftarrow \{v_{start}\}$

- (2) For each neighbor node v of G and depth ≤ 2 :
- (3) Calculate score S'_G if add v into G
- (4) If $\max(S'_G) > S_G$ then:
 - (5) Add the corresponding node v_{max} into G
 - (6) Go back to step (2)
 - (7) Else:
 - (8) For each node v in G except v_{start} :
 - (9) Calculate score S''_G if remove v from G
 - (10) If $\max(S''_G) > S_G$ then:
 - (11) If the corresponding node v'_{max} is removable:
 - (12) Remove v'_{max} from G
 - (13) Go back to step (8).
 - (14) Else:
 - (15) Return G

2.2.4. Parameters

We chose the network's characteristic path length (4.38 for the used PIN) as the default time step ($T = 5$). The second parameter of random walk is the restart ratio r , which weights prior knowledge. As there is no standard criterion to select the restart ratio, we set up the default value as 0.5. Furthermore, we tested iPINBPA with different restart ratios and the corresponding performance is discussed in section 3.4.

2.2.5. Evaluation

To evaluate the performance of iPINBPA, we tested two sets of reported benchmark genes. As shown in a previous study,⁵ if we define association regions (blocks) composed of significant genes (gene-wise p -value ≤ 0.05), there are 665 association blocks containing 1,982 unique genes in the WTCCC2 data set, and 612 blocks containing 1,690 unique genes in the Meta2.5 data set. The size of associated blocks vary from 1 to >100 , thus posing a challenge to quantitatively compare the prediction or enrichment performance for each association block. In this study, we applied iPINBPA to identify sub-networks in a high-confidence PIN and a GWAS data set, and ranked genes using their highest network score in descending order. For genes having the same network score, they were ranked by their gene-wise p -values in ascending order. Based on the ranking, CROC curves²¹ were computed to assess the efficiency of iPINBPA in identifying the benchmark genes. CROC curves use an exponential function ($f(x) = (1 - e^{-\alpha x})/(1 - e^{-\alpha})$) to magnify any relevant portion of the corresponding ROC by an appropriate continuous transformation of the coordinates. CROC curves have been shown to be more effective than ROC curves to measure the ability of methods in drug discovery and gene prediction.²¹ In our case, the early retrieval performance is also adequate, as we only consider the top scored/ranked nodes or sub-networks, and the size of benchmark genes is less than one percent of the total number of genes in the network.

3. Results

Based on its predecessor (PINBPA), iPINBPA introduces node-weighting by means of significant disease-related genes and integrates these weights with gene-based significance into a score, which is further

normalized by network size (see methods). We applied the iPINBPA approach to two independent large-scale GWAS datasets in multiple sclerosis (MS) (*Meta2.5* and *WTCCC2*), and benchmarked its performance against other established methods on the same input data.

In pathway analysis of GWAS, it is necessary to compute gene-wise (rather than SNP-wise) significance. Given that most associations fall outside coding regions, allocating a significant finding to a gene is not always straightforward. One common strategy is to assign the significant association to the closest gene, taking into account recombination hotspots. However, due to linkage disequilibrium (LD), it is not unusual to find that several genes map within the “area of influence” of the lead SNP. While usually the closest gene to the lead SNP is assigned, in reality, patterns of extended LD make it impossible to assign any given gene within that area with certainty.

It is challenging to compare different pathway analysis methods because of the lack of accurate knowledge of complex traits and the incomplete human PIN. Since DAPPLE and NIMMI do not accept a user-defined network and DAPPLE only accepts a short list of SNPs or genes (up to 500), it is not possible to directly compare these methods to iPINBPA. Thus, we compared iPINBPA to PINBPA and to dmGWAS. We performed three different tests: (1) Prediction of *WTCCC2* genes using *Meta2.5* data; (2) Prediction of *iChip* genes using *WTCCC2* data; and (3) Significantly enriched networks from both GWAS data sets.

3.1. Prediction of *WTCCC2* genes using *Meta2.5* data

We first tested the ability of each method to identify the *WTCCC2* genes using *Meta2.5* data. There are 45 *WTCCC2* genes previously identified in GWAS studies (24 of them are represented in our network). *Meta2.5* data are aggregated from seven moderately powered GWAS and one meta-analysis before the completion of *WTCCC2*. *Meta2.5* GWAS data set contains weaker association signals than *WTCCC2* GWAS data set (26 SNPs with $p\text{-value} < 5 \times 10^{-8}$ and 1,690 nominally significant genes in *Meta2.5*, but 57 associated SNPs and 1,982 nominally significant genes in *WTCCC2*).

We measured the fold enrichment of AUC score of each method compared to a random classifier. As shown in Fig. 2A, iPINBPA (*fold enrichment*= 5.858) performs marginally better than PINBPA (*fold enrichment*= 5.386) and significantly better than dmGWAS (*fold enrichment* = 3.646), with $\alpha = 14, f(0.05) = 0.5$.

3.2. Predicting *iChip* genes using *WTCCC2* data

We also tested the ability of each method to identify the latest MS genes reported in a recent study using the ImmunoChip (*iChip*) custom genotyping array¹⁸ using *WTCCC2*. In this test, a total of 135 genes were associated with MS (Although the total number of reported associated loci is 110, some SNPs map to more than one gene). Of these 135 *iChip* MS genes, 42 genes were *WTCCC* genes (23 of them are represented in our network), and thus 93 genes (54 genes represented in our network) found in *iChip* are novel. As shown in Fig. 2B, iPINBPA (*fold enrichment* = 6.22) performs better than PINBPA (*fold enrichment* = 5.211) and dmGWAS (*fold enrichment* = 2.818) in the prediction of *iChip* genes, with $\alpha = 14, f(0.05) = 0.5$.

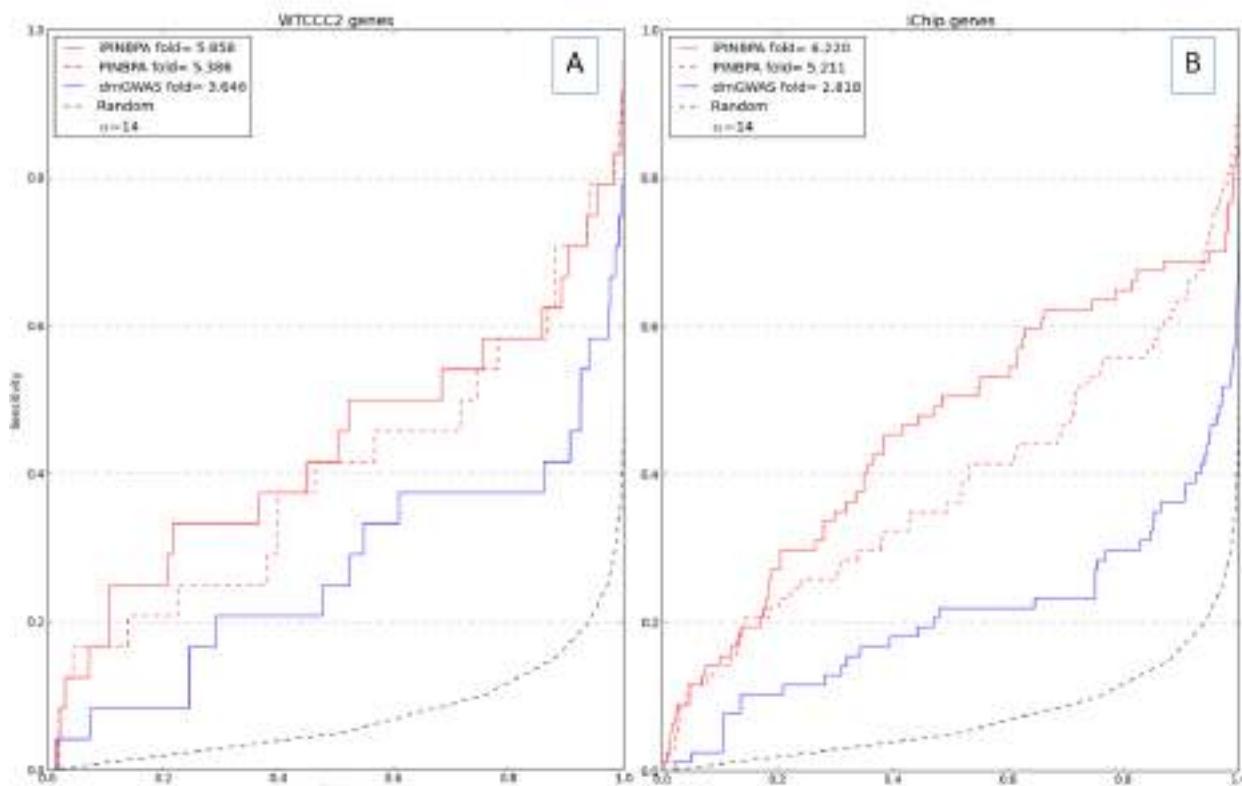


Fig. 2 CROC curves of Meta2.5 and WTCCC GWAS data sets

3.3. Significantly enriched networks

As the primary goal of our approach is to identify the enriched pathways for the given GWAS data set, we selected the top scored sub-networks ($score > 3$ and $size \geq 5$) from each method. For this analysis we also tested NIMMI, which returns sub-networks with p -values. For NIMMI, the sub-networks with p -value < 0.0013 (equivalent z-score to the other methods) were selected. As shown in Table 1, iPINBPA is more sensitive to GWAS signals and identifies smaller networks, resulting in higher precision. By overlapping the selected networks from both *WTCCC2* and *Meta2.5*, iPINBPA identified 1,299 genes (including 17 *WTCCC2* genes and 44 *iChip* genes), PINBPA identified 5,047 genes (including 23 *WTCCC2* genes and 69 *iChip* genes), dmGWAS identified 7,634 genes (including 24 *WTCCC2* genes and 77 *iChip* genes). NIMMI identified 4,832 genes (including 19 *WTCCC2* genes and 49 *iChip* genes). Altogether, iPINBPA achieved the highest precision for both sets of benchmark genes.

To evaluate the biological significance of the 1,299 candidate associated genes reported by iPINBPA, we tested their functional annotation clustering using the online tool DAVID. The KEGG pathways in the cluster with the highest enrichment score (8.94) are listed in Table 2. While the precise etiology of MS is still unclear, it has been consistently described as a T-cell-mediated autoimmune disease. As such, it is not surprising that related KEGG pathways such as allograft rejection, type 1 diabetes mellitus, graft-versus-host disease, and thyroid disease are significantly enriched. This result suggests that genes prioritized by iPINBPA are consistent with the biological functions likely implicated in MS pathogenesis.

Table 1. Stats of top scored sub-networks from iPINBPA, PINBPA and dmGWAS

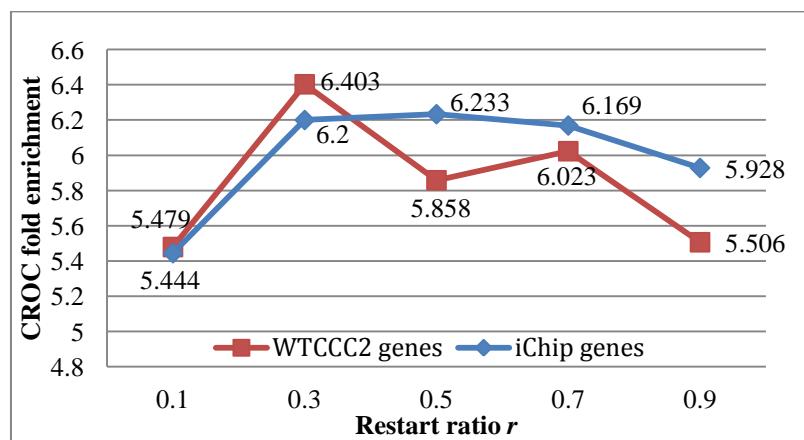
	iPINBPA		PINBPA		dmGWAS		NIMMI	
GWAS data set	WTCCC2	Meta2.5	WTCCC2	Meta2.5	WTCCC2	Meta2.5	WTCCC2	Meta2.5
# networks	1496	1295	4079	4080	7109	7000	402	400
# total nodes	2163	1938	6012	6079	7665	7643	4950	4979
# overlap of nodes	1299		5047		7634		4832	
Precision (# WTCCC2 genes)	0.013 (17)		0.005 (23)		0.003 (24)		0.004 (19)	
Precision (# iChip genes)	0.034 (44)		0.014 (69)		0.01 (77)		0.01 (49)	

Table 2. Functional annotation clusters of 1299 genes selected from iPINBPA in DAVID

KEGG Pathway	Count	P-Value	Benjamini
Allograft rejection	24	4.7E-12	3.5E-11
Type I diabetes mellitus	24	4.0E-10	2.1E-9
Graft-versus-host disease	22	3.5E-9	1.6E-8
Autoimmune thyroid disease	23	2.6E-7	9.8E-7

3.4. Tuning parameters

The restart ratio r in random walk with restart can be tuned by the user. We tested iPINBPA with different restart ratios (0.1, 0.3, 0.5, 0.7, and 0.9) and evaluated its performance as shown in Fig. 3.

Fig. 3 CROC fold enrichments of different values of restart ratio r

In addition, we also tested iPINBPA with different cutoffs of selecting seed genes to start random walks, which controls the sensitivity of iPINBPA indirectly. By default, we used the nominally significant genes (gene-wise $p\text{-value} \leq 0.05$). If a more stringent cutoff is used to select fewer number of seed genes, iPINBPA usually returns smaller sub-networks. Table 3 shows the sizes and precision of top selected sub-networks from iPINBPA with different cutoffs.

Table 3. Stats of top selected sub-networks from iPINBPA with different cutoffs

GWAS data set	$p\text{-value} \leq 0.01$		$p\text{-value} \leq 0.005$		$p\text{-value} \leq 0.001$	
	WTCCC2	Meta2.5	WTCCC2	Meta2.5	WTCCC2	Meta2.5
# networks	1732	1224	1691	1293	1547	1458
# total nodes	2108	1522	2000	1535	1774	1617
Mean of network size (node) (std)	17.25 (15.49)	12.71 (8.78)	12.6 (11.12)	10.24 (4.36)	9.95 (4.52)	7.7 (2.32)
Mean of network size (edge) (std)	26.91 (34.93)	16.22 (16.87)	16.21 (23.8)	12.38 (7.73)	10.68 (6.6)	8.37 (3.67)
# overlap of nodes	1133		1106		1082	
Precision (# WTCCC2 genes)	0.014 (16)		0.013 (14)		0.01 (11)	
Precision (# iChip genes)	0.036 (41)		0.034 (38)		0.028 (30)	

4. Discussion

GWAS have been extremely successful in identifying thousands of associations in hundreds of complex traits. Due to the extensive statistical adjustments needed to avoid type 1 errors, type 2 errors are necessarily a consequence of GWAS studies, thus limiting their effectiveness. Furthermore, typically, only a few markers are replicated in any given GWAS. Effective post-GWAS analysis methods can help prioritize associations using additional sources of evidence and are becoming a useful complementary strategy to the standard analytical pipeline.

Here we introduced a novel network-based pathway analysis strategy for GWAS, which integrates topological connectivity in a PIN and the association signals from GWAS to detect significant sub-networks and also prioritize genes associated with a complex disease. The main feature of iPINBPA is the strategy we employed to identify enriched sub-networks by merging evidence from multiple sources. To our knowledge, this is the first method that integrates node weighting with a greedy search for significant sub-networks. Comparisons with different data sets and methods have demonstrated that our integrative approach dramatically improves the performance in predicting novel associations. The increase of prediction precision comes mostly from the fact that, unlike in the classical approach, potential associations with no biological relationships to statistically confirmed associations are down-weighted in this approach.

Given the multi-dimensional nature of GWAS data, it is not uncommon to see a low precision in prioritizing novel associations through network-based pathway analysis. The identified sub-networks presented here are around nodes with quite significant p-values, thus the overlap of these sub-networks lends additional support to our methods. By incorporating additional information (*e.g.*, regulatory, cell-specific expression, *etc.*), the precision of network-based pathway analysis would be improved gradually.

Unlike dmGWAS, iPINBPA and PINBPA use VEGAS to map SNPs to genes. For the analysis of GWAS data, the mapping of SNPs to genes is an open challenge. In this paper, we focus on the comparison of methodology and performance of different network-based analysis methods. We did not address potential variations emerging from using different strategies of mapping SNPs to genes; the default mapping recommended for each method was utilized.

An inherent limitation of all approaches using protein networks, is that interactions have only been described for a subset of all known proteins. Furthermore, if only high confidence interactions are taken into account as described in this study, approximately only half of all proteins are represented in the network. This necessarily places an upper boundary to the number of successful predictions any of these methods can make. With new and more accurate techniques to determine protein interactions, this limitation may be overcome in the near future. Another potential restriction of these methods is that they use global interactions, when actually tissue specific interactions might be more appropriate. Several efforts are currently underway to develop tissue-specific protein interactions that, together with knowledge about the organ/tissue compromised in a given disease, could be incorporated into network analysis of GWAS in the future. Furthermore, with the incorporation of genome-wide regulatory data (*e.g.*, ENCODE, Epigenomics Roadmap, *etc.*), it will be possible to derive cell specific networks. This will greatly enhance the performance of this approach, as it will enable the incorporation of pathophysiologically relevant and disease-specific data.

The integrative strategy we proposed in this study is generic can be readily applied to any disease or biological datasets, *e.g.*, gene expression datasets and proteomic data, as long as quantitative gene-wise or protein-wise statistical measures and putative disease genes are available.

5. Acknowledgements

This work was partially supported by Natural Sciences and Engineering Research Council of Canada, the Ontario Early Researcher Award, and grants from the National Multiple Sclerosis Society (R01NS049477). SEB is a Harry Weaver Neuroscience Fellow of the National Multiple Sclerosis Society.

References

- 1 T. A. Manolio, *N. Engl. J. Med.*, **363**, 166-176 (2010).
- 2 G. S. Barsh, G. P. Copenhaver, G. Gibson and S. M. Williams, *PLoS Genet.*, **8**, e1002812 (2012).
- 3 K. Wang, M. Li and H. Hakonarson, *Nat. Rev. Genet.*, **11**, 843-854 (2010).
- 4 S. E. Baranzini, N. W. Galwey, J. Wang, et al, *Hum. Mol. Genet.*, **18**, 2078-2090 (2009).
- 5 International Multiple Sclerosis Genetics Consortium, *Am. J. Hum. Genet.*, (2013).
- 6 W. Huang da, B. T. Sherman and R. A. Lempicki, *Nucleic Acids Res.*, **37**, 1-13 (2009).
- 7 P. H. Lee, C. O'Dushlaine, B. Thomas and S. M. Purcell, *Bioinformatics*, **28**, 1797-1799 (2012).
- 8 K. Wang, M. Li and M. Bucan, *Am. J. Hum. Genet.*, **81**, 1278-1283 (2007).
- 9 L. Weng, F. Macciardi, A. Subramanian, et al, *BMC Bioinformatics*, **12**, 99-2105-12-99 (2011).
- 10 B. L. Yaspan, W. S. Bush, E. S. Torstenson, et al, *Hum. Genet.*, **129**, 563-571 (2011).
- 11 T. Ideker, O. Ozier, B. Schwikowski and A. F. Siegel, *Bioinformatics*, **18 Suppl 1**, S233-40 (2002).
- 12 P. Jia, S. Zheng, J. Long, W. Zheng and Z. Zhao, *Bioinformatics*, **27**, 95-102 (2011).
- 13 N. Akula, A. Baranova, D. Seto, et al, *PLoS One*, **6**, e24220 (2011).
- 14 E. J. Rossin, K. Lage, S. Raychaudhuri, et al, *PLoS Genet.*, **7**, e1001273 (2011).
- 15 N. A. Patsopoulos, Bayer Pharma MS Genetics Working Group, Steering Committees of Studies Evaluating IFNbeta-1b and a CCR1-Antagonist, et al, *Ann. Neurol.*, **70**, 897-912 (2011).
- 16 International Multiple Sclerosis Genetics Consortium, Wellcome Trust Case Control Consortium 2, S. Sawcer, et al, *Nature*, **476**, 214-219 (2011).
- 17 J. Z. Liu, A. F. McRae, D. R. Nyholt, et al, *Am. J. Hum. Genet.*, **87**, 139-145 (2010).
- 18 International Multiple Sclerosis Genetics Consortium (IMSGC), A. H. Beecham, N. A. Patsopoulos, et al, *Nat. Genet.*, (2013).
- 19 M. C. Whitlock, *J. Evol. Biol.*, **18**, 1368-1373 (2005).
- 20 S. Köhler, S. Bauer, D. Horn and P. Robinson, *The American Journal of Human Genetics*, **82**, 949-958 (2008).
- 21 S. J. Swamidass, C. A. Azencott, K. Daily and P. Baldi, *Bioinformatics*, **26**, 1348-1356 (2010).

CROWDSOURCING AND MINING CROWD DATA

ROBERT LEAMAN[†]

National Center for Biotechnology Information (NCBI)
8600 Rockville Pike, Bethesda, MD 20894, USA
Email: robert.leaman@nih.gov

BENJAMIN M. GOOD[‡]

Department of Molecular and Experimental Medicine, The Scripps Research Institute
10550 North Torrey Pines Road, La Jolla, CA 92037, USA
Email: bgood@scripps.edu

ANDREW I. SU[‡]

Department of Molecular and Experimental Medicine, The Scripps Research Institute
10550 North Torrey Pines Road, La Jolla, CA 92037, USA
Email:asu@scripps.edu

ZHIYONG LU[†]

National Center for Biotechnology Information (NCBI)
8600 Rockville Pike, Bethesda, MD 20894, USA
Email: zhiyong.lu@nih.gov

1. Introduction

The “crowd” is that body of people that will either respond to an open call for participation (“crowdsourcing”) or who, through the actions they take in public forums, leave behind a trail of information that can be mined to identify new knowledge (“crowd data”). This session considers a variety of approaches utilizing the crowd as a resource to enable biomedical discovery.

While the family of crowdsourcing methodologies is still being actively created, the existing approaches are already diverse [1]. Well-known examples include microtask environments where workers are paid to perform discrete tasks, including Amazon Mechanical Turk [2], games with a purpose such as FoldIt [3], collaborative content creation frameworks like Wikipedia [4], and systems like Twitter that produce repositories of crowd data [5]. The advantages of crowd-driven

[†] Work supported by NIH Intramural Research Program, National Library of Medicine.

[‡] Work supported by National Institute of General Medical Sciences of the National Institutes of Health (R01GM089820 and R01GM083924) and by the National Center for Advancing Translational Sciences (UL1TR001114).

approaches include reduced cost, increased data sizes, and environments closer to those in the real world. These characteristics may ultimately enable research not possible via traditional methods.

Crowdsourcing remains challenging for several reasons, however. The overall problem being addressed must be decomposed into tasks appropriate for a heterogeneous population, typically with minimal training. Suitable incentive strategies need to be devised and implemented. The responses from multiple members of the crowd must be aggregated to produce a high-quality signal. As these are all new challenges, the emerging protocols and resulting data must be validated using robust analyses and evaluation.

2. Session articles

While the six articles accepted to this session address a wide variety of computational tasks within biomedicine, their use of crowdsourcing falls within three primary themes.

The first pair of articles evaluate the ability of novice workers in microtask environments. Irshad et al. use the CrowdFlower microtask platform to annotate images to detect and segment the nuclei of cells. They compare crowdsourced annotations against those performed by pathologists, reporting f-measures as high as 0.885 at a fraction of the time and cost. They report that performance degrades significantly with larger images. Good et al. use Amazon Mechanical Turk to create an annotated text corpus of disease mentions in PubMed abstracts. Their methodology can alternately emphasize precision or recall, or balance cost and quality. They demonstrate an f-measure of 0.872 against gold-standard data, again at a fraction of the cost and time. Both articles conclude that crowdsourcing in a microtask environment can be an effective way to generate annotated datasets.

The second pair of articles stretches the concept of the crowd beyond the more typical lay-public to include focused groups of experts. Both articles involve crowds of experts, but they differ substantially in their approach and problem area. Binder et al. build a collaborative reputation-based system for describing the biology of protein networks, and apply it to curate pathways relevant to chronic obstructive pulmonary disease. Their crowdsourcing experiment resulted in the submission of 885 pieces of evidence, with a validity rate of 77%. Tastan et al. query multiple experts to determine whether protein-protein interactions described in the published literature on HIV represent a physical or indirect interaction. They use a probabilistic latent variable model to jointly estimate the accuracy of each annotation and the probability of each interaction being correct. They evaluate their method with synthetic data, demonstrating significant improvements over the consensus baseline.

The third pair of articles uses the crowd for exploratory analysis. Waldspühl et al. create an online game for structural alignment of non-coding RNA, a computationally challenging problem. Players explore potential alignments via tile-matching, gaining points for finding better alignments, resulting in a casual game similar to Phylo [6]. Odgers et al. note the inadequacy of relying on spontaneous reports of adverse drug events and evaluate whether the search logs of healthcare professionals could be a useful source of signal. Their method provides an AUC of 0.85

for well-known adverse reactions and an AUC of 0.68 for adverse reactions described recently, suggesting the possibility of also detecting novel adverse reactions.

Taken together, the articles presented in this session provide a rich sample of the many emerging efforts to harness the wisdom of the crowd for biomedical research.

Acknowledgments

The authors thank the many anonymous reviewers for their generous assistance and insight as well as the thousands of members of the crowd that collectively made this session possible.

References

1. Good BM, Su AI: **Crowdsourcing for Bioinformatics.** *Bioinformatics* 2013, **29**(16):1925-1933.
2. Burger JD, Doughty E, Khare R, Wei CH, Mishra R, Aberdeen J, Tresner-Kirsch D, Wellner B, Kann MG, Lu Z *et al*: **Hybrid curation of gene-mutation relations combining automated extraction and crowdsourcing.** *Database (Oxford)* 2014, **2014**.
3. Cooper S, Khatib F, Treuille A, Barbero J, Lee J, Beenen M, Leaver-Fay A, Baker D, Popovic Z, Players F: **Predicting protein structures with a multiplayer online game.** *Nature* 2010, **466**(7307):756-760.
4. Finn RD, Gardner PP, Bateman A: **Making your database available through Wikipedia: the pros and cons.** *Nucleic Acids Res* 2012, **40**(Database issue):D9-12.
5. Eysenbach G: **Infodemiology and infoveillance: framework for an emerging set of public health informatics methods to analyze search, communication and publication behavior on the Internet.** *Journal of medical Internet research* 2009, **11**(1):e11.
6. Kawrykow A, Roumanis G, Kam A, Kwak D, Leung C, Wu C, Zarour E, Sarmenta L, Blanchette M, Waldspuhl J: **Phylo: a citizen science approach for improving multiple sequence alignment.** *PloS one* 2012, **7**(3):e31362.

REPUTATION-BASED COLLABORATIVE NETWORK BIOLOGY*

The sbv IMPROVER project team (in alphabetical order): JEAN BINDER¹, STEPHANIE BOUE¹, ANSELMO DI FABIO², R. BRETT FIELDS³, WILLIAM HAYES³, JULIA HOENG^{1†}, JENNIFER S. PARK³, MANUEL C. PEITSCH¹

¹ Philip Morris International R&D, Philip Morris Products S.A., Quai Jeanrenaud 5, 2000 Neuchâtel, Switzerland

² Applied Dynamic Solutions, LLC, 220 Davidson Avenue, Suite 100, Somerset, NJ, 08873, USA

³ Selventa, One Alewife Center, Cambridge, MA 02140, USA

[†]Corresponding author: julia.hoeng@pmi.com

A pilot reputation-based collaborative network biology platform, Bionet, was developed for use in the sbv IMPROVER Network Verification Challenge to verify and enhance previously developed networks describing key aspects of lung biology. Bionet was successful in capturing a more comprehensive view of the biology associated with each network using the collective intelligence and knowledge of the crowd. One key learning point from the pilot was that using a standardized biological knowledge representation language such as BEL is critical to the success of a collaborative network biology platform. Overall, Bionet demonstrated that this approach to collaborative network biology is highly viable. Improving this platform for *de novo* creation of biological networks and network curation with the suggested enhancements for scalability will serve both academic and industry systems biology communities.

1. Introduction

Biological networks represent our knowledge about biological mechanisms as diagrams of nodes (e.g. molecular entities) and edges (relationships between entities). Network biology concerns itself with the building and maintenance of such networks. This requires a great deal of contextual knowledge that is generally beyond the scope of individual biologists. This makes network biology an excellent field for collaborative efforts; such efforts include WikiPathways (<http://wikipathways.org>), BioPax (<http://biopax.org>), and OpenBEL (<http://openbel.org>). Here we propose another model that may be more effective for collaborating in this field, namely, a reputation-based collaborative network biology platform designed to build, edit and verify networks. It can allow more scientists (e.g. subject matter experts) to bring their perspectives to bear on large representations of mechanisms. A reputation-based system can also provide self-moderation, making this a more scalable approach than an assigned moderator-managed collaborative platform. It can also incorporate peer review functions. Here, we present a prototype solution called Bionet and share the usage results from the initial Network Verification Challenge (<http://bionet.sbvimprover.com>) using this platform.

* This work is supported by Philip Morris International.

1.1 Network Verification Challenge (NVC)

The NVC, the third challenge of the sbv IMPROVER project¹, is an effort to validate industrial research approaches and resulting biological networks focused on lung biology and lung diseases such as chronic obstructive pulmonary disease (COPD). The NVC was supported by the creation of a platform for collaborative network biology, called Bionet (<http://bionet.sbvimprover.com>), to help verify and enhance the COPD biological networks. The pilot phase of the NVC consisted of a 5-month open phase during which participants could log into the website and contribute by voting on evidence and edges of fifty biological networks. The open phase was followed by a 3-day in-person Jamboree meeting where the best performers and subject matter experts in the field of lung and COPD biology were invited to discuss and agree on changes to the networks.

1.2 Large-scale Collaboration

Some crowdsourcing efforts, such as the Critical Assessment of Protein Structure Prediction initiative (CASP)², require intense effort and a high level of expertise while others, such as Foldit³, Mechanical Turk (<https://www.mturk.com>) or Wikipedia (<http://wikipedia.com>) require less effort and expertise. The combined level of expertise and effort is generally inversely proportional to the number of people participating in a crowdsourcing effort. Consequently, less intense crowdsourcing efforts which offer relatively strong incentives attracted many more participants. For example, Foldit, a crowd-sourced protein folding game, attracted over 531,000 participants (<http://foldit.com>) while Assemblathon2, a very high effort, high expertise crowdsourcing effort recruited “only” 21 teams⁴.

An important aspect of any crowdsourcing effort is to define the appropriate incentives. In the NVC, access to the resulting networks and network biology (i.e., the possibility to download the networks for further analyses/visualization) was a significant motivator along with a reputation system and associated benefits. Another benefit was an invitation based on earned reputation points to attend the Jamboree review of the resulting networks after the initial phase of network verification and enhancement.

1.3. Reputation System

A reputation system can be used to support self-moderation of a crowd-sourced curation system. Examples of initiatives using reputation systems are ResearchGate (<https://www.researchgate.net>) and StackOverflow (<http://stackoverflow.com>). The ResearchGate reputation score (RGScore) is used only to provide a ranking compared with other researchers, while the StackOverflow reputation score is used in crowd management of the StackOverflow question and answer crowdsourcing site. As members develop their StackOverflow reputations, more powerful moderation features are unlocked; i.e. as members ‘prove’ themselves and become trusted community members, they are given rights and responsibilities of managing the StackOverflow site and participant community. A side effect of the StackOverflow reputation score and associated badges of activity, such as ‘Great Question’ or ‘Guru’, is that these badges are now used as expertise credentials in the software developer community.

2. Materials and Methods

2.1. Biological Networks

In the NVC, fifty networks were made available on the Bionet website for crowd verification. These networks were based on previously constructed non-diseased networks that describe cell proliferation⁵, cell stress⁶, DNA damage, autophagy, cell death and senescence⁷, pulmonary inflammation⁸, and tissue repair and angiogenesis⁹. The networks used in the NVC were enhanced with COPD-relevant mechanisms using a literature and data approach (manuscript in preparation).

2.2. Collaborative Web Platform Functionality

Bionet gives the participants the ability to search for and navigate to a network based on various conditions: name of the network, official name and synonyms of the nodes and edges, and references supporting the edges (PubMed IDs). In the network viewer, the participant can navigate the network by nodes or edges and can use the node and edge lists to quickly view a sorted and filtered list of all elements in the network. By selecting an edge from the edge list, the participant can view a list of published evidence related to the edge, then, by selecting an evidence item, the participant can view the complete details of the evidence. This approach to network navigation can also be applied to the network visualization tool. When a participant selects a node or an edge on the network visualization page, the associated information is presented allowing participants to drill down into the evidence-level detail.

Edge creation and Evidence voting/creation are tied directly to the Reputation system. Participants gain reputation points by verifying and enhancing the networks in various ways: extend networks with new edges, provide additional evidence for edges, and/or approve/reject evidence that has been posted in support of network edges. Participants also gain points if the evidence or edge they added has been approved by other users. To vote on evidence, a participant selects the evidence and is then presented with the option of approving or rejecting the evidence. Based on the type of evidence selected, the participant is asked questions to document the rationale for the approval or rejection. After evidence is submitted, it becomes active, which allows other participants to vote on the newly added evidence. If the evidence is voted on by enough participants, and a majority of participants approve or reject it, the evidence is locked and the edge is marked approved or rejected. If no consensus is reached, the evidence is marked as ambiguous. The network visualization tool reflects the status of the edges using different colors. Moreover, users are able to visualize their own changes in the network viewer.

Participants can create a new edge by selecting a node from the network using the Biological Expression Language (BEL). BEL is a syntax that can represent biological relationships in a standardized computable format. The Bionet application provides a BEL syntax generator to help the participant create a proper BEL statement for the edge. Additionally, a tool to help create one or several BEL statements based on an excerpt from an academic paper (evidence source) is provided in Bionet. When an edge is created, the participant adds evidence to the edge and submits it to the network. After submission, all participants can see it on the network visualization page and can vote on it like any other edge in the application.

Bionet provides a community section that allows participants to see the latest network activity for all users or him/herself and possibly filtered by network. This community area is critical to participation because participants can use it to view the network of interest to them and track the actions taken by other participants. Participants can then vote on the action directly in the activity feed or go to the network to see the action in context.

Leaderboards are used on the site to help participants gauge their level of participation and reputation score in relation to their peers. Points are given for voting and evidence creation and badges are awarded for various actions. Participants can filter the leaderboards by teams, votes, evidence creation, and edge creation.

2.3. Funnel of Participation

The funnel of participation, as coined by Clow, refers to the process of gaining participants for a participatory project¹⁰, which was applicable for this pilot phase of NVC (Figure 1). Awareness was achieved through emails to potential participants based on their research record, presentations at relevant scientific conferences, publications (both peer-reviewed and science news media channels)¹¹ as well as seminars at selected network biology-focused laboratories. In addition, we engaged ‘NVC ambassadors’ who personally called and/or emailed contacts in their scientific network to help increase awareness by notifying and teaching potential participants about the NVC. We estimated that the Bionet Awareness campaign resulted in 1,000,000 impressions on potential participants. From this, there were 1,298 unique visitors to the collaborative website. This resulted in 132 Bionet registered participants from which 26 highly active participants were selected as Best Performers for the Challenge (Figure 1).

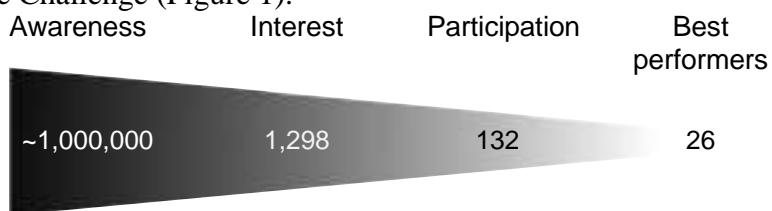


Fig. 1. Funnel of participation for the first NVC.

2.4. Evaluation of Participant Activity

User statistics from Bionet logs were analyzed to calculate a number of metrics related to user participation and network activity. A questionnaire was emailed to participants after the challenge to help understand such factors as motivation and ease of website use.

References associated with networks were counted by searching published scientific literature using Quertle (www.quertle.com) and the name of the network with the word “pathway”. If the network name contained the word “signaling” it was replaced with “pathway” because this is a more specific way of describing molecular events (signaling can refer to electric signals).

2.5. Evaluation of Tissue-relevant Evidence Additions

Crowd-submitted evidence was reviewed to assess the overall degree of tissue relevancy compared with network boundaries. Evidence from primary lung tissue and lung-associated cell types (during

COPD) was deemed within the network boundaries, while additions from non-lung-relevant cell types (e.g. neural progenitor cells) and non COPD-relevant diseases (e.g. colon cancer) were rejected. The number of COPD-relevant evidence additions as a percentage of the total number submitted was taken as an overall assessment of crowd performance.

2.6. Evaluation of Quality of Participant-submitted Causal Biology

A random sampling of 100 pieces of evidence submitted by the participants was independently evaluated by two scientists with expert-level experience in these networks to assess the overall quality of contributions by manually reviewing the primary literature associated with each submission. A random number generator was used to produce 100 numbers within a 1–885 range, corresponding to the total pieces of evidence submitted among all networks. Entries were further blinded by removing all personal participant information stored with the entry (e.g. user name of submitting participant) to prevent bias during the expert evaluation. Key metrics that were evaluated during this process included, 1) relevance of evidence within the individual network, 2) relevance to COPD and/or lung biology, and 3) accuracy of capturing the biological relationship in referenced literature. Evidence meeting all three criteria was rated “Valid” and that deficient in any of the three was rated “Invalid.” For the evaluation, evidence containing minor defects in BEL scripting was not rated “Invalid”.

3. Results

3.1. Evaluation of Participant Activity

A global community of researchers took part in the NVC (Figure 2).



Fig. 2. NVC participant countries.

An analysis of the activity of participants in the NVC Open Phase revealed a range of participant profiles. Some researchers who registered performed only a few actions. In a follow-up questionnaire, the explanations given for low participation were mostly the lack of time and/or

specific interest in the networks. As expected, the “entry level” action of participants with low activity levels was “voting on evidence” (Figure 3, right panel). Distinctive profiles were observed among the best performers (BP) (Figure 3): some spent a lot of effort enhancing specific networks (e.g. BP2), some contributed to many networks by creating new evidence/edges (e.g. BP9), and some dedicated most time verifying (voting on) a number of networks (e.g. BP3). In general, BPs voted on more networks than the number of networks for which they created new evidence (e.g. BP2, BP3).

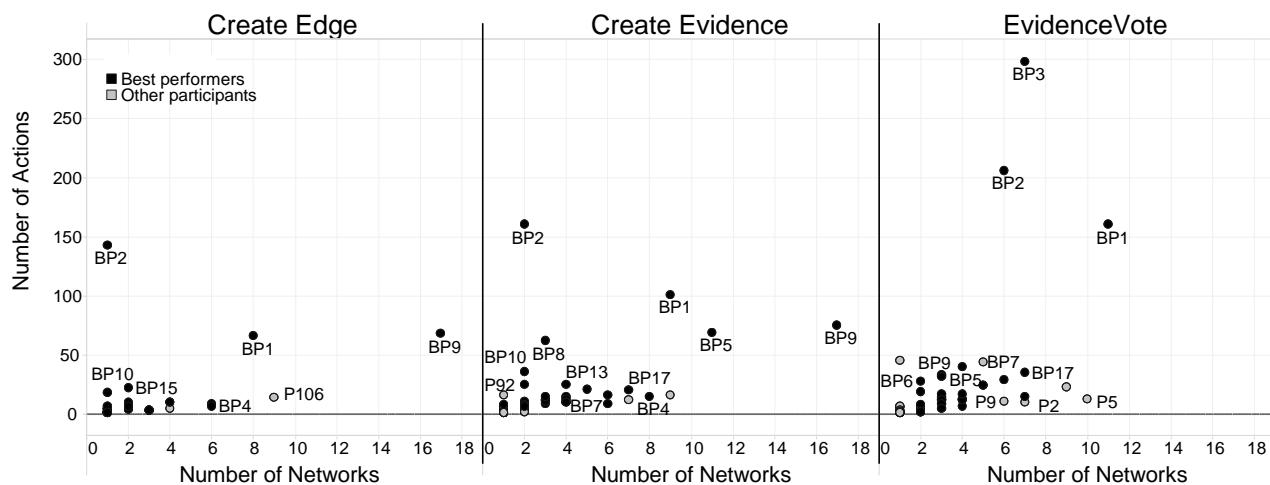


Fig. 3. Participant activity across a number of networks.

Dissecting the activity of participants per network and action type revealed interesting patterns (Figure 4): (i) cell-specific networks and widely studied biological processes (>10,000 associated references) attracted the highest number of participants, especially among those that did not have a lot of activity (non-best performers); (ii) more complex networks were approached more frequently by participants who were more experienced or who spent more time on the NVC (best performers); (iii) in most cases, the number of distinct contributors was highest for the voting actions; and (iv) of the 50 networks, nine attracted at least ten contributors.

3.2. Evaluation of species and tissue-relevant evidence additions

One of the goals of the NVC was to add relevant literature that supported edges in the networks to improve their overall relevance to human COPD. We evaluated the extent of human literature supporting the edges, as well as evidence from lung-relevant experiments added by the participants. In total, the crowd submitted 885 new pieces of evidence, the large majority of which was from human studies (65%) (Figure 5A).

There was great variability on a per-network basis in terms of quality of submissions as well as general activity, with cell-specific network additions conforming more frequently to the tissue boundary conditions. For example, the *Neutrophil Signaling* network received a preponderance of participant submissions with 179 total pieces of evidence submitted (20% of the total submitted). Among the 170 pieces of evidence with tissue metadata, 100% submissions conformed to the network boundary conditions. Similarly, in the *B-cell Signaling* network, 100% of the annotated

submissions fell within the boundary conditions of the network collection (Figure 5B). In contrast, the *Notch* network received the fewest tissue-relevant evidence additions, with only 17% of the annotated submissions falling within the boundary conditions of the network (Figure 5B).

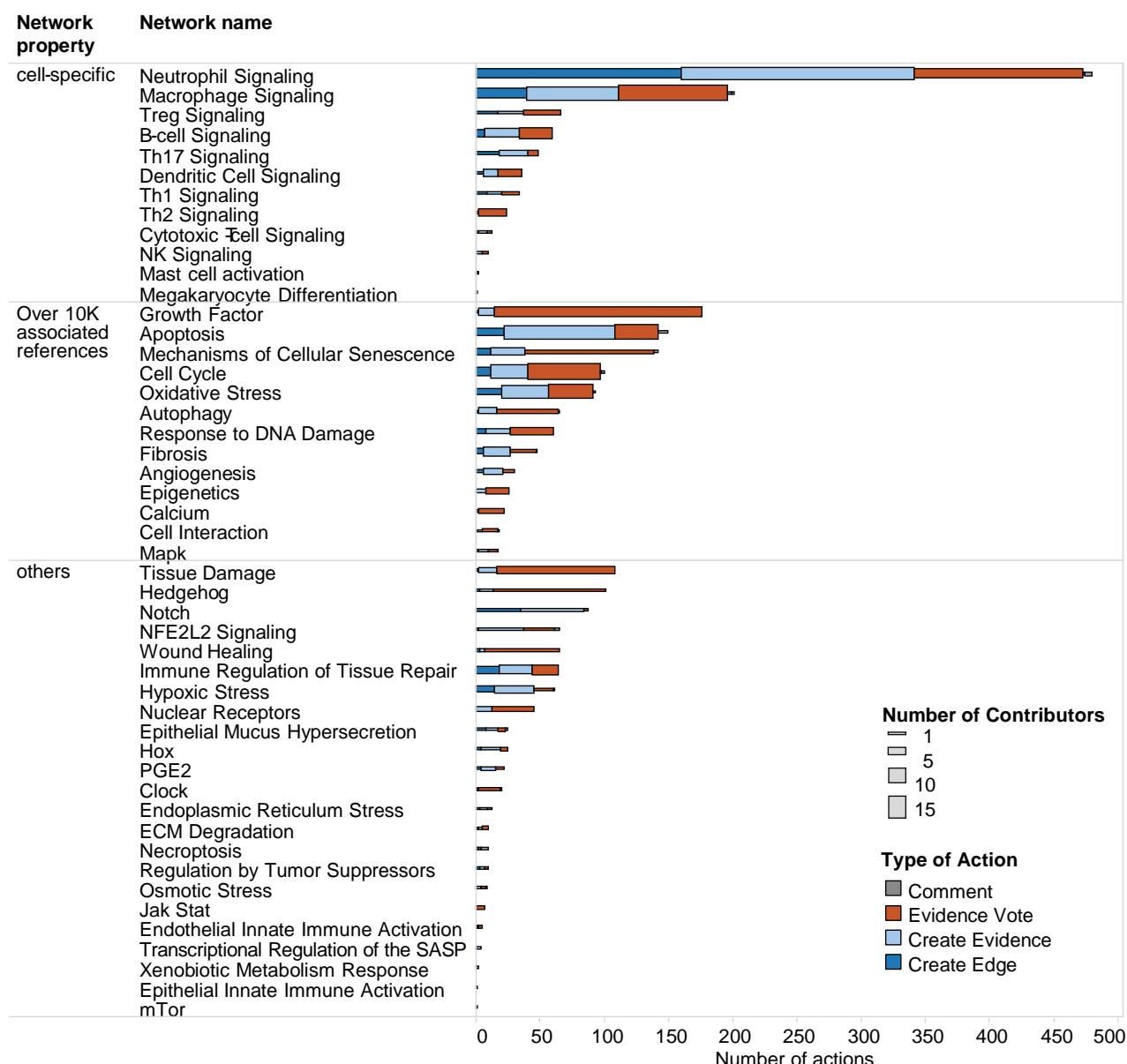


Fig. 4. Participant activity by network.

3.3. Evaluation of quality of participant-submitted causal biology

We assessed the overall quality of the 885 total pieces of evidence submitted by the NVC community using a randomized, independent review process. On average, the quality analysis resulted in a validity rate of 77%, indicating that the majority of additions enhanced the biological

foundation supporting the network connectivity. The majority of evidence deemed invalid was outside the tissue boundaries of the network. However, because the tissue boundaries may not have been obvious to all participants, to evaluate valid biological representation, we calculated the percentage validity again after removing boundaries as a criterion. The average validity rate after disregarding tissue boundaries increased to 88%. The review by one of the scientists yielded 88% valid contributions while the second independent review by another scientist yielded 85% valid contributions, although the same pieces of evidence were not always judged as valid. An assessment of inter-reviewer comparability revealed 83% agreement between the two evaluations. Among the entries where there was disagreement, most cases were caused by subjective interpretation of the primary literature and, by extension, its representation in BEL. Employing a more robust statistical measure of inter-reviewer comparability, Cohen's kappa, produced a coefficient of 0.26, illustrating the subjective nature of assessing the quality of biological submissions, and the importance for several scientists to review the same evidences, as made possible in this initiative.

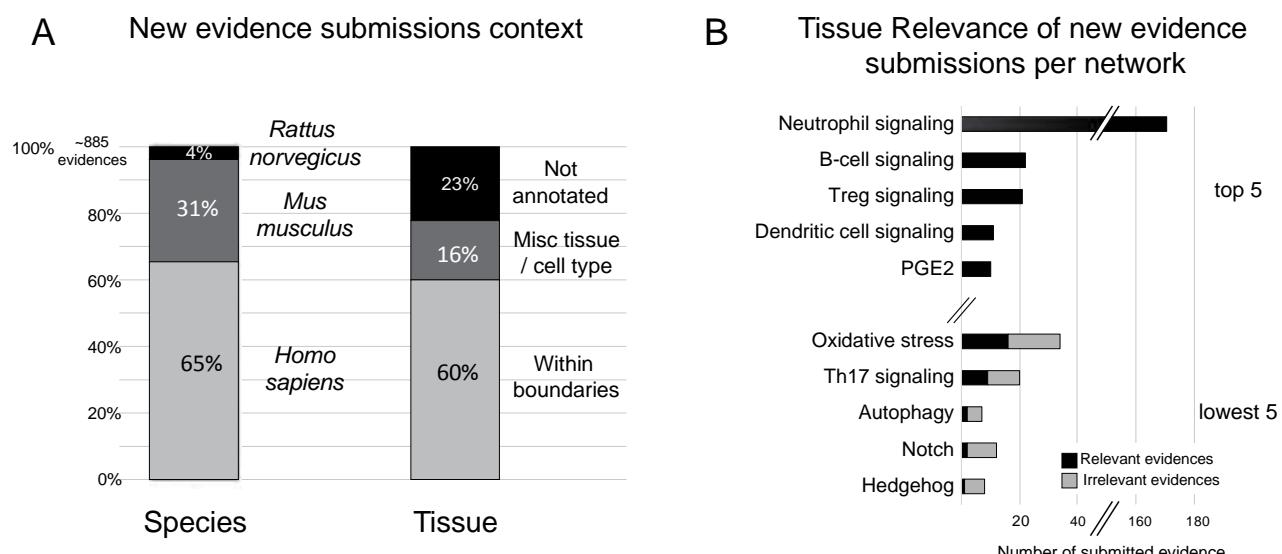


Fig. 5. Relevance of new evidence submissions overall for context (A) and for the top 5 and lowest 5 networks (B).

4. Discussion

4.1. NVC Contributions and Participant Activity

4.1.1 Overall Number of Contributions

In total, 2,456 votes were cast and 885 pieces of evidence were created (including 351 new network nodes) by a relatively small number of participants (~80). Although these numbers are small when compared with the total amount of evidence in the networks (over 180,000 pieces), the actions that were measured took place over a 5-month period in a pilot project while the existence of the NVC was still being disseminated to the scientific community. To verify 50 networks containing thousands of biological connections was overwhelming for the modest number of participants. For

future projects, one approach may be to restrict the number of networks being evaluated to help concentrate participant attention to particular areas. In addition, pilot crowdsourcing ventures with limited adoption can focus more on the creation versus the verification process for enhancing networks. Creation of new evidence and edges could add useful biology even with a small number of participants, while to reach a crowd consensus, verification/voting requires a larger number of participants to ensure a representative sample.

4.1.2. Focused Participant Activity

Overall, the participants in the NVC worked on a small number of networks (1–5) according to their scientific expertise. Participants tended to work on well-studied networks with canonical pathways (*Cell Cycle* and *Apoptosis*) that were reported in the literature, or on cell-specific networks (*Macrophage* and *Neutrophil Signaling*) for which it was straightforward to identify the relevant literature. Networks of high interest and for which there is a lot of information may be more conducive to a crowd-verification approach. However, it is the less-studied networks that might benefit more from crowd review. In the future, the NVC could be restricted to these networks to concentrate attention and effort.

Overall, participants voted rather than created new evidence, with the best performers voting on more networks and creating evidence for fewer networks. This is likely because it is easier to vote (i.e., assess the scientific validity of existing evidence) on diverse topics, whereas it requires significantly more expertise to enhance specific networks by adding new scientific evidence to existing network edges (i.e., identify and extract additional knowledge from the scientific literature) or to create new edges. Indeed, in the participant survey, most researchers scored voting to be “very easy” and rated adding new evidence to be “easy” (data not shown).

4.1.3. Participant Engagement

The NVC was publicized through many different avenues, including conferences, publications, emails, web searches, seminars, advertisements, and “ambassadors”. NVC ambassadors tapped into personal scientific networks to promote the NVC and follow up with one-on-one educational sessions for interested scientists. This mode of promotion was found to be the most effective because of the personal nature of the contact and the opportunity for a tutorial session to ensure that potential participants understand how to use the website and create BEL statements. In fact, a participant survey showed that the majority learned about the NVC through personal contact (data not shown). Because of the success of this method, we are continuing to emphasize this personal ambassador approach among personal scientific networks to further publicize and educate researchers about the next NVC.

NVC participants were motivated by various factors, with the possibility of co-authoring an academic publication being the top motivation according to the participant survey (data not shown). Other motivations included travel and Jamboree invitation rewards, learning about the biology of the networks, the chance to download the networks to use in their research, and the challenge of the verification tasks. For continued participation these types of incentives are important to attract a community of regular participants based on periodic cycles of publications, meetings and latest version networks being released.

4.2. NVC Improved the Relevance and Comprehensiveness of the Networks

4.2.1. Evaluation of Participant Contributions

The new evidence created by participants during this short period represents an improvement in the network comprehensiveness, especially when considering the majority of evidence was concentrated among a small number of networks. The evaluation of tissue relevance for new evidence submissions was one metric used to quantify the value of crowdsourcing for improving biological networks. In this assessment, 23% of submitted evidence contained no contextual annotation, despite providing several entry fields (e.g. tissue, disease, cell type). One way to reduce non-annotated contributions in future projects could be to implement mandatory contextual fields during the submission process. Ultimately, we determined that 60% of evidence submissions conformed to the network boundary conditions as set forth in the platform user tutorials. Although entries from miscellaneous contexts are certainly a significant overall contribution, such entries could be better avoided by publishing the network boundaries more frequently throughout the platform, perhaps once again at point-of-entry for new entries to ensure participants are fully aware of the criteria prior to submission. The boundaries were heavily emphasized during scheduled webinars to promote the NVC and educate users but were less visible and detailed on the website where users probably most needed this reminder.

When the new submission context was assessed on a per-network basis, several networks received an outstanding quality of new evidence because 100% of the submissions conformed to the specified boundary conditions (Figure 5B). For example, for the *Neutrophil Signaling* network, all annotated submissions were sourced from primary literature in which the study was conducted specifically in neutrophils. In contrast, networks detailing more ubiquitous biological pathways (e.g. *Oxidative Stress* and *Notch Signaling*) often received submissions from a broader array of contexts, reflecting the abundance of primary literature among many cellular contexts, but not necessarily relevant to lung. Therefore, we concluded that networks with cell-specificity garnered submissions with better-defined contexts that were more likely to conform to the stated boundary conditions. It may be more reasonable in future challenges to loosen the boundary conditions for the more general biological pathways that are conserved across tissues.

A separate dimension related to the overall quality of crowd-submitted enhancements was assessed by an expert-level review of a random sampling of the submissions. Two independent evaluations of the same data sample revealed a “Validity” rate average of 77%. Because the tissue and disease boundary conditions of the network may not have been apparent to participants, when the validity rate was calculated without regard to these boundary conditions it came out to be 87%. This high validity rate suggested that the participants successfully entered biologically sound mechanistic statements retrieved from the literature into the website to contribute to the networks. Improving communication of the boundaries is an important lesson from this analysis.

Despite the high number of submissions deemed to be of “high quality” by the experts, the calculated Cohen’s kappa statistic of 0.26 revealed a modest degree of comparability between the overall quality assignments, or 83% agreement. This statistic factors in the probability of agreement occurring by chance during a qualitative evaluation, which is heavily influenced by the fact a simple binary rating system (Valid vs. Invalid) was used during the evaluation process. Nevertheless, this

evaluation process illustrates the subjectivity inherent in assessing biological experiments. Some of the differences in the reviewer analyses were related to them not having complete information (unavailable full text) or simply reviewer error (mis-reading the paper). However, in some cases the biology was interpreted in a different way by the scientists. Owing to subjectivity of biological interpretation, in addition to the open phase during the NVC an in-person Jamboree was held to provide a forum where these subjective or controversial items could be discussed. Scientists who contributed to the networks as well as subject matter experts for the biology that each network describes participated in the Jamboree to come to a consensus for finalizing the changes to the networks. Not only was the Jamboree critical for alignment of individual controversial edges within the networks, but it provided a forum to discuss the networks more holistically and edit larger pieces of the network to improve flow, comprehensiveness, and granularity. Overall, the NVC enhancements improved the relevance and comprehensiveness of the networks.

Refined networks have been uploaded to Bionet so that the improvements brought by the crowd can benefit the scientific community. In particular, active participants that have earned the “Download” badge by performing a minimum number of actions may download all networks for further analysis and visualization using their favorite tools.

4.2. Vision for Biological Collaboration

4.2.1. BEL as a Universal Language for Biology

The Bionet platform is designed to require low effort but high subject matter expertise. However, the time that a potential participant needs to invest in learning BEL was a significant challenge that increased the activation effort for a participant and contributed to participant attrition. There is no standard knowledge representation that every network biologist is trained on that is comparable to the chemical reaction language for chemists. Because BEL is new to the biological academic community it has not yet achieved widespread adoption. If BEL, or some other universal network biology knowledge representation, becomes a standard representation for biological relationships, collaboration in biology, especially network biology, can become more effective. A standard representation for biology will greatly increase the viability of a network biology collaboration platform.

4.2.2. A Standard Reputation Platform for Biology

As a long-term, self-moderating platform for the creation and management of biological networks, we propose that the Bionet platform can provide great benefits to network biology. An added benefit is the potential for the reputation points and badges earned in the system based on peer-review to become an important aspect of a network biologist’s *curriculum vitae*. A Reputation Score and associated Badges could become an important credentialing resource for network biologists in the research community, as has StackOverflow reputation score gained importance in the informatics community.

Our goal in creating a network biology-focused collaborative platform is, in the longer term, to provide a reputation system that supports self-management in the same manner as StackOverflow.

In the first iteration of Bionet, the reputation score only ranked one participant against the other participants and is similar to the RGScore. However, the Bionet reputation score was designed to be extended to promote high-reputation participants to moderators of the networks in the same manner as the StackOverflow reputation score. As participants gain reputation points and hit certain targets, they will be able to take on more moderation of the networks on the Bionet platform. This will allow the platform to scale to many thousands of participants and thousands of networks.

The Bionet platform can work well in both public and industry settings. The Bionet reputation-based collaborative network biology platform works well with a small number of active users. The recommended changes that will make it more scalable by automatically promoting high-reputation users to moderators are not required in an industry setting. A single platform that provides online access to networks, allows participants to edit the networks and collaborate easily with other participants regarding the networks, and provides computationally tractable networks using a common knowledge representation will be a great tool for network biology in both the academic and industry sectors.

This pilot project in large-scale collaboration for network biology has highlighted certain aspects that are required for a self-sustaining platform, including a universal biological language and a standardized and therefore valued reputation system. With the insights gained during the NVC, both the Bionet website and biological content will continue to improve and latest version networks are available on Bionet as well as on our Causal Biological Networks database (CBN, causalbionet.com). Bionet is currently open for crowd input during the ongoing second Challenge (NVC2) and associated Jamboree planned for 2015.

5. Acknowledgments

We thank IBM for their help in organizing the NVC and Jamboree, Michael Maria for help in project management, and Sam Ansari, Anita Iskandar, Robin Kleiman, Carine Poussin, Dexter Pratt, Marja Talikka, and Walter Schrage for their scientific input.

References

1. Meyer, P., *et al. Nat Biotechnol* **29**, 811-815 (2011).
2. Moult, J., Fidelis, K., Kryshtafovych, A., Schwede, T. & Tramontano, A. *Proteins: Structure, Function, and Bioinformatics* **82**, 1-6 (2014).
3. Cooper, S., *et al. Nature* **466**, 756-760 (2010).
4. Bradnam, K.R., *et al. GigaScience* **2**, 10 (2013).
5. Westra, J.W., *et al. BMC systems biology* **5**, 105 (2011).
6. Schrage, W.K., *et al. BMC systems biology* **5**, 168 (2011).
7. Gebel, S., *et al. Bioinformatics and biology insights* **7**, 97-117 (2013).
8. Westra, J.W., *et al. Bioinformatics and biology insights* **7**, 167-192 (2013).
9. Park J.S., S.W.K., Frushour B.P., Talikka M., Toedter G. *et al.*, Vol. S12: 002 (2013).
10. Clow, D., Vol. 185-189 (ACM, Leuven, Belgium, 2013).
11. The sbv Improver project team, *et al. Bioinformatics and biology insights* **7**, 307-325 (2013).

MICROTASK CROWDSOURCING FOR DISEASE MENTION ANNOTATION IN PUBMED ABSTRACTS

BENJAMIN M GOOD, MAX NANIS, CHUNLEI WU, ANDREW I SU

Molecular and Experimental Medicine, The Scripps Research Institute, 10550 N. Torrey Pines Rd., La Jolla, CA, 92037, USA

Email: bgood@scripps.edu, max@maxnanis.com, cwu@scripps.edu,asu@scripps.edu

Identifying concepts and relationships in biomedical text enables knowledge to be applied in computational analyses. Many biological natural language processing (BioNLP) projects attempt to address this challenge, but the state of the art still leaves much room for improvement. Progress in BioNLP research depends on large, annotated corpora for evaluating information extraction systems and training machine learning models. Traditionally, such corpora are created by small numbers of expert annotators often working over extended periods of time. Recent studies have shown that workers on microtask crowdsourcing platforms such as Amazon's Mechanical Turk (AMT) can, in aggregate, generate high-quality annotations of biomedical text. Here, we investigated the use of the AMT in capturing disease mentions in PubMed abstracts. We used the NCBI Disease corpus as a gold standard for refining and benchmarking our crowdsourcing protocol. After several iterations, we arrived at a protocol that reproduced the annotations of the 593 documents in the 'training set' of this gold standard with an overall F measure of 0.872 (precision 0.862, recall 0.883). The output can also be tuned to optimize for precision (max = 0.984 when recall = 0.269) or recall (max = 0.980 when precision = 0.436). Each document was completed by 15 workers, and their annotations were merged based on a simple voting method. In total 145 workers combined to complete all 593 documents in the span of 9 days at a cost of \$.066 per abstract per worker. The quality of the annotations, as judged with the F measure, increases with the number of workers assigned to each task; however minimal performance gains were observed beyond 8 workers per task. These results add further evidence that microtask crowdsourcing can be a valuable tool for generating well-annotated corpora in BioNLP. Data produced for this analysis are available at http://figshare.com/articles/Disease_Mention_Annotation_with_Mechanical_Turk/1126402.

1. Background

A large proportion of all biomedical knowledge is represented in text. There are currently over 23 million articles indexed in PubMed, and over one million new articles are added every year. Natural language processing (NLP) approaches attempt to extract this knowledge in the form of structured concepts and relationships such that it can be used for a variety of computational tasks. Just a few of many examples include identifying functional genetic variants [1], identifying biomarkers and phenotypes related to disease [2], and drug repositioning [3].

Research in NLP is largely organized around shared tasks [4]. Periodically, the community settles on a particular challenge (e.g., identifying genes in abstracts [5]), develops manually annotated corpora that reflect the objective of the challenge, and then organizes competitions meant to identify the best computational methods. These shared corpora make it possible for researchers to refine their predictive models (in particular to train models based on supervised learning) and to evaluate the performance of all approaches. These gold standard annotated corpora are generally produced by small teams of well-trained annotators. While this

methodology has been fruitful, the costs inherent to this approach impose limits on the numbers of different corpora as well as the size of individual corpora that can be produced.

Microtask crowdsourcing platforms, such as Amazon’s Mechanical Turk (AMT), facilitate transactions between a ‘requester’ and hundreds of thousands if not millions of ‘workers’. These markets make it possible to harness vast amounts human labor in parallel. Typically a requestor sends a long list of small, discrete “Human Intelligence Tasks” (HITs) to the AMT platform which then distributes the HITs to workers. Workers who choose to work on a given task are paid for each HIT they complete at a rate set by the requestor.

Since their inception, microtask markets have attracted the attention of the NLP community because of the well-known costs of creating annotated corpora [6]. These markets are seen as a way of reducing these costs and dramatically extending the potential size of the datasets needed for training and evaluation [7]. This approach has been particularly useful for tasks that are easy for humans and clearly involve no domain knowledge. For example, a large amount of research is devoted to sentiment analysis (also known as “affect recognition”) [6].

Adoption of these techniques within the biomedical domain has been slower, probably because of the increased complexity of the texts that need to be processed and the concepts that need to be annotated. That being said, BioNLP research groups are now exploring crowdsourcing. Two early studies demonstrated that a gold standard corpus of annotated clinical trials documents could be assembled through microtask crowdsourcing [8, 9]. Recently, others have successfully applied microtasking to validate predicted gene-mutation relations in PubMed abstracts [10] and for medical relation extraction [11]. Here we extend these efforts by implementing and testing a disease mention annotation task on PubMed abstracts using the AMT platform.

2. Task: disease mention annotation

A fundamental step in nearly all NLP tasks is the identification of occurrences of concepts such as diseases, genes, or drugs. Our goal for this work was thus to develop and benchmark a crowdsourcing protocol for annotating PubMed abstracts with concept occurrences. We chose the disease annotation task because a large, expert-crafted gold standard with information about inter-annotator agreement was available for comparison [12] and because we expected that the concept of “disease” would be a tractable place to start testing the ability of “the crowd” to process complex biomedical text.

2.1. Original guidelines for creating the NCBI Disease corpus

To produce the original gold standard disease corpus, annotators were instructed to highlight a span of text, to identify a concept from the Unified Medical Language System (UMLS) that matched the meaning of the highlighted text and to assign it to one of four categories. The categories included: Specific Disease (e.g. “Diastrophic dysplasia”), Disease Class (e.g. “autosomal recessive disease”), Composite Mention (e.g. “Duchenne and Becker muscular dystrophy”), and Modifier (“e.g. colorectal cancer families”). Expert annotators completed this task using the PubTator interface [13]. They followed an annotation guideline document that contained a number of additional rules such as “annotate duplicate mentions” and “do not annotate

overlapping mentions” [12]. The PubTator tool also automatically suggested annotations and provided direct access to a UMLS search tool.

It is worth noting that these guidelines called for the annotation of a number of semantic types that are related to diseases but are not diseases in themselves. For example, the instructions said to highlight mentions of semantic types such as “Sign or Symptom” (e.g. “Back Pain”) and “Acquired Abnormality” (e.g. “Hernia”). They also contained certain criteria that left room for annotator interpretation such as the condition that the highlighted text contain “information that would be helpful to physicians and health care professionals”. Further, the annotation rules occasionally depended on access to the UMLS system. For example the decision about whether to annotate the ‘early onset’ part of phrases like “early onset colorectal cancer” was likely determined by whether or not a relevant concept in the UMLS existed that included the “early onset” qualifier. These aspects of the annotation guidelines, as well as their length (nearly 1000 words including the examples) and complexity, compelled us to write a new instruction set specifically designed for the AMT workers. Other than the wording and examples provided, the largest difference was that our instructions did not call for annotators to identify the type of mention or to identify the concept referred to by the mention in the UMLS.

2.2. Instructions and task description provided to AMT workers

The task posted on the AMT platform was described as: “*You will be presented with text from the biomedical literature which we believe may help resolve some important medically related questions. The task is to highlight words and phrases in that text which are diseases, disease groups, or symptoms of diseases. This work will help advance research in cancer and many other diseases!*”

After agreeing to perform the task, they were presented with the following instructions and asked to take a qualification test. Each rule in the instructions was followed by an animated GIF that illustrated how the correct annotations would look as they were created with our custom highlighting tool. In the tool, users click to highlight single words, click and drag to highlight spans of text, and click again to unhighlight a span. “*Here are some examples of correctly highlighted text. Please study these before attempting to take the qualification test. Please also feel free to refer back to these examples if you are uncertain. Instructions:*”

1 Highlight all diseases and disease abbreviations

“...are associated with Huntington disease (HD)... HD patients received...”

“The Wiskott-Aldrich syndrome (WAS), an X-linked immunodeficiency...”

2 Highlight the longest span of text specific to a disease

“... contains the insulin-dependent diabetes mellitus locus ...”

and not just ‘diabetes’.

“...was initially detected in four of 33 colorectal cancer families...”

and not just ‘cancer’.

“...In inherited breast cancer cases...”

and not just ‘breast cancer’

3 Highlight disease conjunctions as single, long spans.

“...the life expectancy of Duchenne and Becker muscular dystrophy patients..”

“... a significant fraction of breast and ovarian cancer , but undergoes...”

4 Highlight symptoms - physical results of having a disease

“XFE progeroid syndrome can cause dwarfism, cachexia, and microcephaly. Patients often display learning disabilities, hearing loss, and visual impairment.”

5 Highlight all occurrences of disease terms

“Women who carry a mutation in the BRCA1 gene have an 80% risk of breast cancer by the age of 70. Individuals who have rare alleles of the VNTR also have an increased risk of breast cancer”.

6 Highlight all diseases, disease groups and key disease symptoms

“The set of 32 families in which no BRCA1 alterations were detected included 1 breast-ovarian cancer kindred manifesting clear linkage to the BRCA1 region and loss of the wild-type chromosome in associated tumors . Other tumor types found in BRCA1 mutation / haplotype carriers included prostatic, pancreas, skin, and lung cancer, a malignant melanoma, an oligodendrogloma, and a carcinosarcoma”

7 Do not highlight gene names

“... the spastic paraplegia gene (SPG) was found to..”

highlight only the disease mention, not the gene

“...Huntington disease (HD) is caused by variations in huntingtin (HTT)...”

the disease is highlighted, but the related gene is not.

“...Niethold-Alfred syndrome (NAS) is caused by mutation in the gene NAS...” .

In some cases the name of the gene may be the same as the name of the disease. In these cases you should only highlight the text if it is referring to the disease and not to the gene. The first NAS is highlighted while the second is not.

2.3. Worker task flow: Qualification test, 4 training examples, then real tasks

Before gaining access to the HITs, the workers had to pass the qualification test. No other worker filters (e.g. HIT approval %, country, etc.) were employed. This test was a series of true/false questions that assessed their comprehension of the annotation rules based on real examples. We allowed workers to pass with a fairly lenient threshold of 80% correct.

Once a worker earned the qualification, their first HIT was a survey containing questions about gender, age, occupation, education, and motivation. Following the survey, each HIT was to annotate disease mentions in a PubMed abstract. The HITs were completed on a custom website running in an iframe in the context of the AMT site. Workers were paid equally for each HIT (\$0.06), could complete as many as desired, and could leave the site and return at any time. The payment per HIT was arbitrarily selected to be slightly more than in previous studies (e.g. [8] paid \$0.03 for a similar text annotation task).

Following the survey, the next four HITs were manually selected ‘training abstracts’ that provided unambiguous examples of key annotation rules. After submitting their annotations for these abstracts, the system displayed their agreement with the gold standard annotations.

**Title**

MEFV-Gene analysis in armenian patients with **Familial Mediterranean fever**: diagnostic value and unfavorable renal prognosis of the M694V homozygous genotype—genetic and therapeutic implications.

Abstract

Familial Mediterranean fever (FMF) is a **recessively inherited disorder** that is common in patients of Armenian ancestry. To date, its diagnosis, which can be made only retrospectively, is one of exclusion, based entirely on nonspecific clinical signs that result from **serosal inflammation** and that may lead to unnecessary surgery. **Renal amyloidosis**, prevented by colchicine, is the most severe complication of **FMF**, a disorder associated with mutations in the MEFV gene. To evaluate the diagnostic and prognostic value of MEFV-gene analysis, we investigated 90 Armenian **FMF** patients from 77 unrelated families that were not selected through genetic-linkage analysis. Eight mutations, one of which (R408Q) is new, were found to account for 93 % of the 163 independent **FMF** alleles, with both **FMF** alleles identified in 89 % of the patients. In several instances, family studies provided molecular evidence for **pseudodominant transmission** and incomplete penetrance of the disease phenotype. The M694V homozygous genotype was found to be associated with a higher prevalence of **renal amyloidosis** and **arthritis**, compared with other genotypes ($P = .0002$ and $P = .006$, respectively). The demonstration of both the diagnostic and prognostic value of MEFV analysis and particular modes of inheritance should lead to new ways for management of FMF—including genetic counseling and therapeutic decisions in affected families.

- █ Correct annotations
- █ Your annotations

Figure 1: Feedback interface for annotators.

Figure 1 shows an example of the feedback screen that a worker would see after annotating a gold standard document. Workers were shown the annotations that they missed, any that they incorrectly marked, and their F score for that document. Based on earlier studies we found that this feedback was an effective way to train workers to perform the task correctly, measurably improving performance with respect to simply providing the instructions. After the first four training documents were completed, the workers were presented with randomly selected abstracts to annotate. For 10% of the abstracts, feedback based on the gold standard was supplied to them as it was for the original four training documents. If the workers scored below an F measure of 0.5 on three gold standard documents in a row, they were blocked from continuing. When the

workers annotated a non-gold standard document, they were provided similar feedback, but instead of a single set of ‘correct’ annotations, the workers were shown the annotations of up to 5 other workers as additional underlines and no F measure was supplied. The feedback interface informed the workers when they had completed an evaluation case or a regular HIT.

2.4. Data

The experiment described here used the 593 abstracts in the Training Set of the NCBI Disease Corpus [12]. These were divided into three groups: the four training documents, the 10% (60) that were used to provide interspersed gold standard feedback, and the remaining 529 for which no gold standard-based feedback was provided. The results for each document are collected prior to the worker seeing any feedback and they only see each document once so, for the purposes of overall annotation quality assessment, all documents are treated equally.

2.5. Quality statistics

We compared the annotations generated by the AMT with the annotations in the NCBI Disease Corpus using strict matching. An annotation counted as a true positive only if it exactly matched the span of the corresponding gold standard annotation. With exact matches as our comparison metric, we calculated true positives, false positives, and false negatives across all annotations in the dataset and used these to calculate global Precision ($TP/(TP+FP)$), Recall ($TP/(TP+FN)$) and F measure ($2*P*R*/P+R$). (This method, referred to as micro-averaging, essentially treats the corpus as one large document.)

3. Results

3.1. Workers

Out of 346 that took the qualification test, 145 workers passed and completed the annotation of at least one abstract. Work was distributed unevenly, with 23 workers completing over 100 abstracts each, 80 completing over 10, and 65 completing 10 or less. Figure 2 shows the number of abstracts completed per worker.

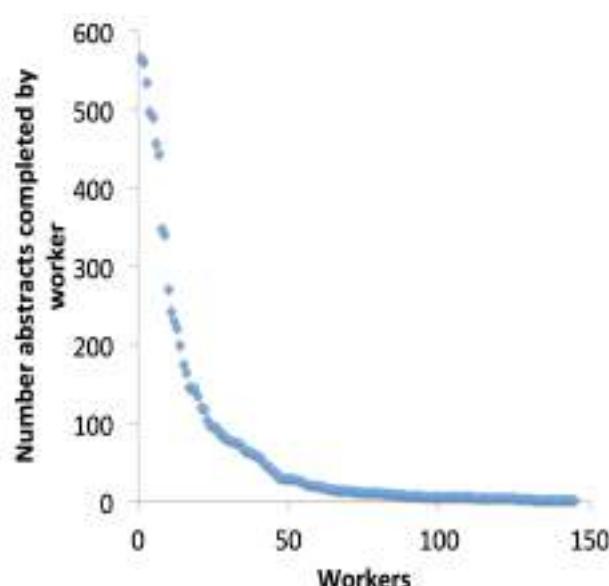


Figure 2: Number of abstracts processed per worker

Figure 3 illustrates that overall, the workers generally performed at a high level for this task. Only one worker was blocked from continuing after repeatedly performing poorly on the embedded gold standard test questions. That worker annotated 24 post-training abstracts with an average F-measure of 0.62 before being blocked. The average F score per worker across all of their annotations was 0.764 with a standard deviation of 0.130. For workers that processed more than ten abstracts, the average F score increased only slightly to 0.767 with the standard deviation decreasing to 0.078. The average F score of workers that completed a hundred or more abstracts was 0.791 with a standard deviation of 0.066. This result generally reiterates the findings in [10], which observed a clear increase in performance for workers that completed more than 100 of their gene-mutation validation tasks. However, the correlation between the number of tasks performed and the average quality of the work was not statistically significant in this experiment.

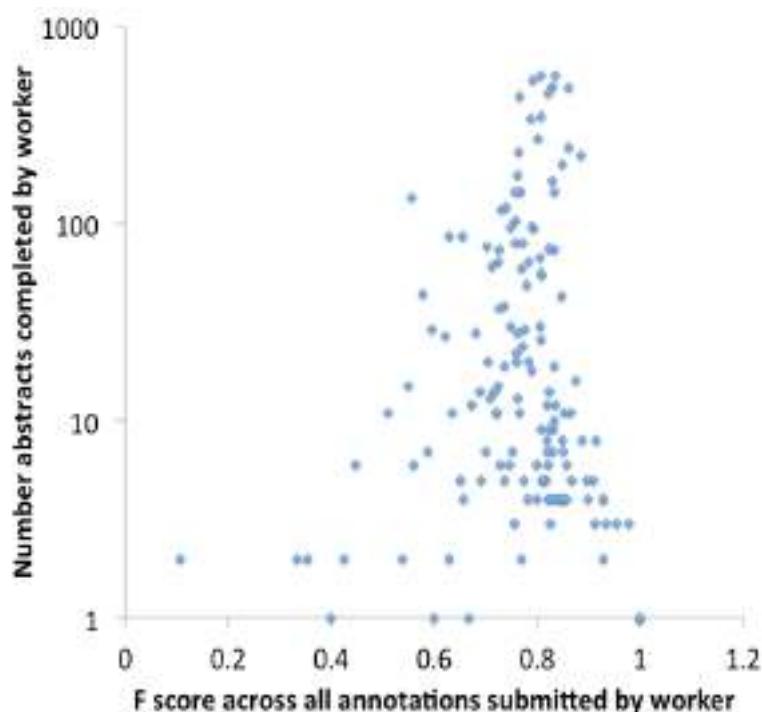


Figure 3: Quality of annotations for each worker compared to number of abstracts processed.

3.2. Wisdom of the crowd aggregation function

For this experiment, we collected annotations from 15 workers for each document (excluding the first four training examples which all workers saw). Following [8] and others, we employed a simple voting strategy to merge the annotations from multiple workers. For each annotation, we counted the number of different workers to produce that annotation and set a threshold K above which the annotation would be kept. We measured the quality of the annotation set generated based on all possible values of K as compared to the original gold standard documents. Figure 4 shows the Precision, Recall and F at each value of K. Precision varies from 0.436 at K=1 to 0.984 at K=15, Recall varies from 0.980 at K=1 to 0.269 at K=15, and the F measure peaks at 0.872 at K=6 (Precision = 0.862, Recall = 0.883).

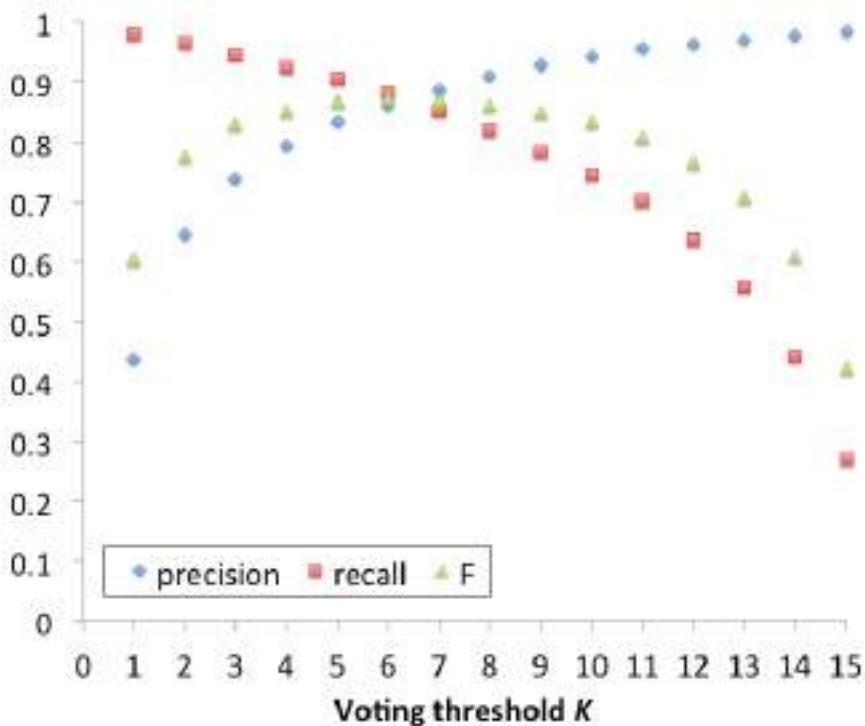


Figure 4: Impact of voting threshold K on the level of agreement with the gold standard.

3.3. Cost

We paid \$.066 per HIT with \$.06 going to the worker and 10% of that (.006) going to Amazon. At 15 workers per document, we paid 99 cents per abstract. In addition, we paid \$.066 for each worker to take the demographic survey and \$.264 for them to complete the first four training documents. 145 workers passed the qualification test and took part in this experiment. The total cost was thus $145 \cdot .33$ (survey and training) + $589 \cdot .99$ (annotation) = \$630.96. All 589 documents were annotated in a span of 9 days. We left the AMT job running until all tasks were completed.

To gauge the impact of annotator redundancy (which equates to cost), we estimated the expected performance of the system with different numbers of workers (referred to as N) per document. (N for the complete collection, whose quality is depicted in Figure 4 was 15.) Using this data, we estimated performance at different values of N by sampling from the workers who completed each document. For example, to estimate performance at $N = 6$, for each document, we randomly selected the annotations from 6 of the workers who annotated that document, calculated the best value for the voting threshold K and recorded the associated F measure. For each value of N , we repeated this random process ten times and recorded the mean and standard deviation for the maximum F measure per N . As Figure 5 illustrates, the range of F scores varies from 0.78 at $N = 1$ to 0.87 at all N greater than 7. The largest increase in performance is between $N = 2$, average $F = 0.78$ and $N = 3$, where average $F = 0.84$. The leveling off of performance increases with increasing workers suggests that, for this task and with this voting strategy, a good balance between cost and quality may be found between 3 and 8 workers (reducing our cost estimates above by more than half).

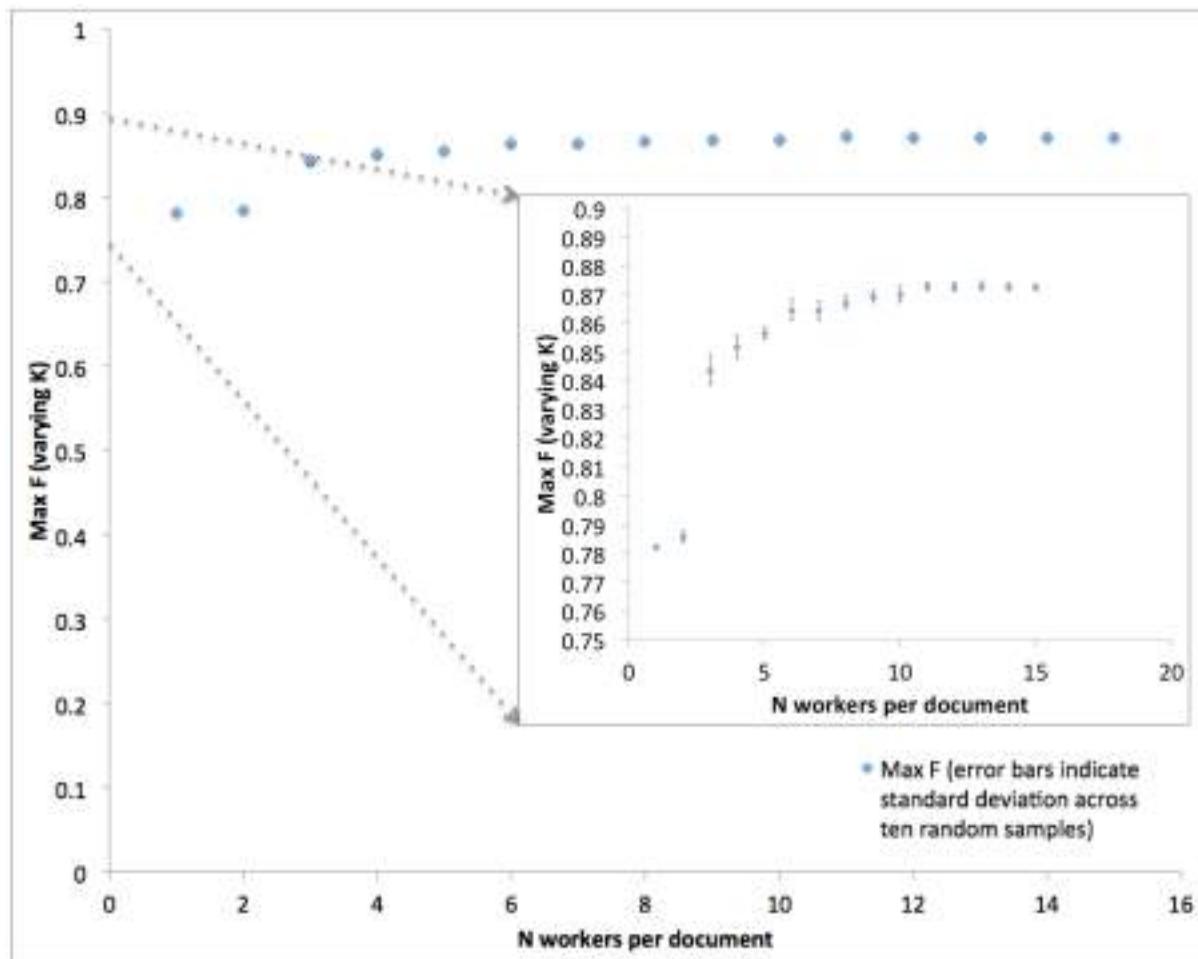


Figure 5: Impact of increasing the number of workers per abstract on annotation quality

3.4. Machine learning

We compared the performance of the BANNER machine learning system [14] given two training sets: a) the original gold standard training set (593 abstracts), and b) the annotations of the 589 abstracts considered here at the voting threshold $K=6$. Training the model with the original gold standard produced an F measure of 0.785 on the 100 abstract test set from the NCBI corpus (Precision 0.808, Recall 0.764). Training the same system on the AMT-generated data produced an F measure of 0.739 (Precision 0.778, Recall 0.703). While this model performed worse on the selected test set than the gold-standard-trained system, it performed substantially better than prior approaches including a BANNER model trained on the original AZDC Disease corpus ($F = 0.35$) and the NCBO Annotator using just the Human Disease Ontology ($F = 0.26$).

3.5. Demographics

According to the pre-task survey, the workers that contributed to this study were 59% female and 41% male. They had a mean age of 32 with 68% between 21 and 35, 19% between 36 and 45, and 11% 46 or older. Regarding education, workers were widely distributed including one worker that reported not finishing high school and four that completed PhD programs. The largest group (28%) had completed a four-year college degree, with an additional 12% completing a masters program. 20% of our AMT workforce reported being unemployed, 15% were students, with the remaining 65% employed in a wide range of occupations with the top categories being “Technical” and “Science”. In general, all of these trends are in agreement with prior demographic studies of the AMT worker population as a whole [15]. Asked why they worked on our HITs in particular, 85% selected “I want to make money”, 69% selected “I want to help science” and 17% selected “Entertainment” (they could select multiple categories). Along with the consistent motivation to earn money, many workers expressed happiness in performing a task that (a) contributed to science and (b) involved them with subject matter that they found educational. This was reiterated in many emails received from the workers.

4. Discussion

One of the key rating-limiting factors in the advancement of the field of NLP is the production of annotated corpora. This is perceived to be even more the case for biomedical applications where the language is complex and full of jargon. This experiment demonstrated how a disease mention annotation corpus with more than 500 PubMed abstracts could be assembled for about \$600 in less than two weeks. Further, we provided evidence that the same system could produce a corpus of the same size with a similar level of quality in much less time and at much lower cost by reducing the number of redundant annotators per document (Figure 5). Previous work demonstrated the potential of crowdsourcing for the creation of gold standard mention annotation corpora with text from clinical trials announcements [8, 9]. Here we extended those results by: testing crowdsourcing for disease annotation on a new text type (PubMed abstracts), benchmarking the impact of increasing the number of annotators per task from 5 to 15, and testing the results of the crowdsourced annotations as input to a machine learning system. Based on these results and on related studies that have achieved good performance for relation verification [10] and extraction [11], we are cautiously optimistic that the crowdsourcing paradigm will be broadly applicable across many BioNLP annotation tasks.

Crowdsourcing-based annotation systems should make it possible to create far more and far larger training sets than would be conceivable with the expert-only approach. But they will be different. Like the dataset generated here, most crowdsourced corpora will not reproduce gold standards exactly. In one sense this is a weakness. Existing NLP systems assume that training and testing data are perfect and binary. A span of text is a disease mention or it is not and there is no grey area. Evaluations of crowd-generated data, such as our assessment of the performance of the crowd-trained BANNER model, may thus show less favorable outcomes. However, in many cases in language the binary premise is likely not reflective of reality. There are always edge cases and ambiguities. Crowdsourced data offer the potential to identify the ambiguities of language and of annotation tasks in a computationally accessible way that could lead to important advances. To illustrate, consider the sentence : “*Significant differences were found between PWS*

patients, SIB controls, and WC controls in the prevalence of febrile convulsions, fever-associated symptoms, and temperature less than 94 degrees F”. In the gold standard, the only annotation captured is “PWS”, an acronym for Prader-Willi syndrome. In the AMT results, 15 of 15 workers highlighted “PWS”, but 13 of 15 workers also highlighted “febrile convulsions” and 11 highlighted “fever-associated symptoms”. Both of these were considered false positives though in other documents highly similar terms such as “benign familial infantile convulsions” and “fever” were part of the gold standard annotations. Another example is provided in the sentence: “*We report the identification of a female patient with the X-linked recessive Lesch-Nyhan syndrome (hypoxanthine phosphoribosyltransferase [HPRT] deficiency)*”. In the gold standard, the phrase “Lesch-Nyhan syndrome” is annotated. However, 11 of 15 AMT workers selected “X-linked recessive Lesch-Nyhan syndrome” instead. (In other gold standard annotations, “X-linked” and “recessive” are often included as modifiers in multi-word disease phrases.) Finally, consider the following title: “*von Willebrand disease type B: a missense mutation selectively abolishes ristocetin-induced von Willebrand factor binding to platelet glycoprotein Ib*”. In the gold standard, both “von Willebrand disease type B” and the second “von Willebrand” are annotated. 13 of 15 workers highlighted the first occurrence correctly yet only 2 highlighted the second, resulting in a false negative. The next sentence of this abstract begins with “*von Willebrand factor (vWF) is a multimeric glycoprotein*” making it questionable whether the second “von Willebrand” was referring to the gene or the disease.

We do not bring these examples up to criticize the NCBI Disease Corpus itself. Arguments could be made for these decisions. The important point is that such edge cases occur in all corpus creation tasks. As a product of the multi-annotator perspective provided by (and in fact required by) crowdsourcing, ambiguity can be quantified directly. Emerging research is showing that such levels of disagreement are signals that can be used both during the training of machine learning models and in more realistic assessments of their predictions [11].

One of the key contributions of the work presented here is to solidify the potential of the crowd of people who will respond to an open call for participation to effectively process biomedical text. Here, we focused on the AMT worker population but other crowds might be tapped for future work. While the AMT workers were shown to be effective, the per-unit cost of this framework does put constraints on the total amount of work that can be performed. Volunteer-based “citizen science” opens up the potential for far more work at lower cost [16]. As long as the collective mind of the crowd continues to outperform computational methods, it may even make sense to tap the crowd to directly perform information extraction tasks in support of biomedical research objectives rather than solely applying them to train computational systems.

Acknowledgments

We would like to thank all of the AMT workers. This research was supported by the National Institute of General Medical Sciences of the National Institutes of Health under award numbers R01GM089820 and R01GM083924, and by the National Center for Advancing Translational Sciences of the National Institute of Health under award number UL1TR001114.

References

1. Jimeno Yepes A, Verspoor K: **Literature mining of genetic variants for curation: quantifying the importance of supplementary material.** *Database (Oxford)* 2014, **2014**:bau003.
2. Trugenberger CA, Walti C, Peregrim D, Sharp ME, Bureeva S: **Discovery of novel biomarkers and phenotypes by semantic technologies.** *BMC Bioinformatics* 2013, **14**:51.
3. Andronis C, Sharma A, Virvilis V, Deftereos S, Persidis A: **Literature mining, ontologies and information visualization for drug repurposing.** *Brief Bioinform* 2011, **12**(4):357-368.
4. Chapman WW, Nadkarni PM, Hirschman L, D'Avolio LW, Savova GK, Uzuner O: **Overcoming barriers to NLP for clinical text: the role of shared tasks and the need for additional creative solutions.** *J Am Med Inform Assoc* 2011, **18**(5):540-543.
5. Yeh A, Morgan A, Colosimo M, Hirschman L: **BioCreAtIvE task 1A: gene mention finding evaluation.** *BMC Bioinformatics* 2005, **6 Suppl 1**:S2.
6. Snow R, O'Connor B, Jurafsky D, Ng AY: **Cheap and fast---but is it good?: evaluating non-expert annotations for natural language tasks.** 2008:254-263.
7. Sabou M, Bontcheva K, Scharl A: **Crowdsourcing research opportunities.** In: *Proceedings of the 12th International Conference on Knowledge Management and Knowledge Technologies - i-KNOW '12*. New York, New York, USA: ACM Press; 2012: 1.
8. Zhai H, Lingren T, Deleger L, Li Q, Kaiser M, Stoutenborough L, Solti I: **Web 2.0-Based Crowdsourcing for High-Quality Gold Standard Development in Clinical Natural Language Processing.** *Journal of Medical Internet Research* 2013, **15**(4):e73.
9. Yetisgen-Yildiz M, Solti I, Xia F, Halgrim SR: **Preliminary experience with Amazon's Mechanical Turk for annotating medical named entities.** In: *CSLDAMT '10 Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*. Stroudsburg, PA, USA: Association for Computational Linguistics; 2010: 180-183.
10. Burger J, Doughty E, Bayer S, Tresner-Kirsch D, Wellner B, Aberdeen J, Lee K, Kann M, Hirschman L: **Validating Candidate Gene-Mutation Relations in MEDLINE Abstracts via Crowdsourcing.** In: *Data Integration in the Life Sciences*. Edited by Bodenreider O, Rance B, vol. 7348: Springer Berlin Heidelberg; 2012: 83-91.
11. Aroyo L, Welty C: **Crowd Truth: Harnessing disagreement in crowdsourcing a relation extraction gold standard.** *WebSci2013 ACM* 2013.
12. Dogan RI, Lu Z: **An improved corpus of disease mentions in PubMed citations.** In: *Workshop on Biomedical Natural Language Processing; Montreal, Canada*. Association for Computational Linguistics 2012: 91-99.
13. Wei CH, Kao HY, Lu Z: **PubTator: a web-based text mining tool for assisting biocuration.** *Nucleic Acids Res* 2013, **41**(Web Server issue):W518-522.
14. Leaman R, Gonzalez G: **BANNER: an executable survey of advances in biomedical named entity recognition.** *Pac Symp Biocomput* 2008:652-663.
15. Ross J, Zaldivar A, Irani L, Tomlinson B: **Who are the turkers? worker demographics in amazon mechanical turk.** *Department of Informatics, University of California, Irvine, USA, Tech Rep* 2009.
16. Good B, Su A: **Crowdsourcing for bioinformatics.** *Bioinformatics* 2013, **29**(16):1925-1933.

CROWDSOURCING IMAGE ANNOTATION FOR NUCLEUS DETECTION AND SEGMENTATION IN COMPUTATIONAL PATHOLOGY: EVALUATING EXPERTS, AUTOMATED METHODS, AND THE CROWD

H. IRSHAD, L. MONTASER-KOUHSARI, G. WALTZ, O. BUCUR, J.A. NOWAK, F. DONG, N.W. KNOBLAUCH and A. H. BECK

*Beth Israel Deaconess Medical Center,
Harvard Medical School, Boston USA*

E-mail: hirshad@bidmc.harvard.edu, lmontase@bidmc.harvard.edu, gwaltz@bidmc.harvard.edu,
obucur@bidmc.harvard.edu, janowak@partners.org, fdong1@partners.org, nknoblau@bidmc.harvard.edu,
abeck2@bidmc.harvard.edu
www.becklab.org

The development of tools in computational pathology to assist physicians and biomedical scientists in the diagnosis of disease requires access to high-quality annotated images for algorithm learning and evaluation. Generating high-quality expert-derived annotations is time-consuming and expensive. We explore the use of crowdsourcing for rapidly obtaining annotations for two core tasks in computational pathology: nucleus detection and nucleus segmentation. We designed and implemented crowdsourcing experiments using the *CrowdFlower* platform, which provides access to a large set of labor channel partners that accesses and manages millions of contributors worldwide. We obtained annotations from four types of annotators and compared concordance across these groups. We obtained: crowdsourced annotations for nucleus detection and segmentation on a total of 810 images; annotations using automated methods on 810 images; annotations from research fellows for detection and segmentation on 477 and 455 images, respectively; and expert pathologist-derived annotations for detection and segmentation on 80 and 63 images, respectively. For the crowdsourced annotations, we evaluated performance across a range of contributor skill levels (1, 2, or 3). The crowdsourced annotations (4,860 images in total) were completed in only a fraction of the time and cost required for obtaining annotations using traditional methods. For the nucleus detection task, the research fellow-derived annotations showed the strongest concordance with the expert pathologist-derived annotations ($F-M = 93.68\%$), followed by the crowd-sourced contributor levels 1,2, and 3 and the automated method, which showed relatively similar performance ($F-M = 87.84\%, 88.49\%, 87.26\%$, and 86.99% , respectively). For the nucleus segmentation task, the crowdsourced contributor level 3-derived annotations, research fellow-derived annotations, and automated method showed the strongest concordance with the expert pathologist-derived annotations ($F-M = 66.41\%, 65.93\%$, and 65.36% , respectively), followed by the contributor levels 2 and 1 (60.89% and 60.87% , respectively). When the research fellows were used as a gold-standard for the segmentation task, all three contributor levels of the crowdsourced annotations significantly outperformed the automated method ($F-M = 62.21\%, 62.47\%$, and 65.15% vs. 51.92%). Aggregating multiple annotations from the crowd to obtain a consensus annotation resulted in the strongest performance for the crowd-sourced segmentation. For both detection and segmentation, crowd-sourced performance is strongest with small images (400 x 400 pixels) and degrades significantly with the use of larger images (600 x 600 and 800 x 800 pixels). We conclude that crowdsourcing to non-experts can be used for large-scale labeling microtasks in computational pathology and offers a new approach for the rapid generation of labeled images for algorithm development and evaluation.

Keywords: Crowdsourcing, Annotation, Nuclei Detection, Nuclei Segmentation, Digital Pathology, Computational Pathology, Histopathology.

1. Introduction

Cancer is diagnosed based on a pathologist's interpretation of the nuclear and architectural features of a microscopic image of a histopathological section of tissue removed from a patient. Over the past several decades, computational methods have been developed to enable pathologists to develop and apply quantitative methods for the analysis and interpretation of histopathological images of cancer.¹ These methods can be used to automate standard methods of histopathological analysis (e.g. nuclear grading),² as well as to discover novel morphological characteristics predictive of clinical outcome (e.g. relational features and stromal attributes), which are difficult or impossible to measure using standard manual approaches.³

Accurate nuclear detection and segmentation is an important image processing step prior to feature extraction for most computational pathology analyses. In the past decade a large number of methods have been proposed for automated nuclear detection and segmentation.⁴ However, despite the generation of a large number of competing approaches for these tasks, the comparative performance of nuclei detection and segmentation methods has not been evaluated rigorously.

A major barrier to rigorous comparative evaluation of existing methods is the time and expense required to obtain expert-derived labeled images. Using traditional approaches, obtaining labeled images requires enlisting the support of a trained research fellow and/or pathologist to annotate microscopic images. Most computational labs do not have access to support from highly trained physicians and research staff to annotate images for algorithm development and evaluation, and even in pathology research laboratories, obtaining high-quality hand-labeled images is a significant challenge, as the task is time-consuming and can be tedious.

These challenges are exacerbated when attempting large-scale image annotation projects of hundreds-to-thousands of images. Further, recent advances in whole slide imaging are enabling the generation of large archives of whole slide images (WSIs) of disease. In contrast to images obtained from a standard microscope camera (which will capture a single region-of-interest (ROI) per image), WSIs are large and capture tissue throughout the entire slide, which typically contains thousands of ROIs and tens-of-thousands of nuclei per WSI.⁵ Thus, it is not feasible to obtain comprehensive annotation from pathologists or research fellows in a single research laboratory for large sets of WSIs.

In this project, we explore the use of crowdsourcing as an alternative method for obtaining large-scale image annotations for nucleus detection and segmentation. In recent years, crowdsourcing has been increasingly used for bioinformatics, with image annotation representing an important application area.⁶ Crowdsourced image annotation has been successfully used to serve a diverse set of scientific goals, including: classification of galaxy morphology,⁷ the mapping of neuron connectivity in the mouse retina,⁸ the detection of sleep spindles from EEG data,⁹ and the detection of malaria from blood smears.^{10,11} To our knowledge, no prior published studies have used non-expert crowdsourced image annotation for nucleus detection and segmentation from histopathological images of cancer. The Cell Slider project by the Cancer Research UK (<http://www.cellslider.net/>) launched in October 2012 is attempting to use crowdsourcing to annotate cell types in histopathological images of breast cancer; however, to our knowledge results of this study have not yet been released.

Here, we provide a framework for understanding and applying crowdsourcing to the annotation of histopathological images obtained from a large-scale WSI dataset. We develop and evaluate this framework in the setting of nucleus detection and segmentation from a set of WSIs obtained from renal cell carcinoma cases that previously underwent comprehensive molecular profiling as part of The Cancer Genome Atlas (TCGA) project.¹²

We performed a set of experiments to compare the annotations achieved by: pathologists, trained research fellows, and non-expert crowdsourced annotators. For the crowdsourced image annotators, we performed additional experiments to gain insight into factors that influence contributor performance, including: assessing the relationship of the contributor's pre-defined skill level with the contributor's performance on the nucleus detection and segmentation tasks; assessing the influence of image size on contributor performance; and comparing performance based on a single annotation-per-image versus aggregating multiple contributor annotations-per-image.

The remainder of the paper is organized as follows. Section 2 describes the dataset used for the study, and the proposed framework for evaluating performance of nucleus detection and segmentation. Experimental results are presented in Section 3, and concluding remarks and proposed future work are presented in Section 4.

2. Method

In this section, we describe the dataset, *CrowdFlower* platform, and design of our experiments.

2.1. Dataset

The images used in our study come from WSIs of kidney renal clear cell carcinoma (KIRC) from the TCGA data portal. TCGA represents a large-scale initiative funded by the National Cancer Institute and National Human Genome Research Institute. TCGA has performed comprehensive molecular profiling on a total of approximately ten-thousand cancers, spanning the 25 most common cancer types. In addition to the collection of molecular and clinical data, TCGA has collected WSIs from most study participants. Thus, TCGA represents a major resource for projects in computational pathology aiming at linking morphological, molecular, and clinical characteristics of disease.^{13,14}

We selected 10 KIRC whole slide images (WSI) from the TCGA data portal (<https://tcga-data.nci.nih.gov/tcga/>), representing a range of histologic grades of KIRC. From these WSIs, we identified nucleus-rich ROIs and extracted 400×400 pixel size images ($98.24\mu\text{m} \times 98.24\mu\text{m}$) for each ROI at 40X magnification. The total number of images per region of interest is 81. Finally, we obtained a total of 810 images from the 10 KIRC WSIs.

2.2. CrowdFlower Platform

We employ the *CrowdFlower* platform to design jobs, access and manage contributors, and obtain results for the nucleus detection and segmentation image annotation jobs. *CrowdFlower* is a crowdsourcing service that works with over 50 labor channel partners to enable access to a network of more than 5 million contributors worldwide. The *CrowdFlower* platform

provides several features aimed at increasing the likelihood of obtaining high-quality work from contributors. Jobs are served to contributors in tasks. Each task is a collection of one or more images sampled from the data set. Prior to completing a job, the platform requires contributors to complete job-specific training. In addition, contributors must complete test questions both before (*quiz mode*) and throughout (*judgment mode*) the course of the job. Test questions serve the dual purpose of training contributors and monitoring their performance. Contributors must obtain a minimum level of accuracy on the test questions to be permitted to complete the job. *CrowdFlower* categorizes contributors into three skill levels (1,2,3) based on performance on other jobs, and when designing a job the job designer may target a specific contributor skill level. In addition, the job designer specifies the payment per task and the number of annotations desired per image. After job completion, *CrowdFlower* provides the job designers with a confidence map for each annotated image. The confidence map is an image in the same dimension as the input image, but the pixel intensity now represents an aggregation of annotations to that image, which is weighted by both the annotation agreement among contributors and each contributor's trust level. Additional information on the *CrowdFlower* platform is available at www.crowdflower.com.

2.3. Job Design

Our study includes two types of image annotation jobs: nucleus detection and segmentation. The contributors used a dot operator (by clicking at the center of a nucleus) for nucleus detection and a polygon operator (by drawing a line around the nucleus) for nuclei segmentation. Each job contains instructions, which provide examples of expert-derived annotations and guidance to assist the contributor in learning the process of nuclear annotation. These instructions are followed by a set of test questions. Test questions are presented to the contributor in one of two modes: quiz mode and judgment mode. Quiz mode occurs at the beginning of a job (immediately following the instructions), while judgment mode test questions are interspersed throughout the course of completing a job. In our experiments, contributors were required to achieve at least 40% accuracy on five test questions in quiz mode in order to qualify for annotation of unlabeled images from the job during judgment mode. In judgment mode, each task consists of four unlabeled images and one test question image, which is presented to the contributor in the same manner as the unlabeled images, such that the contributor is unaware if he/she is annotating an unlabeled image or a test question. The total pool of quiz and judgment mode test questions used in our study was based on 20 images, which had been annotated by medical experts. If the contributor's accuracy decreased to below 40% during judgment mode, the contributor was barred from completion of additional annotations for the job.

There are several additional job design options provided by the *CrowdFlower* platform which may influence annotation performance. The *CrowdFlower* platform divides the contributors into three skill levels based on their performance on prior jobs, and the job designer can target jobs to specific contributor skill levels. In our experiments, we compared performance when targeting jobs to each skill level. The job designer must specify the number of annotations to collect per image. For most of our experiments, we used a single annotation

per image. In addition, we conducted an experiment for the image segmentation job, in which the number of contributors per image ranged from 1 to 3 to 5, and we compared performance across these three levels of redundancy.

In addition to the annotations obtained from the non-expert crowd, we obtained annotations from three additional types of labelers: published state-of-the-art automated nucleus detection and segmentation algorithms;¹⁵ research fellows trained for these specific jobs; and MD-trained surgical pathologists, who have completed residency in Anatomic Pathology.

3. Experiments

3.1. Performance Metrics

Detection Metrics: A detected nucleus was accepted as correctly detected if the coordinates of its centroid were within a range of 15 pixels ($3.75\mu m$) from the centroid of a ground truth nucleus. The metrics used to evaluate nucleus detection include: number of true positives (TP), number of false positives (FP), number of false negatives (FN), sensitivity or true positive rate ($TPR = \frac{TP}{TP+FN}$), precision or positive predictive value ($PPV = \frac{TP}{TP+FP}$) and F-Measure ($F - M = 2 \times \frac{TPR \times PPV}{TPR + PPV}$). The TPR and PPV are presented in Tables 1 - 4 with their 95% Confidence Intervals, computed using the prop.test function in the stats package in R.

Segmentation Metrics: The metrics used to evaluate segmentation annotation include: sensitivity ($TPR = \frac{|A(G) \cap A(S)|}{|A(G)|}$ - proportion of nucleus pixels that are correctly labeled as positive), specificity or true negative rate ($TNR = \frac{|I - (A(G) \cup A(S))|}{|I - A(G)|}$ - proportion of non-nucleus pixels that are correctly labeled as negative), precision ($PPV = \frac{|A(G) \cap A(S)|}{|A(S)|}$), F-Measure, and Overlap = $\frac{|A(G) \cap A(S)|}{|A(G) \cup A(S)|}$; where I is the image, $A(S)$ is the area of the segmented nuclei, $A(G)$ is the area of the ground truth nuclei.

3.2. Detection Results

In the first experiment, we considered pathologist's annotations as ground truth (GT). Pathologists provided annotations on a total of 80 study images. For these 80 images, we assessed the performance of research fellows, the automated method, and non-expert contributors from three skill levels as shown in Table 1. Focusing on the F-M measure (which incorporates both TPR and PPV), we observe the strongest performance for the research fellow, followed by similar performance for the three other annotation groups (FM between 86.99% and 88.49%) as shown in Table 1.

In the second experiment, we considered annotations from research fellows as GT. The research fellows provided annotations for a total of 477 images containing 25,323 annotated nuclei, considered as GT nuclei in this experiment. Thus, the dataset for this evaluation is significantly larger than that for the initial analysis, which used the pathologist annotation as the GT. In this experiment, all four groups showed similar performance, with F-M scores between 83.94% and 85.32%, as shown in Table 2.

In the third experiment, we used the annotations produced by the automated method as the GT. The automated method was run on all 810 study images and detected a total of

44,281 nuclei which were considered as GT nuclei in this experiment. We compared these GT nuclei with the three crowdsourced contributor levels across all 810 images and results are shown in Table 3. Overall, the three contributor levels achieved similar TPR levels, with a significantly higher PPV for Contributor Level 2, resulting in the highest F-M for Contributor Level 2(83.99%), with slightly lower F-M's achieved by Contributor Levels 1 and 2, as shown in Table 3. Visual examples of nucleus detection by different level of contributors are shown in Figure 1.

On the *CrowdFlower* platform interface, individual nuclei are rendered at relatively larger size on smaller images as compared to larger images. Further, smaller images contain fewer nuclei per image. To assess the influence of image size on contributor performance, we performed an experiment in which we extracted images of three different sizes (400×400 , 600×600 and 800×800) from the same ROIs. We collected annotations with Contributor Level 2 and compared the annotations with those obtained with automated methods as shown in Table 4. The image size 400×400 performed significantly better than the larger image sizes. These

Table 1. Detection results on 80 images (Pathologists' annotation as GT) GT nuclei = 4436

Annotations	TP	FN	FP	TPR %	PPV %	F-M %
Research Fellow	4109	327	227	92.63 ± 0.8	94.76 ± 0.7	93.68
Automated Method	3735	701	416	84.20 ± 1.1	89.98 ± 1.0	86.99
Contributor Level 1	3814	622	434	85.98 ± 1.1	89.78 ± 1.0	87.84
Contributor Level 2	4016	420	625	90.53 ± 0.9	86.53 ± 1.0	88.49
Contributor Level 3	3787	649	457	85.37 ± 1.1	89.23 ± 0.9	87.26

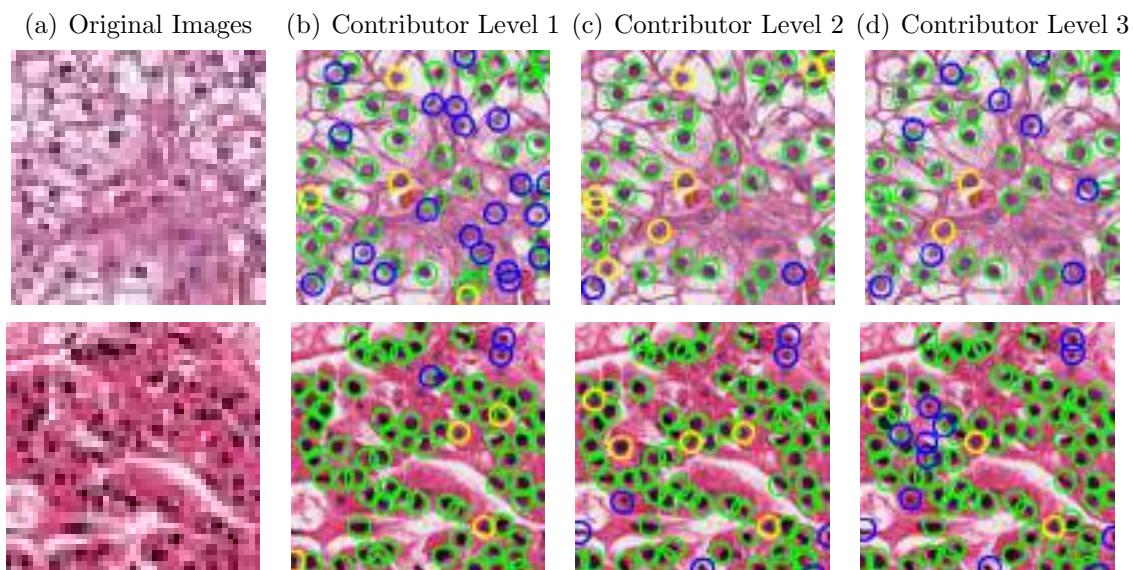


Fig. 1. Examples of nucleus detection results produced by different contributor levels (Green circle indicates TP nuclei, yellow circle indicates FN nuclei and blue circle indicates FP). The automated detected nuclei were used as ground truth.

Table 2. Detection results on 477 images (Research fellows' annotation as GT) GT nuclei = 25323

Annotations	TP	FN	FP	TPR %	PPV %	F-M %
Automated Method	21177	4146	3955	83.63 ± 0.5	84.26 ± 0.5	83.94
Contributor Level 1	21495	3828	3982	84.88 ± 0.4	84.37 ± 0.5	84.63
Contributor Level 2	22488	2835	4904	88.80 ± 0.4	82.10 ± 0.5	85.32
Contributor Level 3	21788	3535	4049	86.04 ± 0.4	84.33 ± 0.5	85.18

Table 3. Detection results on 810 images (Automated Method as GT) GT nuclei = 44281

Annotations	TP	FN	FP	TPR %	PPV %	F-M %
Contributor Level 1	35823	8458	7792	80.90 ± 0.4	82.13 ± 0.4	81.51
Contributor Level 2	36191	8090	5705	81.73 ± 0.4	86.38 ± 0.3	83.99
Contributor Level 3	36125	8156	6874	81.58 ± 0.4	84.01 ± 0.4	82.78

Table 4. Detection results on different image sizes (Automated Method as GT) GT nuclei = 44281

Annotations	TP	FN	FP	TPR %	PPV %	F-M
Image Size 400 × 400 (810 images)	36191	8090	5705	81.73 ± 0.4	86.38 ± 0.3	83.99
Image Size 600 × 600 (380 images)	24870	19411	16993	56.16 ± 0.5	59.41 ± 0.5	57.74
Image Size 800 × 800 (170 images)	12144	32137	21842	27.42 ± 0.4	35.73 ± 0.5	31.03

results suggest that defining a small image size is important for obtaining optimal performance when using crowdsourced microtasks for image annotation for complex and tedious work, such as nucleus detection.

3.3. Segmentation Results

Like nucleus detection, we also performed four experiments for nuclear segmentation. In the first experiment, we considered pathologist's nuclear segmentation as GT segmentation. Pathologists provided annotation on 63 images. We compared those 63 segmented images with the segmentations produced by research fellows, an automated method and three different level of contributors' annotation as shown in Table 5. The strongest performance was achieved by Contributor Level 3, research fellow, and the automated method, which all achieved F-M scores between 65.93% and 66.41%. The Contributor Levels 1 and 2 showed slightly worse performance with F-M scores of 60.9% as shown in Table 5.

In this experiment, we considered the research fellow-derived nuclear segmentation as GT segmentation, which we obtained on 455 images. On these 455 images, all three levels of crowdsourced non-expert annotations significantly outperformed the automated method, as shown in Table 6. Overall, Contributor Level 3 achieved the highest TPR (76.47%), F-Measure (65.15%) and overlap (48.68%).

In our next experiment, we compared the annotations of different contributor levels and used the automated method annotations as the GT across all 810 study images, as shown Table 7. Contributor Level 3 achieved the highest TPR(75.78%), PPV(57.83%), F-Measure(62.10%) and overlap (46.75%), significantly outperforming Contributor Levels 1 and 2. Visual examples of different level of contributor annotations are shown in Figure 2.

Next, we assessed the relationship of contributor performance with image size for the job of nuclear segmentation. As we did for the nucleus detection experiment, we extracted three different image sizes(400×400 , 600×600 and 800×800) from the same ROIs. We collected annotations from Contributor Level 2 and compared them with automated methods as shown in Table 8. As we observed for the nucleus detection job, annotation performance for the nuclear segmentation job was highest for the image size 400×400 and degraded significantly when image size was increased, as shown in Table 8.

In addition to single contributor annotation, we collected multiple contributor annotations-per-image for the segmentation job. As shown in Figure 3, nuclei segmentation performance improved with increasing levels of annotation aggregation. A visual example of nuclei segmentation performance with multiple annotators is shown in Figure 4. Figure 5 shows the

Table 5. Segmentation results on 63 images (Pathologists' annotation as GT)

Annotations	TPR %	PPV %	F-M %	TNR %	Overlap %
Research Fellow	60.40	79.80	65.93	97.68	49.66
Automated Method	76.22	62.26	65.36	96.34	49.87
Contributor Level 1	56.95	71.47	60.87	97.08	44.30
Contributor Level 2	59.02	71.19	60.89	97.04	44.46
Contributor Level 3	67.73	69.07	66.41	96.86	50.14

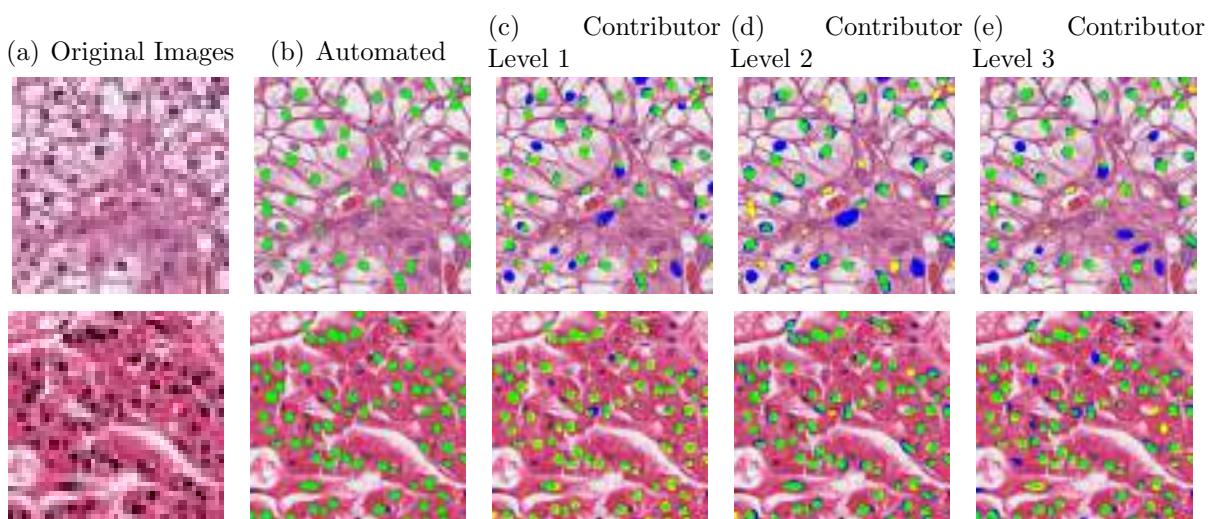


Fig. 2. Examples of nuclear segmentation using an automated method and increasing contributor skill level, ranging from 1 to 3. (Green region indicates TP region, yellow region indicates FN region and blue region indicates FP region). The automated nuclei segmentation used as ground truth.

Table 6. Segmentation results on 455 images (Research Fellows' annotation as GT)

Annotations	TPR %	PPV %	F-M %	TNR %	Overlap %
Automated Method	60.28	48.01	51.92	69.04	40.33
Contributor Level 1	70.13	60.73	62.21	93.58	45.85
Contributor Level 2	68.93	63.98	62.47	94.19	45.95
Contributor Level 3	76.47	59.23	65.15	93.69	48.68

aggregated results of the contributors on a test question for both the nuclei detection and segmentation jobs.

3.4. Cost and Time Analysis, and the Heterogeneity of the Crowd

The Cost and Time analysis aggregated across all images for nucleus detection and segmentation and stratified by Contributor Level are shown on the left-panel in Figure 6, and the time analysis for one image across different contributor levels is shown on the right-panel in Figure 6. These data show that the segmentation job accounts for significantly more time per task and time overall, suggesting that nuclei segmentation is the more complex of the jobs. For the nucleus detection job, the waiting time required for attracting contributors to the job was significantly longer for the higher skill level contributors and the annotation time spent

Table 7. Segmentation results on 810 images (Automated Method's annotation as GT)

Annotations	TPR %	PPV %	F-M %	TNR %	Overlap %
Contributor Level 1	74.17	52.49	57.34	93.10	41.80
Contributor Level 2	74.14	49.31	54.17	91.54	38.97
Contributor Level 3	75.78	57.83	62.10	95.21	46.75

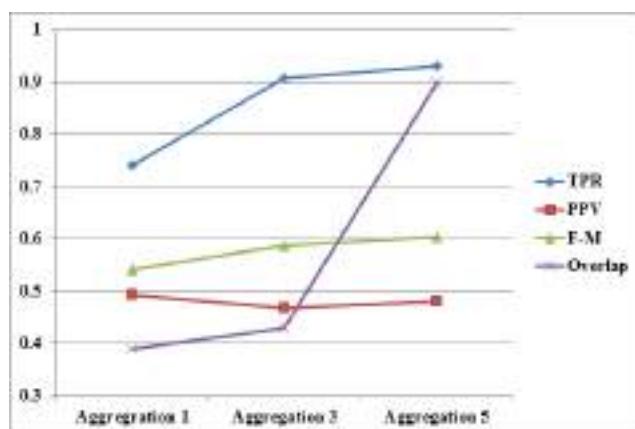


Fig. 3. Graph showing TPR, PPV, F-M and overlap curves for nuclear segmentation results using increasing numbers of aggregated contributor (level 2) annotations (from 1 to 3 to 5). The automated segmentation used as ground truth.

Table 8. Segmentation results on 63 images (Automated Method as GT)

Annotations	TPR %	PPV %	F-M %	TNR %	Overlap %
Image Size 400 × 400	74.14	49.31	54.17	91.54	38.97
Image Size 600 × 600	69.27	30.68	36.75	84.96	24.06
Image Size 800 × 800	44.65	42.10	25.32	80.65	15.87

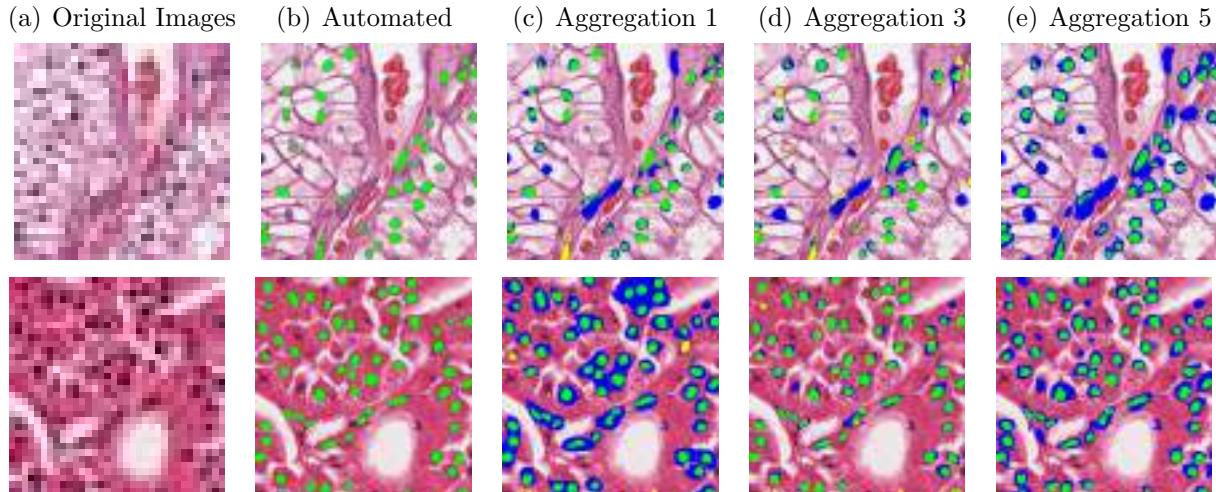


Fig. 4. Examples of nuclear segmentation using an automated method and increasing levels of aggregation from Contributor Level 2, ranging from 1 to 3 to 5. (Green region indicates TP region, yellow region indicates FN region and blue region indicates FP region). The automated nuclei segmentation was used as ground truth.

completing the job was also longer for the more skilled workers. For the segmentation job (which is the more complex job), the overall waiting time and annotation time were shorter for the Contributor Level 3 workers as compared with the Level 1 and 2 workers, likely owing to the fact that the Contributor Level 3 workers may have been more attracted to the higher complexity job. The distribution of contributor judgment *trust level*, which reflects contributor performance on test questions, is displayed in Figure 7. These plots show that although the highest proportion of judgments come from contributors with moderate-to-high trust levels (80% - 90% trust level), there is a wide distribution of contributor trust levels with a significant number of judgments derived from contributors with only moderate-to-poor trust levels. These results suggest the value of targeting specific jobs to specific crowd skill levels, and that by better targeting jobs to the appropriate crowds, we may obtain improvements in performance.

4. Conclusions

Our experiments show that crowdsourced non-expert-derived scores perform at a similar level to research fellow-derived scores and automated methods for nucleus detection and segmentation, with the research fellow annotations showing the strongest performance for detection, and the crowdsourced level 3 scores showing the strongest performance for segmentation. We

conclude that crowdsourced image annotation is a highly-scalable and effective method for obtaining nuclear annotations for large-scale projects in computational pathology. Our results show that performance may be improved further by aggregating multiple crowd-sourced annotations per image, and by targeting jobs to specific crowds based on the complexity of the job and the skill level of the contributors. Ultimately, we expect that large-scale crowdsourced image annotations will lead to the creation of massive, high-quality annotated histopathological image datasets, which will support the improvement of supervised machine learning algorithms for computational pathology and will enable the design of systematic and rigorous comparative analyses of competing approaches, ultimately leading to the identification of top-performing methods, which will power the next generation of computational pathology research and practice.

5. Acknowledgements

Research reported in this publication was supported by the National Library of Medicine of the National Institutes of Health under Award Number K22LM011931. We thank Ari Klein, Nathan Zukoff, Sam Rael and the *CrowdFlower* team for their support, and we thank all the

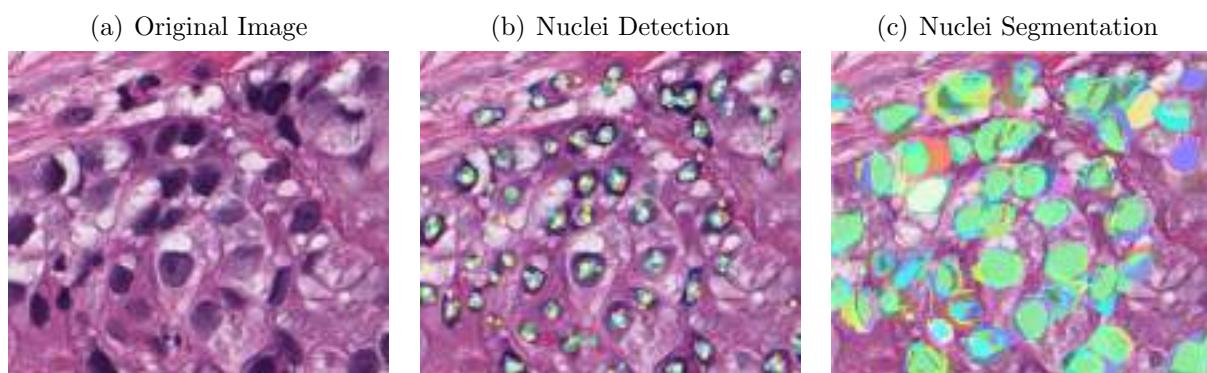


Fig. 5. Aggregation of results of the contributors on a test question. Different colors of dots and region represent different contributor annotations.

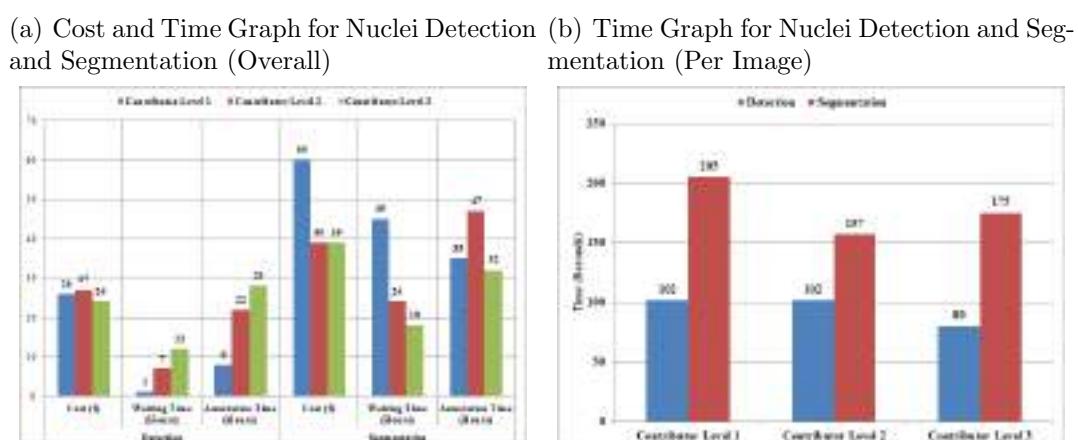


Fig. 6. Time and Cost Analysis for Nucleus Detection and Segmentation for Different Contributor Levels.

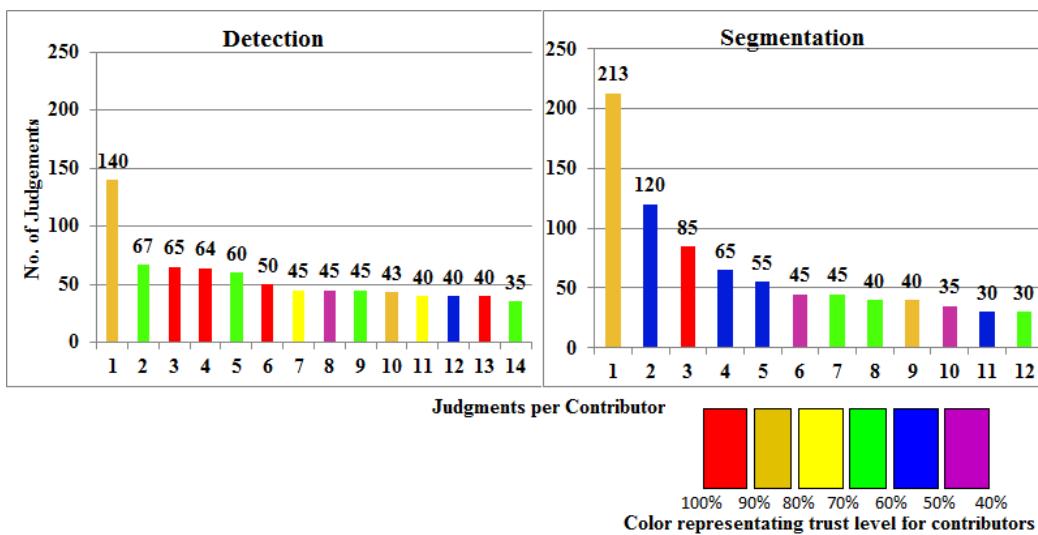


Fig. 7. Distribution of contributor judgments and trust level.

image annotation contributors for making this work possible.

References

1. M. N. Gurcan, L. E. Boucheron, A. Can, A. Madabhushi, N. M. Rajpoot and B. Yener, *Biomedical Engineering, IEEE Reviews in* **2**, 147 (2009).
2. A. Dawson, R. Austin Jr and D. Weinberg, *American journal of clinical pathology* **95**, S29 (1991).
3. A. H. Beck, A. R. Sangoi, S. Leung, R. J. Marinelli, T. O. Nielsen, M. J. van de Vijver, R. B. West, M. van de Rijn and D. Koller, *Science translational medicine* **3**, 108ra113 (2011).
4. H. Irshad, A. Veillard, L. Roux and D. Racoceanu, *Biomedical Engineering, IEEE Reviews in* **7**, 97 (2014).
5. F. Ghaznavi, A. Evans, A. Madabhushi and M. Feldman, *Annual Review of Pathology: Mechanisms of Disease* **8**, 331 (2013).
6. B. M. Good and A. I. Su, *Bioinformatics* , p. btt333 (2013).
7. C. J. Lintott, K. Schawinski, A. Slosar, K. Land, S. Bamford, D. Thomas, M. J. Raddick, R. C. Nichol, A. Szalay, D. Andreeescu *et al.*, *Monthly Notices of the Royal Astronomical Society* **389**, 1179 (2008).
8. J. S. Kim, M. J. Greene, A. Zlateski, K. Lee, M. Richardson, S. C. Turaga, M. Purcaro, M. Balkam, A. Robinson, B. F. Behabadi *et al.*, *Nature* **509**, 331 (2014).
9. S. C. Warby, S. L. Wendt, P. Welinder, E. G. Munk, O. Carrillo, H. B. Sorensen, P. Jenum, P. E. Peppard, P. Perona and E. Mignot, *Nature methods* **11**, 385 (2014).
10. S. Mavandadi, S. Dimitrov, S. Feng, F. Yu, U. Sikora, O. Yaglidere, S. Padmanabhan, K. Nielsen and A. Ozcan, *PLoS One* **7**, p. e37245 (2012).
11. M. A. Luengo-Oroz, A. Arranz and J. Frean, *Journal of medical Internet research* **14** (2012).
12. C. G. A. R. Network *et al.*, *Nature* **499**, 43 (2013).
13. J. Kong, L. A. Cooper, F. Wang, D. A. Gutman, J. Gao, C. Chisolm, A. Sharma, T. Pan, E. G. Van Meir, T. M. Kurc *et al.*, *Biomedical Engineering, IEEE Transactions on* **58**, 3469 (2011).
14. H. Chang, G. V. Fontenay, J. Han, G. Cong, F. L. Baehner, J. W. Gray, P. T. Spellman and B. Parvin, *BMC bioinformatics* **12**, p. 484 (2011).
15. S. D. Cataldo, E. Ficarra, A. Acquaviva and E. Macii, *Computer Methods and Programs in Biomedicine* **100**, 1 (2010).

ANALYZING SEARCH BEHAVIOR OF HEALTHCARE PROFESSIONALS FOR DRUG SAFETY SURVEILLANCE

DAVID J. ODGERS

*Center for Biomedical Informatics, Stanford University, 1265 Welch Road
Stanford CA 94305-5479 USA
Email: djodgers@stanford.edu*

RAVE HARPAZ

*Center for Biomedical Informatics, Stanford University, 1265 Welch Road
Stanford CA 94305-5479 USA
Email: rhpaz@stanford.edu*

ALISON CALLAHAN

*Center for Biomedical Informatics, Stanford University, 1265 Welch Road
Stanford CA 94305-5479 USA
Email: alison.callahan@stanford.edu*

GREGOR STIGLIC

*Faculty of Health Sciences, University of Maribor, Zitna ulica 15
2000 Maribor, Slovenia
Email: gregor.stiglic@um.si*

NIGAM H. SHAH

*Center for Biomedical Informatics, Stanford University, 1265 Welch Road
Stanford CA 94305-5479 USA
Email: nigam@stanford.edu*

Post-market drug safety surveillance is hugely important and is a significant challenge despite the existence of adverse event (AE) reporting systems. Here we describe a preliminary analysis of search logs from healthcare professionals as a source for detecting adverse drug events. We annotate search log query terms with biomedical terminologies for drugs and events, and then perform a statistical analysis to identify associations among drugs and events within search sessions. We evaluate our approach using two different types of reference standards consisting of known adverse drug events (ADEs) and negative controls. Our approach achieves a discrimination accuracy of 0.85 in terms of the area under the receiver operator curve (AUC) for the reference set of well-established ADEs and an AUC of 0.68 for the reference set of recently labeled ADEs. We also find that the majority of associations in the reference sets have support in the search log data. Despite these promising results additional research is required to better understand users' search behavior, biasing factors, and the overall utility of analyzing healthcare professional search logs for drug safety surveillance.

1. Background

Drug safety surveillance is a significant problem given the frequency of post-market adverse drug events (ADEs) that compromise patient health and result in increased costs and burden on the healthcare system [1-4]. The FDA adverse event reporting system (FAERS) is currently the main source for detecting post-market ADEs, but has recognized shortcomings [2, 5-8] such as under-reporting of adverse events (AEs), an issue that is not unique to FAERS [9]. The time required for a sufficient signal when using such sources can delay the release of ADE alerts [10, 11]. As alternatives to spontaneous ADE reports, other sources of data that have been used for detecting drug-AE associations include claims data [12], electronic health records (EHRs) [13-15] and consumer search logs [16-18]. There is a growing trend to use increasingly diverse sources for detecting ADEs, including online social networks. For example, Twitter has been used as a source for mining drug-AE associations [19] with the recognition that more work is required to establish threshold signal levels from such sources towards validating discovered associations. Another data source that is being investigated for ADE detection is the biomedical literature [20-22]. In addition to examining a larger variety of sources, approaches have also been recently developed to combine signals from data sources such as EHRs, claims, biomedical literature, and Internet search logs with signals from FAERS [18, 23, 24].

In this work we explore the potential for using search logs from healthcare professionals as an observational data source for drug safety surveillance. We use two years of search logs from UpToDate, an online source of health information provided by Wolters-Kluwer that includes detailed descriptions of symptoms, diseases, drugs and indications to support evidence based medicine. UpToDate is used on a subscription basis by institutions and any individual who purchases a license. Typically, medical and research institutions purchase licenses for UpToDate, which are then used by members of the institution – this can include physicians, researchers and students. UpToDate use in hospitals is associated with fewer patient complications and adverse events, shorter hospital stays, reduced mortality rates and higher quality performance measures [25, 26]. How medical professionals use UpToDate thus has potentially direct implications for patient health, some of which may be discovered by analyzing the records of this use. Logs of UpToDate use capture the source institution and machine used for a search, the search string entered, the time and date of the search, the type of search, and UpToDate pages visited as a result of the search. UpToDate search logs are thus a rich resource with the potential to enable time-sensitive and context-aware analyses of search behavior. UpToDate search logs are also an important source of observational data for drug safety surveillance because the majority of UpToDate users are health professionals who access UpToDate during their day-to-day practice of providing patient care. In contrast, web search logs and social media capture a broad range of online behavior with unknown context(s).

We investigate an approach to detect drug-AE pairs by first annotating individual UpToDate search logs to identify drugs and events, and then performing association analysis on drug-event pairs identified within a predefined surveillance period. The novelty of our approach is that it capitalizes on a previously untapped source of observational data for detecting drug-AE associations. We assess the performance of our approach by using a reference standard of well-

established associations from the EU-ADR project as well as a reference set of recently labeled ADEs obtained from 2013 FDA product labeling revisions. Our findings are two-fold: (1) the majority of associations from the two reference sets have support in the UpToDate search logs and (2) the approach we investigated can detect ADEs with an accuracy (measured using the area under the receiver operating curve, AUC) of 0.85 for the reference set of well-established ADEs, and an AUC of 0.68 for the reference set of recently labeled ADEs—a result that merits further investigation. It is possible that better methods are required for analyzing healthcare professional search logs to reliably detect more recently reported (and potentially unreported) ADEs.

2. Methods

2.1. Search logs used for analysis

We used two years of UpToDate search logs spanning January 2011 to December 2012 as the data source for exploring associations between drugs and AEs. The log for a single UpToDate search consists of a query string, unique session ID, Internet protocol (IP) address of the computer on which the search was performed and the timestamp of the search. The logs were pre-processed to include only those records for string search queries. Example search logs are shown in Table 1.

Table 1. Three example UpToDate search logs. Institution, session ID and IP addresses have been replaced with fake values but search terms are taken directly from existing records.

<i>Search term</i>	<i>Institution</i>	<i>Session ID</i>	<i>IP address</i>	<i>Time</i>
adrenal insufficiency	A	000018A4C27BDCC	123.456.789	2012-08-01 20:13:23
elevated LDH	B	002679154AE8055C	456.789.199	2012-08-01 15:14:31
effexor xr	C	01B6280B71230987	789.123.456	2012-08-01 11:16:15

The session ID allows us to identify queries performed within a single session, the IP address allows us to identify searches performed on a single machine, and the timestamp enables the temporal reconstruction of the click sequence within a given session.

2.2. Query annotation

We processed the query strings of the search logs using our previously described text processing workflow [27, 28] to annotate the queries with drug and disease terms from biomedical ontologies. Briefly, query strings were processed using MGREP with a lexicon of more than 3 million strings built from biomedical ontologies and terminologies, in which terms and concepts are mapped by synonymy and parent-child relationships. Query strings that match drug names (*e.g.* brand drugs, multi-ingredient drugs, different preparations, dose, form, and salt ingredients) are mapped to the drug’s active ingredient by using RxNorm relationships to obtain all RxNorm codes associated with the given active ingredient. These were then mapped to UMLS concept codes (concept unique identifiers, or CUIs). Query strings that match condition mentions (*i.e.* potential adverse events) are also annotated with UMLS CUIs based on mappings between terms and concepts in UMLS.

2.3. Modeling assumptions

In performing our analysis we make several assumptions about the nature of the UpToDate log data and the manner in which healthcare professionals search for information about drugs and adverse events they are concerned about. We acknowledge that healthcare professionals may search for drug related information beyond the case of adverse events (*e.g.* drug indications), but in this study we restrict our analysis and target only adverse events. We assume that when a healthcare professional is concerned about a potential association between a drug and an AE, and desires to retrieve information about the association, they will perform searches for the drug and event within a short period of time (possibly even in the same query). It is assumed that the longer the time interval between the search for a drug and event, the less likely it is that the searches for drug and event are related to each other or part of the same thought or decision process. We do not assume an ordering on the search terms for the drug and event of interest (either the drug or the AE can be searched for first). We allow that a user may search for the drug and AE multiple times and cycle through searches of the same drug and AE within the same session. We further assume that each session (identified by the session ID) was performed by a single unique healthcare professional (but not the converse), and allow that a unique IP address may be associated with multiple different healthcare professionals. Given these assumptions, we use unique sessions, rather than unique queries or unique IPs as our basic unit for counting and subsequent association analyses.

2.4. Association analysis

We consider three scenarios related to the search order of given drug-event pair of interest: (1) the search for the drug precedes the search for the event; (2) the search for the event precedes the drug; and (3) order is irrelevant. We restrict the amount of time that may elapse between the search of a drug and an event (depending on the ordering) by defining a surveillance period that is indexed (starts) with the first mention of a drug or event (depending on ordering) within a session, and ends at a pre-specified amount of time later within the same session (see Figure 1). For a given drug-event pair of interest, a given pre-specified surveillance period, and a given pre-specified drug-event search ordering, we compute the following 2x2 contingency table

	event	no event
drug	a	b
no drug	c	d

where **a** is the number of unique sessions including search terms associated with the drug and event of interest that falls within the surveillance period indexed by the first mention of a drug (or event), **b** is the number of unique sessions including search terms associated with the drug of interest but not the event of interest that conform with the surveillance period restriction, **c** is the number of unique sessions including search terms associated with the event of interest but not the

drug of interest that conform with the surveillance period restriction, and \mathbf{d} is equal to the total number of unique sessions minus the sum of the cells **a**, **b**, and **c**. Having computed the 2x2 contingency table for a given drug-event pair of interest, we use the Odds Ratio measure, with a zero-cell correction (adding 0.5 to each count in a 2x2 table that contains zero in any of its cells), to quantify the strength of association between the drug and event. As our final association score we use the 5th percentile of the odds ratio distribution, which provides an adjustment for small sample sizes and protects against false associations due to chance. This type of adjustment has been shown to provide greater accuracy than point estimates in a related signal detection study [29] and was also found to provide greater accuracy in this study. We note that by performing a two-dimensional (2x2) analysis we omit to account for potential biasing factors such as temporal trends, media influence, or the search habits of healthcare professionals. We plan to examine these and other types of biases in a follow-up study.

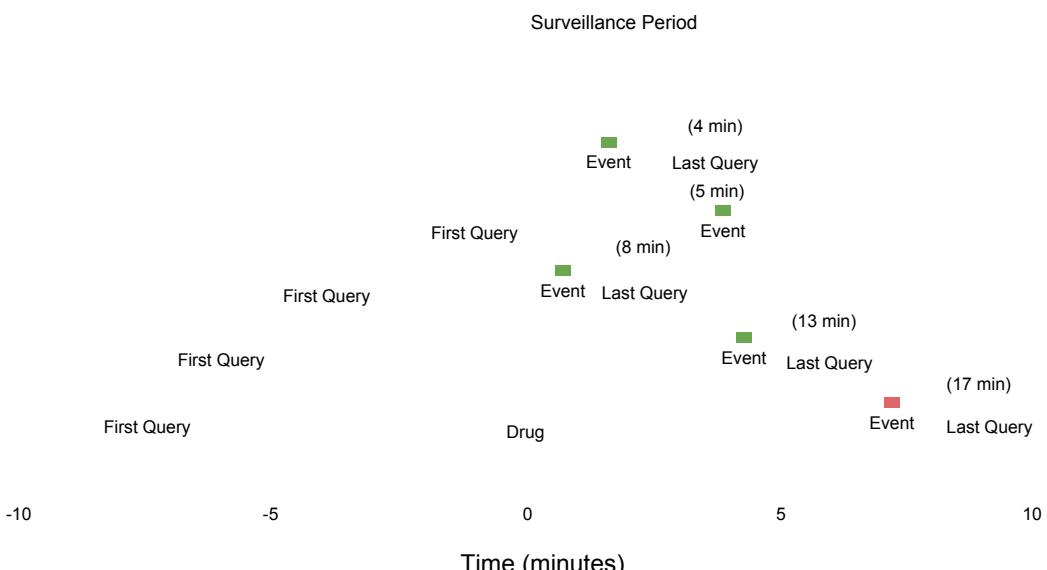


Figure 1. Illustration of a surveillance period relative to UpToDate search sessions that contain occurrences of ‘drug then event’ queries. A single search session is shown as a horizontal line that persists for the time length of the session. Sessions that contain the same drug and event as queries are aligned by assigning the time of drug mention as time 0. Within each session, the instance of drug and event queries with the shortest time difference is compared to the surveillance period. Sessions that have their corresponding shortest drug → event mention time within the surveillance period are counted towards the association analysis (shown as sessions with the event in green). Sessions that have a drug → event time difference that is greater than the surveillance period are not counted (shown as the session with the event in red).

2.5. Evaluation

We evaluate our approach using two reference standards, each consisting of a set of positive test cases of drug-event pairs recognized as truly associated (true ADEs), and a set of negative controls of drug-event pairs that are highly unlikely to be associated. The first reference standard we use was created by the EU-ADR project [30] for the retrospective evaluation of ADE signal detection

approaches based on EHRs and claims data. It includes 44 positive test cases and 50 negative controls spanning ten serious adverse events (*e.g.* acute myocardial infarction, rhabdomyolysis) and 68 unique drugs. The drug-event associations comprising the EU-ADR positive test cases have been well known for a relatively long period of time. This reference set—when used to evaluate methods applied to data representing recent events such as our queries from 2011-2012, which may be influenced by existing information about well-known ADEs—will only provide an indication of the retrospective performance of the method.

Therefore, we used a second reference standard [31] for evaluation that was built specifically for the prospective evaluation of ADE signal detection methods. This time-indexed reference standard was manually curated from all product label updates (*e.g.* warnings) communicated by the FDA in 2013 through monthly summaries posted at FDA's MedWatch website [32] and classified as Boxed Warning, Warning (or Precautions), and Adverse Reactions. In addition, we considered only drugs that are orally administered. Each candidate association extracted from the MedWatch website was manually verified (by reviewing the drug's labeling history obtained from drug@FDA [33]) to ensure that the association is indeed new and is not qualified by a co-existing contraindicated situation or risk factor (this type of association requires non-standard analysis approaches). The drugs underlying each association were normalized to RxNorm active ingredients. Closely related events underlying each association that are synonymous or describe the same clinical syndrome were grouped and given a unique representative name. The final reference set includes 62 positive test cases and 75 negative controls, and covers 44 drugs and 38 events ranging from common and mild to rare and serious. The negative controls were created by randomly pairing the same drugs and events that appear in the set of positive test case, and verifying that these random pairs are not reported in the underlying drug's label as an ADE. The time stamp attached to each positive test case reflects the date on which a label corresponding to a drug was revised to include the adverse event associated with the drug. It is assumed that this date reflects the time by which an association became known to the general public. Consequently, using data that pre-dates the label update year of 2013 (as with the UpToDate logs used in our method) provides an indication of the prospective performance of an association detection method.

Both the EU-ADR and the time-index reference standards define each of the included AEs as sets of UMLS concepts that relate to and are consistent with a description of the overall clinical condition associated with the particular AE (the definitions were created manually by researchers with medical training). For example, the UMLS concepts for bullous eruption (C0235818), erythema multiforme (C0014742), Stevens-Johnson syndrome (C0038325), toxic epidermal necrolysis (C0014518), and bullous dermatosis (C0085932), compose the group of UMLS concepts that define the 'bullous eruptions' AE concept in the EU-ADR reference standard. In our analysis, a query was considered to mention a given AE if any of its search terms (or term sequences) was mapped to one of the UMLS concepts that define the underlying adverse event.

Drugs in the EU-ADR reference standard are specified using bottom-level ATC codes (ATC5), whereas drugs in the time-indexed reference standard are specified using RxNorm codes at the base active ingredient level (IN). The ATC codes were mapped to RxNorm INs. A query was considered to mention a given drug if any of its search terms (or term sequences) was mapped

to one of the UMLS CUIs that represent the drug's active ingredient, thus our analysis and evaluation was performed at the drug ingredient level. Queries that included the drugs and adverse events were tagged with the corresponding drug and adverse event concept identifiers. The tags were then used to compute 2x2 contingency tables and statistical associations as described before. Using the calculated association score for each drug-event pair (from each of the reference standards), the performance of our approach to identify ADEs was quantified using the area under the receiver operating characteristic curve (AUC).

3. Results

3.1. Descriptive statistics

We processed 320,975,093 unique UpToDate search queries associated with 134,217,800 unique sessions and 3,986,773 unique IP addresses. The annotation process resulted in 144,888 unique CUIs which includes 40,416 unique drug concepts and 47,220 unique clinical condition concepts. Sessions constitute our basic unit of count, and so we present several session-related statistics (Table 2). Session length is determined as the difference between the time of the first and last search in the session, and thus sessions that consist of a single search have a recorded time of 0 seconds. An average session lasted 20 minutes and consisted of 5 queries.

Table 2. UpToDate search session statistics.

	<i>Queries Per session</i>	<i>Session Duration (secs)</i>
Min	1	0
Max	304283	17373968
Median	3	0
Mean	5	1202

For the positive test cases of the EU-ADR reference standard, the median time between the query of a drug and event (regardless of ordering) was 621 seconds. For the time-indexed reference standard the median was 1,979 seconds. A possible explanation for this difference is that the EU-ADR positive test cases are better documented in UpToDate due to the fact that they are well known, and thus lead to shorter search sessions. Table 3 displays the top three associations that were searched for regardless of ordering within a surveillance period of five minutes. The table also lists the most frequently occurring search terms associated with each drug and event.

Table 3. Top drug-AE associations from the EU-ADR and time-indexed reference sets, and the most frequently occurring search terms that are evidence for the association.

	<i>Top association (count)</i>	<i>Top search terms (count)</i>
EU-ADR	acetylsalicylic acid; upper gastrointestinal bleed (1683)	Drug: aspirin (1639), Aggrenox (32), Excedrin (7), alka-seltzer (3) Event: gastrointestinal hemorrhage (1652) melena (15), hematemesis (13), upper gastrointestinal hemorrhage (3)
	furosemide; acute renal failure (1026)	Drug: Lasix (639), furosemide (387) Event: acute kidney failure (879), anuria (92), acute kidney tubular necrosis (55)
	simvastatin; rhabdomyolysis (523)	Drug: simvastatin (410), Zocor (92), vytarin (21) Event: rhabdomyolysis (519), myoglobinuria (4)
Time-indexed	ursodeoxycholate; liver damage (2291)	Durg: ursodiol (1844), actigall (248), urso (187), ursofalk (11), urdox (1) Event: liver cirrhosis (933), liver diseases (451), hyperbilirubinemia (335), hepatitis (332), icterus (167)
	levetiracetam; movement disorders (597)	Drug: keppra (496), levetiracetam (101) Event: myoclonus (167), tremor (99), spasm (56), dyskinetic syndrome (44), ataxia (40)
	ketoconazole; liver damage (190)	Drug: ketoconazole (176), nizoral (14) Event: liver diseases (85), hepatitis (56), hepatotoxicity (29), liver cirrhosis (7), icterus (6)

3.2. Performance evaluation

We explored different drug-event orderings and different surveillance periods as defined in the Methods section. We varied the surveillance period from 0 seconds to 100 minutes at intervals of 100 seconds. The surveillance period accounts for the time between the first and second search corresponding to the annotated drug or event in the search logs. A temporal measurement was used to take into consideration the time required to read and process the first search result of a drug/event pair before searching for the corresponding drug or event in the pair. For each combination of a surveillance period length and a drug-event ordering we computed the AUC using the corresponding association scores, and the proportion of test cases that the data supported evaluation of. We define test cases with data support as those test cases that had at least one search session containing the drug and AE ($a > 0$ in the 2×2 contingency table) during the surveillance period. This was used to provide additional insight into the analysis, and is also used to identify the optimal surveillance period as a tradeoff between discrimination accuracy (AUC) and the number of test cases that are supported by the data. The more test cases with data support, the more reliable the performance evaluation. A very small surveillance period that results in only a small portion of test cases that are supported by the data would not be useful in a real setting. Such results cannot be used as an indicator of a method's success in identifying true associations.

Table 4 and Figures 2 and 3 provide a summary of our performance evaluation. Table 4 shows the maximum AUCs obtained by our approach and their corresponding surveillance periods.

Expectedly, retrospective evaluation based on the EU-ADR reference standard yields much better performance (maximum AUC=0.85) than the prospective evaluation based on the time-indexed reference standard (maximum AUC=0.68). The results also suggest that drug-event ordering does not dramatically affect performance, though the best performance is achieved by using the ‘drug *then* event’ ordering for both reference standards. The optimal surveillance period is small (approximately 4-10 minutes), which is consistent with our modeling assumptions.

Table 4. Maximum AUCs and corresponding surveillance period lengths for sessions where a drug was searched for before an event, event before drug, and either ordering.

Search direction	<i>EU-ADR</i>		<i>Time-indexed</i>	
	Max. AUC	Search window	Max. AUC	Search window
Drug → event	0.85	665	0.68	303
Event → drug	0.82	287	0.67	276
Either direction	0.84	245	0.67	296

Figures 2 and 3 display graphs of discrimination performance as a function of surveillance periods (0 seconds to 100 minutes at intervals of 100 seconds) for the EU-ADR and the time-indexed reference standards, respectively. For illustrative purposes we used the ‘drug *then* event’ ordering, and note that similar performance patterns were obtained for the two other drug-event orderings. Each figure displays the AUC (top sub-figure) and the proportion (bottom sub-figure) of both positive and negative test cases supported by the data (test cases that had at least one query within the surveillance period).

Results for both reference standards suggest that performance fluctuates for small surveillance periods and then generally stabilizes, albeit with a decreasing trend as the surveillance period increases. As expected, the figures show that (1) the positive test cases have consistently more data support than the negative test cases; (2) the test cases of the EU-ADR reference standard have more data support than those of the time-indexed reference standard with smaller surveillance periods; and (3) the difference between the proportion of positive test cases and negative test cases with data support is much larger for the EU-ADR reference standard than for time-indexed reference standard. This is consistent with the larger AUCs obtained for the EU-ADR reference standard.

An encouraging result provided by the graphs is that a relatively large proportion (>80%) of positive test cases from the time-indexed reference standard are supported by data with short surveillance periods, around 5 minutes in duration. We hypothesize that this is the time an engaged user would take to read drug information on UpToDate, reason about a connection to an adverse event and then search for the specific event. Almost all the positive test cases are eventually supported by data (queries) with large surveillance periods, though most likely due to chance. The surveillance period at which these proportions start to converge is an alternative point to assess performance since it balances the tradeoff between discrimination accuracy and increasing the test cases that can be identified in the data. For the time-indexed reference standard this optimal point is to the right of the surveillance period yielding the maximum AUC, whereas

for the EU-ADR reference standard it appears to its left. At these points only a modest amount of accuracy is sacrificed, with the benefit that maximum test cases can be identified through the data.

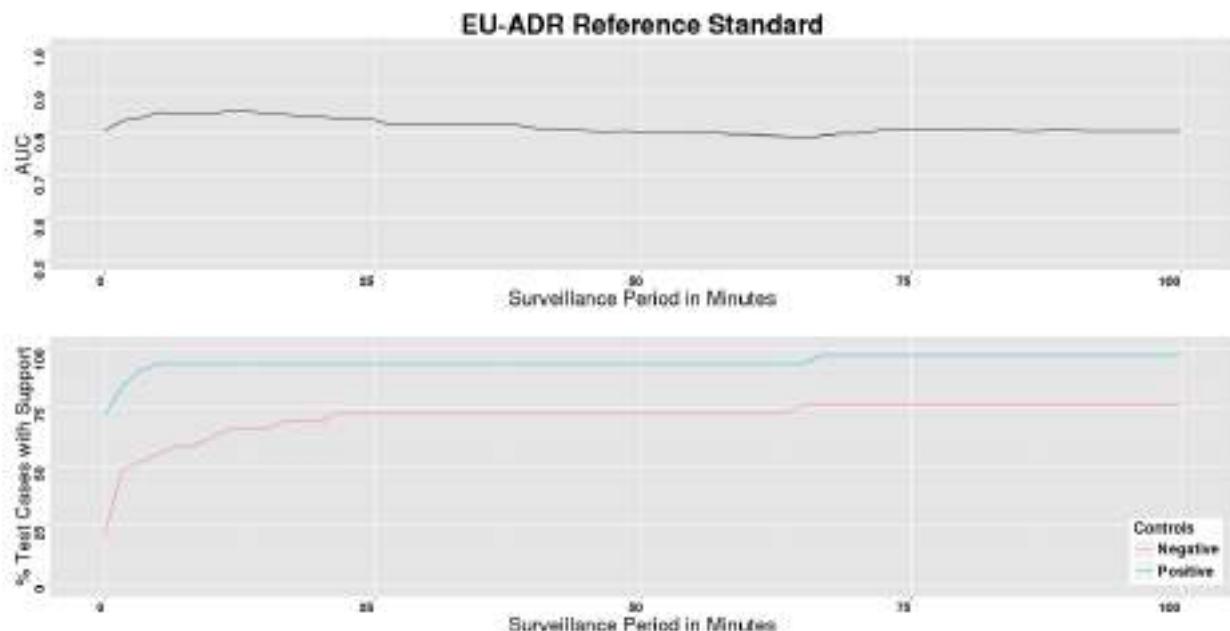


Figure 2. Performance as a function of surveillance period for the EU-ADR reference standard. The top graph shows the change in AUC as surveillance period increases. The bottom graph shows the percent of test cases with data support as surveillance period increases.

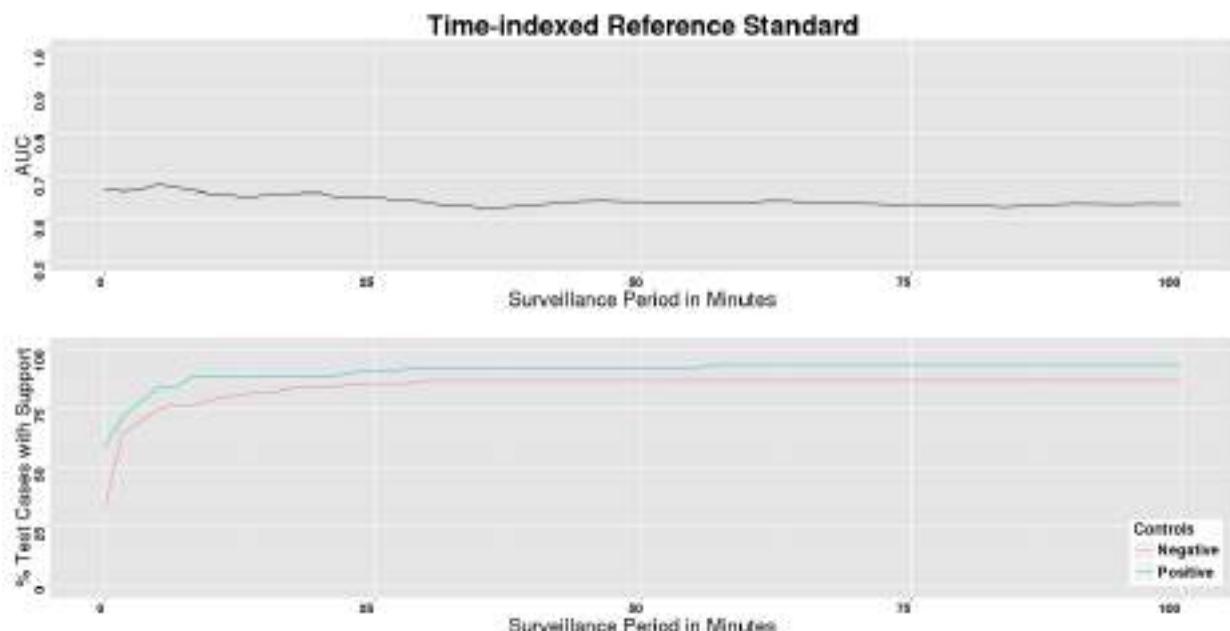


Figure 3. Performance as a function of surveillance period for the time-indexed reference standard. The top graph shows the change in AUC as surveillance period increases. The bottom graph shows the percent of test cases with data support as surveillance period increases.

4. Discussion

This work describes a preliminary investigation into mining drug-AE associations from healthcare professional search logs, a data source that to our knowledge has not been used for drug safety surveillance. In a recent commentary [34], Sarntivijai and Abernethy discuss the merits of using Internet search logs in general as a source for ADE signal detection and emphasize the need for more research in processing and assessing the validity of any discovered drug-AE associations. This work represents such an effort, and our findings suggest that physician search logs could be used as a data source for detecting ADEs.

The performance of our method on the drug-AE pairs (AUC of 0.85) from the EU-ADR project is not surprising, given that these ADEs are likely to be already described in UpToDate and also known to the medical professionals using UpToDate as a reference source; both factors that can lead users to query for the drugs and AEs. The lower AUC of 0.68 for the time-indexed reference standard of more recently labeled ADEs achieved by our method indicates that it is more challenging to detect relatively new or emerging ADEs. Despite the relatively high AUCs achieved by our method, we also observed lower association scores for drug-AE pairs from the time-indexed reference set. This finding warrants further investigation to determine which (if any) of the detected associations may be due to chance or if an odds ratio based association score is a discriminatory enough metric for search log analysis. Given the challenges in detecting relatively new or emerging ADEs in UpToDate logs, this datasource's utility might hinge on development of methods to combine data sources.

The possibility of chance ADE discoveries exists due to the fact that healthcare professionals may search for a given drug and disease within a single UpToDate search session for a variety of reasons, and the existence of multiple terms within a session is not a guarantee that the searches are associated. Caution must therefore be exercised when analyzing such records using data mining methods to avoid detection of spurious associations. Our experiments show that relatively short surveillance periods (and therefore search session length) are optimal for detecting ADEs, and this is consistent with our modeling assumptions regarding search behavior. With long enough surveillance periods, any drug and event may be found to be associated, and so the use of shorter periods can provide some protection against such effects.

The next frontier in drug safety surveillance involves determining time to signal, quantifying the plausibility of potential drug-AE pairs, estimating patients affected by ADEs, prioritizing ADEs by severity, and in combining data sources to improve drug safety surveillance. Our preliminary findings indicate that healthcare professional search logs can be a potential data source for advancing drug-safety research. We have recently demonstrated the performance of combining drug safety signals [23] and future work will build on these efforts by incorporating healthcare professional search logs as a complementary data source. We will also investigate the use of other methods to detect ADEs from physician search logs, develop additional benchmarks, develop more accurate search behavior models (perhaps by profiling or surveying users), and identify the type of events that are appropriate for surveillance via healthcare professional search logs.

Acknowledgements

We acknowledge funding support from NIH grant R01 GM101430. NHS and RH also received funding support from R01 LM011369 and U54 HG004028.

References

1. Classen, D.C., et al., *JAMA*, **277** (4), 301 (1997).
2. Lazarou, J., B.H. Pomeranz, and P.N. Corey, *JAMA*, **279** (15), 1200 (1998).
3. Classen, D.C., et al., *Health Aff (Millwood)*, **30** (4), 581 (2011).
4. Hug, B.L., et al., *Jt Comm J Qual Patient Saf*, **38** (3), 120 (2012).
5. Honig, P.K., *Clin Pharmacol Ther*, **93** (6), 474 (2013).
6. Harpaz, R., et al., *Clin Pharmacol Ther*, **91** (6), 1010 (2012).
7. Platt, R., et al., *N Engl J Med*, **361** (7), 645 (2009).
8. Stang, P.E., et al., *Ann Intern Med*, **153** (9), 600 (2010).
9. Levinson, D.R., Department of Health and Human Services, Available from <http://oig.hhs.gov/oei/reports/oei-06-09-00091.pdf> (2012).
10. Avorn, J. and S. Schneeweiss, *N Engl J Med*, **361** (7), 647 (2009).
11. Hauben, M. and A. Bate, *Drug Discov Today*, **14** (7-8), 343 (2009).
12. Schneeweiss, S., et al., *Epidemiology*, **20** (4), 512 (2009).
13. Coloma, P.M., et al., *Pharmacoepidemiol Drug Saf*, **21** (6), 611 (2012).
14. LePendu, P., et al., *Clin Pharmacol Ther*, **93** (6), 547 (2013).
15. Iyer, S.V., et al., *J Am Med Inform Assoc*, **21** (2), 353 (2014).
16. Yom-Tov, E. and E. Gabrilovich, *J Med Internet Res*, **15** (6), e124 (2013).
17. White, R.W., et al., *J Am Med Inform Assoc*, **20** (3), 404 (2013).
18. White, R.W., et al., *Clin Pharmacol Ther*, **96** (2), 239 (2014).
19. Freifeld, C.C., et al., *Drug Saf*, **37** (5), 343 (2014).
20. Shetty, K.D. and S.R. Dalal, *JAMIA*, **18** (5), 668 (2011).
21. Avillach, P., et al., *J Am Med Inform Assoc*, **20** (3), 446 (2013).
22. Pontes, H., M. Clement, and V. Rollason, *Drug Saf*, **37** (7), 471 (2014).
23. Harpaz, R., et al. *KDD* (2013).
24. Xu, R. and Q. Wang, *BMC Bioinformatics*, **15** (2014).
25. Bonis, P.A., et al., *Int J Med Inform*, **77** (11), 745 (2008).
26. Isaac, T., J. Zheng, and A. Jha, *J Hosp Med*, **7** (2), 85 (2012).
27. Lependu, P., et al., *J Biomed Semantics*, **3 Suppl 1** S5 (2012).
28. Lependu, P., et al., *AMIA Jt Summits Transl Sci Proc*, **2012** 63 (2012).
29. Harpaz, R., et al., *Clin Pharmacol Ther*, **93** (6), 539 (2013).
30. Coloma, P.M., et al., *Drug Safety*, **36** (1), 13 (2013).
31. Harpaz, R., et al., *Nature Scientific Data*, **In press** (2014).
32. U. S. Food and Drug Administration.
<http://www.fda.gov/Safety/MedWatch/SafetyInformation/Safety-RelatedDrugLabelingChanges/default.htm> (2014) [cited 2014 September 28].
33. U. S. Food and Drug Administration.
<http://www.accessdata.fda.gov/scripts/cder/drugsatfda/> (2014) [cited 2014 September 28].
34. Sarntivijai, S. and D.R. Abernethy, *Clin Pharmacol Ther*, **96** (2), 149 (2014).

REFINING LITERATURE CURATED PROTEIN INTERACTIONS USING EXPERT OPINIONS

OZNUR TASTAN^{*,1}, YANJUN QI², JAIME G. CARBONELL³, JUDITH KLEIN-SEETHARAMAN⁴

¹ Department of Computer Engineering, Bilkent University, Cankaya, Ankara, Turkey

² Department of Computer Science, University of Virginia, Charlottesville, VA, USA

³ Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA, USA

⁴ Division of Metabolic and Vascular Health, University of Warwick, Warwick, Coventry, UK

* E-mail: oznur.tastan@cs.bilkent.edu.tr

The availability of high-quality physical interaction datasets is a prerequisite for system-level analysis of interactomes and supervised models to predict protein-protein interactions (PPIs). One source is literature-curated PPI databases in which pairwise associations of proteins published in the scientific literature are deposited. However, PPIs may not be clearly labelled as physical interactions affecting the quality of the entire dataset. In order to obtain a high-quality gold standard dataset for PPIs between human immunodeficiency virus (HIV-1) and its human host, we adopted a crowd-sourcing approach. We collected expert opinions and utilized an expectation-maximization based approach to estimate expert labeling quality. These estimates are used to infer the probability of a reported PPI actually being a direct physical interaction given the set of expert opinions. The effectiveness of our approach is demonstrated through synthetic data experiments and a high quality physical interaction network between HIV and human proteins is obtained. Since many literature-curated databases suffer from similar challenges, the framework described herein could be utilized in refining other databases. The curated data is available at <http://www.cs.bilkent.edu.tr/~oznur.tastan/supp/psb2015/>

Keywords: Protein-protein Interactions, Literature Curated Databases, Crowd-Sourcing

1. Introduction

Literature curated databases^{1–8} extract PPIs from published articles, organize them and make them available online. In addition to supplying prior knowledge for future experiments on a specific protein function, these sets of interactions enable system level analyses of interactomes, serve as benchmark data to quantify error rates in high-throughput experimental assays⁹ or they are used as training/testing data to build predictive models.¹⁰ Such analyses depend critically on the inherent quality of the data.

Obtaining a high quality set of direct physical PPI interactions often presents a challenge. In some cases, databases include a mixture of functional indirect associations and direct physical interactions, without specifying the distinction. For example, the work presented here is motivated by our previous study on predicting the HIV-1-human interactome,¹¹ where we faced the challenge of extracting the subset of direct physical interactions from the NIAID HIV-1, Human Interactions database,^{2,3} which does not readily provide a reliable direct physical interaction set. Since this is a general problem, the International Molecular Exchange (IMEx) consortium^{12–14} has announced that literature-curated databases should provide more detailed information about the experiments using structured ontologies describing the type of the interaction or the experimental techniques employed. In theory, these experimental details might allow the user or the database curator to review each interaction and use their own judgment to decide how reliable a pair is. In practice, this route is time consuming and is not

guaranteed to arrive at good quality datasets. Users either assume all small-scale experiments are of equally high quality¹⁵ or disregard some portion of the interactions based on additional criteria such as the type of experimental technique, or the number of publications that validates the interaction. Few databases provide reliability scores, e.g. the Molecular Interaction Database (MINT)^{16,17} that combines information such as the scale of the experiment, the type of the experiment, the number of publications supporting an interaction, and the presence or absence of ortholog interactions. However, the scoring function blends in several parameters to quantify reliability.^{16,17}

Assessing whether there is enough evidence to conclude that a reported association is a direct physical interaction requires a complex judgment. One has to concurrently account for the methods employed, the proteins under study, and the results of each specific study. Some experimental techniques are more conclusive than others and techniques do not work uniformly well across all proteins. In addition to the variability in the power and limitations of each technique, the conditions under which a study is conducted are important: *in vitro* or *in vivo* environment, the strains used, the nature and positions of mutations or labels introduced. All such parameters should be taken into account when interpreting the results. Such a complex judgment can be best provided by domain experts.

Although domain experts can provide high quality labels for PPIs, different experts might have different opinions, especially when there is not enough evidence accumulated in the literature to give a perfectly conclusive answer. Additionally, disagreements among experts arise because of their biases, expertise and/or stringency levels; e.g., some experts are more difficult to convince with partial evidence or results of certain experimental techniques than others. Despite these limitations it has been demonstrated in several other domains, that harnessing the power of human judgments collectively – although imperfect individually – can be invaluable for solving difficult tasks.¹⁸

A long line of work exists in the biostatistics and epidemiology literature where latent variable models have been used to estimate the observer error rates based on results from multiple diagnostic tests without a ground-truth set.¹⁹ Recently, “learning from crowds” has become a very active research topic in machine learning and has already had several successful applications.^{18,20,21} For example, through the Dialogue on Reverse Engineering Assessment and Methods (DREAM) project, Marbach et al.²² used community-based methods to construct high-confidence gene regulatory networks for *E. coli* and *S. aureus*. Their integration strategy through “learning from crowds” was shown as a powerful and robust tool for the inference of transcriptional gene regulatory networks. Similarly, Lu et al.²³ combines multiple annotations in the Gene Normalization (GN) challenge in BioCreative III.

In this paper, we adopt a crowd-sourcing strategy to revise an otherwise noisy and heterogeneous literature curated HIV-1, human PPI dataset by collecting expert opinions on PPI. We apply a probabilistic approach to estimate experts labeling accuracies in the absence of a benchmark dataset similar to the approach proposed by Raykar et al.²⁰ Using these estimations, the probability of the literature curated pairs to reflect direct interactions is assessed, which results in a high quality set of labels for HIV-1, host PPIs. We verify the utility of this approach through synthetic data experiments as well as performance tests conducted with a

new model that is trained with these high quality labels.

2. Literature Curated HIV-1, Human PPIs and Collected Experts' Labels

2.1. Existing Literature Curated HIV-1, Host PPIs:

We retrieved literature curated HIV-1 and human PPIs from the NIAID HIV-1, human protein interaction database (henceforth referred to as NIAID database).^{2,3} The set includes 2589 HIV-1, human PPIs between 1448 human proteins and 18 HIV-1 proteins. Whether the reported association is a direct physical interaction or not is not provided by the database. Instead, the database describes each interaction by one or more descriptive key phrases such as “interacts with” or “binds”, which are extracted from publications reporting these interactions.

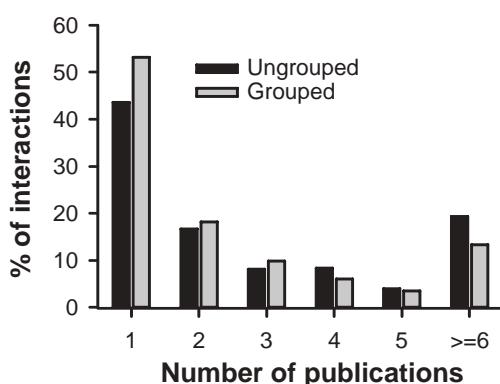


Fig. 1. Distribution of the number of publications supporting each HIV-1, human protein interaction in the NIAID database.^{2,3} The graph depicts two analyses, where publications are counted individually (black) or publications are grouped together if they share a common author (gray).

We retrieve the list of publications supporting each interaction in the NIAID database and conduct an analysis on the number of supporting publications for each PPI. Interestingly, 44% of all the PPIs in the database are reported only in a single publication (Fig. 1). When the publications that share at least one common author are grouped together, the statistics become even more striking; the proportion of PPIs supported by a single group is as high as 53% (Fig. 1). The number of PPIs that are supported by more than five publications constitutes only 19 and 13% of all PPIs when publications are ungrouped and grouped, respectively. The lack of follow-up studies by other labs for a given PPI hints at the possibility that for most interactions, there may not be enough experimental validation for their inclusion in a gold standard dataset. This justifies the need to develop a good way to assign confidences to the associated pairs.

2.2. Collecting Opinions from Experts

To obtain judgments on interactions published, we contacted a number of experts working in the HIV field. 16 experts provided their opinions on the interactions. One of the experts is a PhD student working with HIV-1 experimentally; all others were professors at different universities who have worked several years on one or more HIV-1 proteins experimentally. The experts were asked to annotate only the interactions of the HIV-1 proteins for which they consider themselves experts. For each HIV-1 protein, an Excel file was prepared, in

which interactions with the HIV-1 protein were listed. The file included the human protein interaction partners of the HIV-1 protein, the keywords retrieved from the NIAID database, and hyperlinks to the original publications so the experts could check the articles if necessary. For HIV-1 proteins, where the number of PPIs are ≤ 50 , all the interactions reported in the NIAID database were sent to experts. In cases where > 50 interaction partners were listed for HIV-1 proteins, experts were only provided with the subset of PPIs described with the keywords “interacts with” and “binds”. This was to avoid experts feeling overwhelmed with a long list. We adopted this route to increase the chance of receiving a response. In some cases, experts did not use the files; instead, they provided us with a set of interactions which they thought are real and direct PPIs.

3. Model to Estimate Expert Labeling Qualities and PPI Labels

Presented with the same protein pair and the accompanying published evidence, experts sometimes disagreed on whether to label the PPI pair as a direct interaction or not. When there is not enough evidence accumulated in the literature, disagreements among experts might arise because of their biases, expertise, and/or stringency levels. For these reasons, expert opinions will be noisy and subjective. By asking several experts about the same interaction pair, the reliability of the interaction being a direct interaction can be better assessed. Although having as many expert opinions as possible is beneficial, we were not able to obtain multiple expert opinions on all protein pairs due to time or expertise constraints. Thus, there is variance in the number of expert opinions for some pairs.

Taking these considerations into account, the computational problem becomes the following: given noisy opinions and with possibly varying number of judgments for each case, how can we accurately decide if a protein pair is more likely a “direct physical interaction” and what is the uncertainty of the label. A commonly used strategy in this situation is to decide on labels based on a majority vote of the expert labels. However, majority voting assumes that all experts provide opinions of equal accuracy. When there is subjectivity and noise in the opinions, majority voting cannot be expected to perform as desired. In our model, we account for that with a probabilistic latent variable model. In this model experts’ labeling accuracies and the probability of the association being a direct interaction are estimated jointly.

3.1. A Probabilistic Model for Expert Opinions

Let’s consider N literature reported PPI and let $y_i \in \mathcal{Z}$ indicate the true and hidden label for the i^{th} PPI, where $\mathcal{Z} = \{0, 1\}$ (“direct physical interaction” or “not”, respectively). The labels y_i are unknown. Instead we have multiple expert labels $\mathbf{y}_i = \{y_i^1, y_i^2, \dots, y_i^M\}$ provided by M different experts. We introduce a parameter that represents the true unknown labeling accuracy of the expert j for the label type z . We propose to model each expert’s labeling accuracy using a biased-coin model,²⁰ which is,

$$\mathbf{P}(y_j^j = z | y = z) = \theta_z^j \quad (1)$$

where $z \in \mathcal{Z} = \{0, 1\}$. This model assumes that when the hidden true label is one, the expert j flips a coin with bias θ_1^j , meaning that the expert j has a probability of θ_1^j to assign correct label 1 to this instance. When the hidden true label is 0, the expert j flips a coin with bias θ_0^j , meaning that the expert j has a probability of θ_0^j to assign correct label 0 to the

instance. The above formulation can model the situation of different experts having different error rates in their annotations of “direct interaction” and “not direct interaction” label types. This model assumes that the parameter vector $\theta^j = (\theta_0^j, \theta_1^j)$ does not depend on the instance x . Also, another parameter representing the *prior* probability of different labels is $p_i^z = \mathbf{P}(y_i = z)$. A procedure to estimate $\Theta = \{\theta^1, \theta^2, \dots, \theta^M, p\}$ for all experts is presented in detail throughout the next subsection. Using this model, a *soft probabilistic estimate* of the hidden true label can be calculated as follows:

$$\begin{aligned} g_i(z | \Theta) &\equiv \mathbf{P}(y_i = z | y_i^1, y_i^2, \dots, y_i^M, \Theta) \propto \mathbf{P}(y_i^1, y_i^2, \dots, y_i^M | y_i = z, \Theta) \times \mathbf{P}(y_i = z | \Theta) \quad (2) \\ &= \prod_{j=1}^M \mathbf{P}(y_i^j | y_i = z, \Theta) \times \mathbf{P}(y_i = z | \Theta) = p_i^z \times [\theta_z^j]^{h(y_i^j=z)} \times [(1 - \theta_z^j)]^{1-h(y_i^j=z)} \end{aligned}$$

where h is the indicator function. Note that we assume that decisions by the experts are conditionally independent given the true label. Thus, we use the following equation to predict the most probable label for an interaction:

$$\hat{y}_i = \arg \max_{z \in \mathcal{Z}} g_i(z | \Theta) \quad (3)$$

The uncertainty of this label based on the above equation is defined as:

$$\hat{u}_i(\hat{y}_i) = 1 - \mathbf{P}(y_i = \hat{y}_i | \mathbf{y}_i, \Theta) \quad (4)$$

3.2. Maximum Likelihood Estimator for Experts’ Labeling Qualities

We estimate the parameters Θ through maximum likelihood estimation (MLE):

$$\begin{aligned} \hat{\Theta}^{\text{mle}} &= \arg \max_{\theta} \mathcal{L}(\mathcal{D} | \Theta) \quad (5) \\ \text{where } \mathcal{D} &= \{(y_i^1, y_i^2, \dots, y_i^M)\}_{i=1, \dots, N} \end{aligned}$$

Below, we show how to estimate $\hat{\Theta}^{\text{mle}}$ for the case where every interaction receives opinions from every expert. Later, we refine the model by relaxing this assumption.

Case 1: Every expert provides labels for every example (global annotation case)
The log-likelihood of the observed expert opinions:

$$\mathcal{L}(\mathcal{D} | \Theta) = \sum_{i=1}^N \log \mathbf{P}(\mathbf{y}_i | \Theta) = \sum_{i=1}^N \log \sum_{z=0}^1 \mathbf{P}(\mathbf{y}_i | y_i = z, \Theta) \times \mathbf{P}(y_i = z | \Theta) \quad (6)$$

The last equation marginalizes over the hidden true label, y_i . We assume decisions by the experts are conditionally independent given the true label:

$$\mathcal{L}(\mathcal{D} | \Theta) = \sum_{i=1}^N \log \sum_{z=0}^1 \left(\prod_{j=1}^M \mathbf{P}(y_i^j | y_i = z, \Theta) \mathbf{P}(y_i = z | \Theta) \right) \quad (7)$$

In Eq. 7, $\mathbf{P}(y_i^j | y_i = z, \Theta)$ is the probability of observing expert label $y_{i,j}$ for interaction i , given the true label of that interaction is $y_i = z$ (similar to calculations in Eq. 2):

$$\mathbf{P}(y_i^j | y_i = z, \Theta) = [\theta_z^j]^{h(y_i^j=z)} \times [(1 - \theta_z^j)]^{1-h(y_i^j=z)} \quad (8)$$

where h is the indicator function. $\mathbf{P}(y_i = z) = p_z$ is the prior probability of a potential pair belonging to class z ; it is assumed that this prior probability is the same for all $i = 1 \dots N$. In

order to estimate Θ , first Eq. 8 is inserted into Eq. 7 and next, the expectation-maximization (EM) algorithm²⁴ is applied to maximize a lower bound of this incomplete data likelihood, by considering true label y as the hidden variable:

$$\begin{aligned}\mathcal{L}(\mathcal{D} | \Theta) &= \sum_{i=1}^N \log \sum_{z=0}^1 \mathbf{P}(\mathbf{y}_i, y_i = z | \Theta) \geq \sum_{i=1}^N \sum_{z=0}^1 \log \mathbf{P}(\mathbf{y}_i, y_i = z | \theta) \\ &= \sum_{i=1}^N \sum_{z=0}^1 g_i(z) \log \frac{\mathbf{P}(\mathbf{y}_i, y_i = z | \Theta)}{g_i(z)}\end{aligned}\quad (9)$$

It is iteratively maximized with respect to the probability distribution $g(z)$ and Θ in the expectation and maximization steps, respectively. The derived update equations for step $t+1$ are as follows:

E-step:

$$g_i^{(t+1)}(z') = \mathbf{P}(y_i = z' | \mathbf{y}_i, \Theta^{(t)}) = \frac{\prod_{j=1}^M \mathbf{P}(y_i^j | y_i = z', \Theta^{(t)}) \mathbf{P}(y_i = z' | \Theta^{(t)})}{\sum_{z=0}^1 \mathbf{P}(y_i = z | \Theta^{(t)}) \prod_{j=1}^M \mathbf{P}(y_i^j | y_i = z, \Theta^{(t)})} \quad (10)$$

M-step:

$$[\theta_{z'}^j]^{(t+1)} = \frac{\sum_{i=1}^N g_i^{(t)}(z') \times h(y_i^{j'} = z')}{\sum_{i=1}^N g_i^{(t)}(z')} \quad (11)$$

The prior probability p^z is an estimate of the class distribution, is derived from majority vote labels based on the MLE solution. The above procedure is repeated until convergence is attained.

Case 2: Experts only provide labels for a subset of examples (subset annotation case) Not every expert might be available to provide labels for each potential interaction due to expertise and time limitations. Thus, the assumption that every expert can report labels for every instance may not hold. In this section, we provide a solution that works when this assumption is relaxed. Let the group of labelers for the interaction i be a subset, $A_i \subset \{1, \dots, m\}$. It is required that each interaction will have received at least one opinion from at least one expert. In this case, the EM update equations are modified as follows:

E-step:

$$g_i^{(t+1)}(z') = \mathbf{P}(y_i = z' | \mathbf{y}_i, \Theta^{(t)}) = \frac{\prod_{\substack{j=1 \\ j:j \in A_i}}^M \mathbf{P}(y_i^j | y_i = z', \Theta^{(t)}) \mathbf{P}(y_i = z' | \Theta^{(t)})}{\sum_{z=0}^1 \mathbf{P}(y_i = z | \Theta^{(t)}) \prod_{\substack{j=1 \\ j:j \in A_i}}^M \mathbf{P}(y_i^j | y_i = z, \Theta^{(t)})} \quad (12)$$

M-step:

$$\theta_{z',j'}^{(t+1)} = \frac{\sum_{\substack{i=1 \\ i:j' \in A_i}}^n g_i^{(t)}(z') h(y_i^{j'} = z')}{\sum_{\substack{i=1 \\ i:j' \in A_i}}^n g_i^{(t)}(z')} \quad (13)$$

Once the expert labeling accuracies are obtained from the above procedure, they can be plugged into Eqs. 3 and 4 to identify the most probable label and the uncertainty associated with it.

4. Synthetic Data Experiments

Since there are no real data with opinions and expert labels, synthetic data experiments were carried out to test the effectiveness of the method. The synthetic data was generated as follows: Given a prior distribution of label types, a set of true labels was first generated randomly. Meanwhile, each expert's true labeling quality on each label type, $\theta_{j,z}$, was assigned uniformly at random. The underlying rationale is that experts are likely to produce better-than-random answers. In the next step, to simulate an expert's opinion on an instance, true labels for data points were randomly converted to incorrect label types with a probability that follows the expert's error rate ($1-\theta_{j,z}$).

Two scenarios were considered: i) in the *global annotation* scenario, every expert produces labels for every example and ii) in the *subset annotation* scenario, each instance receives a label from a subset of labelers (see above). To realize the second scenario, a probability, $\gamma_{j,z}$, is assigned to each expert and label type which defines how often the expert provides labels for label type z . Each $\gamma_{j,z}$ is drawn uniformly at random from the interval [0,1]. $\gamma_{j,z} = 1$ indicates expert j labels all instances contained in class z and $\gamma_{j,z} = 0$ indicates the expert never labels that label type.

4.1. Baseline Estimators

The most probable label estimation is compared to the four following estimators; each labels the interaction as a 'direct interaction' if:

- (1) the majority of the experts label them as "direct interaction" (*Majority voting*).
- (2) there is at least one expert that thinks it is a "direct interaction" (*Single voting*)
- (3) there is at least two experts voting for "direct interaction" (*Double voting*)
- (4) all the experts agree on the "direct interaction" label (*All voting*)

4.2. Evaluation of Synthetic Data Experiments

The synthetic experiments were repeated $n = 300$ times. To measure how accurately the maximum likelihood estimator can recover the true expert labeling accuracies and uncertainties, the average mean squared error (AMSE) was calculated:

$$AMSE(\hat{\theta}^{\text{mle}}) = \frac{1}{n} \frac{1}{2m} \sum_{z=1}^2 \sum_{j=1}^m (\hat{\theta}_{z,j}^{\text{mle}} - \theta_{z,j})^2, \quad AMSE(\hat{u}) = \frac{1}{n} \left(\hat{u}_i(\hat{y}_i, \hat{\theta}^{\text{mle}}) - u_i(\hat{y}_i, \theta) \right)^2 \quad (14)$$

In order to assess whether the accuracy of the final label is correctly assigned, precision and recall rates are reported. The precision is defined as the fraction of true direct interactions that are identified by the method as “direct interaction”. On the other hand, recall is the fraction of the correctly identified “direct interaction” pairs among all the pairs that are direct interactions:

4.3. Performance Results

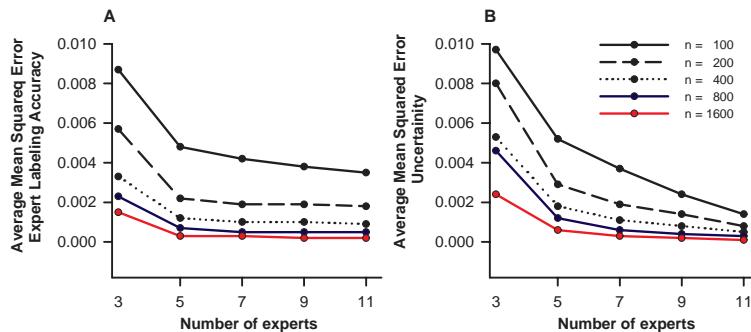


Fig. 2. The average mean squared errors in estimating a) expert labeling qualities and b) uncertainties are plotted as a function of the number of experts for different numbers of pairs to be annotated.

In order to understand the method’s robustness for the number of examples and experts that are present, the error rates were measured as a function of the number of experts and number of pairs annotated. Fig. 2 A displays the results for estimated expert label accuracies for the global annotation case. Not surprisingly, the error decreases as more experts are included and more data are provided. Nevertheless, the average MSE of expert labeling accuracies, θ , is $0.0087(\pm 0.0121)$, even for the case with only 3 experts and $n = 100$ data points. Similarly, the error in estimating the uncertainties of data points is not more than 0.010 (see Fig. 2 B). Comparison of error curves for different n reveals that the gain in accuracy decreases in different data regimes. For example, the error in estimating θ decreases by an amount of 0.003 when n is ramped from 100 to 200 and there are 3 experts. This difference is only 0.0008 for cases $n = 800$ and $n = 1600$. A similar trend was observed for the number of experts; the largest gain in accuracy occurs when the number of experts increases from 3 to 5.

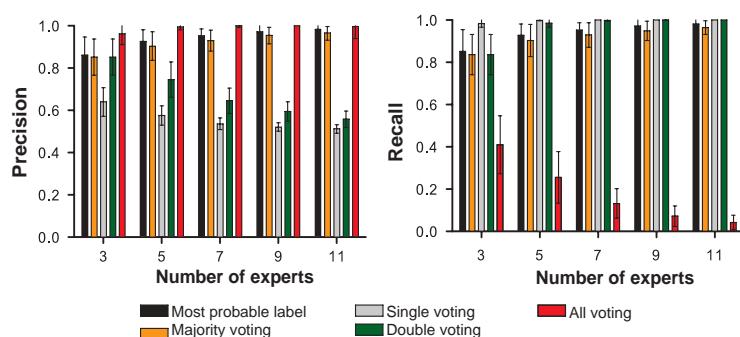


Fig. 3. Precision and recall rates of different labeling strategies.

To assess how well the method can identify true direct interactions, the precision and

recall rates of the estimator are calculated. The precision and recall of the MLE estimator were compared to four other estimators (described in Section 4.1). Fig. 3 displays the precision and recall for different numbers of labelers for the experiments described above and shown in Fig. 2. As can be seen in Fig. 3, single-voting would cover the largest quantity of the true interactions correctly, but it would also consider many incorrect interactions as real physical interactions thus exhibiting a low precision and high recall rate. A similar observation is valid for the double-voting scenario. In both cases the precision becomes worse as the number of labelers increase, since the probability of any pair of labelers producing an incorrect label increases. The opposite trend is true for the all-voting case; the precision is high since the criterion to label a pair as direct interaction is very strict: an agreement between all experts is sought. However, this estimator suffers from low recall rates. In summary, the all-voting strategy results in high confidence sets, but disregards a large portion of the available data; whereas single-voting or double-voting lead to sets with high coverage but both suffer from high false positive rates. The majority-voting method is a robust one; both the precision and recall rates are high, and additionally, as the number of experts increases, the performance improves too. Nevertheless, the maximum likelihood estimator is the best method for both precision and recall rates for all numbers of experts. This is because the noise is taken into account in our probabilistic framework.

5. Refined Literature Curated HIV-1, Human Protein Interactome

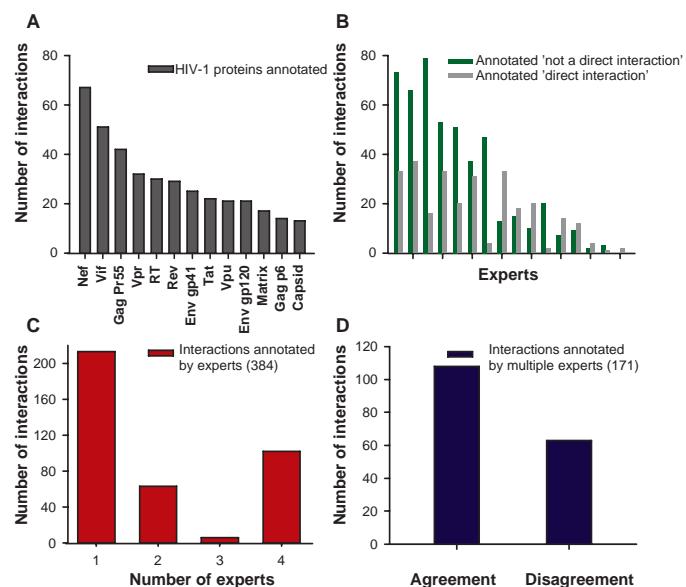


Fig. 4. a) Number of interactions annotated for each HIV-1 protein b) Number of interactions annotated as “direct interaction” (green) and “not a direct interaction” (gray) by each of the 16 HIV-1 experts c) the number of experts annotating each interaction d) Counts of agreed and disagreed interactions when multiple experts assign labels.

In order to estimate expert labeling accuracies, only the interactions that are multiply annotated were considered. For various HIV-1 proteins, different numbers of interactions are annotated; the HIV-1 protein nef has 67 interactions annotated, whereas capsid has only 13 (Fig. 4a). Each HIV-1 expert provided different number of labels (Fig. 4b). The majority of interactions received only one expert opinion (213/384), whereas the rest (171/384) received

multiple expert comments (Fig. 4c). For interactions for which there are multiple opinions from different experts, disagreements among the experts were observed: for 37% of interactions (63/171) experts disagree on the label type (Fig. 4d). Of all the expert annotated interactions, 299 of them are described by the keywords “interacts” with and/or “binds” and these have the most potential to be direct interactions. However, at least two experts still annotated 73 out of 299 as “not a direction interaction” with no disagreements. These results highlight the necessity of reviewing the published interactions with community opinion.

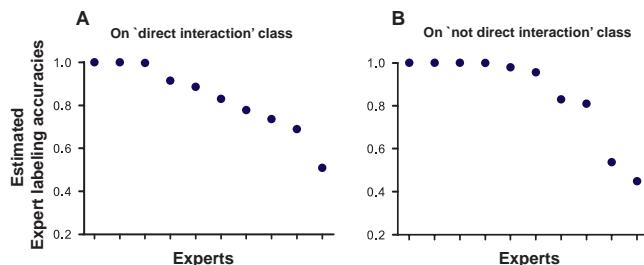


Fig. 5. The estimated labeling accuracies of experts on annotating the PPIs for a) “direct interaction” class and b) “not direct interaction” class.

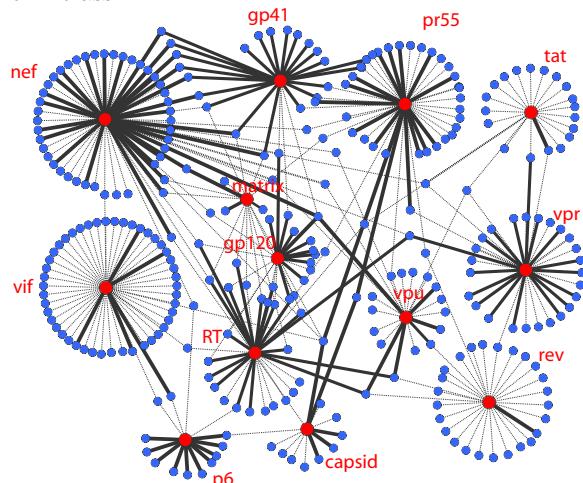


Fig. 6. Refined HIV-1, human protein interaction network based on the estimated confidence scores using HIV-1 expert opinions. Nodes indicate HIV-1 proteins (red) and human proteins (blue). An edge indicates an interaction at least one expert having provided an opinion about it. The thicker the edge, the higher the probability of the pair being a direct interaction according to the estimates. The solid lines represent the interactions for which $P(y = \text{direct interaction}) > 0.5$, whereas dashed lines represent cases for which this probability is < 0.5 . The HIV-1 proteins’ names are placed next to its subnetwork of interactions.

Using the interactions that receive more than one expert opinion (171 interaction pairs), the experts’ labeling accuracies were assessed. There were 16 experts in total, but not all experts provided data for both label types. The estimated label accuracies for the experts are plotted in Fig. 5. 7 out of 10 experts have a labeling accuracy of more than 75 % accuracy on the “direct interaction” class and 8 out of 10 retain this level of accuracy for the “not direct interaction” class. Using the expert labeling accuracies, the most probable class label was calculated for all the annotated interactions. For 147 (out of 384) reported interactions, there was enough evidence to conclude that they have a direct interaction. Fig. 6 displays the resulting network.

6. Validation of Derived Labels with a Supervised Model

Table 1. Performance of the three Random Forest models.

Trained with	MAP avg	MAP stderr	AUC avg	AUC stderr
expert positives + expert negatives	0.7144	0.0083	0.7818	0.0052
expert positives + random selected negatives	0.5466	0.0077	0.6392	0.0084
Baseline	0.4298	0.0075	0.4966	0.0098

The estimated labels obtained in this paper provide a high quality set of interaction labels, which include 158 positive examples and 226 negative examples. In the ensuing discussion, we will be designating this set as “expert-labeled”. To validate these labels, we explore a supervised prediction strategy. We built models to classify an interaction into two classes of “direct interaction” and “not direct interaction” based on biological features of interacting proteins ^a. A Random Forest classifier was learned using the the expert-curated labels in a cross-validation setting. This model achieves a 71% MAP score, which is significantly better than the baseline whose MAP score is 43% (Table 1). The baseline classifier was trained on the same set of examples while the labels of the training examples were randomly permuted. This indicates that the expert-derived labels correlate with the features more strongly.

The subset of interactions estimated to have the “not direct interaction” label type is especially valuable for the prediction task as there is no negative set of PPI interactions readily available. A common way to circumvent this difficulty is to create negative datasets by selecting randomly paired examples that are not in the positive set.²⁵ As the randomly selected negative labels are likely to be far away from the class boundary, they are more easily classified and more likely to give optimistic estimates favoring prediction success; that is the decision boundary learned from them might be too far away from the real decision boundary. Conversely, the expert-labeled negatives from our study are more likely to be in close proximity of the class boundary because they are functional associations. Therefore, our expert-labeled negative examples will define a better decision boundary.

In order to judge how our expert-labeled negative data contributes to the supervised prediction model, a third Random Forest classifier²⁶ was trained. This model uses the expert labeled positive set and a negative label set comprising random pairs that are not reported in the NIAID database. This model performs better than the baseline, 0.55 (± 0.0077) (compare to 43%), but it performs worse than the first model which uses the expert labeled negative examples, (compared to 71% MAP). The AUC values are also ranked similarly (See Table 1). All three models were tested on the expert labeled data in a 3-fold cross validation (CV) scenario. The CV procedure is repeated 10 times, where at each repeated run the splitting of the data is different and random. These empirical results strongly indicate that the curated data is highly valuable.

7. Conclusions

In this paper, a set of expert opinions on literature curated HIV-1, host interactions were collected. To account for noise and subjectivity in expert opinions, the expert labeling accuracies

^aEach PPI pair is described by 42 features derived from diverse set of biological information with details described in our earlier work.¹¹

were estimated and these estimates were used to compute reliability scores that rank interactions with their likelihood to be direct physical interactions. A Random Forest model trained with the derived labels validates the quality of the collected data. The scope of the method presented here is not limited to HIV data, but is applicable to other bodies of literature-curated databases, where noisy labels from multiple experts are available and there is no benchmark data to estimate labeler qualities.

Acknowledgement

We are grateful to the 16 HIV-1 experts for sharing their opinions about HIV-1, human dataset and Prof. Ziv-Bar Joseph for valuable discussions. We also thank Pittsburgh Center for HIV Protein Interactions, especially Prof. Chris Aiken and Dr. Teresa Brosenitsch, for expert advice and helping us in reaching out experts. The work has been supported in part by NIH, NIAID P50GM082251, NSF CCF-1144281, NIH-NLM 2RO1LM007994-05, EraSysBio+ grant and BMBF, SHIPREC and EU Marie Curie Actions 626470 MPFP FP7-PEOPLE-2013-IIF. O.T. acknowledges support from Bilim Akademisi - The Science Academy, Turkey under the BAGEP program and the support from L’Oreal-UNESCO under the National Fellowships Programme for Young Women in Life Sciences.

References

1. B. J. Breitkreutz, C. Stark, T. Reguly *et al.*, *Nucleic Acids Res* **36**, D637 (2008).
2. R. G. Ptak, W. Fu, B. E. Sanders-Bear *et al.*, *AIDS Res Hum Retroviruses* **24**, 1497 (2008).
3. W. Fu, B. E. Sanders-Bear, K. S. Katz *et al.*, *Nucleic Acids Res* **37**, D417 (2009).
4. A. Chatr-aryamontri, A. Ceol, L. M. Palazzi *et al.*, *Nucleic Acids Res* **35**, D572 (2007).
5. I. Xenarios, E. Fernandez, L. Salwinski *et al.*, *Nucleic Acids Res* **29**, 239 (2001).
6. S. Kerrien, Y. Alam-Faruque, B. Aranda *et al.*, *Nucleic Acids Res* **35**, D561 (2007).
7. T. S. Keshava Prasad, R. Goel, K. Kandasamy *et al.*, *Nucleic Acids Res* **37**, D767 (2009).
8. G. D. Bader, I. Donaldson, C. Wolting *et al.*, *Nucleic Acids Res* **29**, 242 (2001).
9. H. Yu, P. Braun, M. A. Yildirim *et al.*, *Science* **322**, 104 (2008).
10. B. A. Shoemaker and A. R. Panchenko, *PLoS Comput Biol* **3**, p. e43 (2007).
11. O. Tastan, Y. Qi, J. G. Carbonell *et al.*, *Pac Symp Biocomput* , 516 (2009).
12. S. Orchard, H. Hermjakob and R. Apweiler, *Proteomics* **3**, 1374 (2003).
13. S. Orchard, S. Kerrien, P. Jones *et al.*, *Proteomics* **7 Suppl 1**, 28 (2007).
14. S. Orchard, L. Salwinski, S. Kerrien *et al.*, *Nat Biotechnol* **25**, 894 (2007).
15. Editorial, *Nature Methods* **6**, p. 2 (2009).
16. A. Chatr-Aryamontri, A. Ceol, L. Licata *et al.*, *Trends Biochem Sci* **33**, 241 (2008).
17. L. Licata, L. Briganti, D. Peluso, L. Perfetto *et al.*, *Nucleic Acids Research* **40**, D857 (Jan 2012).
18. L. von Ahn, B. Maurer, C. McMillen *et al.*, *Science* **321**, 1465 (2008).
19. P. Albert and L. Dodd, *Biometrics* **60**, 427 (2004).
20. V. C. Raykar, S. Yu, L. H. Zhao, G. H. Valadez, C. Florin, L. Bogoni and L. Moy, *The Journal of Machine Learning Research* **11**, 1297 (2010).
21. P. G. Ipeirotis, F. Provost and J. Wang, Quality management on amazon mechanical turk, in *Proceedings of the ACM SIGKDD workshop on human computation*, 2010.
22. D. Marbach, J. C. Costello, R. Küffner, Vega *et al.*, *Nature methods* **9**, 796 (2012).
23. Z. Lu, H.-Y. Kao, C.-H. Wei *et al.*, *BMC Bioinformatics* **12**, p. S2 (2011).
24. A. Dempster, N. Laird and D. Rubin, *Journal of the Royal Statistical Society, Series B* **39**, p. 138 (1977).
25. A. Ben-Hur and W. S. Noble, *BMC Bioinformatics* **7 Suppl 1**, p. S2 (2006).
26. L. Breiman, *Mach. Learn.* **45**, 5 (October 2001).

CROWDSOURCING RNA STRUCTURAL ALIGNMENTS WITH AN ONLINE COMPUTER GAME

JÉRÔME WALDISPÜHL and ARTHUR KAM

*School of Computer Science, McGill University,
Montreal, QC H3A 0E9, Canada*

E-mail: jeromew@cs.mcgill.ca

http://csb.cs.mcgill.ca

PAUL P. GARDNER

*Biomolecular Interaction Centre, School of Biological Science, University of Canterbury,
Christchurch, New Zealand
E-mail: paul.gardner@canterbury.ac.nz*

The annotation and classification of ncRNAs is essential to decipher molecular mechanisms of gene regulation in normal and disease states. A database such as Rfam maintains alignments, consensus secondary structures, and corresponding annotations for RNA families. Its primary purpose is the automated, accurate annotation of non-coding RNAs in genomic sequences. However, the alignment of RNAs is computationally challenging, and the data stored in this database are often subject to improvements. Here, we design and evaluate Ribo, a human-computing game that aims to improve the accuracy of RNA alignments already stored in Rfam. We demonstrate the potential of our techniques and discuss the feasibility of large scale collaborative annotation and classification of RNA families.

Keywords: Crowd-sourcing, Human-computing game, RNA alignment

1. Introduction

Non-coding RNAs (ncRNAs) are functional RNA molecules that are not translated into proteins. They play key roles in aspects of gene transcription, protein transport, molecular assembly and regulatory processes (e.g. riboswitches and microRNAs).^{1,2} The annotation and classification of ncRNAs is essential to decipher molecular mechanisms of gene regulation in normal and disease states. The Rfam database maintains alignments, consensus secondary structures, and corresponding annotations for RNA families. Its primary purpose is the automated, accurate annotation of non-coding RNAs in genomic sequences.³ However, the alignments stored in this database are often subject to improvements. In fact, the Rfam consortium recently released a open call for participation, asking to its users to submit new or improved RNA alignments (http://rfam.sanger.ac.uk/submit_alignment).

Initiatives such as OpenStreetMap and crowdcrafting have proven that crowd-sourcing and human-computing techniques are valuable ways to both analyze and annotate large datasets that require human expertise as well as to solve problems that are difficult to treat with classical computer algorithms. Scientific games like Foldit⁴ and our previous contribution Phylo⁵ illustrate the potential of these techniques for studying, mining, and processing molecular biology data. More recent applications of scientific games to molecular biology problems include Dizeez,⁶ The Cure,⁷ EteRNA,⁸ Fraxinus,⁹ and Nanocrafter.¹⁰ Currently, people collectively

spend an estimated 3 billions hours per week playing computer games. By tapping into this tremendous source of human attention and effort, human-computing games have the potential to bring massive resources to bear on solving complex problems arising in genomics.¹¹

This call emphasizes the potential impact of new tools such as **Phylo** in genomics. The number of new RNA discoveries has accelerated thanks to new sequencing technologies and computational tools,^{12–18} consequently the **Rfam** curators are now overwhelmed by the sheer number of ncRNAs that require their attention.¹⁹ This is largely driven by the discovery of thousands of ncRNAs using RNA-seq datasets.^{18,20} Functional validation is also carried out using high-throughput approaches such as transposon mutagenesis^{15,21} and large-scale genome project provides useful evolutionary conservation dimension.^{22,23} The small resources of the **Rfam** consortium are significantly stretched with keeping the database up to date with new families, revisiting existing families with new information and maintaining the core **Rfam** resources such as the website, MySQL database and the different data queries. Some of this has been alleviated by contributions from the community via the RNA Families track with the journal *RNA Biology*.^{24,25} However, additional inputs from the research community such as crowd-sourcing and human-computing techniques could be valuable to maintain the quality of the data.

One of the difficulties for RNA analysis is that RNA sequences are generally poorly conserved whereas RNA structures are generally conserved.²⁶ The bulk of these structures are determined by secondary structure interactions.²⁷ These are formed by hydrogen bonding interactions between nucleotides (A-U, C-G and G-U) and basepair stacking interactions. Consequently, the tools for aligning homologous RNAs need to take structure into account in order to make accurate predictions. However, this is an NP-complete optimization problem²⁸ and to-date no ideal heuristic solution has been implemented.²⁹ Therefore, the “gold standard” for RNA sequence analysis remains the manual refinement of RNA alignments which have produced highly accurate structure predictions. In extreme cases, 97-98% of manually inferred structures were validated by crystallographic methods.³⁰

Our group was the first to bring citizen science to the field of comparative genomics when, in 2010, we released **Phylo** (<http://phylo.cs.mcgill.ca>), a human-computing framework to solve the multiple sequence alignment (MSA) problem. The key idea of **Phylo** is to convert the MSA problem into a casual puzzle game that can be played by ordinary web users with a minimal prior knowledge of the biological context. In our original study,⁵ the puzzles were extracted from a 44-species MSA stored at the UCSC genome browser, and the best solutions have been re-inserted at their original locations to produce a higher quality MSA. One of the main innovations of **Phylo** was to push the gamification aspect at its limits. Unlike **Foldit** that require each new player to learn the basics of the biophysics of protein folding through a detailed tutorial before starting to play, **Phylo** is a true casual game, requiring absolutely no knowledge of genomics. Indeed, the latter is an intuitive Tetris-like game where players have to match colored blocks. As a consequence, the game is accessible to a broader audience and can benefit of the workforce provided by crowds composed of ordinary web users with a minimal prior knowledge of the biological context.

Here, we design and experiment with **Ribo**, a human-computing game that aims to improve the accuracy of RNA alignments. ncRNAs are characterized by a conserved secondary structure

associated with their function. Therefore, RNA alignments require to simultaneously align sequences and secondary structures. We propose to develop a game inspired from *Phylo* for this specific case. We introduce new types of blocks representing the base-pairs of the secondary structure. Our working prototype uses left- and right-handed triangles to represent open and closing base-pairs of the bracket notation. In addition, unlike *Phylo*, this game is not using a phylogenetic tree and thus is easier to understand for non-experienced players. We evaluate the quality of the alignments calculated by the players with the *Infernal* package.³¹ In particular, we show that the solutions we collected through *Ribo* enabled us to build covariance models with better overall homolog recognition performances than the ones built from the initial *Rfam* alignments. This work suggests that the use of human-computing games has the potential to become a valuable resource to maintain RNA alignment databases. *Ribo* is available at <http://ribo.cs.mcgill.ca>.

2. Methods

2.1. *RNA secondary structures and RNA alignments*

Ribonucleic acids (RNAs) are versatile biomolecules that are involved in a diverse number of biological functions. For example, as messenger RNA it encodes genes, as microRNA it regulates genes and as ribosomal RNA it translates genes. To achieve their functions, non-coding RNAs (ncRNAs) use sophisticated structures that can be described at two levels. First, the secondary structure is the set of all canonical base-pairing interactions found in the native conformation of the molecule. The canonical base pairs include Watson-Crick interactions between Adenine (A) and Uracil (U) or Guanine (G) and Cytosine (C), as well as Wobble interactions between Guanine and Uracil bases. Contiguous canonical base pairs form secondary structure elements called helices (or stems) connected together through various types of loops (E.g. hairpins, bulges, internal loops and multi-loops). The majority of secondary structures can be represented as planar graphs. Moreover, less than 5% of the secondary structures found in the *Rfam* database³ contain crossing interactions also called pseudo-knots. Hence, many secondary structures can be conveniently represented using the dot-bracket notation that is illustrated in Figure 1. Then, the secondary structure elements are assembled together via numerous van der Waals contacts and specific hydrogen bonds into the tertiary structure (or 3D structure). RNA folding is hierarchical. The secondary structures of RNA form rapidly, these act as a scaffold for the slower formation of tertiary structures.²⁷ For this reason, secondary structures provide a relatively accurate signature of the molecular function, as illustrated by the strong evolutionary conservation of RNA secondary structures.³² For instance, the secondary structures of tRNAs adopt a typical cloverleaf shape.^{33,34} The evolution of homologous ncRNAs is constrained by these functional structures, and their alignments generally comply with this information too. Thus, a ncRNA alignment is associated with a consensus secondary structure that is representative of the functional family. To date, the *Rfam* database is the most popular repository of structured ncRNA alignments.³ These alignments can then be used to build covariance models, which are widely used for the functional annotation of ncRNA sequences with unknown functions.

2.2. Scoring scheme

The scoring scheme that we currently use for evaluating how well the RNA sequences and structures are aligned is based upon the nucleotide sequence scoring scheme derived using a Markovian transition model (States et al. 1991), this approach is how the PAM (Point Accepted Mutation, for proteins) matrices were derived.³⁵ The nucleotide scores suggest that for approximately 65% sequence identity, a score between 1.40 and 1.34 should be used for matches and between -1.15 and -1.04 for mismatches. The ratios between these scores convert to approximately the integers +5/-4 for match/mismatch scores. This scoring scheme has been shown to work relatively well for the RNA homology search problem.³⁶ We selected relatively low gap-open and gap-extend penalties of -5 and -2 respectively, as indels are thought to be relatively frequent.^{37,38} We add bonuses for matching base-pairs. These bonuses should exceed the penalty for any double mismatches (i.e. > 8) resulting from structure-neutral variation (e.g. *A·U* to *G·C*) as well as tolerate the indels that are required to explain base-pair conservation. Therefore a bonus of +12 was selected for aligning base-pairs. At present, covarying sites are not awarded bonuses,³⁹ nor are inconsistent and contradicting base-pairs penalized.⁴⁰

2.3. Datasets

We used the 5S rRNA multiple sequence alignment (**Rfam** ID RF00001) from the last release of the **Rfam** database³ to perform our experiments. This alignment contains 712 sequences and has 231 columns. The 5S rRNA is a component of the large ribosome subunit, and therefore is an essential and ubiquitous RNA, but it has been difficult for **Rfam** to get this alignment correct (data not shown). Although it is a well-known and heavily studied RNA, its alignment remains an open-problem. The 5S rRNA tertiary structure is essential to ribosome assembly and function; hence it is strongly conserved across species. This structure has been experimentally determined as part of the complete ribosome,^{41–43} and this has been used as a reference to obtain the current **Rfam** alignment.

Currently, the grid of the game allows us to represent up to 10 sequences. Thus, we aimed to improve MSA of similar height. We extracted a set of sub-alignments MSA-ref with 4, 6, 8 and 10 sequences. We selected sequences with low average sequence similarity. In our dataset, the average sequence similarity vary from 36% to 58%. This value can be compared with the average sequence similarity of the complete **Rfam** alignment, which is 60%. A full description of the dataset is available in Table 1. This metric is important because sequences with low sequence similarity are hard to align.

We built the test sets with the sequences from the original **Rfam** alignment not used in the reference sub-alignment set MSA-ref. Hence, the discriminative power of a sequence of 6 sequences has been estimated on a benchmark set containing the other 706 sequences from the **Rfam** alignment.

2.4. Puzzle construction

We built two sets of puzzles with 25 and 35 columns, labeled as “Easy” and “Hard”, from the **Rfam** sub-alignments MSA-ref described above. The choice of the sizes has been determined to maximize the use of the grid of 50 columns currently used in our game, and at the same

time to give the players enough room to explore the configuration space.

Because the experimentally determined structure may not always be available, we ignored the consensus structure that is available in the original **Rfam** alignment. Instead, we predicted a secondary structure using the maximum expected accuracy (MEA) secondary structure predicted by **RNAfold**^{44,45} for each individual sequence. Therefore, it is important to keep in mind that in this study, the **Rfam** alignments benefit of an information not used by the players. We removed empty columns (i.e. columns containing only gaps) from the sub-alignments, and extracted all continuous regions of 25 and 35 columns. Then, we removed from each region the base-pairs that were not included within this region. In other words, if a nucleotide has a predicted interaction with another nucleotide outside of the region of interest, this base-pair is ignored.

We sorted all regions according to the total number of base-pairs in the region (i.e. the sum of the number of valid base-pairs predicted by **RNAfold** within the region for each sequence of the sub-alignment). Regions without any base-pair were ignored. Finally, we selected first the region with the larger number of base-pairs, then the next one provided that it does not overlap with the region previously selected, until the queue is left empty. In the end, we generated 27 puzzles that are described in Table 1. In this table, we report the number of columns (width) and sequences (height) of the puzzle, and its ID in the game. We also report the average sequence similarity of the **Rfam** sub-alignment used to create the puzzle, as well as the average sequence similarity of the puzzle. Finally, we report the percentage of nucleotides involved in a base pairing interaction, and the percentage of gaps found in the initial configuration of the puzzle.

2.5. *Benchmarking methodology*

The quality of the alignments was evaluated using **Infernal**.³¹ **Infernal** is the software suite used to build the covariance models from **Rfam** seed alignments and search for homologs (available in the full **Rfam** alignment).

For each submission (i.e. a puzzle solved by a player), we substituted the original alignment of the region used to build the puzzle, with the solution provided by the player. Then, we calculated a covariance model for each of these alignments (the original one and the one built using the submission) with the program **cmbuild** of the **Infernal** package. Finally, we calibrated the covariance models with the program **cmcalibrate**, and used the program **cmsearch** to compute a fitness score evaluating the likelihood of the covariance model on each sequence of the test set (the set of the **Rfam** seed sequences not used in the original alignment).

The fitness is estimated with the E-value calculated by **cmsearch**. In our experiments, we report the average E-value of all sequences in the test set. Among all solutions collected for a given sub-alignment, only the best values are reported. Indeed, as in **Phylo** the purpose of our system is to generate a sparse set of potential solutions in which we have high probability to find a configuration improving the original one.

Table 1. Ribo puzzle data set

Width	Height	ID	Average sequence similarity	Percentages in puzzles	
			Rfam alignment	Ribo puzzle	base pairs
25	4	1	45	56	34
25	4	2	45	33	28
25	6	3	52	73	36
25	6	4	52	58	61
25	8	5	51	61	27
25	8	6	51	54	61
25	10	7	52	68	31
25	10	8	52	47	40
25	4	17	36	33	32
25	4	18	36	40	40
25	6	19	40	47	46
25	8	20	45	56	33
25	10	21	43	48	32
35	4	9	58	73	37
35	4	10	58	65	55
35	6	11	53	59	39
35	6	12	53	51	44
35	8	13	52	67	30
35	8	14	52	44	38
35	10	15	57	68	35
35	10	16	57	59	47
35	4	22	36	34	37
35	4	23	36	40	38
35	6	24	40	34	32
35	6	25	40	43	40
35	8	26	45	51	34
35	10	27	43	49	38

3. Results

3.1. Game design

Ribo is inspired from our previous contribution Phylo. We abstract a sequence alignment into a tile-matching game, where nucleotides are represented with coloured bricks that can be moved horizontally on a grid. The objective of the players is to align the nucleotides of similar colours within the same columns, in order to reveal similarities between sequences.

Nonetheless, RNA alignments have a major difference with DNA alignments. The conservation of the native (functional) structure is often more important than the conservation of the primary sequence. As in Rfam, the molecular structure are represented by secondary structures. Hence, RNA alignments aim to conserve base-pairs. For instance, if a base-pair occurs between indices (i_1, j_1) in one sequence and another one between indices (i_2, j_2) in a second sequence, then the alignment of nucleotides at index i_1 and i_2 in the same column must be, as much as possible, associated with the alignment of nucleotides at index j_1 and j_2 . Since the conservation of base-pairing properties is essential for RNA alignments, we need to design new mechanisms to represent this information and enable users to use it in the game.

RNA secondary structures encompass the maximal set of stem and stem-loops formed by

canonical base-pairing interactions (Watson-Crick and Wobble). Each nucleotide can be involved in at most one base-pair (i.e. no base-triples) and crossing interactions are forbidden (i.e. no pseudo-knots).

In **Ribo**, we chose to adapt the bracket notation frequently used to represent RNA secondary structures. An open parenthesis indicates that the nucleotide is paired with the first available nucleotide associated with a close parenthesis on its right. Dots represent unpaired nucleotides. The bricks used in **Ribo** merge the sequence and structural information into a single token. As in **Phylo**, the colour of the brick encodes the type of the nucleotide (i.e. A, C, G or U). In addition, we use now a new set of bricks with different shapes to encode the base-pairing properties. Hence, a triangle pointing to the right indicates that the nucleotide is paired with another nucleotide on its right (i.e. the equivalent of the open parenthesis), while a triangle pointing to the left indicates the opposite. Unpaired nucleotides are represented using a squared brick. Figure 1 illustrates our encoding.

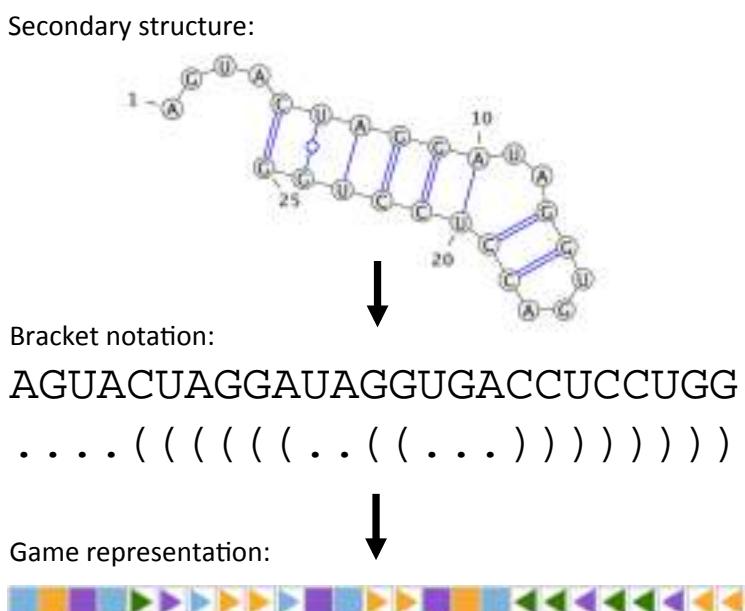


Fig. 1. Encoding of RNA sequence and secondary structure in **Ribo**. The secondary structure is drawn with VARNA⁴⁶

The scoring scheme used in **Ribo** significantly diverges from the one used in **Phylo**. Since the phylogenetic tree is unknown (as it is often the case in RNA alignments), we use a sum-of-pair scoring function. In other words, the total score of a multiple sequence alignment is the sum of the alignment scores of each pair of sequences. Here, matches (bricks with identical colours aligned together) receive a bonus of +5, and mismatches (bricks with different colours) receive a penalty of -4. The opening of a gap costs -5 and their extension only -2. Finally, the alignment of a base-pair receive a bonus a +12. We motivate these choices in section 2.2. As it could be the case in practical bioinformatics applications, in **Ribo** we do not penalize misaligned or contradicting alignments of base-pairs.⁴⁷ We argue that including such penalties

would affect the design of the game and diminish the engagement of players. By contrast, the bonuses assigned to matched base-pairs create an extra incentive to players to explore the configuration space, and is sufficient to serve our purpose to align secondary structures.

We show a screenshot of the game in Figure 2. The game board uses some successful element designs previously developed with **Phylo** such as the score bar of top indicating the current score, best score achieved during this session and score to beat (i.e. the Par). We also added new features such as the locks on the left side. The latter enables the players to “lock” a row and move the full sequence as whole, preserving gaps between the bricks. This feature aims to facilitate the playability of the game on tablets and mobile devices. Moreover, the visual identification of long-range base-pairs can be difficult. To address this issue, we implemented a highlight mechanism that shows the base-paired brick every time that the cursor overlaps with a brick. Finally, we do not need to represent a phylogenetic tree as it was the case in **Phylo**. Thus, we have more space to display the game board. With **Ribo**, we decided to increase the grid to 50 columns (instead of 25 with **Phylo**). This upgrade is essential because base-pairs can involve nucleotides that are very distant in the sequence. **Ribo** is available at <http://ribo.cs.mcgill.ca>.

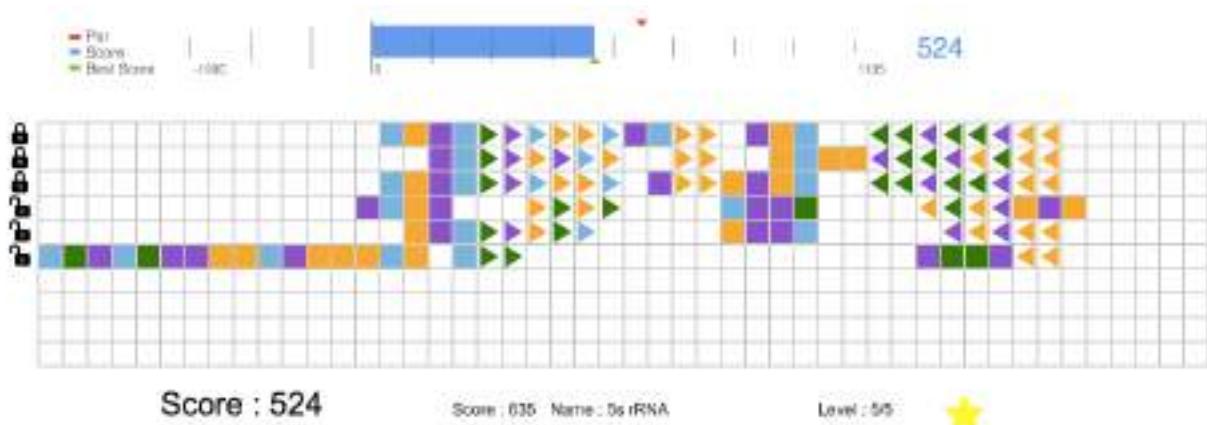


Fig. 2. Screenshot of Ribo

The progression of the player within the game is similar to the one used in **Phylo**. First, the user starts with two sequences and tries to find an alignment with a score that is at least as good as the one found in the original alignment. Once this milestone is reached, the player can access the next stage and add one more sequence to the game. The game is complete when all sequences have been added and when the player managed to beat or match the score of the original alignment.

3.2. Game statistics

Approximately 15 players recruited from undergraduate and graduate students in computational biology at McGill and University of Canterbury participated to the study. We collected 115 submissions (i.e. puzzles completed) whose distribution is detailed in Table 2. The “easiest” puzzles (least numbers of rows and columns) have been significantly more played than the

others. It had to be expected since all participants were beginners and thus needed to learn the rules and train on easy instances first. Nonetheless, it is worth noting that the majority of participants was not familiar with RNA alignments before starting to play.

Table 2. Number of solutions collected

	number of columns	number of sequences				Total
		4	6	8	10	
number of columns	25	36	16	9	8	69
number of columns	35	16	11	5	14	46
Total		52	27	14	22	

3.3. Performance

In Figure 3(a), we report the average E-values obtained on puzzles with the same number of sequences. The decrease of E-values observed on the alignments improved by the gamers tends to validate our approach. An exception is for alignments with 6 sequences. This discrepancy is most likely due to incorrectly predicted base-pairs that resulted in alignment of worse quality. We also note a trend toward higher average E-values when the number of sequence increases. This phenomenon could be an artifact of the small sample set, but could also reflect a real phenomenon. Since all the sequences are quite diverse, higher numbers of sequences in the alignments result in lower the probabilities for each state in the covariance model. Consequently the E-values are likely to be higher.

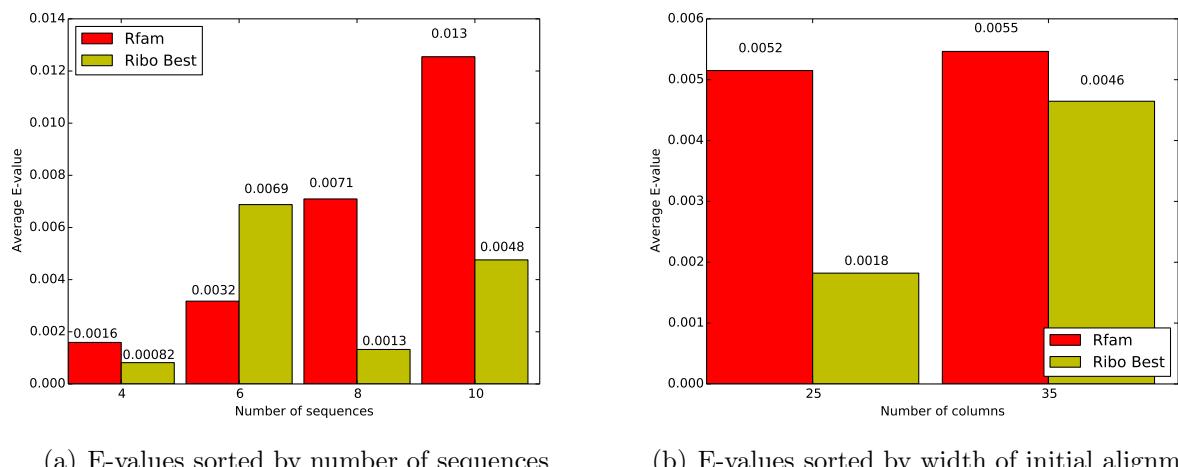


Fig. 3. Average E-values of the covariance model on sequences in the test set. Average E-values calculated with the covariance model obtained from the Rfam alignments are shown in red, and average E-values calculated with the covariance model obtained from Ribo alignments are displayed in yellow. The left panel shows the results obtained when the puzzles are sorted by number of sequences, while the right panel shows the same data when the puzzles are sorted by number of columns (i.e. defined as difficulty in the game).

Similar observations can be done when the puzzles are sorted by difficulty in Figure 3(b) (i.e. the number of columns of the regions used to build in them). The average E-values

decrease for the Easy (25 columns) and Hard puzzles (35 columns). However, the magnitude of the improvement is less pronounced for the largest puzzles. A lower number of submissions (See Table 2) as well as higher difficulties to solve these puzzles can justify this difference. The data presented above suggest that overall our game has the potential to improve RNA alignments. Nonetheless, the distribution of E-values also needs to be considered to understand the real performance of this methodology. Indeed, our data show that E-values obtained with the covariance models calculated from the alignments generated by gamers were better (i.e. lower) on lowest ranked sequences (i.e. sequences with the worst fit to the covariance model), than the E-values obtained on the same sequences with the covariance model calculated from the original alignment. By contrast, the E-values obtained with the covariance model obtained from the original alignment are better than those obtained with the alignment improved by the game on highest ranked sequences (i.e. the sequences with the best fit to the covariance model). Therefore, the new covariance models appear to outperform the ones built from **Rfam** sub-alignments for recognizing distant homologs, but may lack the specificity of the latter to identify sub-families.

4. Discussion

The accurate alignment of sequences for structural RNAs remains a challenging problem. The “gold standard” remains the manual construction of alignments. In fact, the accuracy of careful manual comparisons of sequences were shown to be 100% accurate when evaluated against structures derived from crystallographic data.³² However, this approach is very time consuming and requires highly trained and committed individuals.

We have shown the potential for “crowd sourcing” the RNA multiple sequence alignment problem. Alignments can be broken into a series of sub-sequences and sub-alignments. Crowd-sourced solutions to these can be stitched together, thus building up reasonable solutions to computationally challenging problems. This paper is a proof-of-concept that crowd-sourcing techniques can be used to maintain and improve public RNA alignment repositories.

Feedbacks from our players collected after the benchmark suggest that the limit of the human-computing system have not been reached yet. In particular, we could increase the number of sequences and columns. Nonetheless, such upgrades will also require the development of new GUI features to help the player to deal with large data sets, and help them to efficiently explore the conformational space. For instance, advanced visualization tools to display long-range base-pairing interactions.

We argue that more sophisticated strategies to build the puzzles have the potential to increase the performance of our crowd-computing system. Indeed, the puzzles used in this study are built from continuous regions of an **Rfam** alignment with 35 columns. This strategy prevents us to use long-range interactions between nucleotides that are separated by more than 35 positions. This is an important issue if we wish to use **Ribo** to align multi-loop regions of RNA with sophisticated secondary structures. To address this problem, we suggest building puzzles from discontinuous regions of a full alignment. For instance, we can concatenate a region with 20 columns with another region of 20 columns that contains the nucleotides predicted to base-pair with those of the first region.

Although relatively rare, pseudo-knotted can carry important functions. Currently, 89 Rfam families among 2208 have pseudo-knot annotations. To handle these families, a second set of parenthesis could be used to represent interleaved interactions.

Finally, due to the broad interest of the scientific community in obtaining accurate RNA alignments, we can envision the use of Ribo as a web widget on research web sites to promote the understanding on RNA research to a broad public and engage citizen scientists. The deployment of an open crowd-computing platform such as Open-Phylo⁴⁸ is also scheduled.

5. Availability

Ribo can be played at <http://ribo.cs.mcgill.ca>. The source code and data used in the project are also freely accessible at <http://jwgitlab.cs.mcgill.ca/arthurkam/rna-phylo>.

References

1. S. R. Eddy, *Nat Rev Genet* **2**, 919 (Dec 2001).
2. T. R. Cech and J. A. Steitz, *Cell* **157**, 77 (Mar 2014).
3. S. W. Burge, J. Daub, R. Eberhardt, J. Tate, L. Barquist, E. P. Nawrocki, S. R. Eddy, P. P. Gardner and A. Bateman, *Nucleic Acids Res* **41**, D226 (Jan 2013).
4. S. Cooper, F. Khatib, A. Treuille, J. Barbero, J. Lee, M. Beenen, A. Leaver-Fay, D. Baker, Z. Popović and F. Players, *Nature* **466**, 756 (Aug 2010).
5. A. Kawrykow, G. Roumanis, A. Kam, D. Kwak, C. Leung, C. Wu, E. Zarour, Phylo players, L. Sarmenta, M. Blanchette and J. Waldspühl, *PLoS One* **7**, p. e31362 (2012).
6. S. Loguercio, B. M. Good and A. I. Su, *PLoS One* **8**, p. e71171 (2013).
7. B. M. Good, S. Loguercio, O. L. Griffith, M. Nanis, C. Wu and A. I. Su, *JMIR Serious Games* **2**, p. e7 (2014).
8. J. Lee, W. Kladwang, M. Lee, D. Cantu, M. Azizyan, H. Kim, A. Limpaecher, S. Yoon, A. Treuille, R. Das and EteRNA Participants, *Proc Natl Acad Sci U S A* **111**, 2122 (Feb 2014).
9. D. Maclean, *Elife (Cambridge)* **2**, p. e01294 (2013).
10. U. o. W. Center for Game Science, Nanocrafter (2014).
11. B. M. Good and A. I. Su, *Bioinformatics* **29**, 1925 (Aug 2013).
12. M. Guttman, I. Amit, M. Garber, C. French, M. F. Lin, D. Feldser, M. Huarte, O. Zuk, B. W. Carey, J. P. Cassady, M. N. Cabili, R. Jaenisch, T. S. Mikkelsen, T. Jacks, N. Hacohen, B. E. Bernstein, M. Kellis, A. Regev, J. L. Rinn and E. S. Lander, *Nature* **458**, 223 (Mar 2009).
13. Z. Weinberg, J. X. Wang, J. Bogue, J. Yang, K. Corbino, R. H. Moy and R. R. Breaker, *Genome Biol* **11**, p. R31 (2010).
14. A. Bateman, S. Agrawal, E. Birney, E. A. Bruford, J. M. Bujnicki, G. Cochrane, J. R. Cole, M. E. Dinger, A. J. Enright, P. P. Gardner, D. Gautheret, S. Griffiths-Jones, J. Harrow, J. Herrero, I. H. Holmes, H. D. Huang, K. A. Kelly, P. Kersey, A. Kozomara, T. M. Lowe, M. Marz, S. Moxon, K. D. Pruitt, T. Samuelsson, P. F. Stadler, A. J. Vilella, J. H. Vogel, K. P. Williams, M. W. Wright and C. Zwieb, *RNA* **17**, 1941 (Nov 2011).
15. L. Barquist, G. C. Langridge, D. J. Turner, M. D. Phan, A. K. Turner, A. Bateman, J. Parkhill, J. Wain and P. P. Gardner, *Nucleic Acids Res* **41**, 4549 (Apr 2013).
16. N. A. Siegfried, S. Busan, G. M. Rice, J. A. Nelson and K. M. Weeks, *Nat Methods* **11**, 959 (Sep 2014).
17. R. Andersson, C. Gebhard, I. Miguel-Escalada, I. Hoof, J. Bornholdt, M. Boyd, Y. Chen, X. Zhao, C. Schmidl, T. Suzuki, E. Ntini, E. Arner, E. Valen, K. Li, L. Schwarzfischer, D. Glatz, J. Raithel, B. Lilje, N. Rapin, F. O. Bagger, M. Jørgensen, P. R. Andersen, N. Bertin, O. Rackham, A. M.

- Burroughs, J. K. Baillie, Y. Ishizu, Y. Shimizu, E. Furuhata, S. Maeda, Y. Negishi, C. J. Mungall, T. F. Meehan, T. Lassmann, M. Itoh, H. Kawaji, N. Kondo, J. Kawai, A. Lennartsson, C. O. Daub, P. Heutink, D. A. Hume, T. H. Jensen, H. Suzuki, Y. Hayashizaki, F. Müller, C. FANTOM Consortium, A. R. Forrest, P. Carninci, M. Rehli and A. Sandelin, *Nature* **507**, 455 (Mar 2014).
18. S. Lindgreen, S. Ugur Umu, A. Sook-Wei Lai, H. Eldai, W. Liu, S. McGimpsey, N. Wheeler, P. J. Biggs, N. R. Thomson, L. Barquist, A. M. Poole and P. P. Gardner, *ArXiv e-prints* (June 2014).
 19. W. A. Baumgartner, Jr, K. B. Cohen, L. M. Fox, G. Acquaah-Mensah and L. Hunter, *Bioinformatics* **23**, i41 (Jul 2007).
 20. Z. Wang, M. Gerstein and M. Snyder, *Nat Rev Genet* **10**, 57 (Jan 2009).
 21. L. Barquist, C. J. Boinett and A. K. Cain, *RNA Biol* **10**, 1161 (Jul 2013).
 22. C. 1000 Genomes Project Consortium, G. R. Abecasis, A. Auton, L. D. Brooks, M. A. DePristo, R. M. Durbin, R. E. Handsaker, H. M. Kang, G. T. Marth and G. A. McVean, *Nature* **491**, 56 (Nov 2012).
 23. W. Nasser, S. B. Beres, R. J. Olsen, M. A. Dean, K. A. Rice, S. W. Long, K. G. Kristinsson, M. Gottfredsson, J. Vuopio, K. Raisanen, D. A. Caugant, M. Steinbakk, D. E. Low, A. McGeer, J. Darenberg, B. Henriques-Normark, C. A. Van Beneden, S. Hoffmann and J. M. Musser, *Proc Natl Acad Sci U S A* **111**, E1768 (Apr 2014).
 24. P. P. Gardner and A. G. Bateman, *RNA Biology* **6**, 2 (2009).
 25. P. P. Gardner, *RNA Biology* (2012).
 26. M. P. Hoeppner, P. P. Gardner and A. M. Poole, *PLoS Comput Biol* **8**, p. e1002752 (2012).
 27. I. Tinoco and C. Bustamante, *J Mol Biol* **293**, 271 (Oct 1999).
 28. L. Wang and T. Jiang, *J Comput Biol* **1**, 337 (1994).
 29. P. P. Gardner, A. Wilm and S. Washietl, *Nucleic Acids Res* **33**, 2433 (2005).
 30. R. R. Gutell, *RNA Biol* **11**, 254 (Mar 2014).
 31. E. P. Nawrocki and S. R. Eddy, *Bioinformatics* **29**, 2933 (Nov 2013).
 32. R. R. Gutell, J. C. Lee and J. J. Cannone, *Curr Opin Struct Biol* **12**, 301 (Jun 2002).
 33. R. W. Holley, J. Apgar, G. A. Everett, J. T. Madison, M. Marquisee, S. H. Merrill, J. R. Penswick and A. Zamir, *Science* **147**, 1462 (Mar 1965).
 34. J. T. Madison, G. A. Everett and H. Kung, *Science* **153**, 531 (Jul 1966).
 35. M. O. Dayhoff and B. C. Orcutt, *Proc Natl Acad Sci U S A* **76**, 2170 (May 1979).
 36. E. K. Freyhult, J. P. Bollback and P. P. Gardner, *Genome Res* **17**, 117 (Jan 2007).
 37. A. Löytynoja and N. Goldman, *Science* **320**, 1632 (Jun 2008).
 38. A. Löytynoja and N. Goldman, *Proc Natl Acad Sci U S A* **102**, 10557 (Jul 2005).
 39. E. Freyhult, V. Moulton and P. Gardner, *Appl Bioinformatics* **4**, 53 (2005).
 40. S. Lindgreen, P. P. Gardner and A. Krogh, *Bioinformatics* **22**, 2988 (Dec 2006).
 41. V. Ramakrishnan, *Cell* **108**, 557 (Feb 2002).
 42. J. Harms, F. Schluenzen, R. Zarivach, A. Bashan, S. Gat, I. Agmon, H. Bartels, F. Franceschi and A. Yonath, *Cell* **107**, 679 (Nov 2001).
 43. N. Ban, P. Nissen, J. Hansen, P. B. Moore and T. A. Steitz, *Science* **289**, 905 (Aug 2000).
 44. R. Lorenz, S. H. Bernhart, C. Höner Zu Siederdissen, H. Tafer, C. Flamm, P. F. Stadler and I. L. Hofacker, *Algorithms Mol Biol* **6**, p. 26 (2011).
 45. I. Hofacker, W. Fontana, P. F. Stadler, S. Bonhoeffer, M. Tacker and P. Schuster, *Monatshefte f. Chemie* **125**, 167 (1994).
 46. K. Darty, A. Denise and Y. Ponty, *Bioinformatics* **25**, 1974 (Aug 2009).
 47. P. P. Gardner and R. Giegerich, *BMC Bioinformatics* **5**, p. 140 (Sep 2004).
 48. D. Kwak, A. Kam, D. Becerra, Q. Zhou, A. Hops, E. Zarour, A. Kam, L. Sarmenta, M. Blanchette and J. Waldspühl, *Genome Biol* **14**, p. R116 (2013).

PERSONALIZED MEDICINE: FROM GENOTYPES, MOLECULAR PHENOTYPES AND THE QUANTIFIED SELF, TOWARDS IMPROVED MEDICINE

JOEL T DUDLEY

Icahn School of Medicine at Mount Sinai, 1425 Madison Ave., New York, NY

Email: joel.dudley@mssm.edu

JENNIFER LISTGARTEN

Microsoft Research, One Memorial Drive, Cambridge, MA, 02142

Email: jennl@microsoft.com

OLIVER STEGLE

European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, United Kingdom

Email: oliver.stegle@ebi.ac.uk

STEVEN E BRENNER

Department of Plant & Microbial Biology, 111 Koshland Hall, University of California, Berkeley 94720

Email: brenner@compbio.berkeley.edu

LEOPOLD PARTS

University of Toronto, Donnelly Centre for Cellular and Biomolecular Research, 160 College Street,

Toronto, ON M5S 3E1, Canada

Email: leopold.parts@utoronto.ca

Advances in molecular profiling and sensor technologies are expanding the scope of personalized medicine beyond genotypes, providing new opportunities for developing richer and more dynamic multi-scale models of individual health^{1,2}. Recent studies demonstrate the value of scoring high-dimensional microbiome³, immune⁴, and metabolic⁵ traits from individuals to inform personalized medicine. Efforts to integrate multiple dimensions of clinical and molecular data towards predictive multi-scale models of individual health and wellness are already underway⁶⁻⁸. Improved methods for mining and discovery of clinical phenotypes from electronic medical records⁹ and technological developments in wearable sensor technologies present new opportunities for mapping and exploring the critical yet poorly characterized "phenome" and "envirome" dimensions of personalized medicine^{10,11}. There are ambitious new projects underway to collect multi-scale molecular, sensor, clinical, behavioral, and environmental data streams from large population cohorts longitudinally to enable more comprehensive and dynamic models of individual biology and personalized health¹². Personalized medicine stands to benefit from inclusion of rich new sources and dimensions of data. However, realizing these improvements in care relies upon novel informatics methodologies, tools, and systems to make full use of these data to advance both the science and translational applications of personalized medicine.

Genotyping and large-scale molecular phenotyping are already available for large patient cohorts and may soon become available for many patients. Exome or complete genome sequences are increasingly being collected, and in some cases are now covered by insurance. Prenatal diagnosis has been improved by genotyping fetal DNA circulating in mother's blood tissue. Robust statistical and computational methods for analyzing these data will be critical to realizing the promise of personalized medicine. The challenges span from accurate low-level analyses of high throughput datasets to high-level synthesis of mechanisms of action, and identification of causal links between different abstract layers of molecular information, before, finally, incorporating them into health-care such as diagnostics. Important analysis problems include accurate phenotypic characterization, identifying and correcting for latent structure, dealing with missing data, deciding at what level to test (e.g., within genomes, whether to use single base pair values, sets of polymorphisms, exonic regions, etc.), data heterogeneity, the problem of multiple testing, integrating various modalities, deducing functional consequences *in silico*, addressing data quality, and making sense of new data types as they become available.

The path from genotype to disease state goes through intermediate phenotypes. To modulate the disease risk or trait, one of the molecular intermediates must be changed in a controlled way using small molecules or changes in environment. Finding the right intermediate molecule to target for these interventions remains a key challenge. A first level of understanding should come from genetic mapping studies – that is, to determine to which extent do the loci responsible for heritable disease risk affect intermediate traits. Much progress has been made on this front over the last years, especially for genetic control of RNA levels^{13,14}—so-called “eQTL” analysis, but also for protein, metabolite and epigenetic modification abundances¹⁵⁻¹⁸, with much remaining to be done. The next task is distinguishing the actual drivers of ailment from traits that do respond to genotype, but do not cause disease. Causal models, such as those based on Mendelian randomization and mediation analysis, will play a crucial role in separating out the molecular causes of disease from the high-dimensional state of the organism^{19,20}.

Medicine is gradually moving away from the traditional model of reactive sick-care towards wellness and all-time learning healthcare systems that aim to prevent individuals from perturbing their individual biology towards states of disease^{1,2}. Personalized medicine aims to soon allow dynamic, quantitative representation of an individual patient's health "GPS coordinates" estimated from multiple modalities of personal health data¹. Still, much work is required in all areas, from basic discovery of molecular mechanisms of disease pathology, to statistical methods of causality and publicly available computational infrastructure to deliver on the promise of genetic and other personalized information in the clinic and beyond.

Session contributions

Dr. Nathan Price gives the invited talk. Dr. Price, along with colleagues at the Institute for Systems Biology, is spearheading the innovative Wellness 1K program that aims to take personalized medicine from “sick care” to maintenance of wellness by way of democratized healthcare².

The electronic medical record (EMR) captures clinical phenotype information and is being used increasingly as an important source of research data for precision medicine discovery. In our session, **Glicksberg et al.** discuss a novel integrative method combining genetic and EMR data for data-driven

discovery of disease relationships. The authors integrate disease-associated variants reported in the literature with EMR data from a large metropolitan hospital. The method evaluates statistical overlaps between patients sharing disease diagnoses in the EMR and disease phenotypes sharing overlapping associated loci. The results identify 19 putatively novel disease pairs supported by both EMR and genetic data that suggest possible shared etiological factors or novel risk factors.

Equipping clinical investigators with the ability to perform large-scale analysis of integrated clinical and molecular data is a key challenge in precision medicine. Clinical scientists sit at the interface of patient care and medical research and thereby serve as critical translators of clinical needs into specific research questions. Integrative methods combining genomic and clinical data offer powerful high-dimensional approaches for clinical hypothesis testing and patient cohort exploration. However, clinical investigators often lack the technical skills required to build, manage, and query integrated genomic and clinical data. In our session, **Hinterberg *et al.*** present the Phenotype-Expression Association eXplorer (PEAX) software enabling interactive data exploration of relationships between multivariate phenotype models and gene expression. The PEAX software interface enables visual, interactive definition of sub-phenotyping using clinical parameters and the system performs background statistical analysis to identify and plot gene expression correlates of sub-phenotype definitions. The PEAX software implementation uses open-source frameworks and source code is made available for download.

Also in our session, **Diggans *et al.*** describe a translational bioinformatics study identify and validate pre-operative mRNA based diagnostic test for V600E DNA mutations in thyroid nodules. A machine learning approach was applied in the discovery phase to identify a predictive 128-gene linear support vector machine from a feature space 3,000 transcripts measured from 716 thyroid fine needle aspirate biopsies (FNABs). The authors evaluate the 128-gene predictor against qPCR data in an independent test set and observe high positive and negative percent agreement with the qPCR test set. The results provide support for further clinical validation of the predictor and the potential for a first-of-a-kind diagnostic test for an unmet clinical need in thyroid cancer.

Efficient methods for inferring causal relationships across multiple scales of molecular traits are critical for modeling the complexity of biological systems. In our session, **Chang *et al.*** describe a novel method using Bayesian belief propagation for inferring the responses of perturbation events on molecular traits given a hypothesized graph structure. The method is not constrained by the conditional dependency arguments that limit the ability of statistical causal inference methods to resolve causal relationships within sets of graphical models that are Markov equivalent. The authors infer causal relationships from synthetic microarray and RNA sequencing data, and also apply their method to infer causality in real metabolic network with v-structure and feedback loop. Their approach is found to recapitulate the causal structure and recover the feedback loop given only steady-state data.

Accurate detection and modeling of tumor heterogeneity is a central challenge in understanding tumorigenesis and individual patient tumor characteristics. In our session, **Sengupta *et al.*** present a novel approach for modeling tumor heterogeneity (TH) using next-generation sequencing (NGS) data. The authors take a Bayesian approach that extends the Indian buffet process (IBP) to define a class of nonparametric models. Instead of partitioning somatic mutations into non-overlapping clusters with similar cellular prevalences, the authors do not assume somatic mutations with similar cellular prevalence

must be from the same subclone and allow overlapping mutations shared across subclones. The authors argue that this representation is closer to the underlying theory of phylogenetic clonal expansion, where somatic mutations occurred in parent subclones should be shared across the parent and child subclones. Their method yields posterior probabilities of the number, genotypes, and proportions of subclones in a tumor sample, thereby providing point estimates as well as variabilities of the estimates for each subclone. The method is implemented in a software package called BayClone that is made available for download.

Technological advances and increased public availability of data offer new opportunities to gain insights into the complexity of the eukaryotic transcriptome. Alternative cleavage of 3' UTRs has numerous functional consequences pertaining to the stability, transport, and translocation of transcripts. 3' UTR cleavages site analysis is also important clinically, particularly in cancer, where proto-oncogene can be activated by mRNA isoforms having shorter cleaved 3' UTRs. Thus both biological investigations and clinical applications benefit from more accurate methods for cleavage site analysis from transcriptional profiling data. In our session, **Birol et al.** describe KLEAT, a novel analysis tool that uses de novo assembly of RNA-sequencing data to search for and prioritize cleavage sites in poly(A) tails. The authors apply KLEAT to RNA-sequence data from ENCODE cell lines for which RNA-PET libraries are also available to compare predicted and actual 3' poly(A) signatures. The authors find that KLEAT exhibits > 90% positive predictive value when there are at least three RNA-sequencing reads supporting a poly(A) using the validation criteria of a minimum of three RNA-PET reads mapping within 100 nucleotides. The KLEAT software may accelerate biological and clinical applications of 3' UTR cleavage site analysis by enabling accurate analysis from more standard RNA-sequencing data and obviating the need for specialized wet lab techniques or sequencing libraries.

Though thousands of genes are implicated as underlying factors of disease it remains challenging to identify highest-value targets for novel drug development. In our session, **Gao et al.** address the question whether genes affected by strong genetic or environmental effects present better proxy therapeutic drug targets. To address this question, the authors propose a modeling approach that recovers both regulatory networks and estimates of environmental and genetic effects on gene expression. They apply their method to a gene expression data measured from blood samples from monozygotic and dizygotic twins and use the Connectivity Map database to assess whether genetic or environmental effects are more informative of gene's competency as a proxy target. The study findings suggest that a gene with strong genetic effects is more likely to act as a proxy target than a gene with strong environmental effects. This raises the intriguing hypothesis that diversity of a gene's expression across a genetically diverse population that makes it a suitable proxy rather than its sensitivity to environmental effects.

Finally, **Fan-Minogue et al.** evaluate the effectiveness of the differential expression (DE), disease-associated single nucleotide polymorphisms (SNPs) and combination of the two in recovering known therapeutic targets across 56 human diseases. They find that the performance of each feature varies across diseases and generally the features have more recovery power than predictive power. The systematic study results offer compelling evidence that the combination of the two features has more predictive power than each feature alone.

References

- 1 Topol, E. J. Individualized medicine from prewomb to tomb. *Cell* **157**, 241-253, doi:10.1016/j.cell.2014.02.012 (2014).
- 2 Hood, L. & Price, N. D. Demystifying disease, democratizing health care. *Science translational medicine* **6**, 225ed225, doi:10.1126/scitranslmed.3008665 (2014).
- 3 Schloissnig, S. *et al.* Genomic variation landscape of the human gut microbiome. *Nature* **493**, 45-50, doi:10.1038/nature11711 (2013).
- 4 Gaudilliere, B. *et al.* Clinical recovery from surgery correlates with single-cell immune signatures. *Science translational medicine* **6**, 255ra131, doi:10.1126/scitranslmed.3009701 (2014).
- 5 Kuehnbaum, N. L., Gillen, J. B., Gibala, M. J. & Britz-McKibbin, P. Personalized metabolomics for predicting glucose tolerance changes in sedentary women after high-intensity interval training. *Scientific reports* **4**, 6166, doi:10.1038/srep06166 (2014).
- 6 Chen, R. *et al.* Personal omics profiling reveals dynamic molecular and medical phenotypes. *Cell* **148**, 1293-1307, doi:10.1016/j.cell.2012.02.009 (2012).
- 7 Stanberry, L. *et al.* Integrative analysis of longitudinal metabolomics data from a personal multi-omics profile. *Metabolites* **3**, 741-760, doi:10.3390/metabo3030741 (2013).
- 8 Li-Pook-Than, J. & Snyder, M. iPOP goes the world: integrated personalized Omics profiling and the road toward improved health care. *Chemistry & biology* **20**, 660-666, doi:10.1016/j.chembiol.2013.05.001 (2013).
- 9 Newton, K. M. *et al.* Validation of electronic medical record-based phenotyping algorithms: results and lessons learned from the eMERGE network. *Journal of the American Medical Informatics Association : JAMIA* **20**, e147-154, doi:10.1136/amiajnl-2012-000896 (2013).
- 10 Ozdemir, A. T. & Barshan, B. Detecting falls with wearable sensors using machine learning techniques. *Sensors* **14**, 10691-10708, doi:10.3390/s140610691 (2014).
- 11 Viventi, J. *et al.* A conformal, bio-interfaced class of silicon electronics for mapping cardiac electrophysiology. *Science translational medicine* **2**, 24ra22, doi:10.1126/scitranslmed.3000738 (2010).
- 12 Barr, A. in *Wall Street Journal* (New York, N.Y., 2014).
- 13 Lappalainen, T. *et al.* Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501**, 506-511, doi:10.1038/nature12531 (2013).
- 14 Parts, L. *et al.* Extent, causes, and consequences of small RNA expression variation in human adipose tissue. *PLoS genetics* **8**, e1002704, doi:10.1371/journal.pgen.1002704 (2012).
- 15 Johansson, A. *et al.* Identification of genetic variants influencing the human plasma proteome. *Proceedings of the National Academy of Sciences of the United States of America* **110**, 4673-4678, doi:10.1073/pnas.1217238110 (2013).
- 16 Kettunen, J. *et al.* Genome-wide association study identifies multiple loci influencing human serum metabolite levels. *Nature genetics* **44**, 269-276, doi:10.1038/ng.1073 (2012).
- 17 Quon, G., Lippert, C., Heckerman, D. & Listgarten, J. Patterns of methylation heritability in a genome-wide analysis of four brain regions. *Nucleic acids research* **41**, 2095-2104, doi:10.1093/nar/gks1449 (2013).
- 18 McRae, A. F. *et al.* Contribution of genetic variation to transgenerational inheritance of DNA methylation. *Genome biology* **15**, R73, doi:10.1186/gb-2014-15-5-r73 (2014).
- 19 Gagneur, J. *et al.* Genotype-environment interactions reveal causal pathways that mediate genetic effects on phenotype. *PLoS genetics* **9**, e1003803, doi:10.1371/journal.pgen.1003803 (2013).
- 20 Fall, T. *et al.* The role of adiposity in cardiometabolic traits: a Mendelian randomization analysis. *PLoS medicine* **10**, e1001474, doi:10.1371/journal.pmed.1001474 (2013).

KLEAT: CLEAVAGE SITE ANALYSIS OF TRANSCRIPTOMES*

INANÇ BIROL, ANTHONY RAYMOND, READMAN CHIU, KA MING NIP, SHAUN D JACKMAN, MAAYAN KREITZMAN, T RODERICK DOCKING, CATHERINE A ENNIS, A GORDON ROBERTSON, ALY KARSAN

Canada's Michael Smith Genome Sciences Centre, British Columbia Cancer Agency

Vancouver, BC, V5Z 4S6, Canada

Email: ibirol@bcgsc.ca

In eukaryotic cells, alternative cleavage of 3' untranslated regions (UTRs) can affect transcript stability, transport and translation. For polyadenylated (poly(A)) transcripts, cleavage sites can be characterized with short-read sequencing using specialized library construction methods. However, for large-scale cohort studies as well as for clinical sequencing applications, it is desirable to characterize such events using RNA-seq data, as the latter are already widely applied to identify other relevant information, such as mutations, alternative splicing and chimeric transcripts. Here we describe KLEAT, an analysis tool that uses *de novo* assembly of RNA-seq data to characterize cleavage sites on 3' UTRs. We demonstrate the performance of KLEAT on three cell line RNA-seq libraries constructed and sequenced by the ENCODE project, and assembled using Trans-ABYSS. Validating the KLEAT predictions with matched ENCODE RNA-seq and RNA-PET libraries, we show that the tool has over 90% positive predictive value when there are at least three RNA-seq reads supporting a poly(A) tail and requiring at least three RNA-PET reads mapping within 100 nucleotides as validation. We also compare the performance of KLEAT with other popular RNA-seq analysis pipelines that reconstruct 3' UTR ends, and show that it performs favourably, based on an ROC-like curve.

* This work is supported by Canadian Institutes of Health Research, Genome Canada, Genome British Columbia, British Columbia Cancer Foundation and the National Institutes of Health under Award Number R01HG007182. The content of this work is solely the responsibility of the authors and does not necessarily represent the official views of any of our funding agencies.

1. Introduction

The section of an mRNA transcript that is translated into protein sequence is flanked by 5' and 3' untranslated regions (UTRs). These UTRs play a number of important biological roles. The 3' end of an mRNA molecule (the 3' UTR) helps to regulate its stability and localization, hence the amount of corresponding protein that is produced [1-4]. Over 50% of human genes produce two or more transcript isoforms via alternative polyadenylation (APA) of the 3' UTRs [5]. APA is recognized as playing a role in cancer biology [6-9].

A number of direct sequencing protocols have been developed for characterizing polyadenylated (poly(A)) tails of 3' UTRs and APA [9-15]. A cost-effective alternative to these direct sequencing protocols would be high throughput transcriptome sequencing (RNA-seq) [16], coupled with a validated bioinformatics pipeline to detect 3' UTR cleavage sites (CS).

RNA-seq is a central data type for many studies, including the ENCODE (ENCyclopedia Of DNA Elements) project, whose goal is to identify all functional elements in the human genome sequence [17]. Using various sequencing protocols, an ENCODE study [18] identified over 100,000 transcripts, about 60,000 of which were protein coding, and reported that transcript expression levels span six orders of magnitude. This is remarkable, as it speaks to the sensitivity of the RNA-seq technology. The lower range of the reported expression levels of 10^{-2} RPKM in that study implies that RNA-seq can detect a transcript expressed by 1 in 100 cells [16]. This resolution of RNA-seq data can be leveraged to identify 3' UTR ends of transcripts. An earlier study [19] inferred 3' UTR switching using sudden changes in expression profiles near cleavage sites, but did not utilize the direct evidence of observed poly(A) sequences.

In this report, we introduce KLEAT, a post-processing tool for characterizing 3' UTRs in assembled RNA-seq data through direct observation of poly(A) tails. While we developed KLEAT as an extension to the Trans-ABySS analysis pipeline [20, 21], it can also accept contigs from other transcriptome assembly tools, as we demonstrate below. It analyses the structures of assembled transcripts for poly(A) tails, filters 3' UTR cleavage site (CS) candidates using several evidence types within RNA-seq reads, and gathers and reports metrics that can be used in downstream post-processing, such as for filtering calls by their levels of read support.

2. Methods

The key technology KLEAT uses in detecting 3' UTR ends is *de novo* transcriptome assemblies. Compared to genome assembly, a successful transcriptome assembly has to address some particular challenges. These include robust assembly of transcripts from a wide range of transcript abundance levels, and resolution of transcripts from alternative isoforms and gene families. There are several specialized *de novo* assembly tools, including Trans-ABySS [21], Trinity [22] and Oases [23] that successfully address these challenges.

The KLEAT pipeline (Figure 1) uses Trans-ABySS by default. Using the raw reads and assembled contigs, it performs two levels of alignments in parallel: (1) reads to contigs; and (2) contigs to reference genome. It processes these alignment results to identify *tail*, *bridge*, and *link* evidence (Figure 2), and collates the evidence to predict cleavage sites.

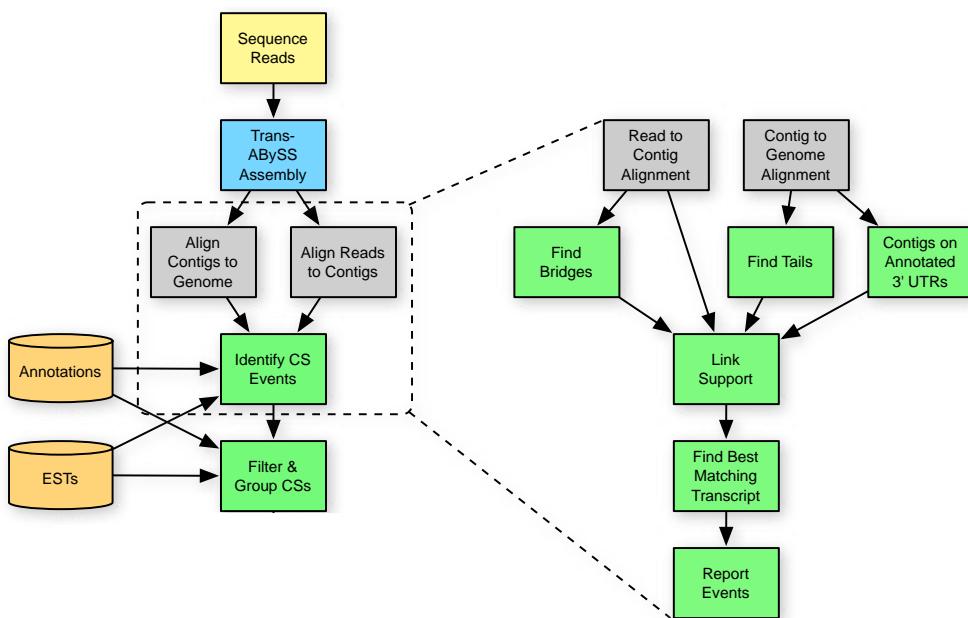


Fig. 1. Flowchart of the KLEAT pipeline. Two shades of yellow flowchart elements designate raw and external input to the pipeline; blue and grey indicate existing internal and external tools, respectively; green denotes new tools developed specifically for KLEAT.

2.1. Tail

Contig sequences that end in a poly(A) stretch represent high-confidence candidates. We filter these candidates to identify true poly(A) tails by aligning the flagged contigs to a reference genome. Accounting for the direction of transcription, we classify contigs with untemplated poly(A) sequence (a stretch of poly(A) sequence not observed in the reference genome) at their 3' ends as *tail* type events. For a transcript that is sufficiently abundant, this would be the expected default event type.

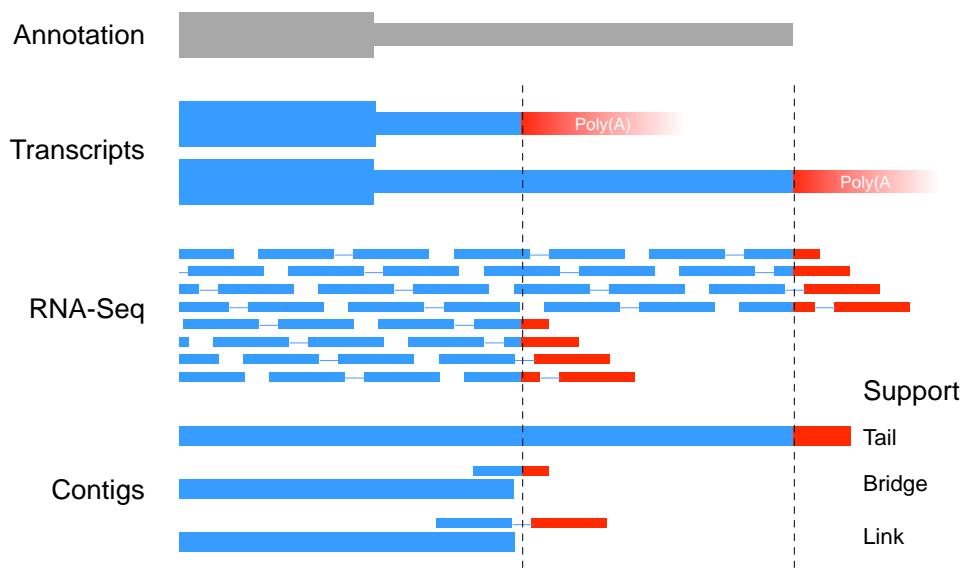


Fig. 2. Three types of support for detecting cleavage sites using RNA-seq data. The gene annotation (grey) indicates a single 3' UTR isoform, while the sample expresses two APA (red) variants. RNA-seq data capture the presence of these two alternatives with reads that end in poly(A) sequence (red). Contigs with supporting evidence have either a poly(A) "tail", an overhanging read that is "bridge" to a poly(A) sequence, or a read that has a "link" to a pair with poly(A) sequence.

2.2. Bridge

Expressed alternative long and short 3' UTRs present alternative paths for contig extensions during *de novo* assembly. In such cases, if the graph indicates a branch that does not extend to a poly(A) sequence, the alternative branch with the poly(A) sequence is removed by an assembly quality assurance stage within ABySS, in an operation called trimming [24]. While this is a desirable behaviour in general, it creates a particular challenge in assembling contigs with poly(A) tails. However, the information removed during this step can be recovered later by aligning reads to contigs, then assessing the sequences of partial read alignments at the contig edges. When an overhanging read alignment represents an untemplated poly(A) sequence, we infer the presence of a cleavage site. We call such cases *bridge* type evidence.

2.3. Link

The sequence complexity of 3' UTRs may drop substantially near their 3' ends, where the region is dominated by AU-rich sequence [25], and this may affect contig extensions due to loss of specificity of read-to-read overlaps. When this happens near a cleavage site, the corresponding contig may fail to present a tail type evidence, and may terminate extension before a read with a poly(A) tail can bridge it to beyond the cleavage site. However, the 3' UTR end may be within a typical sequencing fragment length, and if we identify read pairs linking the end of a contig to an untemplated poly(A) sequence, we classify the corresponding contig as having *link* type evidence.

Some cleavage sites may have supporting evidence from a combination of these evidence types, and even multiple observations from the same evidence type. The latter is partly due to the fuzzy definition of a cleavage site, where the end of a 3' UTR may fluctuate by about ± 30 nucleotides (nt) between mRNA molecules of the same transcript species [26]. Accordingly, we cluster cleavage sites predicted from multiple contigs if they fall within a certain window, label them as being representatives of the same cleavage site, and tally the counts presented by each evidence type in a given cluster to score the strength of our prediction.

We note that read-to-contig alignments performed in the pipeline have unique requirements. Although we demonstrate our results using BWA [27] – an established general-purpose sequence alignment tool – we recognize that detecting cleavage sites is most effective when reads are aligned to contigs with a tool that is capable of handling alignments with overhangs, that is, when a read aligns to the end of a contig with its sequence extending beyond the boundary(ies) of the contig. Because many high-throughput general-purpose sequence alignment tools are developed by the explicit or implicit assumption of a reference sequence that is composed of a small number of long contigs (i.e. chromosomes), they may suffer from accuracy and performance issues when the reference sequence is in many short pieces (as in an assembled transcriptome). Alignments near contig or scaffold edges are particularly challenging for general-purpose alignment software. We call this the *edge effect*, and address it by an FM-index based aligner within the ABySS genome assembly package [24], as an alternative. This aligner weighs edge alignments that are shorter but with fewer mismatches more favourably than longer alignments with more mismatches to provide local alignments.

KLEAT compares putative cleavage sites to annotation and EST databases to characterize and annotate them with other supporting observations, if any. Again using the annotation and EST databases, KLEAT groups, classifies and filters the putative events.

For method validation, we used the RNA-PET protocol [10] as our gold standard. We quantified the concordance between the “putative” (KLEAT) and “real” (RNA-PET) cleavage sites (CS) using the following definitions:

- false positive* | A called CS not within a certain window of an RNA-PET cluster
- true positive* | An RNA-PET cluster with at least one called CS in a window
- false negative* | An RNA-PET cluster without a called CS in a window
- true negative* | Cannot be defined

One way to gauge the performance of a detection tool is to study its receiver-operator-characteristic (ROC) curve, where a stringency parameter is varied to plot the true positive rate, TPR (the ratio of the true positive count to the total number of events) versus the false positive rate, FPR (the ratio of the false positive count to the total number of negatives). Note that, because *true negative* is undefined in this context, FPR cannot be defined either. A common practice in such cases is to use the false discovery rate, FDR (defined as the ratio of the false positive count to the total number of calls) as surrogate for the FPR.

3. Results and Discussion

For validating our method, we employed experimental data collected by the ENCODE project [18]. The ENCODE consortium characterized transcript ends using the RNA-PET protocol [10], and generated RNA-seq data for some of the same samples. We considered three cell lines (H1-hESC, A549 and MCF-7) for which RNA-PET and RNA-seq data were available (Table 1).

For RNA-PET coverage data, we applied an expression level threshold (default, 3 reads), and clustered observations that occurred within a certain distance (default, 100 nt) of each other. We used the resulting clusters as ‘true’ events.

We used Trans-ABySS v1.4.7 [20, 21] to assemble the RNA-seq reads; aligned the assembled transcripts to the human genome reference hg19 using GMAP 2012-12-20 [28]; and aligned reads back to the assembled contigs using BWA-SW v0.6.2-r126 [27]. We processed the results to identify the tail, bridge and link evidence, and clustered the CS calls. We also used Trinity Release 2013-02-25 [22] for RNA-seq assembly, and used the same pipeline to identify CS calls. Table 2 summarizes the performance of KLEAT on the assembled transcriptomes (with Trans-ABySS and Trinity contigs) for the three cell lines.

Table 1. Three ENCODE cell lines used for validation. All libraries were prepared using the long polyA+ RNA fraction protocol. RNA samples represent the whole cell transcriptome. RNA-PET reads were sequenced at 2x36 nt. H1-hESC cell line RNA-seq reads were sequenced at 2x78 nt, and A549 and MCF-7 cell lines at 2x76 nt. All data were generated and made publicly available by the ENCODE project [18].

Cell Line	# RNA-PET Read Pairs (million)	# RNA-seq Read Pairs (million)
H1-hESC	50.2	78.3
A549	181.7	70.5
MCF-7	174.0	87.4

Table 2. Summary statistics on *de novo* assembly of the RNA-seq data and KLEAT calls. Assembly figures are for contigs longer than 500 nt in length. The total number of cleavage sites called by KLEAT, and the average number of alternative polyadenylation sites per gene, are shown in the last two columns. Two sub-columns for the number of cleavage sites and APA per gene represent total and true positive (TP) calls, at a support threshold of three reads.

Cell Line	Assembler	# Contigs	N50 (nt)	(Mnt)	Total	# Cleavage Sites	APA per gene	
					Reconstruction	All	TP	All
H1-hESC	Trans-ABySS	879,277	1,049	309.1	11,975	8,998	2.54	2.49
	Trinity	159,627	2,250	188.3	9,896	7,565	2.37	2.30
A549	Trans-ABySS	788,688	1,352	299.9	13,984	12,940	2.83	2.80
	Trinity	149,880	2,791	197.4	12,249	11,369	2.62	2.57
MCF-7	Trans-ABySS	1,002,984	1,171	391.5	15,240	13,308	3.09	3.05
	Trinity	237,875	2,925	323.6	12,845	11,387	2.82	2.76

Using contigs from either transcriptome assembly tool as input, we observed the number of cleavage site calls to be the lowest for H1-hESC, and highest for MCF-7. The number of APA sites per gene also follows this pattern, ranging from roughly 2.5 to 3.0 APA isoforms, on average. Interestingly, while the fraction of true positive cleavage site calls range from 75 to 93%, the average number of APA isoforms per gene is insensitive to filtering for true positives.

We compared the use of contigs from *de novo* transcriptome assembly to detect cleavage sites, with the use of transcripts reconstructed by aligning the reads to a reference genome. We ran the Cufflinks pipeline v2.1.1 [29] on the same dataset, and used the reconstructed 3' UTR ends of predicted transcripts to measure its accuracy in detecting poly(A) tails. Cufflinks takes RNA-seq read alignments to a reference genome as input, and builds those alignments into a parsimonious set of transcripts with or without annotation support. We ran the pipeline with annotation support, allowing for transcript discovery.

Figure 3 depicts the ROC curves for these two paradigms for the three cell lines used. Curves closer to the top-left corner indicate better performance. These results suggest that to identify 3' UTR cleavage sites an assembly-first approach (using either Trans-ABySS or Trinity) may be preferred over the alignment-first approach implemented in the Cufflinks pipeline. We note that the Trans-ABySS and Trinity results are similar, while the Trans-ABySS assemblies perform marginally yet consistently better. This may be due to the difference between the total reconstruction figures of the two tools (Table 2).

Although the magnitudes of the reported TPR figures are low (<10%), we note that this reflects our simple but relaxed definition of the ground truth, with no distinction between 3' and 5' UTR ends, and applying none of the filtering suggested in the ENCODE report. These choices would inflate the denominator of TPR, and lead to underestimates in the reported figures. However, neither of these would change the relative performance of the analysis tools we present.

We also compared the concordance between these three sets of results (Figure 4). Our analysis indicates that Trans-ABySS and Trinity contigs are largely concordant in their reconstruction of cleavage sites, identifying roughly 9,000 to 13,000 and 7,000 to 11,000 true positive calls, respectively (for an RNA-PET threshold of 3 reads) and agreeing on about 80 to 90% of the calls.

In contrast, Cufflinks would identify 7,000 to 10,000 true positive calls with the same RNA-PET threshold, agreeing with the assembly-first results only about 25 to 30% of the time, meaning that it identifies a smaller yet largely distinct set of events.

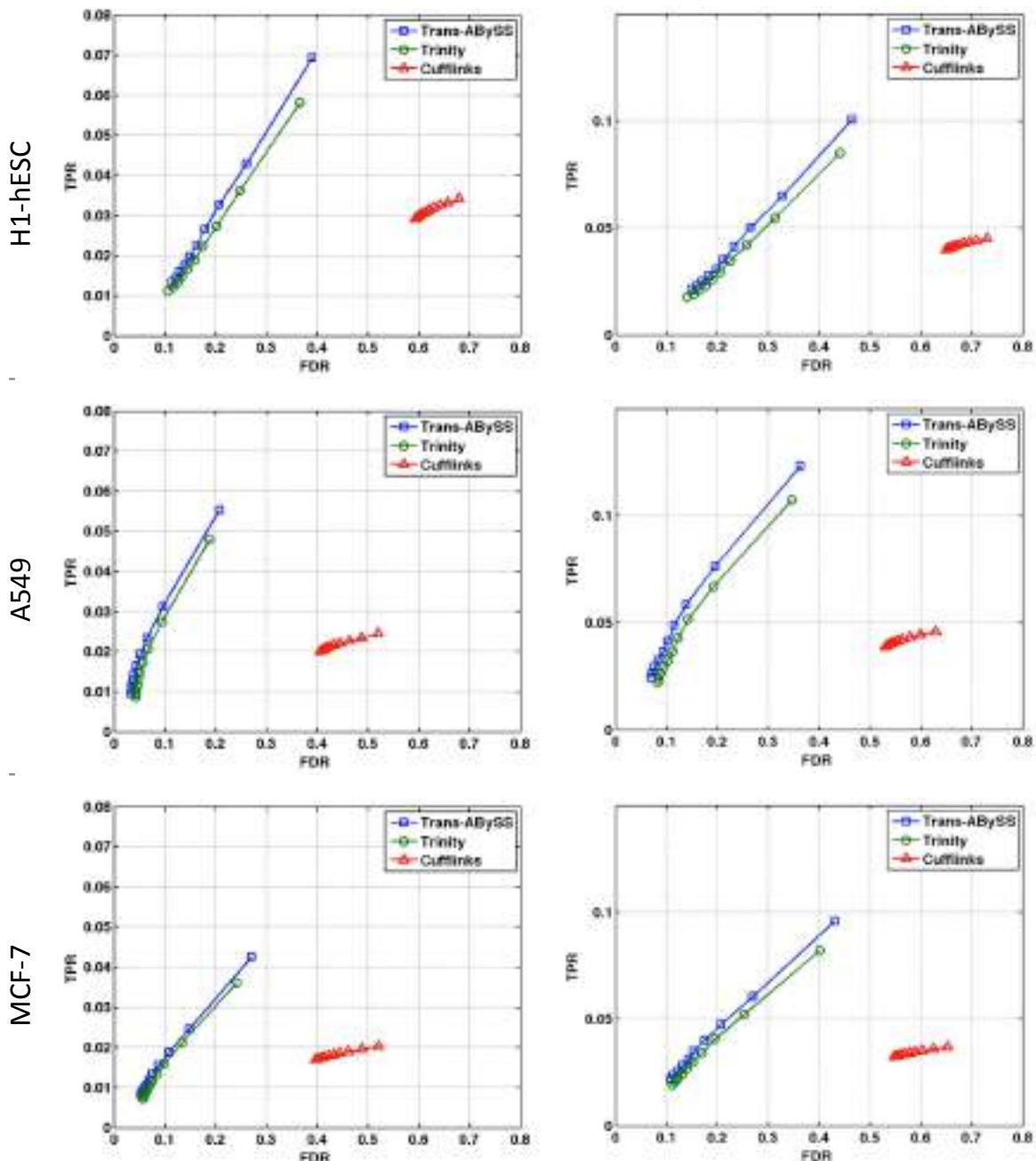


Fig. 1. Performance of KLEAT on three ENCODE cell lines. Curves represent the true positive rate (TPR) as a function of the false discovery rate (FDR), for KLEAT with Trans-ABYSS (blue) and KLEAT with Trinity (green) and Cufflinks (red). RNA-PET evidence is considered as the gold standard at two support cutoffs: left column: 3 or more, right column: 10 or more RNA-PET reads.

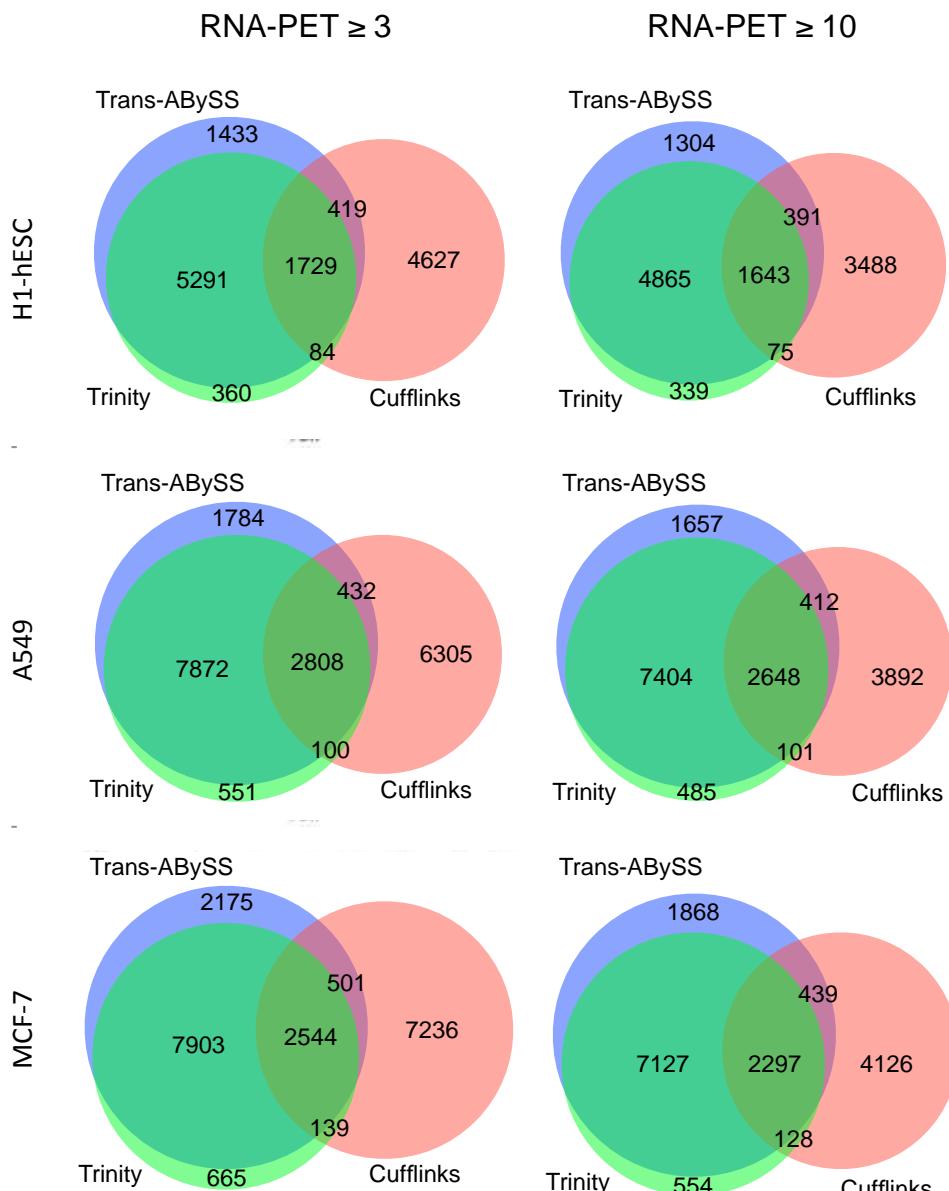


Fig. 4. Concordance between three methods. Blue, green and red sets indicate events detected by KLEAT/Trans-ABYSS, KLEAT/Trinity and Cufflinks, respectively, that are supported by at least 3 or 10 RNA-PET reads, as indicated.

We note that in this dataset the Cufflinks pipeline is more sensitive for detecting weakly expressed transcripts. This observation is supported by the performance metrics of the three tools when we increased the RNA-PET threshold from three to 10 reads. At the increased threshold, KLEAT with Trans-ABYSS and Trinity loses about 6 to 10% of its true positive calls, while Cufflinks loses about 10 to 33% of such calls.

Further supporting this observation, a coverage histogram of the expressed genes in the A549 cell line, as detected by Cufflinks, is depicted in Figure 5. Here the two x-axes represent expression levels in units of average coverage and FPKM on logarithmic scales. Cufflinks reports a major peak for transcripts represented at 1- to 10-fold average coverage, also reconstructing a

large number of transcripts with less than 1-fold coverage, some of which would report cleavage sites also observed by RNA-PET data. In contrast, *de novo* assembly methods would typically reconstruct transcripts over 10-fold coverage. This apparent difference in target expression levels for transcript reconstruction explains the lack of concordance between KLEAT (in conjunction with Trans-ABySS or Trinity) and Cufflinks, as reported in Figures 3 and 4.

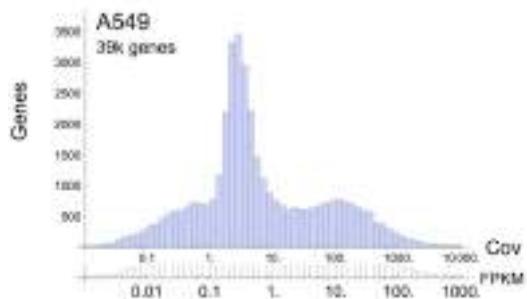


Fig. 5. Gene level coverage histogram for the A549 cell line RNA-seq data, as reported by Cufflinks. The histogram is presented with two logarithmic scales on the x-axis, average coverage (Cov) and FPKM, to show the correspondence between the two units for the sequencing depth of the experimental data.

4. Conclusions

In this study, we introduced KLEAT as an analysis tool for detecting 3' UTR cleavage sites using *de novo* assembled RNA-seq reads. We validated our method using data from the ENCODE project [30]. We measured the accuracy of KLEAT using two transcriptome assembly tools (Trans-ABySS [20, 21] and Trinity [22]), and compared its performance to results from an alignment-based analysis tool (Cufflinks [29], the method of choice in the ENCODE study).

Our results demonstrate that one can reliably detect around 10,000 poly(A) tails per sample using RNA-seq data, at a sequencing depth of 70 million read pairs. The depth of sequencing data will certainly affect the number of transcripts observed, hence the number of poly(A) tails detected. Therefore, although we suggest that detecting on the order of 10,000 features will already provide important biological insights for highly expressed transcripts, if one wants to observe more features, one approach might be to sequence a library to greater depth (albeit with diminishing returns). With sequencing throughput on the Illumina platform pushing beyond 250 million read pairs per lane, experimental design (such as pooling multiple samples per lane) reflects a balance between cost and value, and that balance is determined by the particular experimental goals and the budget of a study.

We also note that overlapping sense/anti-sense gene annotations can potentially confuse the poly(A) tail calls. Using a strand-specific RNA-seq protocol should help mitigate this issue.

Surveying 15 human cell lines, the ENCODE study reports a total of 128,824 poly(A) sites mapping within annotated Gencode transcripts [30]. This observation puts the average number of polyadenylation sites in this dataset to 2.5 per gene. Interestingly, before this landmark publication, the APA multiplicity was estimated to be around 1.1. Our analysis of RNA-seq data from three ENCODE cell lines (APA per gene statistics in Table 2) are in agreement with this ENCODE estimate.

There is a growing appreciation of 3' UTRs, their molecular assembly, mechanistic roles and variants [9-15]. Many of these studies developed novel wet lab techniques to build specialized sequencing libraries, and applied them to interrogate a particular biological condition.

KLEAT offers an alternative analysis method to characterize 3' UTRs and APA from RNA-seq data at a nucleotide scale resolution. We anticipate the tool to be an enabling technology for many applications, including large-scale disease studies and clinical genomics, and to provide added value to the large volume of sequencing data already generated using the data type. KLEAT is available at www.bcgsc.ca/platform/bioinfo/software, and is offered free for academic use.

Acknowledgments

The authors thank, Canadian Institutes of Health Research, Genome Canada, Genome British Columbia and British Columbia Cancer Foundation for their generous support. The work is also partially funded by the National Institutes of Health under Award Number R01HG007182. The content of this work is solely the responsibility of the authors and does not necessarily represent the official views of any of our funding agencies. We also thank the ENCODE project for enabling the validation work presented by making their datasets publicly available.

References

- [1] J. D. Keene, "RNA regulons: coordination of post-transcriptional events," *Nature reviews. Genetics*, vol. 8, no. 7, pp. 533-43, Jul, 2007.
- [2] K. K. To, Z. Zhan, T. Litman, and S. E. Bates, "Regulation of ABCG2 expression at the 3' untranslated region of its mRNA through modulation of transcript stability and protein translation by a putative microRNA in the S1 colon cancer cell line," *Molecular and cellular biology*, vol. 28, no. 17, pp. 5147-61, Sep, 2008.
- [3] Z. Ji, J. Y. Lee, Z. Pan, B. Jiang, and B. Tian, "Progressive lengthening of 3' untranslated regions of mRNAs by alternative polyadenylation during mouse embryonic development," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 106, no. 17, pp. 7028-33, Apr 28, 2009.
- [4] S. Muller, L. Rycak, F. Afonso-Grunz, P. Winter, A. M. Zawada, E. Damrath, J. Scheider, J. Schmeh, I. Koch, G. Kahl, and B. Rotter, "APADB: a database for alternative polyadenylation and microRNA regulation events," *Database (Oxford)*, vol. 2014, 2014.
- [5] B. Tian, J. Hu, H. Zhang, and C. S. Lutz, "A large-scale analysis of mRNA polyadenylation of human and mouse genes," *Nucleic Acids Res*, vol. 33, no. 1, pp. 201-12, 2005.
- [6] P. Singh, T. L. Alley, S. M. Wright, S. Kamdar, W. Schott, R. Y. Wilpan, K. D. Mills, and J. H. Gruber, "Global changes in processing of mRNA 3' untranslated regions characterize clinically distinct cancer subtypes," *Cancer Res*, vol. 69, no. 24, pp. 9422-30, Dec, 2009.
- [7] Y. Hamaya, S. Kuriyama, T. Takai, K. Yoshida, T. Yamada, M. Sugimoto, S. Osawa, K. Sugimoto, H. Miyajima, and S. Kanaoka, "A distinct expression pattern of the long 3'-untranslated region dicer mRNA and its implications for posttranscriptional regulation in colorectal cancer," *Clin Transl Gastroenterol*, vol. 3, pp. e17, 2012.
- [8] E. Devany, X. Zhang, J. Y. Park, B. Tian, and F. E. Kleiman, "Positive and negative feedback loops in the p53 and mRNA 3' processing pathways," *Proc Natl Acad Sci U S A*, vol. 110, no. 9, pp. 3351-6, Feb, 2013.
- [9] Y. Fu, Y. Sun, Y. Li, J. Li, X. Rao, C. Chen, and A. Xu, "Differential genome-wide profiling of tandem 3' UTRs among human breast cancer and normal cells by high-throughput sequencing," *Genome Res*, vol. 21, no. 5, pp. 741-7, May, 2011.

- [10] P. Ng, C. L. Wei, W. K. Sung, K. P. Chiu, L. Lipovich, C. C. Ang, S. Gupta, A. Shahab, A. Ridwan, C. H. Wong, E. T. Liu, and Y. Ruan, "Gene identification signature (GIS) analysis for transcriptome characterization and genome annotation," *Nat Methods*, vol. 2, no. 2, pp. 105-11, Feb, 2005.
- [11] Y. Ruan, H. S. Ooi, S. W. Choo, K. P. Chiu, X. D. Zhao, K. G. Srinivasan, F. Yao, C. Y. Choo, J. Liu, P. Ariyaratne, W. G. Bin, V. A. Kuznetsov, A. Shahab, W. K. Sung, G. Bourque, N. Palanisamy, and C. L. Wei, "Fusion transcripts and transcribed retrotransposed loci discovered through comprehensive transcriptome analysis using Paired-End diTags (PETs)," *Genome Res*, vol. 17, no. 6, pp. 828-38, Jun, 2007.
- [12] P. J. Shepard, E. A. Choi, J. Lu, L. A. Flanagan, K. J. Hertel, and Y. Shi, "Complex and dynamic landscape of RNA polyadenylation revealed by PAS-Seq," *RNA*, vol. 17, no. 4, pp. 761-72, Apr, 2011.
- [13] X. Ruan, and Y. Ruan, "Genome wide full-length transcript analysis using 5' and 3' paired-end-tag next generation sequencing (RNA-PET)," *Methods Mol Biol*, vol. 809, pp. 535-62, 2012.
- [14] D. Hafez, T. Ni, S. Mukherjee, J. Zhu, and U. Ohler, "Genome-wide identification and predictive modeling of tissue-specific alternative polyadenylation," *Bioinformatics*, vol. 29, no. 13, pp. i108-16, Jul, 2013.
- [15] R. Minasaki, D. Rudel, and C. R. Eckmann, "Increased sensitivity and accuracy of a single-stranded DNA splint-mediated ligation assay (sPAT) reveals poly(A) tail length dynamics of developmentally regulated mRNAs," *RNA Biol*, vol. 11, no. 2, Feb, 2014.
- [16] A. Mortazavi, B. A. Williams, K. McCue, L. Schaeffer, and B. Wold, "Mapping and quantifying mammalian transcriptomes by RNA-Seq," *Nat Methods*, vol. 5, no. 7, pp. 621-8, Jul, 2008.
- [17] ENCODE Project Consortium, "The ENCODE (ENCyclopedia Of DNA Elements) Project," *Science*, vol. 306, no. 5696, pp. 636-40, Oct, 2004.
- [18] S. Djebali, C. A. Davis, A. Merkel, A. Dobin, T. Lassmann, A. Mortazavi, A. Tanzer, J. Lagarde, W. Lin, F. Schlesinger, C. Xue, G. K. Marinov, J. Khatun, B. A. Williams, C. Zaleski, J. Rozowsky, M. Roder, F. Kokocinski, R. F. Abdelhamid, T. Alioto, I. Antoshechkin, M. T. Baer, N. S. Bar, P. Batut, K. Bell, I. Bell, S. Chakrabortty, X. Chen, J. Chrast, J. Curado, T. Derrien, J. Drenkow, E. Dumais, J. Dumais, R. Duttagupta, E. Falconnet, M. Fastuca, K. Fejes-Toth, P. Ferreira, S. Foissac, M. J. Fullwood, H. Gao, D. Gonzalez, A. Gordon, H. Gunawardena, C. Howald, S. Jha, R. Johnson, P. Kapranov, B. King, C. Kingswood, O. J. Luo, E. Park, K. Persaud, J. B. Preall, P. Ribeca, B. Risk, D. Robyr, M. Sammeth, L. Schaffer, L. H. See, A. Shahab, J. Skancke, A. M. Suzuki, H. Takahashi, H. Tilgner, D. Trout, N. Walters, H. Wang, J. Wrobel, Y. Yu, X. Ruan, Y. Hayashizaki, J. Harrow, M. Gerstein, T. Hubbard, A. Reymond, S. E. Antonarakis, G. Hannon, M. C. Giddings, Y. Ruan, B. Wold, P. Carninci, R. Guigo, and T. R. Gingeras, "Landscape of transcription in human cells," *Nature*, vol. 489, no. 7414, pp. 101-8, Sep 6, 2012.
- [19] W. Wang, Z. Wei, and H. Li, "A change-point model for identifying 3'UTR switching by next-generation RNA sequencing," *Bioinformatics*, vol. 30, no. 15, pp. 2162-70, Aug 1, 2014.
- [20] I. Birol, S. D. Jackman, C. B. Nielsen, J. Q. Qian, R. Varhol, G. Stazyk, R. D. Morin, Y. Zhao, M. Hirst, J. E. Schein, D. E. Horsman, J. M. Connors, R. D. Gascoyne, M. A. Marra,

- and S. J. M. Jones, “De novo transcriptome assembly with ABySS,” *Bioinformatics*, vol. 25, no. 21, pp. 2872-2877, Nov 1, 2009.
- [21] G. Robertson, J. Schein, R. Chiu, R. Corbett, M. Field, S. D. Jackman, K. Mungall, S. Lee, H. M. Okada, J. Q. Qian, M. Griffith, A. Raymond, N. Thiessen, T. Cezard, Y. S. Butterfield, R. Newsome, S. K. Chan, R. She, R. Varhol, B. Kamoh, A.-L. Prabhu, A. Tam, Y. Zhao, R. A. Moore, M. Hirst, M. A. Marra, S. J. M. Jones, P. A. Hoodless, and I. Birol, “De novo assembly and analysis of RNA-seq data,” *Nature Methods*, vol. 7, no. 11, pp. 909-U62, NOV 2010, 2010.
- [22] M. G. Grabherr, B. J. Haas, M. Yassour, J. Z. Levin, D. A. Thompson, I. Amit, X. Adiconis, L. Fan, R. Raychowdhury, Q. Zeng, Z. Chen, E. Mauceli, N. Hacohen, A. Gnirke, N. Rhind, F. di Palma, B. W. Birren, C. Nusbaum, K. Lindblad-Toh, N. Friedman, and A. Regev, “Full-length transcriptome assembly from RNA-Seq data without a reference genome,” *Nat Biotechnol*, vol. 29, no. 7, pp. 644-52, Jul, 2011.
- [23] M. H. Schulz, D. R. Zerbino, M. Vingron, and E. Birney, “Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels,” *Bioinformatics*, vol. 28, no. 8, pp. 1086-92, Apr 15, 2012.
- [24] J. T. Simpson, K. Wong, S. D. Jackman, J. E. Schein, S. J. M. Jones, and I. Birol, “ABySS: A parallel assembler for short read sequence data,” *Genome Research*, vol. 19, no. 6, pp. 1117-1123, Jun, 2009.
- [25] T. Zhang, V. Kruys, G. Huez, and C. Gueydan, “AU-rich element-mediated translational control: complexity and multiple activities of trans-activating factors,” *Biochem Soc Trans*, vol. 30, no. Pt 6, pp. 952-8, Nov, 2002.
- [26] E. Beaudoin, S. Freier, J. R. Wyatt, J. M. Claverie, and D. Gautheret, “Patterns of variant polyadenylation signal usage in human genes,” *Genome Res*, vol. 10, no. 7, pp. 1001-10, Jul, 2000.
- [27] H. Li, and R. Durbin, “Fast and accurate short read alignment with Burrows-Wheeler transform,” *Bioinformatics*, vol. 25, no. 14, pp. 1754-60, Jul 15, 2009.
- [28] T. D. Wu, and C. K. Watanabe, “GMAP: a genomic mapping and alignment program for mRNA and EST sequences,” *Bioinformatics*, vol. 21, no. 9, pp. 1859-75, May 1, 2005.
- [29] C. Trapnell, B. A. Williams, G. Pertea, A. Mortazavi, G. Kwan, M. J. van Baren, S. L. Salzberg, B. J. Wold, and L. Pachter, “Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation,” *Nat Biotechnol*, vol. 28, no. 5, pp. 511-5, May, 2010.
- [30] J. Harrow, A. Frankish, J. M. Gonzalez, E. Tapanari, M. Diekhans, F. Kokocinski, B. L. Aken, D. Barrell, A. Zadissa, S. Searle, I. Barnes, A. Bignell, V. Boychenko, T. Hunt, M. Kay, G. Mukherjee, J. Rajan, G. Despacio-Reyes, G. Saunders, C. Steward, R. Harte, M. Lin, C. Howald, A. Tanzer, T. Derrien, J. Chrast, N. Walters, S. Balasubramanian, B. Pei, M. Tress, J. M. Rodriguez, I. Ezkurdia, J. van Baren, M. Brent, D. Haussler, M. Kellis, A. Valencia, A. Reymond, M. Gerstein, R. Guigó, and T. J. Hubbard, “GENCODE: the reference human genome annotation for The ENCODE Project,” *Genome Res*, vol. 22, no. 9, pp. 1760-74, Sep, 2012.

CAUSAL INFERENCE IN BIOLOGY NETWORKS WITH INTEGRATED BELIEF PROPAGATION

RUI CHANG* and JONATHAN R KARR and ERIC E SCHADT*

*Department of Genetics and Genomic Sciences
Ichan School of Medicine, Mount Sinai
NY, NY 10029, USA
*E-mail: rui.r.chang@mssm.edu
eric.schadt@mssm.edu*

Inferring causal relationships among molecular and higher order phenotypes is a critical step in elucidating the complexity of living systems. Here we propose a novel method for inferring causality that is no longer constrained by the conditional dependency arguments that limit the ability of statistical causal inference methods to resolve causal relationships within sets of graphical models that are Markov equivalent. Our method utilizes Bayesian belief propagation to infer the responses of perturbation events on molecular traits given a hypothesized graph structure. A distance measure between the inferred response distribution and the observed data is defined to assess the 'fitness' of the hypothesized causal relationships. To test our algorithm, we infer causal relationships within equivalence classes of gene networks in which the form of the functional interactions that are possible are assumed to be nonlinear, given synthetic microarray and RNA sequencing data. We also apply our method to infer causality in real metabolic network with v-structure and feedback loop. We show that our method can recapitulate the causal structure and recover the feedback loop only from steady-state data which conventional method cannot.

Keywords: Causal inference, Top-down & bottom-up modeling, Predictive network modeling, Causal network learning

1. Introduction

One of the primary objectives of biomedical research is to elucidate the networks of molecular interactions underlying complex human phenotypes such as cancer and Alzheimer's disease. Over the past few years, a wave of advanced biotechnologies has swept over the life sciences landscape to enable more holistic profiling of biological systems. Whole genome sequencing, RNA sequencing, methylation profiling, and mass spectroscopy and NMR based metabolite and protein profiling technologies have been applied to a wide range of biological problems and have contributed to discoveries relating to the complex network of biochemical processes as well as to the reconstruction of gene networks underlying living systems¹ and common human diseases.^{2,3}

State-of-the-art statistical learning methods assume a Markov condition for gene network reconstructions as a way to reduce the complexity of the joint probability distribution that graphical network structures represent. It is well-known that algorithms based on Markov conditions can learn the correct causal relationships up to Markov equivalence given a large enough sample size.⁴ However, because the structures represented within a given equivalence class are statistically indistinguishable from one another, it is not possible to further resolve the correct causal relationships within a class without introducing perturbations that break the symmetry giving rise to this equivalence. Given the recovery of accurate mechanistic networks

is a crucial first-step to understanding the pathophysiology of human disease and how best to diagnose and treat it, methods to accurately infer causal relationships are critical.

To at least partially address this problem, we had previously proposed a Bayesian network learning framework¹ to improve causal inference within Markov equivalence classes by integrating genotypic data associated with molecular phenotypes (e.g., expression quantitative trait loci, or eQTL) and disease traits as an asymmetric and systematic source of perturbations. This approach has been effective in helping untangle the causality in gene networks given it leverages the propagation of structural asymmetry required to break Markov equivalence.

In this paper, we propose a statistical method that infers causal structures from multivariate datasets (e.g. gene expression data), reducing the need to integrate additional data such as eQTLs to accurately infer causal relationships. Our proposed method is complementary to more integrative approaches, given it will enable accurate causal inferences in cases in which integrative approaches are not possible or not well powered. For example, even when eQTL data are available, inferring a complete causal network structure remains challenging given the conventional top-down Bayesian network learning approach decomposes the joint probability function representing a given graph into the product of local conditional probabilities based on d-separation, so that the effect of causal information stemming from eQTL-controlled root nodes will not always effectively propagate through the entire network, leaving the issue of equivalence classes unresolved at distal local structures. This issue is exacerbated as the number of nodes in the network increases, given the super-exponential rate of growth of the space of possible networks and the fact that trans-acting eQTL effects (eQTL that act on genes that are distal to the physical location of the eQTL) are difficult to detect. Therefore, the development of methods to infer causality among structures in equivalence classes remains a fundamental objective for reconstructing accurate probabilistic causal network structures. To demonstrate the utility of our causal inference procedure, we apply it to simulated gene expression data that reflects the type of noise structures common in these high-throughput biological experiments as well as the biological relationships we seek to represent. In this context we demonstrate the ability to resolve Markov equivalent structures across a variety of general assumptions regarding the nature of gene-gene interactions.

2. Developing a Method to Infer Causality from Associative Data

A new class of methods referred to as Information Geometric Causal Inference (IGCI) methods⁵ defines classic measures of independence among variables in terms of orthogonal components of the joint probability distribution of these variables, as a way to leverage more information regarding the relationships among them. This approach stands in contrast to traditional approaches in which data informing on the relationships among variables of interest are extracted using only conditional independencies. In the IGCI methods, to infer whether X causes Y , orthogonality is computed between the conditional distribution $P_{Y|X}$ and P_X , which are then compared to the values computed for $P_{X|Y}$ and P_Y . If the relationship "X causes Y" is true, then the orthogonality metric is such that the causal hypothesis Y causes X is implausible. Remarkably, this asymmetry between cause and effect becomes particularly simple if X and Y are deterministically related. In the case of a nonlinear relationship between

X and Y , the nonlinearity in the function defining the relationship between the cause X and effect Y , i.e. $Y=f(X)$, can also be considered for causal inference in the presence of additive noise.⁶ The nonlinearity provides information on the underlying causal model and thus allows more aspects of the true causal mechanism to be identified. An alternative approach, referred to as functional causal modeling (a.k.a. structural causal or nonlinear structural equation modeling), involves a joint distribution function that along with a graph satisfies the causal Markov assumption.⁷ This functional form can also allow one to distinguish between $X \rightarrow Y$ and $X \leftarrow Y$.

Our proposed method sits squarely within the class of functional causal modeling approaches,⁷ where we utilize the inherent probabilistic inference capability of the Bayesian network framework to generate predictions of hypothesized child (response) nodes using the observed data of the hypothesized parent (causal) nodes. By defining a distance metric in probability space that assesses how well the predicted distribution of child nodes matches the distribution of their observed values, we can evaluate the different graphical structures within an equivalence class to determine the one best supported by the data. One key advantage of this modeling approach is that it enables the propagation of the effects of a parent node to child nodes that can be greater than a path length of one from the parent, thereby making it possible to infer causality in a chain of nodes or in a more complex network structures.

To describe our approach, we begin by reformatting the general posterior probability corresponding to any given graphical structure as defined in conventional Bayesian network approaches.⁸ Here we let X represent the vector of variables represented as nodes in the network; E is the evidence; D denotes the observed data; G is the graphical network structure to infer; and θ is a vector of model parameters. From this we can write the posterior probability as $P(G|D) = P(D|G)P(G)/P(D)$, where the marginal probability $P(D|G)$ can be expressed as an integral over the parameters, given a particular graphical structure G : $P(D|G) = \int_{\theta} P(D|G, \theta)P(\theta|G)d\theta$. Unlike the traditional Bayesian Dirichlet score, D in our case contains continuous values, and thus, the likelihood of the data is not derived from a multinomial distribution, but rather a continuous density function whose form is estimated using a kernel density estimation procedure. In addition, the parameter prior $P(\theta|G)$ does not follow a Dirichlet distribution but rather is either described by a set of non-parametric constraints in parameter space or is sampled from a uniform distribution defined in its range (the approach used herein). Given this, the above integral has no analytical solution. We will optimize the data likelihood by estimating θ using maximum-a-posteriori (MAP) estimation:

$$P(D|G) \cong P(D|G, \hat{\theta}) \quad (1)$$

where $\hat{\theta} = \text{argmax}_{\theta} \{P(D|G, \theta)P(\theta|G)\}$. We use a Monte Carlo sampling procedure to efficiently sample θ from $P(\theta|G)$, evaluating the likelihood for each parameter sample.

2.1. Deriving a Data Likelihood Score

To calculate and optimize the data likelihood $P(D|G, \theta)$, we incorporate belief propagation as a subroutine in the causal inference procedure to predict the marginal probabilities of all response variables given the observed data for the predictor variables for a given causal structure (G). In this instance, the marginal probability of X given G and the sampled parameter θ is calculated via belief propagation.⁹ In what follows we discuss how to use the Bayesian belief inference $P(X|E, G, \theta)$ to calculate the data likelihood $P(D|G, \theta)$ given in Eq. 1, where

E and D represent the observed data on the parent and child nodes, respectively. First, to avoid confusion, we introduce the notion X_b to describe the binary variable in the probability space mapped from the continuous variable $X \in R$. Second, we rescale the original observation data so that it falls in the interval $[0,1]$ (see discussion below). Third, we introduce a hidden variable H to fully specify the data likelihood as

$$P(D|G, \theta) = \int_H P(D|H)P(H|G, \theta) \quad (2)$$

Given G and θ , the soft evidence enters $P(X_b|E, G, \theta)$ as the observed, rescaled data D , which effectively "clamps" (or fixes) the marginal probability of the parent nodes, from which the marginal probabilities of the child nodes are predicted via belief propagation in the Bayesian network.⁹ These marginal probabilities are then used to define the hidden data H , which are used to construct the marginal data likelihood in Eq. 2. In probability space, the belief inference is deterministic, i.e. given a causal structure G , a specific set of parameters θ , and evidence E , $P(X_b|E, G, \theta)$ is uniquely determined. In Eq. 2, when $H=P(X_b|E, G, \theta)$, $P(H|G, \theta)=1$ and 0 otherwise, and as a result the data marginal likelihood in Eq. 2 can be re-written as

$$P(D|G, \theta) = P(D|H(G, \theta)) = P(D|P(X_b|E, G, \theta)) \quad (3)$$

The inner probability describes the marginal belief of the binary variable X_b in probability space to which the original continuous variable X has been mapped. This belief probability is a linear function between the child and parent marginal probabilities, multiplied by the conditional probabilities determined by sampling over the uniform distribution the parameters θ are assumed to follow:

$$P(X_b^{chd} = k) = \sum_i [P(X_b^{chd} = k | X_b^\pi = C_i)P(X_b^\pi = C_i)] \quad (4)$$

with X_b^π representing X_b for the parent node, X_b^{chd} for the child node, and where C_i represents the i -th configuration of the parent nodes. For example, given the hypothesis "gene A activates gene B", the marginal belief is calculated as

$$P(B) = P(B|A)P(A) + P(B|\bar{A})(1 - P(A)) = [P(B|A) - P(B|\bar{A})]P(A) + P(B|\bar{A}) \quad (5)$$

In this instance, the conditional probability distribution $\theta=\{P(B|A), P(B|\bar{A})\}$ is the parameter sampled from the uniform distribution on $[0,1]$. We note that this belief propagation can be considered as a linear regression of the form $P(B)=\beta P(A)+C$, where $\beta=(P(B|A) - P(B|\bar{A}))$ and $C=P(B|\bar{A})$. Given probability measures are constrained to be between 0 and 1, $\beta \in [-1, 1]$ and $C \in [0,1]$. These constraints, which follow naturally from probability theory, give rise to the asymmetric basis of our approach for causal inference (see section 2.4). We further note that in our approach we implicitly assume binary variables in probability space, where the belief probability of a binary variable is defined as the level of belief on that variable observed in its maximal state (i.e., $P(X_b = 1)$) or minimal state (i.e., $P(X_b = 0)$). When X is equal to its minimum (or maximum) value in the real valued space D , in probability space X_b is observed in its minimum (or maximum) state, and therefore, this observed sample will correspond to $Pr(X_b = 0) = 1$ (or $Pr(X_b = 1) = 0$) in probability space. As the value of X varies between its minimum and maximum values, the belief probability of the binary mapping of this variable will vary between $[0,1]$. The Bayesian interpretation of the belief probability allows us to compare the inferred belief probability to the real-valued observed data by implicitly assuming

that the original data D and the marginal probabilities of H are positively correlated, though the precise kinetics of this correlation is unknown. To make such a comparison, we rescale $D \in R$ to $D \in [0,1]$.

Since the exact function mapping between real values and their belief probabilities is unknown, we employ a non-parametric metric, i.e. Kullback-Liebler (KL) divergence, to compare the distribution of real observations to the distribution of predicted marginal probabilities. If the predicted and observed distribution of the child nodes match well, we can conclude that the predictions based on G and θ well reflect the observed data D , which results in a smaller value of the KL-divergence. To force the KL divergence to behave as a true probability measure, we make symmetry and normalization modifications to this function defined on D and H such that $k(D, H)=1-\exp[-(KL(D||H)+KL(H||D))/2]$. The data likelihood function in Eq. 3 can be defined by any normalized monotonic decreasing function on the kernel. For model selection, we set $S = -\log(K(D, H))$ to represent the posterior score of the model, which is negatively correlated with the kernel value. To optimize the model, we maximize this score, which is equivalent to minimizing the KL-divergence.

2.2. Piecewise Regression Fitting to D by Integrated Probability Inference

The calculation of the KL-divergence involves comparing the real-valued observed data D to the inferred belief probabilities H given a particular causal hypothesis G and θ . The original interaction between parent(s) X_π and child X_{chd} nodes in D can be described by an arbitrary function $X_{chd}=\mu(X_\pi)$ plus some observation noise. Depending on the nature of the causal relationships to be modeled, $\mu()$ can take various forms, including linear, non-linear, monotonic, non-monotonic, concave, convex, step and periodic functions. In the biology domain, a direct causal interaction between two proteins or between a protein and DNA molecule often take the form of a hill function, a step function, or a more general non-monotonic, nonlinear function.

One way to derive the belief inference given in Eq. 4 is to represent the relationship between the parent and child nodes as a cubic spline, which can well approximate general nonlinear relationships. However, a more straightforward alternative to splines would be regressing the marginal belief of X_{chd} onto X_π , assuming a linear relationship. If we subdivide the range of the parent nodes into L segments based upon the behavior we expect in a causal relationship between two nodes, linear regressions can be carried out in each segment.¹⁰ In the case $|\pi|=1$, our problem is to regress onto a single variable whose range has been divided into L segments, whereas when $|\pi|>1$ we are regressing onto multiple variables divided into a $|\pi|$ -dimensional grid comprised of $L^{|\pi|}$ components. Here we focus only on inferring causality between equivalent class structures in which $|\pi|=1$, but our approach easily extends to the more general case. To derive a procedure for fitting the belief inference equation to the data in a piecewise fashion, we first define some terms. Let \mathbf{X} denote a vector of predictor(s)/parent node(s) (in this pair-wise causality setting \mathbf{X} represents just a single variable for any given Markov equivalent structure being considered) and let Y represent the response variables. Let $D \in R^n$ represent the original noisy observed data and $D_0^l \in [0, 1]^n$ denote the rescaled observed data in the l -th segment/grid element, which is comprised of the observed values over the parent and child nodes in the given segment/grid element, i.e. $D_0^l=\{D_{\mathbf{X}}^l, D_Y^l\}$. Similarly, let the predicted data in the l -th segment be given by $H^l=\{H_{\mathbf{X}}^l, H_Y^l\}$. Given this, we pre-define a total of K bins that are evenly distributed in $[0,1]$, $I^k=[k/K, (k+1)/K], k=[0, K-1]$, and then for each bin we count the number of occurrences of the inferred marginal probability of Y falling in each

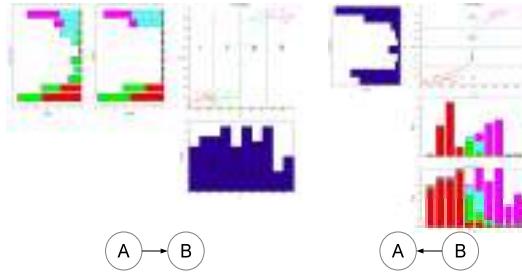


Fig. 1. Pairwise Causal Inference

of the l segments, i.e. H_Y^l falls in the k -th bin $I^k \in [0,1]$. We denote the number of occurrences for the k -th bin and the l -th segment/grid element by M_k^l . The counterpart of this number in D_Y^l , with respect to the observed data, is denoted by N_k^l . The frequency for the predicted data is then calculated as $p_k^l = M_k^l / \sum_k M_k^l$ for H_Y^l and similarly for the observed data, D_Y^l , $q_k^l = N_k^l / \sum_k N_k^l$. These counts and frequencies are used to compute the KL-divergence kernel, maximize the likelihood score, and identify the maximum LR model $\hat{\theta}$ in Eq. 1 per segment as described below. To maximize the data likelihood function $P(D|G, \hat{\theta})$ in Eq. 1 in each segment, we identify the parameter $\hat{\theta}$ that minimizes the KL-divergence for the current causal hypothesis G , which is defined as the symmetrized KL-divergence between the predicted belief and the rescaled observed data for every segment:

$$\begin{aligned} \hat{\theta}^l &= \underset{\theta}{\operatorname{argmax}} \{P(D_0^l | G, \theta) P(\theta | G)\} \propto \underset{\theta}{\operatorname{argmax}} \{P(D_Y^l | P(Y|E = D_X^l, G, \theta))\} \propto \underset{\theta}{\operatorname{argmax}} \{P(k(D_Y^l, H_Y^l(G, \theta)))\} \\ &\propto \underset{\theta}{\operatorname{argmax}} \{-\log(k(D_Y^l, H_Y^l(G, \theta)))\} \propto \underset{\theta}{\operatorname{argmin}} \left\{ \sum_{k=0}^K p_k^l \ln(p_k^l/q_k^l) + \sum_{k=0}^K q_k^l \ln(q_k^l/p_k^l) \right\} \end{aligned} \quad (6)$$

where $P(\theta | G) = 1/M$ for θ sampled uniformly in $[0,1]$. The statistical counts of the predicted probability, p_k^l for l -th segment in k -th bin, is a function of G and θ . The optimal statistical count of the fitted model for the l -th segment and k -th bin is \hat{M}_k^l . The overall fitted linear regression model $(G, \hat{\theta}^l | l = 1, \dots, L)$ with the counts across all segments in k -th bin of $[0,1]$ is obtained by summing \hat{M}_k^l over the total L segments, i.e. $\hat{\mathbf{M}}_k = \sum_{l=1}^L \hat{M}_k^l$. According to Eq. 1 and Eq. 3, the final optimized estimation of the data likelihood is then equal to

$$P(D|G, \hat{\theta}) \propto -\log(1 - \exp(-\frac{1}{2} \sum_{k=0}^K \hat{\mathbf{p}}_k \ln(\hat{\mathbf{p}}_k/\mathbf{q}_k) + \sum_{k=0}^K \mathbf{q}_k \ln(\mathbf{q}_k/\hat{\mathbf{p}}_k))) \quad (7)$$

For simplicity, in the experiment section below, to obtain the L segments in $[0,1]$, we simply divide the range of each parent node evenly into L segments.

2.3. Pair-wise Causality Example

To illustrate our causal inference procedure, we apply it to a pair of variables, leaving to the following section our application to more complicated causal networks such as triple-node equivalence classes. To begin we generate synthetic data given true pair-wise relationships as depicted in Fig. 1. We assume the observed data for the parent nodes is drawn from a uniform distribution and use a hill function to describe the common interactions between the parent and child nodes. We add Gaussian noise to the synthetic data to model uncertainty inherent in

the measurement data, and without loss of generality we set $L=4$. In practice we must exercise some care in the selection of L , since if L is set too high, the power of the likelihood score to distinguish the true causal direction from the null hypothesis can be significantly decreased. On the other hand, if L is set too low, the fit of the data could be poor, leading to likelihood scores for the true and null causal models that may not achieve statistical significance.

In Fig. 1, panels (a: A→B) and (b: A←B) show the fit of the regression model in the true and false causal directions, respectively. Each color denotes a different segment. The linear regression (LR) model is depicted by the line in each segment. The distribution of the predictor variables (parent nodes), is plotted in blue and the distribution of the response variables (child nodes) are shown by the stack plot, with the different colors corresponding to the different segments. In panel (a), A is the parent and its belief probability is clamped according to the observed and fitted distributions of the predicted values of the child node B. If the predicted values well match the observations, the likelihood score for (a) will be increased relative to the likelihood score for (b), which represents the opposite relationship (B as the parent node and A as the child). We note that in (b) for the 'flat' regions (I&IV) of the interaction function, the fitted LR does not cover the full range of values A can take on in these segments, which results in a truncation of the distribution of A at the ends of these two segments, resulting in a worse likelihood score compared to the true causal direction.

2.4. Performance Analysis

The asymmetric performance in predicting values along the true and false causal directions results from the constraints ($[-1,1]$ in Eq. 4) defined for our pairwise causality test, which constrains the slope coefficient to fall $\in [-1,1]$ and the intercept coefficient to fall $\in [0,1]$ for each segment. These constraints enforce an asymmetry in the fit of the regression model between the true and false causal directions needed to infer the true direction. We have also assumed that any nonlinear curve can be well approximated by a piecewise linear regression (LR) model. That is, in each segment the LR is good enough when fitting along the true causal direction. When a segment is fit along the correct causal direction (i.e., the segment is assumed to lie in the dimension of the parent node and is mapped to the child node via the regression function), the length to set for the segment should be determined by the degree of noise in the data. If noise levels are low (high) along the true causal direction, then the size of the segment can be longer (shorter) with a smaller (larger) number of segments. In our case, given we constrain the slope and intercept coefficients in the LR, the belief propagation is able to fit the distribution of noisy observations well so long as the segment size is small enough. However, if the segment is fit along the wrong causal direction, the length of the segment will no longer help scatter the observed data into different segments, given the distribution of observed values of the child nodes in the flat or U shaped regions of the distribution of the parent nodes are not be completely captured, but instead are truncated as discussed above. As a result, there will be a high probability of the observed values on the child nodes falling in the same segment no matter how small the segment size is or how low the noise level is. In this case, the ideal length of the segment will be determined by the shape of the interaction function.

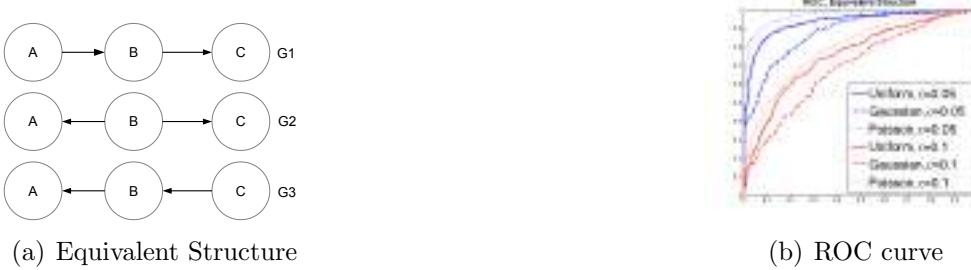


Fig. 2. Synthetic Causal Network Structure

We note that while we have made the argument that smaller-sized segments will not significantly improve the fit along the wrong causal dimension, we have not ruled out the possibility that the distribution of the true parent node given the child node (wrong causal direction) is well approximated across the different segments (i.e., the truncation of the ends of the distribution discussed above goes away), when an extremely small segment size is chosen. If such a case were to arise, the likelihood score would be equally good in the true and false causal directions. Similarly, if the segment size chosen is too big, the possibility exists that even along the correct causal dimension, the coverage of the predicted distribution of the child node may not be complete, making the likelihood score in the true and false causal directions equally bad. Although there are some existing algorithms on choosing the optimal number of segments, e.g. Multivariate Adaptive Regression Splines,¹³ we leave further investigation of choosing optimal segment sizes and positions as an interesting topic for future research.

3. Inferring Causality Among Markov Equivalent Structures

We next formally tested our causal inference procedure on synthetic data simulated to represent relationships that are common in biological systems, and on data generated from a dynamical systems model to recover metabolic network models. For the simulation experiment, we generated a large number synthetic datasets based on different nonlinear functions from the more difficult triple-node structures that are Markov equivalent. To infer known metabolic networks, we applied our method to yeast data generated from a metabolic model to demonstrate the ability to recover known metabolic networks.

3.1. Markov Equivalent Structures

For this problem, we test our method using the nonlinear functions $y=ax^3+\sin(k\pi x)$ and $z=by^2+\sin(k\pi y)$ to model equivalent structures. These functions represent a flexible framework for representing nonlinear biological relationships such as activation/inhibition and feedback control relationships, with the parameters a , b and k controlling the nonlinear features of such relationships: $a,b \in [-2,+2]$ and $k \in [0,2]$. To demonstrate the broad applicability of our method to general nonlinear data, we allow the parameters to vary across a wide range of values. For the simulation component of our study, we generated 1000 datasets for each of the scenarios depicted in Fig. 2, with each dataset comprised of 100 samples (a typical sample size in biological experiments). The data were simulated based on the ground truth structure G_1 shown

in Fig. 2. The data were simulated from different distributions: Uniform(U), Gaussian(G) and Poisson(P), to mimic microarray and RNA-sequencing gene expression data. For each simulation, the parent node A (x) was sampled from the U, G and P distributions, and the child nodes B (y) and C (z) were then generated according to the above nonlinear function. Before adding Gaussian noise, we firstly rescale the observation of A, B and C into [0,1]. Next, Gaussian noise ($0, \sigma^2$) was also simulated to reflect technological variation inherent in the types of measures made in biology; Finally, this noise was added to the values of A, B and C and the added value is rescaled to [0,1] to complete the generation of our observation data D . The graphical structures depicted in Fig. 2a are Markov equivalent and so in the context of conventional Bayesian networks they all give rise to the same data likelihood, and thus, are statistically indistinguishable from one another. We evaluated the learning performance by generating receiver operator characteristic (ROC) curves (Fig. 2b). Here we see that our method infers the correct causal relationships among the Markov equivalent structures, given the explicit assumptions made on the nature of the interaction between the parent and child nodes and on the distribution of these nodes.

3.2. Inferring Metabolic Signaling Pathways in Yeast

We next applied our method to metabolic data generated from a yeast model to assess whether we could recover a known metabolic pathway, trehalose biosynthesis. Trehalose functions as a carbohydrate reservoir and has recently been shown to play a crucial role in stabilizing proteins and cellular membranes under stress conditions such as heat shock. The metabolic pathway that produces trehalose is believed to regulate glucose uptake, particularly when the cell exists in a high-stress environment. It has also been shown that trehalose 6-phosphate (T6P), an intermediate of trehalose biosynthesis, plays a key role in the control of glycolytic flux. The kinetic values of this dynamical model have been identified experimentally¹² and are represented in the BioModels Database¹¹ (BIOMD0000000380).

We generated data for the trehalose biosynthetic pathway (in-silico) simulating the kinetic model for this model represented in the BioModel database. This model is comprised of a cycle of reactions between 11 different metabolites: external glucose (GLX), cellular glucose (GLC), glucose 6-phosphate (G6P), fructose 6-phosphate (F6P), glucose 1-phosphate (G1P), uridine triphosphate (UTP), uridine diphosphate (UDP), uridine diphosphate-glucose (UDG), diphosphate (Pi), trehalose 6-phosphate (T6P), and trehalose (TRH). These metabolites can be divided into two groups, the primary metabolites ($M=7$) whose concentrations vary as a result of reactions in this pathway, and extracellular glucose and boundary metabolites whose concentrations are fixed but they impact the reaction rates. Fig. 3 depicts the core causal signaling network we attempted to recover with our causal inference procedure. Represented in this network is a v-structure and feedback loop, structures containing Markov equivalent components that cannot be unambiguously resolved using classic Bayesian network approaches.

To infer causality along each undirected (bold) edge, we generated a dataset by sampling 100 (a typical sample size in biological experiments) starting concentrations of extracellular glucose¹² (changing the medium) X_{glx}^0 from the interval [0,100] (To ensure the system exhibits nonlinearity response to perturbations, we choose a wide range for the extracellular



Fig. 3. Core Causal Signaling Pathway for Trehalose Synthesis in Yeast

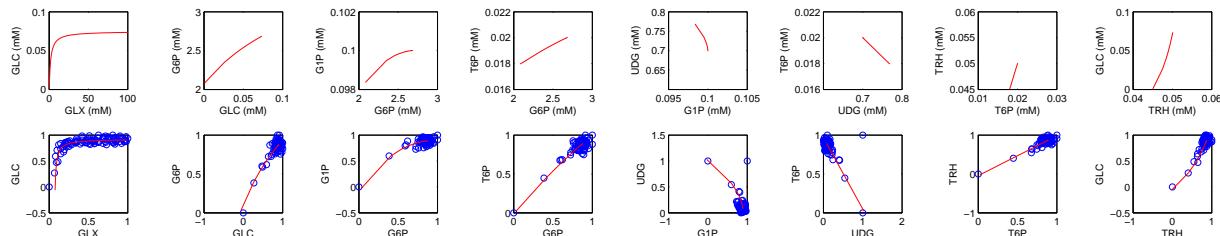


Fig. 4. Steady-state concentration given variations in starting extracellular glucose concentrations

glucose level). For each starting condition we let the dynamic system representing the trehalose biosynthetic pathway evolve to its new steady-state, generating a vector of steady state values for every primary metabolite, for each of the 100 starting conditions. This procedure resulted in a 100×7 data matrix of primary metabolite steady-state concentrations. In addition, we simulated Gaussian noise and added these noise components to the data matrix. Our final observation dataset is shown in Fig. 4. Each subplot describes the relationship of the steady-state concentrations between two (undirected) neighbor nodes in the pathway given the 100 different external glucose concentration starting conditions. The upper row shows these steady-state values before adding noise, while the lower row shows the rescaled values with the noise terms added, which represent the data used for the causal inference.

Given the connectivity structure of this network, we sought to resolve the edge direction by applying our method by calculating the causal structure score (Eq. 7) for each of the possible causal configurations. Given there are a total 8 edges in this network and that each edge can be oriented in one of two possible directions, there are 256 possible causal configurations to consider. The causal structure with the highest score was selected as the most likely causal structure supported by the data. In table 1 we list the top three inferred causal structures and their causality scores. We can see that our inferred top structure is the true causal network in 3. We note that the correct causal structure was inferred by considering the global structure of this network, as opposed to resolving the structure using pairwise causal relationships. One of the unique features of the modeling approach we developed is the ability to propagate information through the entire network. As a result, our global causal inference approach can leverage the correctly inferred causal relationships at between a given pair of nodes to infer the appropriate causal relationships among other nodes. This feature of our modeling approach is demonstrated by the causal inference of the feedback loop, i.e. TRH \rightarrow GLC. With existing causal inference procedures, the inferred causal relationship would be estimated as GLC \rightarrow TRH, whereas our method appropriately leveraged the global structure to correctly infer this edge, given the fitness of GLC, G1P and UDG is improved when we consider the

Inf. Str. 1st.(top) rank	GLX,TRH→GLC	GLC→G6P	G6P→G1P	G1P→UDG	G6P,UDG→T6P	T6P→TRH	Total Score
	3.41012	4.47932	5.3111	4.2021	4.3202	4.21400	25.93696
Inf. Str. 2nd. rank	GLX→GLC	GLC,T6P→G6P	G6P→G1P	G1P,T6P→UDG	TRH→T6P	GLC→TRH	Total Score
	3.40264	4.47932	5.27015	4.08511	4.04915	4.04646	25.33285
Inf. Str. 3rd. rank	GLX→GLC	GLC→G6P	G6P→G1P	G1P,T6P→UDG	G6P→T6P	GLC→TRH	Total Score
	3.40264	4.564723	4.99965	4.09345	4.16278	4.04646	25.26973

feedback in the top structure, compared to the other competing structures.

4. Conclusion and Discussion

In the life and biomedical sciences the technology now exists to score molecular and higher order phenotypes and genotypes on a massive scale, producing rich patterns of associations among molecular and higher order features that have the potential to elucidate the complexity of living systems. However, missing in biology is knowledge of the comprehensive set of pathways that operate in living systems, the structure of these pathways, how they interact with each other, how they change over time in response to different biological contexts etc. Even what are considered as canonical pathways are routinely shown to be incomplete and even inaccurate in different contexts. Therefore, methods that can help infer the causal relationships among the vast sea of phenotypes that can be scored are needed to better focus the type of hypotheses that can be experimentally pursued in a laboratory setting. Here we have attempted to address one significant limitation towards this end by developing a method to infer causality from correlation-based data by utilizing a Bayesian belief inference framework that is capable of distinguishing between Markov equivalent structures. By assuming different functional forms of the interactions that are possible among molecular features, observed data can be fit to probabilistic models of these relationships to assess which model best predicts the observed data. Our method is able to achieve good power in resolving causality by appropriately constraining the parameters of the probabilistic model in a way that allows the putative deterministic relationship between two variables to be assessed in a probabilistic framework. We applied our algorithm to multiple synthetic gene expression and RNA sequencing datasets to demonstrate that our approach can accurately infer causality under different biologically realistic assumptions regarding interaction types and noise structures.

Perhaps the biggest advantage of our approach is that it enables causal inference in a more complex network setting compared to previous methods that are limited to assessing pairwise causal relationships. This generality is achieved by leveraging the marginal probabilistic inference in a Bayesian network setting. We believe this advantage has significant importance given conventional top-down Bayesian network approaches can be systematically combined with our causal inference approach to form a integrated learning-inference framework. That is, our approach can enable a unified bottom-up and top-down modeling approach. Of course there are a number of ways in which the modeling approach we propose can be improved beyond our initial proof concept. Because our approach infers causality by maximizing the data likelihood that is based on a symmetrized KL-divergence measure between predicted and observed probabilities, implemented using a piece-wise linear regression framework, optimizing

the selection of the segment size and number is necessary to achieve maximal power and accuracy. Further, the causality inference can be embeded in structure search engine to search for optimal causality given initial graph. Finally, integrating our approach with a conventional structure-based learning approach, e.g. Bayesian network, has the potential to provide a very flexible framework that can model biological systems in a more comprehensive and accurate fashion, providing a way to incorporate bottom up modeling in a top-down framework to maximally leverage not only existing data, but knowledge derived from such data.

5. Acknowledgement

We thank the grant R01MH097276 of The National Institute of Mental Health (NIMH) and R01AG043076 of the Nation Institute of Aging (NIA) at National Institute of Health for their support to this work. JRK was supported by a James S. McDonnell Foundation Postdoctoral Fellowship Award in Complex Systems.

References

- Eric E Schadt, John Lamb, Xia Yang, Jun Zhu, (17 others). (2005) An integrative genomics approach to infer causal associations between gene expression and disease. *Nature Genetics* **37** 710 - 717.
- Emilsson V, Thorleifsson G, Leonardson AS, (29 others), Stefansson K, Schadt EE. (2008) Genetics of gene expression and its effect on disease. *Nature* **452**:423-428.
- Chen Y, Zhu J, Lum PY, Yang X, (16 others), Schadt EE. (2008) Variations in DNA induce changes in molecular network states that in turn lead to variations in obesity and related metabolic traits. *Nature* **452**:429-435.
- Nir Friedman, Daphne Koller. (2003) Being Bayesian About Network Structure. A Bayesian Approach to Structure Discovery in Bayesian Networks. *Machine Learning* Volume 50, Issue 1-2.
- Dominik Janzing, Joris Mooij, Kun Zhang, Jan Lemeire, Jakob Zscheischler, Povilas Daniusis, Bastian Steudel, Bernhard Schölkopf. (2012) Information-geometric approach to inferring causal directions. *Artificial Intelligence* Volumes 182-183.
- Patrik O Hoyer, Dominik Janzing, Joris M Mooij, Jonas Peters, Bernhard Schölkopf. (2008) Nonlinear causal discovery with additive noise models. *Neural Information Processing Systems*
- Bernhard Schölkopf, Dominik Janzing, Jonas Peters, Eleni Sgouritsa, Kun Zhang, Joris Mooij. (2012) On Causal and Anticausal Learning. *International Conference on Machine Learning*
- David Heckerman. (1996) A Tutorial on Learning With Bayesian Networks. *Technical Report*.
- Judea Pearl. (2009) *Causality: Models, Reasoning, and Inference; 2nd Edit.* Cambridge Univ. Press.
- Trevor Hastie, Robert Tibshirani, Jerome Friedman. (2009) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction; 2nd Edit.* Springer
- Li Chen, Donizelli Marco, Rodriguez, Nicolas, Dharuri Harish, Endler Lukas, Chelliah Vijayalakshmi, Li Lu, He Enuo, Henry Arnaud, Stefan Melanie, Snoep Jacky, Hucka Michael, Le Novere Nicolas, Laibe Camille. (2010) BioModels Database: An enhanced, curated and annotated resource for published quantitative kinetic models. *BMC Systems Biology* Volume 4-1.
- Kieran Smallbone, Naglis Malys, Hanan L. Messiha, Jill A. Wishart, Evangelos Simeonidis, Chapter eighteen - Building a Kinetic Model of Trehalose Biosynthesis in *Saccharomyces cerevisiae*, In: Daniel Jameson, Malkhey Verma and Hans V. Westerhoff, Editor(s), *Methods in Enzymology, Academic Press, 2011, Volume 500, Pages 355-370.*
- Jerome H Friedman. (1991) *Multivariate Adpative Regression Splines. The Annals of Statistics Volume 19, No. 1,1-141.*

MACHINE LEARNING FROM CONCEPT TO CLINIC: RELIABLE DETECTION OF BRAF V600E DNA MUTATIONS IN THYROID NODULES USING HIGH-DIMENSIONAL RNA EXPRESSION DATA

JAMES DIGGANS¹, SU YEON KIM¹, ZHANZHI HU¹, DANIEL PANKRATZ¹, MEI WONG¹, JESSICA REYNOLDS¹, ED TOM¹, MORAIMA PAGAN¹, ROBERT MONROE¹, JUAN ROSAI², VIRGINIA A. LIVOLSI³, RICHARD B. LANMAN¹, RICHARD T. KLOOS¹, P. SEAN WALSH¹, AND GIULIA C. KENNEDY¹

1. *Veracyte, Inc., South San Francisco, California, USA (email: giulia@veracyte.com)*

2. *Centro Diagnostico Italiano, Milan, Italy*

3. *Department of Pathology, Perelman School of Medicine
University of Pennsylvania, Philadelphia, Pennsylvania, USA*

The promise of personalized medicine will require rigorously validated molecular diagnostics developed on minimally invasive, clinically relevant samples. Measurement of DNA mutations is increasingly common in clinical settings but only higher-prevalence mutations are cost-effective. Patients with rare variants are at best ignored or, at worst, misdiagnosed. Mutations result in downstream impacts on transcription, offering the possibility of broader diagnosis for patients with rare variants causing similar downstream changes. Use of such signatures in clinical settings is rare as these algorithms are difficult to validate for commercial use. Validation on a test set (against a clinical gold standard) is necessary but not sufficient: accuracy must be maintained amidst interfering substances, across reagent lots and across operators. Here we report the development, clinical validation, and diagnostic accuracy of a pre-operative molecular test (Afirma BRAF) to identify BRAF V600E mutations using mRNA expression in thyroid fine needle aspirate biopsies (FNABs). FNABs were obtained prospectively from 716 nodules and more than 3,000 features measured using microarrays. BRAF V600E labels for training ($n=181$) and independent test ($n=535$) sets were established using a sensitive quantitative PCR (qPCR) assay. The resulting 128-gene linear support vector machine was compared to qPCR in the independent test set. Clinical sensitivity and specificity for malignancy were evaluated in a subset of test set samples ($n=213$) with expert-derived histopathology. We observed high positive- (PPA, 90.4%) and negative (NPA, 99.0%) percent agreement with qPCR on the test set. Clinical sensitivity for malignancy was 43.8% (consistent with published prevalence of BRAF V600E in this neoplasm) and specificity was 100%, identical to qPCR on the same samples. Classification was accurate in up to 60% blood. A double-mutant still resulting in the V600E amino acid change was negative by qPCR but correctly positive by Afirma BRAF. Non-diagnostic rates were lower (7.6%) for Afirma BRAF than for qPCR (24.5%), a further advantage of using RNA in small sample biopsies. Afirma BRAF accurately determined the presence or absence of the BRAF V600E DNA mutation in FNABs, a collection method directly relevant to solid tumor assessment, with performance equal to that of an established, highly sensitive DNA-based assay and with a lower non-diagnostic rate. This is the first such test in thyroid cancer to undergo sufficient analytical and clinical validation for real-world use in a personalized medicine context to frame individual patient risk and inform surgical choice.

1. Background and Significance

Thyroid nodules are solid or cystic growths found with increasing frequency with age. These nodules are evaluated using ultrasound-guided fine-needle aspirate biopsy (FNAB) because some nodules are malignant, and in 62 to 85%¹ of cases are diagnosed benign by cytopathology. The remainder of cases have an indeterminate or malignant cytopathology diagnosis and, historically, have undergone diagnostic surgery to remove part- (hemithyroidectomy) or all (total thyroidectomy) of the thyroid gland. Although the cytopathologically malignant nodules are almost always confirmed as cancer post-operatively, in upwards of 75% of operated nodules with indeterminate cytopathology², the nodule is found to be benign yet these patients have borne the risks and costs of diagnostic surgery and are relegated to a lifetime of thyroid hormone replacement therapy (HRT) to replace the missing organ. Conversely, for patients found to have cancer after an initial hemithyroidectomy, many must return for a completion thyroidectomy, difficult in a neck scarred from the initial surgery, to remove the rest of the thyroid tissue so that post-operative radioiodine ablation of remnant cancer will be effective.

Deciding on the extent of surgery in the initial operation on a cytologically indeterminate thyroid nodule remains a vexing question. Physicians must weigh the risk of missing active cancer or performing incomplete surgery against risks of overtreatment that compromise patients' long-term quality of life when making this choice.

Here, molecular diagnostics have a powerful role to play in providing personalized estimated risks of malignancy, thereby enabling physicians to accurately balance risk and reward in selecting a treatment strategy. These diagnostics can be categorized as either 'rule-out' tests with high sensitivity and negative predictive value (NPV, providing a confident declaration of benignity) or 'rule-in' tests with high specificity and positive predictive value (PPV, providing a confident declaration of malignancy³).

One example of a rule out test enabling observation in lieu of surgery on cytologically indeterminate but genetically benign FNABs is the Afirma GEC⁴. The GEC makes use of the gene expression of 167 genes in the cells of an FNAB to preoperatively predict whether a given FNAB is from a benign or malignant nodule². Given the high NPV and moderate PPV of this test, a negative result is reported as 'benign' while a positive result is reported as 'suspicious' rather than malignant.

Several DNA mutations and gene fusions have been well-studied in thyroid cancer and used as 'rule-in' markers (i.e. their presence is highly specific to malignancy although they are not sensitive due to their low prevalence in thyroid cancers). Among the most widely studied of these are mutations in BRAF, a member of the mitogen-activated protein kinase (MAPK) cascade involved in cell signaling and proliferation^{5,6}. The most common activating mutation (comprising 97% of BRAF mutations in thyroid carcinomas⁷) results in a thymine to adenine transversion at nucleotide 1,799 (1799T>A) resulting in a substitution of valine (V) at codon 600 with glutamate (E). This V600E mutation is highly specific for papillary thyroid carcinoma (PTC) diagnosis but has low sensitivity (i.e. V600E absence is not itself diagnostic of benignity).

The presence or absence of BRAF V600E in FNABs is usually assessed using DNA via PCR- or sequencing-based methods but these approaches all share three major limitations. These include (1) they traditionally have low analytical sensitivity requiring that a large proportion (up to 20%) of a given nodule have the relevant mutation before detection is possible. In addition, (2) reliance upon a single, well-studied mutation cannot detect patients with alternate, lower-frequency mutations that result in the same pattern of pathway activation. Finally, (3) PCR-based approaches with high analytical sensitivity (i.e. <5% mutant allele) often require a large amount of DNA that is frequently difficult to isolate from the small number of cells in an FNAB. This requirement leads to a high proportion of non-diagnostic FNABs, forcing patients to return to their physician for additional sample collection.

Gene expression signatures have been used to predict the presence or absence of point mutations or rearrangements in DNA in several cancers^{8,9} but these studies were performed on cell cultures or on blood, collection methods not directly relevant to solid tumor assessment. A gene expression signature detecting BRAF V600E in a small cohort of PTC nodules has previously been reported¹⁰ but the classifier was built on tissue samples rather than FNABs, the generalization of these classifiers to independent test sets was not evaluated and analytical verification studies were not performed. In the current work, we demonstrate the analytical and clinical validity of a gene expression signature in accurately classifying BRAF V600E mutation status in thyroid nodule FNABs. We also show that mRNA-based methods can improve upon all three of the shortcomings of DNA methods and accurately detect the presence of BRAF V600E with high analytical sensitivity using input amounts consistently recovered from FNABs. In addition, we show that at least one low prevalence mutation in BRAF results in the same gene expression pattern and is detected by Afirma BRAF (and is not detected by 1799T>A-specific assays).

2. Methods

FNABs were obtained prospectively from 716 patients as either part of a previously-reported collection² (n=360) or from de-identified samples consecutively referred to the Veracyte CLIA-certified clinical laboratory for GEC testing (n=356). Institutional Review Board (IRB) approvals were obtained from all applicable local or central IRBs including consent for validation of the GEC and additional molecular testing research. For the CLIA-certified laboratory samples, review and IRB-exempt status was obtained (Liberty IRB, DeLand, FL).

Each patient had a slide prepared from an FNAB and read by a cytopathologist. FNABs collected spanned Bethesda cytopathology categories¹¹ II through VI (II: Benign, III: Atypia of Undetermined Significance, IV: Follicular Neoplasm or Suspicious for Follicular Neoplasm, V: Suspicious for Malignancy and VI: Malignant). A second FNAB for molecular testing was collected from the same nodule. RNA and DNA from FNABs were extracted using the AllPrep Micro kit (QIAGEN) per manufacturer's instructions. Total RNA was amplified, hybridized to a custom microarray, and gene expression measured as previously described².

A Competitive Allele-Specific TaqMan PCR (castPCR™, Life Technologies, Carlsbad, CA) assay specific to the BRAF 1799T>A mutation was used to determine the percent mutation (% MUT) of BRAF 1799T>A-derived V600E present in each DNA sample as previously reported¹². Training samples with % MUT greater than 2.5% were labeled BRAF V600E-positive (BRAF-positive) and samples with % MUT of 2.5% or less were labeled BRAF V600E-negative (BRAF-negative). This threshold for the analytical sensitivity of the castPCR assay in FNAB-derived thyroid DNA was established to minimize unreliable training class labels due to stochastic effects on amplification in low copy-number samples.

Table 1: Sample counts by Bethesda cytology category. VERA001: samples prospectively collected in a previously-reported study; CLIA: samples from patients consecutively referred to the Veracyte CLIA laboratory. Risk of malignancy increases with increasing Bethesda category ranging from benign (Bethesda II) to malignant (Bethesda VI). Training set labels derived from castPCR results at a threshold of 2.5%; independent test set labels shown using a threshold of 5% (although results were evaluated at 0%, 2.5% and 5%; see Table 3).

Cytology	Source	Training Set			Independent Test Set		
		BRAF-	BRAF+	Prevalence	BRAF-	BRAF+	Prevalence
Bethesda II	All Samples	18	1	5.3%	32	1	3.0%
	CLIA	0	0	-	0	0	-
	VERA001	18	1	5.3%	32	1	3.0%
Bethesda III/IV	All Samples	37	4	9.8%	298	3	1.0%
	CLIA	12	2	14.3%	131	2	1.5%
	VERA001	25	2	7.4%	167	1	0.6%
Bethesda V	All Samples	34	27	44.3%	61	28	31.5%
	CLIA	17	14	45.2%	41	21	33.9%
	VERA001	17	13	43.3%	20	7	25.9%
Bethesda VI	All Samples	25	35	58.3%	29	83	74.1%
	CLIA	17	19	52.8%	20	60	75.0%
	VERA001	8	16	66.7%	9	23	71.9%
Total		114	67	37.0%	420	115	21.5%
			181			535	

2.1. Classifier training and validation

Samples were randomized into training and independent test sets to ensure Bethesda cytology category-specific representation in both training and test performance evaluation. Patient age and gender, nodule size, cytology sub-type (PTC, etc.) and % MUT were evaluated for homogeneity between sets after randomization. Investigators responsible for test set scoring were not involved in randomization and were blind to test set castPCR results.

Training of the Afirma BRAF RNA classifier was carried out using Robust Multichip Average (RMA)-normalized transcript cluster-level gene expression summaries and 10-fold cross-validation (CV) across a variety of classification methods and gene counts. Gene selection occurred within each CV loop via *limma*¹³ to identify genes distinguishing BRAF-positive from BRAF-negative samples. Classifiers were evaluated for positive- (PPA) and negative percent agreement (NPA)¹⁴ with castPCR-derived training set labels. PPA and NPA are utilized when a surrogate comparison is made to results from a second test (in this case, castPCR) *in lieu* of a clinical reference standard. They are computed identically to sensitivity and specificity, respectively. The highest scoring classification method and gene set were then used in a final round of model building with all 181 training samples resulting in the Afirma BRAF RNA classifier.

As use of this classifier in a ‘rule in’ context prioritized specificity over sensitivity, a series of simulations were conducted using training set scores (under 10-fold CV) over a range of assumed levels of run-to-run variability. For each level of variability, 5,000 simulated technical replicates of the training data were generated and, in each, the resulting number of false positives and false negatives were counted. The classifier decision threshold was then adjusted to minimize the probability of false positives (maximizing specificity and PPV) while maintaining acceptable false negative risk.

The classifier and this adjusted decision threshold were then locked prior to scoring the test set and evaluating performance against castPCR. To strike a balance between assay analytical sensitivity and clinical relevance of predictions, we evaluated the PPA and NPA of Afirma BRAF calls with castPCR at % MUT thresholds ranging from 0% to 10%. Additional experiments characterized the accuracy, reproducibility (inter-laboratory and inter- and intra-run), and robustness of the Afirma BRAF classifier.

For a subset (n=213) of FNABs in the test set for which GEC and castPCR results were previously reported¹² and for which expert-derived histopathology was available, the histopathology served as a clinical gold standard and was used to evaluate the clinical sensitivity and specificity of both Afirma BRAF and castPCR to detect malignancy via detection of the BRAF V600E mutation or gene expression signature.

In order to evaluate the underlying biological pathways affected by the V600E mutation, over/under-representation analyses (ORA) were performed using GeneTrail¹⁵ with either Afirma BRAF signature genes or all genes differentially expressed between BRAF-negative and –positive samples (n=2,502, false discovery rate (FDR) < 0.1 by *limma*) as the ORA test sets. The ORA reference set included all human genes (n=44,829) and annotation in the KEGG pathways database¹⁶. Significance was evaluated via Fisher’s exact test with a corrected FDR threshold of p < 0.05.

3. Results

3.1. Classifier comparison to castPCR

We computed PPA and NPA under 10-fold CV (using the training set) and found that 128 transcripts in a linear support vector machine¹⁷ (SVM) maximized the area under the receiver-operator characteristic (ROC, see Figure 1) curve (AUC) while minimizing run-to-run score variability. The linear SVM outperformed SVMs using radial basis function or polynomial kernels as well as regularized logistic regression. Only 11 of the final 128 transcripts are also used in the Afirma GEC indicating that these two models are detecting relatively distinct signals. Simulated technical replicates at varying levels of run-to-run score variability resulted in adjustment of the decision threshold from 0 to 0.45 to minimize the risk of false positives and target a specificity on the independent test set of at least 95% (see Figure 2).

The locked Afirma BRAF classifier and associated decision threshold were then used to score the test set and agreement between Afirma BRAF and castPCR was assessed across a range of castPCR label thresholds. Maximal PPA and NPA for all cytology categories were observed when the threshold for BRAF-positive status was $\geq 5\%$ MUT. We interpret this result as demonstrating the effective analytical sensitivity of Afirma BRAF to be equivalent to 5% MUT by castPCR. This 5% threshold represents a conservative lower bound on the analytical sensitivity of Afirma BRAF given that we did not observe any Afirma BRAF-positive samples with non-zero castPCR %MUT values less than 5% with the exception of the false positives (0% MUT) discussed below.

At 5% analytical sensitivity, Afirma BRAF demonstrates a PPA with castPCR of 90.4% (95% exact binomial confidence interval [CI] 83.5-95.1%) and an NPA of 99% (95% CI 97.6-99.7%) (Table 2). NPA was not significantly different across cytology categories but PPA appears lower in Bethesda V samples ($p=0.059$). Neither PPA nor NPA was significantly different between training and test sets overall or within each cytology category. We observed two samples in the training set and four in the test set that were Afirma BRAF positive but unambiguously 0% MUT by castPCR. This disagreement may have been due to technical variability in either assay or could be due to mutations other than the V600E mutation that cause similar gene expression changes.

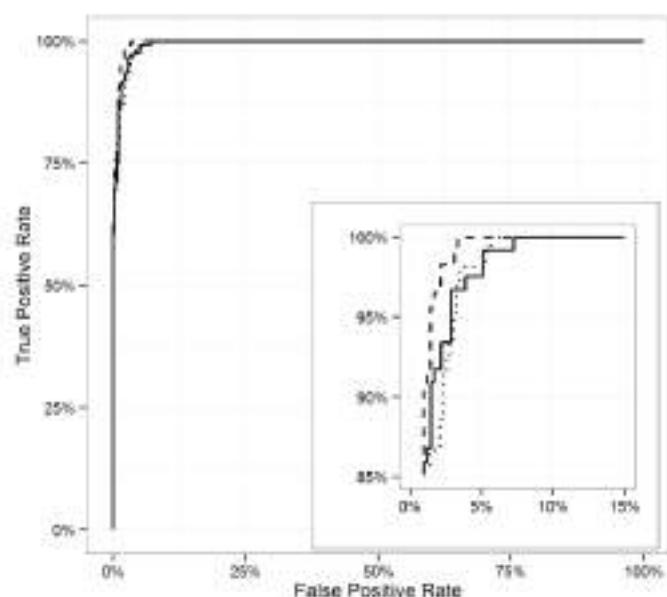


Fig. 1: ROC curves for Afirma BRAF performance on the test set at three different thresholds for BRAF V600E-positivity by castPCR. Inset plot shows more detail of the upper-left hand corner of the ROC curve.

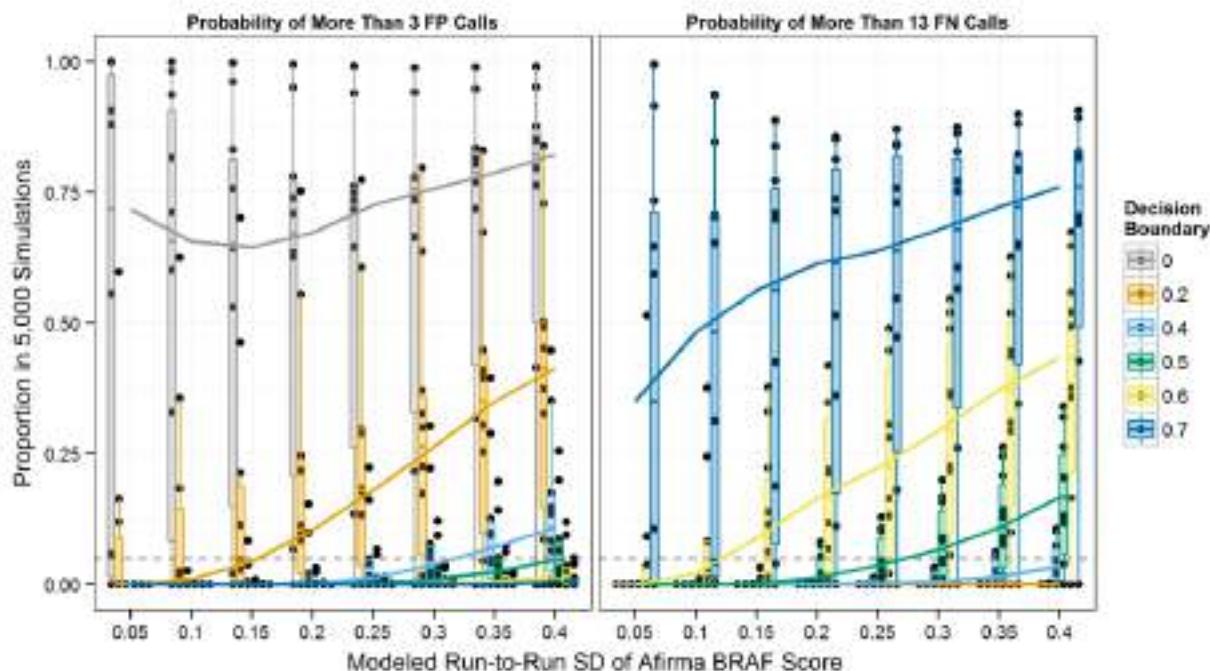


Fig. 2: Distribution of the proportion of 5,000 technical replicate simulations (y-axis) at varying levels of score reproducibility (x-axis) with more than three false positives (left) and more than 13 false negatives (right) at each of several candidate decision boundary values. The horizontal dotted line indicates a risk threshold of 5%.

Table 2: Positive percent agreement (PPA), negative percent agreement (NPA) and area under the ROC curve (AUC) for training (under cross validation) and test sets. AUCs for Bethesda II and III/IV cohorts were all equal to 1 in training and test but due to the small number of BRAF-positive samples, we report AUCs only for the remaining cytology cohorts.

	Cytology	PPA	NPA	AUC
Training	Bethesda II	100% [2.5%-100%]	100% [81.5%-100%]	-
	Bethesda III/IV	100% [39.8%-100%]	100% [90.5%-100%]	-
	Bethesda V	85.2% [66.3%-95.8%]	100% [89.7%-100%]	0.996 [0.987-1]
	Bethesda VI	88.6% [73.3%-96.8%]	96.0% [79.6%-99.9%]	0.982 [0.958-1]
	Overall	88.1% [77.8%-94.7%]	99.1% [95.2%-100%]	0.993 [0.986-1]
Test	Bethesda II	100% [2.5%-100%]	100% [89.1%-100%]	-
	Bethesda III/IV	100% [29.2%-100%]	100% [98.8%-100%]	-
	Bethesda V	75.0% [55.1%-89.3%]	96.7% [88.7%-99.6%]	0.975 [0.951-1]
	Bethesda VI	95.2% [88.1%-98.7%]	93.1% [77.2%-99.2%]	0.980 [0.955-1]
	Overall	90.4% [83.5%-95.1%]	99.0% [97.6%-99.7%]	0.997 [0.994-0.999]

We evaluated these samples via deep, targeted DNA sequencing of the BRAF gene along with several other true BRAF-positive and BRAF-negative samples to serve as controls (data not shown). We found one of these six discrepant samples to have a double mutation at nucleotide positions 1798 (T>A) and 1799 (T>A), leading, via codon degeneracy, to the same valine to glutamate amino acid change found in the most common BRAF mutation. We found no mutations within BRAF in the other five discrepant samples. All samples positive by Afirma BRAF with 0% MUT by castPCR were called ‘suspicious’ by the Afirma GEC so the positive finding by Afirma BRAF is consistent with an elevated risk for malignancy. In addition, two samples negative by Afirma BRAF and castPCR were found to have identical mutations in NRAS, 182A>G (Q61R), previously reported in melanoma¹⁸. An additional Afirma BRAF/castPCR-negative sample was found to have a mutation in KRAS, 35G>T (G12V), previously reported in colorectal cancer¹⁹. That all three samples were negative by Afirma BRAF suggests a lack of cross-reactivity with mutations in other genes upstream of BRAF in the MAPK pathway.

3.2. Clinical performance

We assessed the diagnostic value of BRAF V600E status for evaluation of nodules with Bethesda III-VI cytopathology using a subset of samples with associated gold-standard histopathology truth as previously described². Expert pathologists were blinded to the molecular results. Both Afirma BRAF and castPCR called all histopathologically benign samples as BRAF V600E-negative (specificity 100%, 95% CI 97.4%-100%), recapitulating the previously reported high specificity of the BRAF V600E mutation^{12,20-25}. Of the 73 histopathologically malignant samples, and at castPCR thresholds ranging from 0 to 2.5%, both assays identified a total of 32 as BRAF-positive (sensitivity 43.8%, 95% CI 32.2%-55.9%, Table 3). Sensitivity was not significantly different between the two assays across Bethesda cytology sub-classes. While both Afirma BRAF and castPCR identified 32 malignant samples as BRAF-positive, two samples called BRAF-positive by castPCR (with 4.2% and 20.2% MUT detected) were Afirma BRAF-negative. Two additional samples were called positive by Afirma BRAF but showed 0% MUT by castPCR. All four of these samples were malignant by histopathology.

Table 3: Performance of Afirma BRAF and castPCR (at various thresholds in analytical sensitivity) in predicting malignancy (as defined by histology after resection) by cytology category. NPV and PPV are calculated using study prevalence (34.3%, 73 malignant nodules in 213 total nodules).

	Sensitivity	Specificity	NPV	PPV	AUC
Afirma BRAF	43.8% [32.2%-55.9%]	100% [97.4%-100%]	77.30%	100%	0.840 [0.779-0.901]
castPCR (0%)	43.8% [32.2%-55.9%]	100% [97.4%-100%]	77.30%	100%	0.719 [0.662-0.776]
castPCR (2.5%)	43.8% [32.2%-55.9%]	100% [97.4%-100%]	77.30%	100%	0.719 [0.662-0.776]
castPCR (5.0%)	42.5% [31%-54.6%]	100% [97.4%-100%]	76.90%	100%	0.719 [0.662-0.776]

3.3. Reproducibility and analytical specificity

Intra- and inter-run reproducibility of the classifier was evaluated using 9 FNABs and three tissue controls selected from among training samples with high (BRAF-positive) or low (BRAF-negative) classifier scores and scores near the classifier decision boundary. Each FNAB and tissue was processed from total RNA in triplicate in each of three different runs across days, operators and reagent lots. The intra-assay standard deviation (SD) of Afirma BRAF scores is 0.171 (95% CI 0.146-0.204). Of the 106 Afirma BRAF calls produced (two arrays failed quality control requirements), 106 resulted in concordant calls across all three runs (100% concordance). The inter-assay SD of scores is 0.204 (95% CI 0.178-0.237) for scores measured on a six point scale.

FNABs often contain lymphocytes, blood or benign thyroid tissue that may interfere with or dilute BRAF-positive cells. To evaluate the impact of this dilution on Afirma BRAF signal, an Afirma BRAF-positive PTC sample was mixed in silico (using a previously reported mixture model²⁶) with increasing proportions of diluent samples. These in silico mixtures included dilution with samples of lymphocytic thyroiditis (LCT), pure blood, or benign thyroid tissue. BRAF-positive

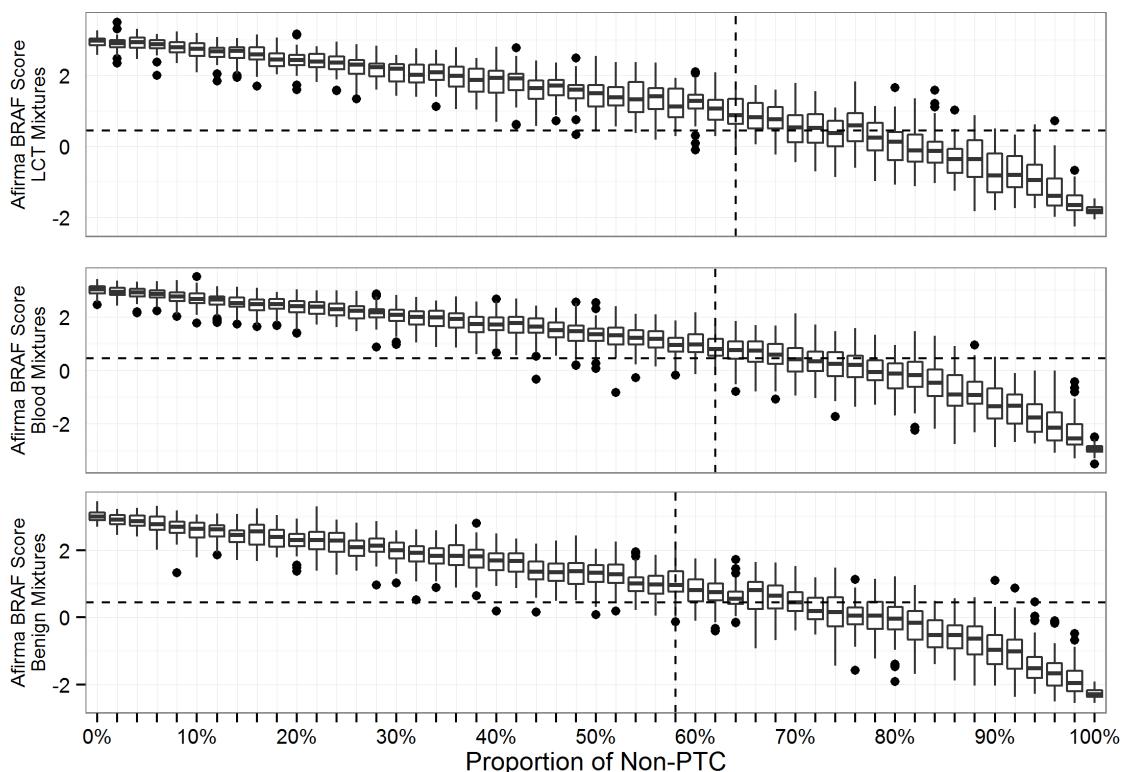


Fig 3: Afirma BRAF score versus the proportion of blood, LCT or benign nodule FNAB mixed in silico. Each box plot summarizes the results of fifty simulated mixtures of one BRAF-positive PTC sample with one non-malignant sample. Vertical dotted lines indicate the highest proportion of interferents (down to 36%, 38%, and 42% PTC for LCT, pure blood and benign FNAB, respectively) at which at least 80% of simulations call the mixture BRAF-positive.

samples were called correctly at least 80% of the time in mixtures representing 36%, 38% and 42% BRAF-positive PTC content, respectively. Afirma BRAF results for the pure blood, LCT and benign thyroid tissue samples were all BRAF-negative and all BRAF-negative FNAB mixtures were correctly called BRAF-negative regardless of mixture proportion, thus the presence of diluents commonly encountered in thyroid FNABs does not result in Afirma BRAF false positives.

4. Discussion

The current work is the first to describe the analytical verification and clinical validation of an RNA-based BRAF expression signature using a sample type relevant for clinical assessment of solid tumors. We developed an mRNA-based classifier that detects the gene expression signature of the BRAF V600E mutation in FNABs with high diagnostic accuracy. The classifier demonstrates both high PPA and NPA in comparison with a sensitive DNA-based assay for the BRAF V600E mutation. Clinical validation of Afirma BRAF using a cohort of samples with expert-derived post-surgical truth found no false positives identified in a cohort of 140 histopathologically benign nodules. The sensitivity of Afirma BRAF on this cohort was identical to that of castPCR, and both assays have clinical sensitivity for thyroid malignancy (43.8%) limited by the prevalence of BRAF V600E, as not all malignant nodules harbor this mutation.

Afirma BRAF had decreased PPA and NPA with castPCR for samples with less than 5% MUT indicating that castPCR is a slightly more analytically sensitive assay. The clinical relevance of low level BRAF mutation in thyroid nodules is unclear and the therapeutic benefit of early, aggressive treatment of such lesions is not well-defined^{27,28}. Given the equivalent performance of Afirma BRAF and castPCR on the clinical validation set, analytical sensitivity at less than 5% MUT may not translate into more accurate prediction of clinical outcome and may only contribute to rare false positives.

Indeed, one challenge in using increasingly sensitive PCR-based assays to detect individual mutations like BRAF V600E is the risk of an analytical true positive that has no clinical significance at the time of resection²⁹. Highly sensitive BRAF mutation assays (down to 0.1%) may find mutations in 80% of papillary thyroid microcarcinomas³⁰, even though these generally do not behave like cancers and may regress spontaneously²⁸.

We observed six samples that were Afirma BRAF positive but 0% MUT by castPCR. Since Afirma BRAF detects gene expression patterns associated with V600E, we considered whether a sample can exhibit a BRAF-positive-like profile caused by non-T1799A DNA alterations and indeed observed a sample with a double mutation at positions 1798-1799 in the BRAF gene. Due to the primer design of the castPCR assay, such double mutants would not be detected as BRAF V600E-positive by castPCR even though the resulting protein still contains glutamate at position 600. This provides evidence that the Afirma BRAF classifier correctly identified the downstream transcriptional effects of the mutated BRAF protein. Another advantage to using RNA-based analysis over DNA-based testing is that Afirma BRAF had a significantly lower non-diagnostic rate

due to sample insufficiency compared to castPCR (7.6% vs. 24.5, $p<0.001$), thus allowing a reportable result for more samples.

Additionally, it is important to consider that gene expression is a better approximation of the biological functions of relevance to thyroid malignancy, and is downstream of possible epigenetic regulatory mechanisms (e.g. gene silencing or allele-specific expression) that may prevent the expected phenotypic expression of a DNA mutation. We also hypothesize that the Afirma BRAF classifier may potentially recognize non-canonical cell signaling with an expression signature similar to BRAF activation. Conversely, epistatic down-regulation of the V600E expression signature by other mutations or signaling pathways remains a formal possibility. In such cases, the Afirma BRAF classifier may register a result consistent with the absence of an active V600E expression signal.

Previous studies have found that 1.3-8.3% of cytology benign nodules may harbor BRAF V600E mutations (range 1.3%-8.3%)^{27,31,32}. In the cohort reported here, we also found that 2 of 52 (3.8%, 95% CI 0.5%-13.2%) cytology benign FNABs which were malignant by histology were positive by both castPCR and Afirma BRAF.

Analytical validity studies of Afirma BRAF show that the test is accurate and precise and are reported in accordance with the STARD (STAndards for Reporting of Diagnostic Accuracy) guidelines. These studies demonstrate that Afirma BRAF has low intra- and inter-run variability and is highly robust to diluents potentially encountered in routine clinical testing. Taken as a whole, these studies meet Evaluation of Genomic Applications in Practice and Prevention (EGAPP) level 1 for analytical verification (inter-laboratory comparison) and EGAPP level 1 for clinical validity (well-designed longitudinal cohort studies)³³. To our knowledge, this is the first mRNA expression-based multivariate classifier to meet these STARD and EGAPP levels of evidence for accurate identification of a DNA mutation.

Pathway analysis of Afirma BRAF classifier genes reveals enrichment of tight junction, cell adhesion, and ECM-receptor molecules. These molecules are not only involved in apico-basal architectural changes³⁴, but are also increasingly implicated as mediators in cancer signaling³⁵⁻³⁷. A broader analysis using all differentially expressed genes on the array identifies pathways involved in MAPK, ErbB, Wnt, and p53 cancer signaling as overrepresented in BRAF V600E nodules.

Preoperative treatment decisions that may be affected by the presence of BRAF V600E may include extent of thyroidectomy (hemi- versus total), performance of central neck dissection, and administration of radioactive iodine. The ability of Afirma BRAF to accurately detect V600E status may assist physicians in making these treatment decisions and potentially improve patient care.

Acknowledgments

The authors would like to thank Sharlene Velichko, Julie Mathison and Matthew Muller for technical assistance, Lyssa Friedman for protocol, accrual and regulatory assistance, and Emma Caoili and Tami Wong for sample accessioning.

References

1. Aschebrook-Kilfoy, B., Ward, M. H., Sabra, M. M. & Devesa, S. S. *Thyroid*. **21**, 125 (2011).
2. Alexander, E. K. *et al.* *NEJM*. 705 (2012).
3. Xing, M., Haugen, B. R. & Schlumberger, M. *Lancet*. **381**, 1058–69 (2013).
4. National Comprehensive Cancer Network Thyroid Carcinoma Guidelines. (2013).
5. Reuter, C. W., Catling, A. D., Jelinek, T. & Weber, M. J. *J. Biol. Chem.* **270**, 7644–55 (1995).
6. Weber, C. K., Slupsky, J. R., Kalmes, H. A. & Rapp, U. R. *Cancer Res.* **61**, 3595–8 (2001).
7. Barollo, S. *et al.* *Thyroid*. (2014).
8. Kote-Jarai, Z. *et al.* *Clin. Cancer Res.* **12**, 3896–901 (2006).
9. Van Vliet, M. H. *et al.* *Genet. Test. Mol. Biomarkers*. **17**, 395–400 (2013).
10. Giordano, T. J. *et al.* *Oncogene*. **24**, 6646–56 (2005).
11. Cibas, E. S. & Ali, S. Z. *Am. J. Clin. Pathol.* **132**, 658–65 (2009).
12. Kloos, R. T. *et al.* *J. Clin. Endocrinol. Metab.* **98**, E761–8 (2013).
13. Smythe, G. (eds. Gentleman, R. *et al*) (Springer, 2005).
14. FDA CDRH. Statistical Guidance on Reporting Results from Studies Evaluating Diagnostic Tests. 1–39 (2007).
15. Backes, C. *et al.* *Nucleic Acids Res.* **35**, W186–92 (2007).
16. Kanehisa, M. *et al.* *Nucleic Acids Res.* **38**, D355–60 (2010).
17. Cortes, C. & Vapnik, V. *Mach. Learn.* **20**, 273–297 (1995).
18. Curtin, J. A. *et al.* *N. Engl. J. Med.* **353**, 2135–47 (2005).
19. Faulkner, N. *et al.* *ASCO Mol. Markers*. (2010).
20. Zeiger, M. A. & Schneider, E. B. *Ann. Surg. Oncol.* **20**, 3–4 (2013).
21. Howell, G. M. *et al.* *Ann. Surg. Oncol.* **20**, 47–52 (2013).
22. Joo, J.-Y. *et al.* *J. Clin. Endocrinol. Metab.* **97**, 3996–4003 (2012).
23. Colanta, A. *et al.* *Acta Cytol.* **55**, 563–9 (2011).
24. Cañadas-Garre, M. *et al.* *Ann. Surg.* **255**, 986–92 (2012).
25. Xing, M. *et al.* *JAMA*. **309**, 1493–501 (2013).
26. Chudova, D. *et al.* *J. Clin. Endocrinol. Metab.* **95**, 5296–304 (2010).
27. Rossi, M. *et al.* *J. Clin. Endocrinol. Metab.* **97**, 2354–61 (2012).
28. Ross, D. S. & Tuttle, R. M. *Thyroid*. **24**, 3–6 (2014).
29. Dilorenzo, M. M. *et al.* *Endocr. Pract.* **20**, e8–e10
30. Lee, S.-T. *et al.* *J. Clin. Endocrinol. Metab.* **97**, 2299–306 (2012).
31. Nikiforov, Y. E. *et al.* *J. Clin. Endocrinol. Metab.* **94**, 2092–8 (2009).
32. Cantara, S. *et al.* *J. Clin. Endocrinol. Metab.* **95**, 1365–9 (2010).
33. Teutsch, S. M. *et al.* *Genet. Med.* **11**, 3–14 (2009).
34. Gardiol, D. *et al.* *Int. J. Cancer*. **119**, 1285–90 (2006).
35. Rangel, L. B. A. *et al.* *Clin. Cancer Res.* **9**, 2567–75 (2003).
36. Nagano, M. *et al.* *Int. J. Cell Biol.* **2012**, 310616 (2012).
37. Lu, P., Weaver, V. M. & Werb, Z. *J. Cell Biol.* **196**, 395–406 (2012).

A SYSTEMATIC ASSESSMENT OF LINKING GENE EXPRESSION WITH GENETIC VARIANTS FOR PRIORITIZING CANDIDATE TARGETS

HUA FAN-MINOGLUE*

*Division of System Medicine, Department of Pediatrics, Stanford University
Stanford, CA 94305, USA
Email: fminogue@stanford.edu*

BIN CHEN*

*Division of System Medicine, Department of Pediatrics, Stanford University
Stanford, CA 94305, USA
Email: binchen1@stanford.edu*

WERONIKA SIKORA-WOHLFELD

*Division of System Medicine, Department of Pediatrics, Stanford University
Stanford, CA 94305, USA
Email: wsikora@stanford.edu*

MARINA SIROTA

*Division of System Medicine, Department of Pediatrics, Stanford University
Stanford, CA 94305, USA
Email: msirota@stanford.edu*

ATUL J BUTTE[†]

*Division of System Medicine, Department of Pediatrics, Stanford University
Stanford, CA 94305, USA
Email: abutte@stanford.edu*

Gene expression and disease-associated variants are often used to prioritize candidate genes for target validation. However, the success of these gene features alone or in combination in the discovery of therapeutic targets is uncertain. Here we evaluated the effectiveness of the differential expression (DE), the disease-associated single nucleotide polymorphisms (SNPs) and the combination of the two in recovering and predicting known therapeutic targets across 56 human diseases. We demonstrate that the performance of each feature varies across diseases and generally the features have more recovery power than predictive power. The combination of the two features, however, has significantly higher predictive power than each feature alone. Our study provides a systematic evaluation of two common gene features, DE and SNPs, for prioritization of candidate targets and identified an improved predictive power of coupling these two features.

* co-first author

[†] corresponding author

1. Introduction

A major goal of biomedical research is to identify disease genes to guide drug discovery that aims to improve the disease outcomes (1). Genes are defined as disease genes when they carry disease-causing aberrations (2). To identify an aberration of a gene, or a gene feature, and prove it as a causal link between the gene and a disease involves experimental testing and is time consuming. The advancement in high-throughput experimental techniques has facilitated this process by enabling rapid generation of vast amount of data for disease-associated gene features. Those techniques include the gene expression microarray, which allows the study of differential gene expression (DE) between disease and control samples; and high-throughput genotyping and next generation sequencing, which allows the study of disease-associated single nucleotide polymorphisms (SNPs) by comparing disease and control populations. However, these disease-associated features could be assigned to thousands of candidate genes. Prioritizing genes by incorporating these features for further experimental testing of causal relation is therefore necessary to narrow down the search space and increase the effectiveness of translating these candidates (3).

DE is often considered when prioritizing candidate genes, largely because it has been widely used to discover differentially regulated genes and deregulated molecular mechanisms (4). However, it has also been shown that DE genes might not perform well for specific diseases, where highly differentiated genes were not directly related to diseases (5). Yet, whether it can be generalized for all diseases is not clear and most researchers still use DE genes as their primary choice for seeking molecular explanations of biological phenotypes. SNPs to phenotype associations from genome-wide association studies provide unbiased screens of common variant associations. Using disease-associated SNPs to prioritize candidate genes are on the rise, especially as the sequencing technology is getting cheaper and more comprehensive computational tools have been developed to facilitate the process of the raw sequencing data. However, disease-associated SNPs derived from a defined population could fail in a larger or different population (6) and how SNPs perform across different disease conditions is largely unknown.

Increasing effort has been put to link different types of gene features from different sources to improve the performance of each individual feature. As an example, highly differentially expressed genes were found more likely to harbor disease-associated SNPs (7). However, how this feature combination would affect the candidacy of the gene for target validation has not been studied. More comprehensive integration of genetic variants with other types of genomic and biological data has been performed in individual disease condition (8). Although it showed great promise of using genetics to guide drug discovery, whether this can be generalized for other disease conditions is not clear.

An objective assessment of the performance of DE genes and disease-associated SNPs alone or in combination in different disease conditions will help understand the utility of these features and provide guidance to the application of them for target prioritization. However, that

type of assessment is currently lacking, mainly because it will require multiplex data collection and incorporation between features across disease conditions.

In this study, we integrated gene expression with disease-associated SNPs and therapeutic target data sets across a diverse set of 56 diseases in 12 disease categories (Figure 1). We systematically evaluated how successful DE genes, disease-associated SNPs or the combination of both can recover known disease targets, and how well they can predict the known targets by comparing with random sampling of these features. We demonstrate that the performance of DE genes, disease-associated SNPs or the combination of both varies across diseases. We observe that both DE genes and disease-associated SNPs have more recovery power than predictive power. The combination of the two features, however, has more predictive power than each feature alone. This suggests linking DE genes with disease-associated SNPs improves the accuracy of prioritizing candidate targets.

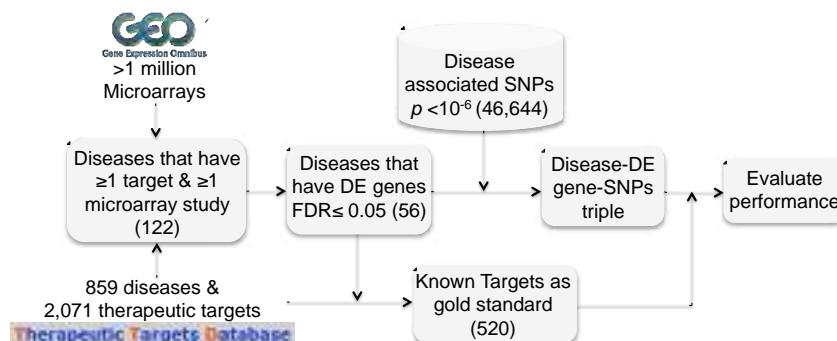


Figure 1. The schematic diagram of the work flow.

2. Methods

2.1 Selection of diseases

To examine the relation between gene expression and disease targets, we focused our study on diseases that have at least one gene expression microarray study and one known target.

To identify diseases and their associated microarray data, we utilized a text mining approach using previously published methods (9,10). Briefly, Gene Expression Omnibus (GEO) experiments that are relevant to human diseases and measure both normal and disease states were collected by an automated annotation and mapping between the Medical Subject Heading (MeSH) terms of the experiment associated publications and the disease concepts in the Unified Medical Language System (UMLS). Disease annotations and the associated microarray datasets were manually reviewed in a post-processing step to ensure accuracy. The resulting datasets included 238 disease concepts and 8,435 microarray samples.

To identify diseases that have at least one known target, we used the Therapeutic Target Database (TTD) (11), which provides manually annotated information about known therapeutic targets, their targeted disease conditions and corresponding drugs. It had 897 disease conditions

and 2,071 therapeutic targets (accessed in Aug. 2013), which include targets that are successful, in clinical trials, in pre-clinical research or discontinued. We extracted the UniProt IDs of the targets and mapped them to human Gene IDs. Then we converted the disease conditions of successfully mapped targets to UMLS concept IDs using the MetaMap (12). The maximum confidence score of 1000 was used as a cutoff for the successful mapping. The resulting dataset consists of disease-known target pairs that are represented by the disease concept ID and the target gene ID across 859 diseases and 2,071 therapeutic targets.

Next, we mapped the disease concept IDs between the disease-microarray and the disease-target datasets, which resulted in 122 diseases (Figure 1). These diseases that have at least one known therapeutic target and one gene expression microarray study were further analyzed for their DE genes.

2.2 Determination of DE genes and disease-associated SNPs

We used the method of significance analysis of microarray (SAM) (13) and its Bioconductor R package (*siggene*) to identify DE genes for each of the 122 diseases. For diseases that have multiple associated studies, the study that has the largest sample size was chosen. For genes with multiple probes, the expression level of the probe that had the highest absolute value was used. The raw data of the microarrays were processed and normalized as described in our previous publication (14). With a false discovery rate (FDR) < 0.05, 56 diseases were found to have at least one DE gene, which includes a total of 17,409 unique DE genes across all the diseases.

To identify the associated SNPs of these 56 diseases, we utilized a human disease-SNP association database (VARIMED) (15,16). In a recent release (Sep. 2013), we have manually-curated over 466,000 disease-associated SNPs across about 6,600 associated diseases and related phenotypes from 17,088 publications. To evaluate the performance of using SNPs for recovering known targets, we used a cutoff $p < 10^{-6}$ and obtained 46,644 disease-associated SNPs from VARIMED. The SNP associated disease names were then mapped to concept IDs of the 56 diseases that have at least one DE genes and at least one known target. Thirty-eight diseases were assigned with at least one SNPs. Unassigned diseases were marked as having 0 SNPs in Table 1. SNPs associated genes were obtained from the dbSNP138 database. Linkage disequilibrium (LD) effect was not counted for selecting disease-associated SNPs to obtain a general pool of SNPs.

By combining the disease-DE genes dataset with the disease-SNPs dataset, we built 129,905 triples between the 56 diseases, their DE genes and associated SNPs. The resulted dataset was mapped with the gold standard of disease targets, which allowed us to examine how often DE genes and associated SNPs alone or in combination can recover and predict the known targets of each disease.

2.3 Determination of the gold standard for disease targets

Targets of the 56 diseases that have at least one DE genes were extracted from the disease-target dataset derived from TTD. Total 520 targets were selected and used as the gold standard for the

evaluation. These targets are primary targets, which are directly responsible for the efficacies of the corresponding drugs that were confirmed by strong experimental evidence (11).

2.4 Evaluation

To evaluate how often the DE genes and disease-associated SNPs can recover and predict the known targets in each disease, we calculated the percentage of targets that have each feature (recall) and the percentage of each feature that are associated with targets (precision). We also calculated the percentage of targets that have both features and the percentage of having both features and being targets for each disease. This allowed us to evaluate the combinatory effect of differential expression and genetic variants on recovering and predicting known targets. To obtain the expectation of the performance of these features, we randomly sampled (1,000 times) the same amount of genes and SNPs against the total gene sets in the microarray and the whole dbSNP138 pool, respectively. The precision and recall of the random samples were then calculated the same way as above. The q value was calculated as the percentage of the precision or recall of random sampling that is better than the original. The known targets of each disease were used as the gold standards. For comparing the performance between features, the precision and recall of each feature for all diseases were plotted (Figure 3).

2.4.1 Precision

For each disease, the precision of DE genes, disease-associated SNPs and both are calculated using the following formulas:

$$\text{Precision (DE genes)} = \frac{\text{Number of DE genes that are targets}}{\text{Number of DE genes}} \quad (1)$$

$$\text{Precision (SNPs)} = \frac{\text{Number of disease-associated SNPs in targets}}{\text{Number of disease-associated SNPs}} \quad (2)$$

$$\text{Precision (DE genes & SNPs)} = \frac{\text{Number of DE genes harboring SNPs that are targets}}{\text{Number of DE genes harboring SNPs}} \quad (3)$$

2.4.2 Recall

For each disease, the recall of DE genes, disease-associated SNPs and both are calculated using the following formulas:

$$\text{Recall (DE genes)} = \frac{\text{Number of targets that are DE genes}}{\text{Total number of targets}} \quad (4)$$

$$\text{Recall (SNPs)} = \frac{\text{Number of targets that harbor SNPs}}{\text{Total number of targets}} \quad (5)$$

$$\text{Recall (DE genes & SNPs)} = \frac{\text{Number of targets that are DE genes & harbor SNPs}}{\text{Total number of targets}} \quad (6)$$

3. Results

3.1 Statistics of the diseases studied

Overall we studied 56 diseases (Figure 2). According to Human Disease Ontology, they consisted of 16 cancers, 7 nervous system diseases, 6 metabolic diseases, 5 gastrointestinal system diseases,

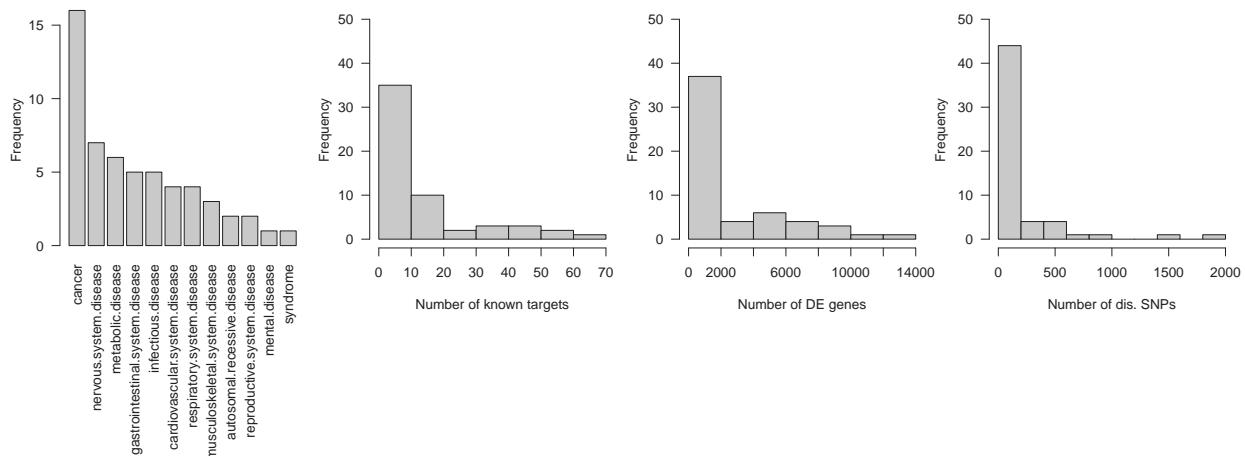


Figure 2. Histogram of disease categories, known targets, DE genes, and disease SNPs of the 56 diseases studied.

5 infectious diseases, 4 cardiovascular system diseases, 4 respiratory system diseases, 3 musculoskeletal system diseases, 3 autosomal recessive diseases, 2 reproductive system diseases, 1 mental disease, and 1 syndrome (Figure 2). These diseases had total 520 unique known targets, 17,409 unique DE genes and 8,235 unique disease-associated SNPs. About 2/3 of them had fewer than 10 targets; 2/3 of them had under 2,000 DE genes; and about 80% of them had fewer than 200 associated SNPs (Figure 2). On average, these diseases had 13.6 known targets, with obesity (64), prostate cancer (59) and breast cancer (51) having the largest number of known targets (Table 1). With a FDR<0.05, the average DE genes these diseases had was 2320. Spinal muscular atrophy and breast cancer had the largest number of DE genes, which were 12,648 and 10,314 respectively. Given the 10^{-6} p-value cutoff, the average number of disease-associated SNPs was 163. Rheumatoid arthritis (RA) and type 1 diabetes mellitus had the largest number of disease-associated SNPs, which were 1,826 and 1,456 respectively. However, 18 out of the 56 diseases did not have any associated SNPs with $p<10^{-6}$ (# of dis. SNPs=0, Table 1).

3.2 Recovering known targets by DE genes and disease-associated SNPs

Next, we asked for each disease how often the known targets were differentially expressed, harbored disease-associated SNPs or had both gene features. Those are essentially the true positive rates (recall) of using DE genes, disease-associated SNPs or both to recover known targets. We also calculated how often DE genes, disease-associated SNPs or both were associated with known targets, which are the positive predictive values (precision) of using these gene features to predict targets (Table 1).

Table 1. Statistics of the 56 diseases in the study

Disease Name	# of known targets	# of DE genes (FDR<0.05)	# of dis. SNPs (p<10 ⁻⁶)	Recall			Precision		
				% of targets being DE genes	% of targets harboring dis. SNPs	% of targets being DE genes & harboring dis. SNPs	% of DE genes being targets	% of dis SNPs in targets	% of DE genes harboring SNPs that are targets
Obesity	64	1	507	0	3.1*	0	0	0.4*	NA
Prostate Cancer	59	1030	407	23.7**	1.7	0	1.4**	0.5*	0
Breast Cancer	51	10314	189	82.4**	2.0	2.0	0.4*	0.5	3.6**
Asthma	48	2754	348	12.5	4.2	0	0.2	1.7**	0
Rheumatoid Arthritis	43	858	1826	11.6	7.0*	4.7**	0.6	0.5**	15.4**
Type 2 Diabetes Mellitus	41	26	647	0	9.8**	0	0	1.4**	0
Alzheimer's Disease	39	4682	416	30.8	7.7**	0	0.3	2.2**	0
Atherosclerosis	35	93	0	0	0	0	0	NA	NA
Hypertension	32	71	161	0	3.1*	0	0	0.6*	NA
Parkinson's Disease	25	1235	899	8.0	0	0	0.2	0	0
Multiple Sclerosis	22	131	435	9.1*	4.6	4.6**	1.5*	0.2	25.0**
Inflammatory Bowel Disease	18	4682	39	50.0*	0	0	0.2	0	0
Non-small Cell Lung Cancer	17	7834	4	52.9	0	0	0.1	0	0
Hypercholesterolemia	17	7179	1	64.7	0	0	0.2	0	NA
Malignant Melanoma	17	7901	78	82.4	5.9*	5.9*	0.2	3.9**	6.2**
Myocardial Infarction	17	4	132	0	0	0	0	0	NA
Osteoarthritis	15	131	59	6.7	0	0	0.8	0	0
Lymphoma	12	888	14	33.3	0	0	0.5	0	NA
Crohn's disease	11	5590	352	54.5	9.1*	0	0.1	0.3*	0
Glaucoma	11	140	28	0	0	0	0	0	NA
Chronic Obstructive Pulmonary Disease	11	23	42	0	0	0	0	0	NA
Acute Myeloid Leukemia	10	2097	0	40.0	0	0	0.2	NA	NA
Malaria	10	194	27	0	0	0	0	0	NA
Erectile Dysfunction	10	1	3	0	0	0	0	0	NA
Sepsis	10	29	0	0	0	0	0	NA	NA
Colon Cancer	9	2547	235	11.1	0	0	0	0	0
Irritable Bowel Syndrome	8	69	0	0	0	0	0	NA	NA
Ulcerative Colitis	7	2587	119	0	0	0	0	0	0
Cystic Fibrosis	7	4	0	0	0	0	0	NA	NA
Type 1 Diabetes Mellitus	7	35	1456	0	0	0	0	0	NA
Small Cell Carcinoma of Lung	7	8118	0	28.6	0	0	0	NA	NA
Bacterial Infection	6	233	0	0	0	0	0	NA	NA
HIV	6	356	350	16.7	16.7*	0	0.3	0.3*	NA
Chronic Lymphocytic Leukemia	6	5996	40	66.7	0	0	0.1	0	0
Amyotrophic Lateral Sclerosis	5	2	84	0	0	0	0	0	NA
Skin Squamous Cell Carcinoma	5	877	2	0	0	0	0	0	NA
Cancer of the Stomach	5	1846	55	0	0	0	0	0	0
Gastro-esophageal Reflux Disease	4	2	0	0	0	0	0	NA	NA
Huntington's Disease	4	8100	15	75.0	0	0	0	0	0
Pulmonary Hypertension	4	10	0	0	0	0	0	NA	NA
Endometriosis	3	9	33	0	0	0	0	0	NA
Acute Promyelocytic Leukaemia	3	2	0	0	0	0	0	NA	NA
Macular Degeneration	3	721	0	0	0	0	0	NA	NA
Pulmonary Fibrosis	3	14	1	0	0	0	0	0	NA
Cervical Cancer	3	66	0	0	0	0	0	NA	NA
Myelodysplastic Syndrome	2	760	0	0	0	0	0	NA	NA
Alpha-1 Anti-trypsin Deficiency	2	5	0	0	0	0	0	NA	NA
Sickle Cell Anemia	1	6429	7	100.0	0	0	0	0	0
Urothelial Carcinoma	1	8767	0	100.0	0	0	0	NA	NA
Cardiomyopathy, Dilated	1	5728	19	100.0	0	0	0	0	0
Hepatic Cirrhosis	1	118	6	0	0	0	0	0	NA
Spinal Muscular Atrophy	1	12648	0	100.0	0	0	0	NA	NA
Vitamin A Deficiency	1	88	0	0	0	0	0	NA	NA
Idiopathic Fibrosing Alveolitis	1	241	18	0	0	0	0	0	NA
Testis Cancer	1	5280	61	0	0	0	0	0	0
Severe Acute Respiratory Syndrome	1	359	0	0	0	0	0	NA	NA
<i>Average</i>		13.6	2319.7	162.8	20.7	1.3	0.3	0.1	2.4

NA: either the total # of dis. SNPs is 0 or the total # of DE genes harboring SNPs is 0. *: $q < 0.1$, **: $q < 0.05$

The average recall by DE genes was 20.7%. Thirty-two of the 56 diseases (57.1%) had no targets that were DE genes, or 0% recall. Four diseases (urothelial carcinoma, spinal muscular atrophy, sickle cell anemia and dilated cardiomyopathy) had 100% recall, because they had only one known target and that target was a DE gene. Compared to random sampling, DE genes did not perform better in most of the diseases, except for prostate cancer, breast cancer, multiple sclerosis, and inflammatory bowel disease ($q < 0.1$ or $q < 0.05$). The average recall by disease-associated SNPs was 1.3% and 44 of them (78.6%) were 0%, where no targets of those diseases harbored disease-associated SNPs. HIV and type 2 diabetes mellitus had the highest recalls, 16.7% and 9.8% respectively. For diseases with non-zero recall, SNPs of most of them performed better than

random sampling, except prostate cancer, breast cancer, asthma, and multiple sclerosis. The average recall by both DE genes and disease-associated SNPs was 0.3% and 52 of them (92.9%) were 0%. Only 4 diseases (malignant melanoma, rheumatoid arthritis, multiple sclerosis and breast cancer) had targets that were DE genes and harbored disease-associated SNPs, which had a recall of 5.9%, 4.7%, 4.6% and 2%, respectively. Compared to random sampling, the combination performed better in all four diseases, except breast cancer.

On the other hand, the average precision by DE genes was 0.1%, where 39 diseases (69.6%) had no DE genes that were targets, or 0% precision. Multiple sclerosis and prostate cancer had the best precision, 1.5% and 1.4% respectively. Similarly, DE genes did not predict better than random sampling in most of the diseases. Since 18 of the 56 diseases had no associated SNPs (NA in the second to the last column, Table 1), the precision by disease-associated SNPs was calculated for 38 diseases. The average precision by disease-associated SNPs was 0.3% and 26 of them (68.4%) were 0%, where no SNPs of those diseases occurred in targets. Malignant melanoma and Alzheimer's disease had the best precision, 3.9% and 2.2% respectively. For most of the diseases, SNPs also predicted better than random sampling. Thirty-five diseases had no DE genes that also harbored disease-associated SNPs (NA in the last column, Table 1), thus the precision by DE genes and SNPs were calculated for 21 diseases. The average precision by both DE genes and disease-associated SNPs is 2.4% and 17 of them (80.9%) were 0%, where no DE genes that harbored SNPs were targets. The four diseases that had DE genes that harbored disease-associated SNPs and were targets were multiple sclerosis, rheumatoid arthritis, malignant melanoma, and breast cancer, with a precision of 25%, 15.4%, 6.2% and 3.6% respectively. The combined features of all four diseases predicted better than random sampling ($q < 0.05$).

To compare the performance between features, we plotted the precision and recall of each feature for all diseases (Figure 3). Although it was not the common precision and recall curve for

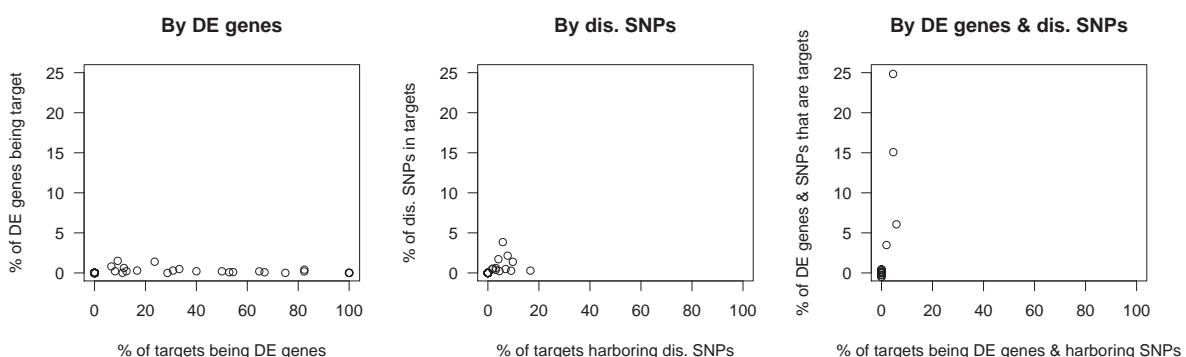


Figure 3. Performance of recovering known targets of each disease by DE genes, disease SNPs or both. Each point indicates the values of one disease.

evaluating the performance of a classifier, it showed how the performance of each feature varied in different diseases and allowed the comparison of performance between features. The performance of each feature varied greatly between diseases. When using DE genes, the recall values ranged from 0-100%, while the precision values were all below 3%. In other words, DE

genes could recover all known targets of a disease, but most of the DE genes were not disease targets. When using disease-associated SNPs, the recall values were between 0-20%, while the precision values were below 5%. In other words, disease-associated SNPs could recover a small portion of the known targets, but most of them were not in known targets. When using both DE genes and disease-associated SNPs, the recall values were below 6%, yet the precision values ranged from 0-25%. In other words, although the combination of both features could hardly recover any known targets, if a gene was differentially expressed and contained disease-associated SNPs, it would have higher chance to be a target for that disease.

When comparing between features, we found that DE genes gave better recall than disease-associated SNPs and disease-associated SNPs gave better recall than the combination of both. On the other hand, disease-associated SNPs gave better precision than DE genes, and the combination of both gave the best precision. We also identified the genes that were differentially expressed, harbored disease-associated SNPs and were targets. They were MOG (Myelin-oligodendrocyte glycoprotein) of multiple sclerosis, C5 (Complement C5) and TNF (Tumor necrosis factor) of rheumatoid arthritis, MC1R (Melanocyte-stimulating hormone receptor) of malignant melanoma, and ERBB4 (Receptor tyrosine-protein kinase erbB-4) of breast cancer (Table 2).

Table 2. Diseases that have targets that are DE genes and harbor dis. SNPs

Disease Name	All Known Targets (DE genes in <i>italic</i> , harboring SNPs in bold , DE genes & harboring SNPs in <i>bold italic</i>)
Multiple Sclerosis	ADRB2, CASP3, CNP, CRH, LPAR1, CXCR3, ICAM1, IFNAR2, IFNG, ITGA4, KCNA3, LEP, MMP9, MOG , MPO, PPARG, KLK6, CFLAR, NR1I2, CCR2, <i>SPP1</i>
Rheumatoid Arthritis	C5 , <i>CD4</i> , CD80, CCL2, CCR2, CD86, CFLAR, CTSK, F2RL1, FGF2, IKBKB, IKBKE, IL12A, IL13, IL15, IL17A, IL1R1, IL4, <i>IL6ST</i> , ITGA4, ITGB1, LTA , ITGB7, JAK3, JUN, LIF, LTB4R, MAPK11, MAPK12, MAPK14, MIF, MMP8, MMP9, <i>MYD88</i> , OSM, PTGES, PTGS2, SYK, TLR9, TNF , TNFRSF1B, TRBV7-9, VEGFB
Malignant Melanoma	<i>ALOX12</i> , <i>BIRC5</i> , <i>BRAF</i> , <i>CDH2</i> , CTLA4, CTSL1, <i>DCT</i> , <i>EDNRB</i> , <i>FN1</i> , HDAC4, <i>HSP90AA1</i> , <i>IFNAR2</i> , JUN, <i>MAP3K4</i> , MC1R , <i>PLAU</i> , <i>TXNIP</i>
Breast Cancer	AKT1, ANGPT2, CDH2, CYP1B1, <i>BRCA2</i> , CCND1, <i>CDC25A</i> , CLU, COPS5, CTSD, CXCL12, CXCR4, <i>CYP19A1</i> , DNMT3B, EGFR, EPHA2, ERBB2, ERBB4 , ESR2, ESRRA, FOS, HSD17B1, JUN, LHCGR, MAP2K1, <i>MAP3K4</i> , MDM2, MGFE8, MMP2, MUC1, NCOA3, NRG1, PGR, PLAUR, PRL, PRLR, PTGS2, PTK6, PTN, SERPINB5, SNCG, SRC, STC1, TPBG, VDR, <i>HSP90AA1</i> , MAP2K5, SCGB2A2, STS, TYMP

4. Discussion

Gene expression and genetic variants are the two most commonly measured and used features for selecting the best candidate genes for target validation. Their efficiency in target prioritization is often studied in specific disease conditions and their performance between diseases is largely unknown. Here we incorporated three diverse datasets from GEO microarray database, VARIMED disease-associated SNPs database and TTD target database, and systematically evaluated each feature and the combination of them in recovering and predicting known targets of 56 human diseases.

We found that the performance of each feature varied between diseases, which indicates that each feature could have different therapeutic utility for different diseases. However, overall, both DE genes and SNPs had lower precision than recall, which suggests that the DE or disease-associated SNP feature by itself is not good at predicting a target. The combination of being DE genes and harboring disease-associated SNPs had significantly improved precision ($q < 0.05$) compared to each feature alone (Figure 3). This implies that genes that are differentially expressed and harbor disease-associated SNPs are more likely to be targets. Indeed, for example, TNF (Table 2) is a successful target for RA validated by others (17) and carries risk variants via genome-wide association studies (18). Thus this combinatory feature could be used as a new criterion for prioritizing candidate genes for target validation.

In this study, DE genes, disease-associated SNPs or the combination of them was directly evaluated to allow objective assessment of their performance in target prioritization. Although the combination of DE and SNPs showed increased predictive power, it was still not great (< 25%). Optimizing the two features may improve their performance in prioritizing targets. A common alternative way to prioritize DE genes is their fold change (fc). Disease-associated SNPs can be ranked by how often they are associated with DE genes (%SNPs), since genetic variants associated with disease traits are likely to influence gene expression (1). Then the rank sum of fc and %SNPs can be used combinatorially. Many other prioritization methods can be incorporated with each feature, including the use of protein-protein interaction network, pathway involvement, literature and ontology. However, their effect on the performance may not necessarily improve the overall performance and need to be evaluated on a disease-by-disease basis.

There are limitations in this study that should be recognized. First, the microarrays used to derive the DE genes were from the study with the largest sample size, which could be the reason for the over 10,000 DE genes in some diseases. Meta-analysis of all microarray studies of each disease might result in more robust set of DE genes and a better disease signature (19). Likewise, meta-analysis of genome-wide studies for the same disease, as well as accounting for LD structure among the associated variants, may increase the reliability of disease-SNPs pairs. In this work, we used stringent thresholds (i.e., FDR < 0.05 and p value $< 10^{-6}$), changing which can alter the number of DE genes and disease-associated SNPs that will affect the precision and recall. Second, the known targets of each disease were extracted from the TTD database. Other databases may help derive more known targets, such as the DrugBank (20) and PharmGKB (21). However, DrugBank does not provide direct relations between targets and diseases, while PharmGKB has more pharmacogenomic information than drug-therapeutic targets relations. It is also important to recognize that all of these databases capture the current knowledge, which is not complete or perfect. As we discover more therapeutic targets and evaluate their efficacy, these resources will become more comprehensive and serve as a better gold standard.

Our study revealed a baseline performance of the two most common gene features, DE and SNPs, on prioritizing candidate targets, and identified an increased predictive power of the combination of the two features than that of each feature alone.

5. Acknowledgments

We thank Dr. Hyojung Paik for the constructive discussion. The research reported in this publication was supported by the National Institute of General Medical Sciences of the National Institutes of Health under award number R01GM079719. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The full data from these analyses is available on request from the authors (fminogue@stanford.edu).

References

1. Plenge, R. M., Scolnick, E. M., and Altshuler, D. (2013) Validating therapeutic targets through human genetics. *Nature reviews. Drug discovery* **12**, 581-594
2. Bromberg, Y. (2013) Chapter 15: disease gene prioritization. *PLoS computational biology* **9**, e1002902
3. Moreau, Y., and Tranchevent, L. C. (2012) Computational tools for prioritizing candidate genes: boosting disease gene discovery. *Nature reviews. Genetics* **13**, 523-536
4. Murray, D., Doran, P., MacMathuna, P., and Moss, A. C. (2007) In silico gene expression analysis--an overview. *Molecular cancer* **6**, 50
5. Hudson, N. J., Dalrymple, B. P., and Reverter, A. (2012) Beyond differential expression: the quest for causal mutations and effector molecules. *BMC genomics* **13**, 356
6. Thomas, G., Jacobs, K. B., Yeager, M., Kraft, P., Wacholder, S., Orr, N., Yu, K., Chatterjee, N., Welch, R., Hutchinson, A., Crenshaw, A., Cancel-Tassin, G., Staats, B. J., Wang, Z., Gonzalez-Bosquet, J., Fang, J., Deng, X., Berndt, S. I., Calle, E. E., Feigelson, H. S., Thun, M. J., Rodriguez, C., Albanez, D., Virtamo, J., Weinstein, S., Schumacher, F. R., Giovannucci, E., Willett, W. C., Cussenot, O., Valeri, A., Andriole, G. L., Crawford, E. D., Tucker, M., Gerhard, D. S., Fraumeni, J. F., Jr., Hoover, R., Hayes, R. B., Hunter, D. J., and Chanock, S. J. (2008) Multiple loci identified in a genome-wide association study of prostate cancer. *Nature genetics* **40**, 310-315
7. Chen, R., Morgan, A. A., Dudley, J., Deshpande, T., Li, L., Kodama, K., Chiang, A. P., and Butte, A. J. (2008) FitSNPs: highly differentially expressed genes are more likely to have variants associated with disease. *Genome biology* **9**, R170
8. Okada, Y., Wu, D., Trynka, G., Raj, T., Terao, C., Ikari, K., Kochi, Y., Ohmura, K., Suzuki, A., Yoshida, S., Graham, R. R., Manoharan, A., Ortmann, W., Bhangale, T., Denny, J. C., Carroll, R. J., Eyler, A. E., Greenberg, J. D., Kremer, J. M., Pappas, D. A., Jiang, L., Yin, J., Ye, L., Su, D. F., Yang, J., Xie, G., Keystone, E., Westra, H. J., Esko, T., Metspalu, A., Zhou, X., Gupta, N., Mirel, D., Stahl, E. A., Diogo, D., Cui, J., Liao, K., Guo, M. H., Myouzen, K., Kawaguchi, T., Coenen, M. J., van Riel, P. L., van de Laar, M. A., Guchelaar, H. J., Huizinga, T. W., Dieude, P., Mariette, X., Bridges, S. L., Jr., Zhernakova, A., Toes, R. E., Tak, P. P., Miceli-Richard, C., Bang, S. Y., Lee, H. S., Martin, J., Gonzalez-Gay, M. A., Rodriguez-Rodriguez, L., Rantapaa-Dahlqvist, S., Arlestig, L., Choi, H. K., Kamatani, Y., Galan, P., Lathrop, M., consortium, R., consortium, G., Eyre, S., Bowes, J., Barton, A., de Vries, N., Moreland, L. W., Criswell, L. A., Karlson, E. W., Taniguchi, A., Yamada, R., Kubo, M., Liu, J. S., Bae, S. C., Worthington, J., Padyukov, L., Klareskog, L., Gregersen, P. K., Raychaudhuri, S., Stranger, B. E., De Jager, P. L., Franke, L., Visscher, P. M., Brown, M. A., Yamanaka, H., Mimori, T., Takahashi, A., Xu, H., Behrens, T. W., Siminovitch, K. A., Momohara, S., Matsuda, F., Yamamoto, K., and Plenge, R. M. (2014) Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature* **506**, 376-381
9. Butte, A. J., and Chen, R. (2006) Finding disease-related genomic experiments within an international repository: first steps in translational bioinformatics. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, 106-110
10. Dudley, J., and Butte, A. J. (2008) Enabling integrative genomic analysis of high-impact human diseases through text mining. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 580-591
11. Zhu, F., Shi, Z., Qin, C., Tao, L., Liu, X., Xu, F., Zhang, L., Song, Y., Liu, X., Zhang, J., Han, B., Zhang, P., and Chen, Y. (2012) Therapeutic target database update 2012: a resource for facilitating target-oriented drug discovery. *Nucleic acids research* **40**, D1128-1136

12. Phan, J. H., Young, A. N., and Wang, M. D. (2012) Robust microarray meta-analysis identifies differentially expressed genes for clinical prediction. *TheScientificWorldJournal* **2012**, 989637
13. Tusher, V. G., Tibshirani, R., and Chu, G. (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences of the United States of America* **98**, 5116-5121
14. Dudley, J. T., Tibshirani, R., Deshpande, T., and Butte, A. J. (2009) Disease signatures are robust across tissues and experiments. *Molecular systems biology* **5**, 307
15. Chen, R., Davydov, E. V., Sirota, M., and Butte, A. J. (2010) Non-synonymous and synonymous coding SNPs show similar likelihood and effect size of human disease association. *PloS one* **5**, e13574
16. Ashley, E. A., Butte, A. J., Wheeler, M. T., Chen, R., Klein, T. E., Dewey, F. E., Dudley, J. T., Ormond, K. E., Pavlovic, A., Morgan, A. A., Pushkarev, D., Neff, N. F., Hudgins, L., Gong, L., Hodges, L. M., Berlin, D. S., Thorn, C. F., Sangkuhl, K., Hebert, J. M., Woon, M., Sagreiya, H., Whaley, R., Knowles, J. W., Chou, M. F., Thakuria, J. V., Rosenbaum, A. M., Zarnek, A. W., Church, G. M., Greely, H. T., Quake, S. R., and Altman, R. B. (2010) Clinical assessment incorporating a personal genome. *Lancet* **375**, 1525-1535
17. Criscione, L. G., and St Clair, E. W. (2002) Tumor necrosis factor-alpha antagonists for the treatment of rheumatic diseases. *Current opinion in rheumatology* **14**, 204-211
18. International, M. H. C., Autoimmunity Genetics, N., Rioux, J. D., Goyette, P., Vyse, T. J., Hammarstrom, L., Fernando, M. M., Green, T., De Jager, P. L., Foisy, S., Wang, J., de Bakker, P. I., Leslie, S., McVean, G., Padyukov, L., Alfredsson, L., Annese, V., Hafler, D. A., Pan-Hammarstrom, Q., Matell, R., Sawcer, S. J., Compston, A. D., Cree, B. A., Mirel, D. B., Daly, M. J., Behrens, T. W., Klarekog, L., Gregersen, P. K., Oksenberg, J. R., and Hauser, S. L. (2009) Mapping of multiple susceptibility variants within the MHC region for 7 immune-mediated diseases. *Proceedings of the National Academy of Sciences of the United States of America* **106**, 18680-18685
19. Chen, R., Khatri, P., Mazur, P. K., Polin, M., Zheng, Y., Vaka, D., Hoang, C. D., Shrager, J., Xu, Y., Vicent, S., Butte, A. J., and Sweet-Cordero, E. A. (2014) A meta-analysis of lung cancer gene expression identifies PTK7 as a survival gene in lung adenocarcinoma. *Cancer research* **74**, 2892-2902
20. Knox, C., Law, V., Jewison, T., Liu, P., Ly, S., Frolkis, A., Pon, A., Banco, K., Mak, C., Neveu, V., Djoumbou, Y., Eisner, R., Guo, A. C., and Wishart, D. S. (2011) DrugBank 3.0: a comprehensive resource for 'omics' research on drugs. *Nucleic acids research* **39**, D1035-1041
21. Hernandez-Boussard, T., Whirl-Carrillo, M., Hebert, J. M., Gong, L., Owen, R., Gong, M., Gor, W., Liu, F., Truong, C., Whaley, R., Woon, M., Zhou, T., Altman, R. B., and Klein, T. E. (2008) The pharmacogenetics and pharmacogenomics knowledge base: accentuating the knowledge. *Nucleic acids research* **36**, D913-918

DRUG-INDUCED mRNA SIGNATURES ARE ENRICHED FOR THE MINORITY OF GENES THAT ARE HIGHLY HERITABLE

TIANXIANG GAO^{1*}, PETTER BRODIN^{2,3*}, MARK M DAVIS^{3,4}, VLADIMIR JOJIC^{1*}

¹ Department of Computer Science, University of North Carolina at Chapel Hill,
Chapel Hill, NC 27599, USA

² Science for Life laboratory, Department of Medicine, Solna Karolinska Institutet
SE-171 76 Stockholm, SWEDEN

³ Department of Microbiology and Immunology, Stanford University School of Medicine,

⁴ The Howard Hughes Medical Institute, Stanford University School of Medicine,

* E-mail: {tgao,vjojic}@cs.unc.edu, petter.brodin@ki.se

The blood gene expression signatures are used as biomarkers for immunological and non-immunological diseases.¹ Therefore, it is important to understand the variation in blood gene expression patterns and the factors (heritable/non-heritable) that underlie this variation. In this paper, we study the relationship between drug effects on the one hand, and heritable and non-heritable factors influencing gene expression on the other. Understanding of this relationship can help select appropriate targets for drugs aimed at reverting disease phenotypes to healthy states. In order to estimate heritable and non-heritable effects on gene expression, we use Twin-ACE model on a gene expression dataset MuTHER,² measured in blood samples from monozygotic and dizygotic twins. In order to associate gene expression with drug effects, we use CMap^{3,4} database. We show that, even though the expressions of most genes are driven by non-heritable factors, drugs are more likely to influence expression of genes, driven by heritable rather than non-heritable factors. We further study this finding in the context of a gene regulatory network. We investigate the relationship between the drug effects on gene expression and propagation of heritable and non-heritable factors through regulatory networks. We find that the decisive factor in determining whether a gene will be influenced by a drug is the flow of heritable effects supplied to the gene through regulatory network.

1. Introduction

In this paper, we examine a general question: whether a drug aiming to perturb a disease phenotype should target genes whose expression is dominated by heritable or non-heritable factors?

The expression level of a gene is determined by both heritable effects and non-heritable effects. We estimated those effects for expressions of 3245 genes from MuTHER twin database (in Figure 1) and found that the expression of most genes are driven by non-heritable effects.⁵ At first, we would expect that a gene that is significantly impacted by non-heritable effects would be more likely to be affected by a drug – drugs could take advantage of such gene’s environmentally driven variability.

Naturally, heritable factors also play an important role in drug response.^{6–8} In our study, we find that the strong heritable effects on a gene’s expression are predictive of whether drugs can influence this gene. A simple experiment in Section 3.2 uses CMap database to show that genes robust to non-heritable effects – hence, strongly driven by heritable effects – are more likely to be part of a drug influenced gene expression signature.

The result of this experiment led us to examine the first question in a broader context of gene regulatory network. Previous studies show that genes can pass the drug influence through

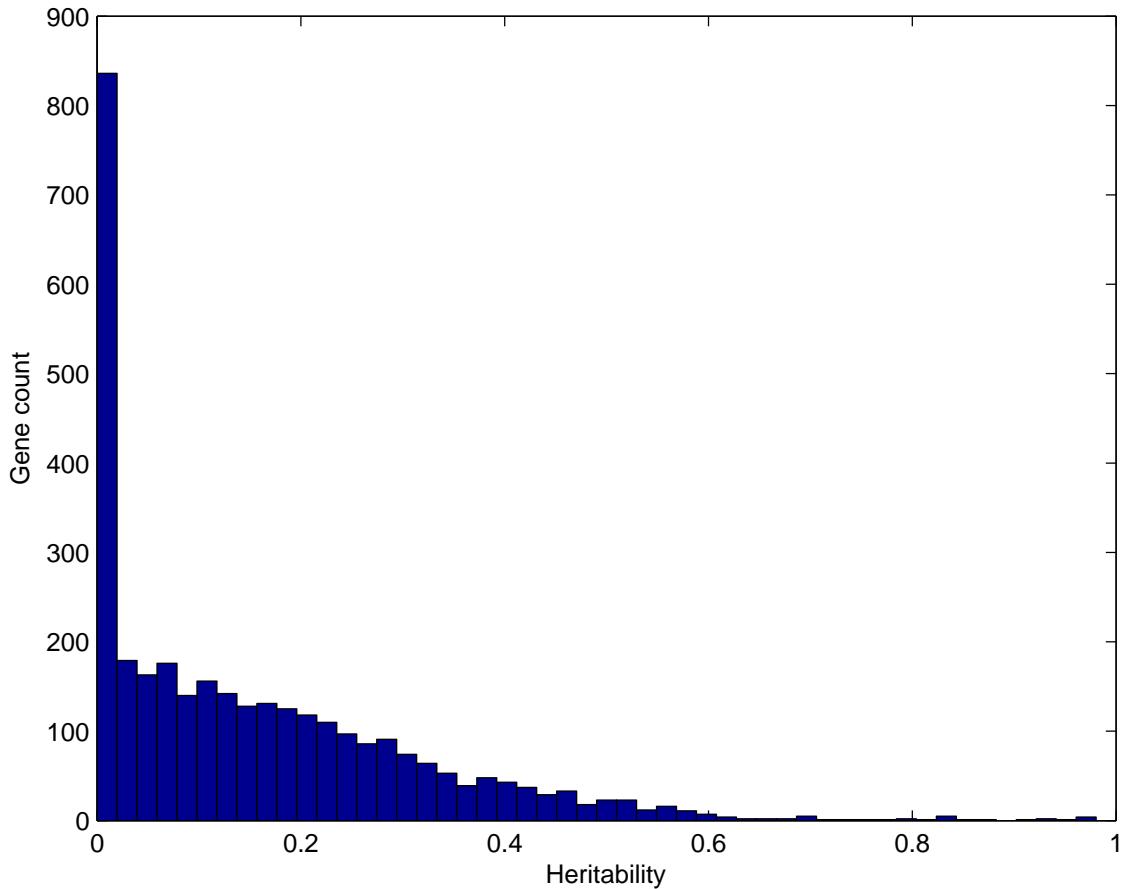


Fig. 1. The heritability estimation for 3245 genes from MuTHER twin database. Heritability is the ratio between variance driven by heritable effects and total variance.

the biological network to induce change in a target gene.^{9,10} We deem such genes source genes, as they can be seen as sources of drug effect propagation in the network. Hence, we wondered whether we can predict the flow of drug influence through a regulatory network from source to target genes.

We say that there exists a **regulation flow** between a source gene and a target gene if there is a sequence of strong regulatory relationships leading from the source gene to the target gene. Naturally, highly variable source gene with a strong flow to target gene will induce variance in the target gene. We can estimate how much of the heritably or non-heritably driven variance is propagated from source to target. Consequently, we introduced quantitative measures of strength of heritable or non-heritable flow between source and target genes.

We call a regulation flow between a source and a target gene **drug influence flow** if there exists a drug influencing both genes. Hence we can pose a question: Does a strong heritable flow between source and a target imply existence of a drug influence flow between these two genes?

An experiment in Section 3.3 reveals the fact: strong heritable flows are predictive of drug influence flows. Equally importantly, strong non-heritable flows are *not* predictive of drug

influence flows. Hence, a drug influence flow leading to particular target gene is best identified through the regulation flows propagating substantial heritable effects.

Here, we provide a simple example of this relationship in Figure 2. In this example, both the source gene PI4KB and PIK3R5 have regulation flow to target gene INPP5F. However, heritable flow strength to INPP5F from PI4KB (red) is higher than heritable flow strength from PIK3R5, even though they have similar regulation flow strength. Our validation in CMap shows that there is no drug influencing both PIK3R5 and INPP5F; while the drug “phthalylsulfathiazole” is known to influence both PI4KB and INPP5F.

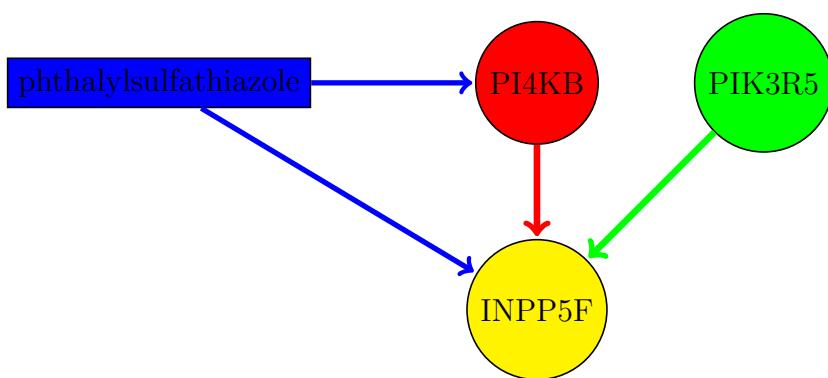


Fig. 2. An example of choosing the drug influence flow from our results. Yellow colored gene INPP5F is the target gene that we want to influence. The red regulation flow from PI4KB has regulation flow strength 0.3 and heritable flow strength of 0.1. The green regulation flow from PIK3R5 has regulation flow strength 0.3 and heritable flow strength of 0. The blue nodes is one of the drugs that is known to influence both PI4KB and INPP5F.

The rest of this paper is organized as follows: we introduce methods to find the drug influence flows in a specific network and Twin-ACE model for heritable effects estimation in Section 2. Supportive experiments and results are discussed in Section 3 and 4.

2. Methods

In this section, we discuss the methods that we use to recover the regulatory network and estimate heritable and non-heritable effects. In Section 2.1, we first introduce the directed acyclic graph gene regulatory network and the parameter estimation of a given network using linear model. In the second part, we introduce Twin-ACE model that we are using to estimate the impact of heritable and non-heritable effects on genes’ expression, and the quantitative measures of the heritable and non-heritable flow strength in a regulatory network. Twin-ACE model in combination with regulatory network model can be seen as a mixed model of the joint data.[?]

2.1. DAGRN and its estimation

Directed Acyclic Gene Regulatory Network(DAGRN) Our method operates on a DAGRN, as this representation enables a straightforward way of calculating the regulatory

effects using linear regression of expression. Specifically, DAGRN is a graph of P nodes with no loops or undirected edges. Each node t in the graph is a random variable x_t represents the measurement of a gene expression value. It has the following conditional probability density:

$$x_t | \mu_t, \mathbf{W}, \sigma_t^2 \sim \mathcal{N}(\mu_t + \sum_{s \in pa(t)} w_{s,t} x_s, \sigma_t^2), \quad (1)$$

where $pa(t)$ is the set of parent nodes that link to node t . \mathbf{W} is an adjacency matrix of size $P \times P$. Each entry $w_{s,t}$ indicates the strength of edge $s \rightarrow t$. μ_t is the local mean. We call σ_t^2 the residual variance of x_t .

Linear model estimation Given a gene expression matrix \mathbf{X} of size $N \times P$, where N is the number of samples and P is the number of genes, we can estimate the parameters $\boldsymbol{\mu} = [\mu_1, \dots, \mu_P]$, $\boldsymbol{\sigma} = [\sigma_1, \dots, \sigma_P]$ and \mathbf{W} .

Let us look at a specific gene. We use a vector \mathbf{y} to denote the N observed samples for that gene. We denote regulators' gene expression matrix as \mathbf{R} . This is a matrix of size $N \times R$. Each of R columns is an observed expression levels corresponding to a parent node, a regulator. We can write the joint distribution for all the samples as:

$$p(\mathbf{y} | \boldsymbol{\mu}, \mathbf{w}, \boldsymbol{\sigma}) = \prod_{i=1}^N \mathcal{N}(\mu_i + \mathbf{r}_i \mathbf{w}, \sigma_i^2), \quad (2)$$

and \mathbf{r}_k denotes k^{th} row of matrix \mathbf{R} . μ is the local mean for \mathbf{y} , σ^2 is the residual variance. We can derive the maximum likelihood estimation (MLE) update for parameters in (2) as :

$$\mu^{\text{MLE}} = \frac{1}{N} \sum_{i=1}^N y_i \quad (3)$$

$$\sigma^{\text{MLE}} = \sqrt{\frac{1}{N} \|\tilde{\mathbf{y}} - \mathbf{R}\mathbf{w}\|_2^2} \quad (4)$$

$$\mathbf{w}^{\text{MLE}} = \underset{\mathbf{w}}{\operatorname{argmin}} \|\tilde{\mathbf{y}} - \mathbf{R}\mathbf{w}\|_2^2. \quad (5)$$

We write $\tilde{\mathbf{y}} = \mathbf{y} - \mu^{\text{MLE}} \mathbf{1}_N$ for convenience. The optimization problem (5) is a linear regression problem.

Stimulation-Response Matrix After acquiring the regulation weights for the network, we need to calculate the total regulation influence from any source node to any target node. Therefore, we introduce the Stimulation-Response matrix. To differentiate the notation from μ_t , which is the local mean of a gene expression x_t , we define $\alpha_t \triangleq \mathbb{E}[x_t]$ as the global mean of x_t . Suppose the drug treatment changes mean of the source gene x_s by $\Delta\alpha_s$, we want to know the response change $\Delta\alpha_t$ in the target gene. For convenience, we will call $\Delta\alpha_s$ **stimulation change** and $\Delta\alpha_t$ **response change**.

As our model is a Linear-Gaussian model, the stimulation change and response change are also linear:¹¹ $\Delta\alpha_t = S_{s,t} \Delta\alpha_s$. We call \mathbf{S} Stimulation-Response matrix. It is a $P \times P$ matrix, where entry (s, t) indicates the response change $\Delta\alpha_t$ in the target gene t given unit stimulation change ($\Delta\alpha_s = 1$) in the source gene s . We can use Algorithm 1 to calculate the Stimulation-Response matrix. The computational complexity for this method is $O(P^2V)$, V is the total edge number.

```

input :  $W$ : adjacency matrix,  $P$  : total node number
output:  $\mathbf{S}$  : stimulation-response matrix
initialize  $\mathbf{S} \leftarrow \mathbf{0}_{P \times P}$ ;
for  $x \leftarrow 1$  to  $P$  do
     $S_{x,x} \leftarrow 1$ ;
    for  $y \leftarrow \text{nodes has directed edges link from } x$  do
        for  $j \leftarrow 1$  to  $P$  do
             $S_{j,y} \leftarrow S_{j,y} + S_{j,x} * W_{x,y}$ ;
        end
    end
end

```

Algorithm 1: The algorithm for calculating Stimulation-Response matrix

We call the value in entry (s,t) of \mathbf{S} as the **regulation flow strength** of the regulation flow $s \rightarrow t$. We only consider a regulation flows between source gene s and target gene t for which $S(s,t) > T$, where T is a certain threshold.

2.2. Twin-ACE model and heritable/non-heritable effect estimation

We are using the Twin-ACE model to estimate the heritable and non-heritable effects on the phenotypes. The standard ACE model for twin studies is based on the assumption that identical twins share their genes while fraternal twins share approximately half of their polymorphic gene sequences. ACE study design assumes presence of both monozygotic and dizygotic twins. In order to model the relatedness of the gene expression measurements in twins, we utilize the standard ACE model.¹² The main assumption underlying the ACE models is that the covariance of the gene expression in a twin pair can be decomposed into three contributions: 1) Additive genetic component 2) Common environmental component, and 3) twin-specific Environmental component.

The additive genetic component is the only component that is dependent on whether the twins are identical (monozygotic, MZ) or fraternal (dizygotic, DZ). Identical twins share same genetic material and hence differences in their gene expressions are attributable to environmental factors. In contrast, fraternal twins, on average, share only half of their genetic sequences and hence differences in their gene expressions can be attributed to heritable or non-heritable factors. This observation motivates parametrization of phenotypic covariance in terms of additive components $\mathbf{A}_{\text{MZ}} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$ and $\mathbf{A}_{\text{DZ}} = \begin{bmatrix} 1 & \frac{1}{2} \\ \frac{1}{2} & 1 \end{bmatrix}$ reflecting the expectation of higher covariance among mono-zygotic twins to the extent the gene expression is heritable. In terms of notation, to indicate zygosity of a twin pair t_1 and t_2 we will use $\text{zyg}(t_1, t_2)$, naturally $\text{zyg}(t_1, t_2) \in \{\text{MZ}, \text{DZ}\}$.

In addition to additive genetic effects, we also model potential environment effects. We denote them as \mathbf{C} and \mathbf{E} . These effects are assumed to be independent of twins' genomes. Furthermore, the common environmental effects are assumed to be affecting both of the twins

in a family, and hence off-diagonal covariance terms are 1. The twin-specific environmental effects are assumed to have an independent effect on each of the twins, so the off-diagonal terms are 0. Hence, we have $\mathbf{C} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$ and $\mathbf{E} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$.

Estimation of ACE parameters In order to use the relatedness of the gene expression measurements in twins to estimate the heritable and non-heritable effects, we use Twin-ACE model. Suppose we have F families and each family has two twins. For a given twin pair (t_1, t_2) we will use $\text{zyg}(t_1, t_2)$ to indicate whether the twin pair is monozygotic (MZ) or dizygotic (DZ). Similarly, we have the gene expression measurement \mathbf{y} for $N = 2F$ twins. The joint distribution for all samples is:

$$p(\mathbf{y}|\mu, a, c, e) = \prod_{(t_1, t_2)} \mathcal{N}\left(\begin{bmatrix} \mu \\ \mu \end{bmatrix}, \Sigma_{\text{zyg}(t_1, t_2)}(a, c, e)\right), \quad (6)$$

where

$$\Sigma_{\text{zyg}(t_1, t_2)}(a, c, e) = a^2 \mathbf{A}_{\text{zyg}(t_1, t_2)} + c^2 \mathbf{C} + e^2 \mathbf{E}.$$

We call a, c, e ACE parameters for gene expression vector \mathbf{y} . We will abbreviate covariances $\Sigma_{\text{MZ}}, \Sigma_{\text{DZ}}$ while acknowledging their dependence on parameters a, c, e .

In this model, μ is the population mean of the \mathbf{y} . To estimate the parameters a, c, e , we need to solve the follow optimization:

$$(a, c, e)^{\text{MLE}} = \underset{a, c, e}{\text{argmax}} \left[-\frac{1}{2} (\mathbf{y} - \mu)^T \Sigma^{-1} (\mathbf{y} - \mu) - \frac{1}{2} \log |\Sigma| \right]$$

$$\Sigma = \begin{bmatrix} \Sigma_{\text{zyg}(t_1, t_2)} & 0_{2 \times 2} & 0_{2 \times 2} & \dots & 0_{2 \times 2} \\ 0_{2 \times 2} & \Sigma_{\text{zyg}(t_3, t_4)} & 0_{2 \times 2} & \dots & 0_{2 \times 2} \\ 0_{2 \times 2} & 0_{2 \times 2} & \Sigma_{\text{zyg}(t_5, t_6)} & \dots & 0_{2 \times 2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0_{2 \times 2} & 0_{2 \times 2} & 0_{2 \times 2} & \dots & \Sigma_{\text{zyg}(t_{2F-1}, t_{2F})} \end{bmatrix},$$

This optimization problem can be solved by Newton's method.¹³ We define the heritable effects on a gene as a^2 , and the non-heritable effects on a gene as $c^2 + e^2$.

Heritable/non-heritable flow strength The total variance propagated through a regulation flow can be decomposed into heritable flow and non-heritable flow. We define the **heritable flow strength** of a regulation flow $s \rightarrow t$ as: $G(s, t) = a_s^2 \times S(s, t)$. $G(s, t)$ indicates how much of heritably driven variance is propagated by the regulation flow from source gene s to target gene t . Similarly, we define **non-heritable flow strength** $E(s, t) = (c_s^2 + e_s^2) \times S(s, t)$ as the amount of non-heritably driven variance propagated by the regulation flow.

Note that the total variance propagated by regulation flow is $G(s, t) + E(s, t) = (a^2 + e^2 + c^2) \times S(s, t)$. As we standardized all the gene expressions in all our experiments, we have $a^2 + e^2 + c^2 = 1$, so $G(s, t) + E(s, t) = S(s, t)$. $S(s, t)$ is different for each regulation flow, so we cannot directly compute $E(s, t)$ from $G(s, t)$.

3. Experiment and Result

In this section, we performed two experiments. An overview of the data source and experiment information is shown in Figure 3. The pre-processing of the data is discussed at first. The following experiment “Drug target preference” is a significance test for the overlap between genes influenced by drugs and genes driven by strong heritable effects. We found there are more drugs preferring genes driven by strong heritable effects than non-heritable effects.

In the second experiment “Drug influence flows identification”, we treat the problem of predicting drug influence flows as a classification problem using heritable, non-heritable and regulation flow strength as the classifier features. Our result shows that heritable flow strength is the best feature for drug influence flow prediction.

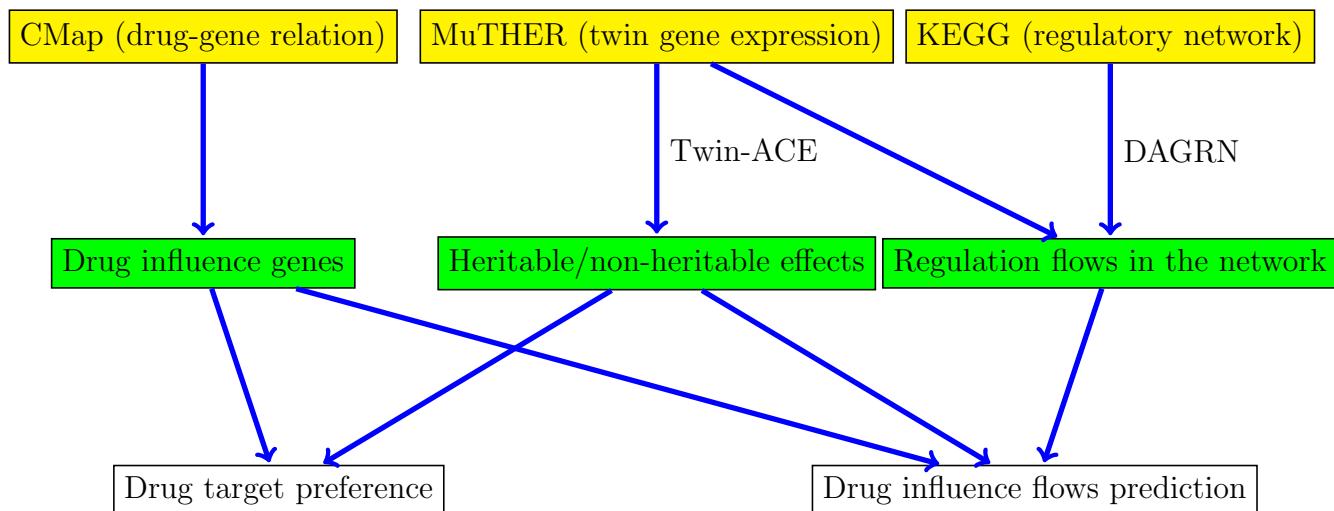


Fig. 3. A process flow chart of the relationship between data and experiments. Yellow nodes are the dataset we used. Green nodes are the information we extracted from the dataset. White nodes are the experiment we conducted.

3.1. Data pre-process

CMap database^{3,4} is used to extract effective gene-drug relationships. The CMap database maintains a rank matrix for gene's differential expression under influence of 6100 drugs. For each drug d , if gene x 's expression fold change is in the 99th percentile, we say that gene x is influenced by drug d .

Furthermore, in a regulation flow $s \rightarrow t$, if both gene s and t are influenced by a drug d , we deem the flow $s \rightarrow t$ a drug d 's influence flow.

Our twin gene expression data is acquired from MuTHER² database. The data was measured on blood samples from 276 monozygotic and 442 dizygotic twins. All the gene expression data is centered and standardized. Hence, the expression vector of each gene has mean zero and unit variance. Heritable and non-heritable effects for each gene are estimated using Twin-ACE model.

We combined a joint gene regulatory network (GRN) from 257 human signal pathways in KEGG pathway^{14,15} database. We selected the sub-network from GRN that contains genes from MuTHER and converted it into a DAGRN. We sorted the nodes based on their children count. We then removed edges conflicting with this order, that is to say edges pointing from lower ranked nodes to higher ranked nodes under the order. The transformed DAGRN from KEGG pathway contains 28600 directed edges and 3245 genes.

3.2. Drug target preference

We test the significance of the overlap between genes influenced by drugs and genes driven by strong heritable effects. 6100 drug instances from CMap database were used for the test. We calculate the heritability of each gene using the ACE parameters as $h = \frac{a^2}{a^2+c^2+e^2}$. Hence, heritability is a value between 0 and 1 indicates how much the percentage of heritable effect in total variance of the gene. The genes with heritability over 0.5 are deemed heritable genes; genes with heritability under 0.2 are deemed non-heritable genes. There are 120 heritable genes and 2199 non-heritable genes. We performed a hypergeometric significance test. There are total $U = 3249$ genes, $M = 120$ of them are heritable. For each drug with $N = 32$ (top 1%) genes influenced, where K of them are heritable genes, we calculate the p-value as the probability of having K or more heritable genes in randomly chosen N samples from total U genes. If p-value is smaller than a certain threshold, we deem the overlap between the genes influenced by the drug and heritable genes significant. We call this kind of drugs “heritable-gene-targeting drugs”. We also performed the same significance test to identify “non-heritable-gene-targeting drugs”.

The result is shown in Figure 4. When we select p-value threshold as 0.001, there are 4 heritable-gene-targeting drugs and 1 non-heritable-gene-targeting drug. The names of the drugs and the constitution of the genes influenced by the drugs are shown in Figure 5.

Figure 4 shows dramatic difference in counts of heritable-gene-targeting drugs and non-heritable-gene-targeting drugs. Hence, current drugs are more likely to target a gene driven by strong heritable effects rather than strong non-heritable effects.

3.3. Drug influence flows prediction

We estimated the regulatory network built from DAGRN using the twin gene expression data from MuTHER database. Stimulate-Response matrix is calculated for 3245 genes in the regulatory network. We selected the threshold $T = 0.3$ to be the threshold for regulation flows and extracted 233 regulation flows. There are 212 different target genes and 164 different source genes. From the CMap, we found 77 regulation flows are true drug influence flows.

To validate our assumption that the regulation flow with high heritable flow strength is more likely to be a drug effective flow, three features are compared here as a classifier of drug influence flow:

- 1. Heritable flow strength of the regulation flow: $G(s, t)$.
- 2. Non-heritable flow strength of the regulation flow: $E(s, t)$.
- 3. Regulation flow strength of the regulation flow: $S(s, t)$.

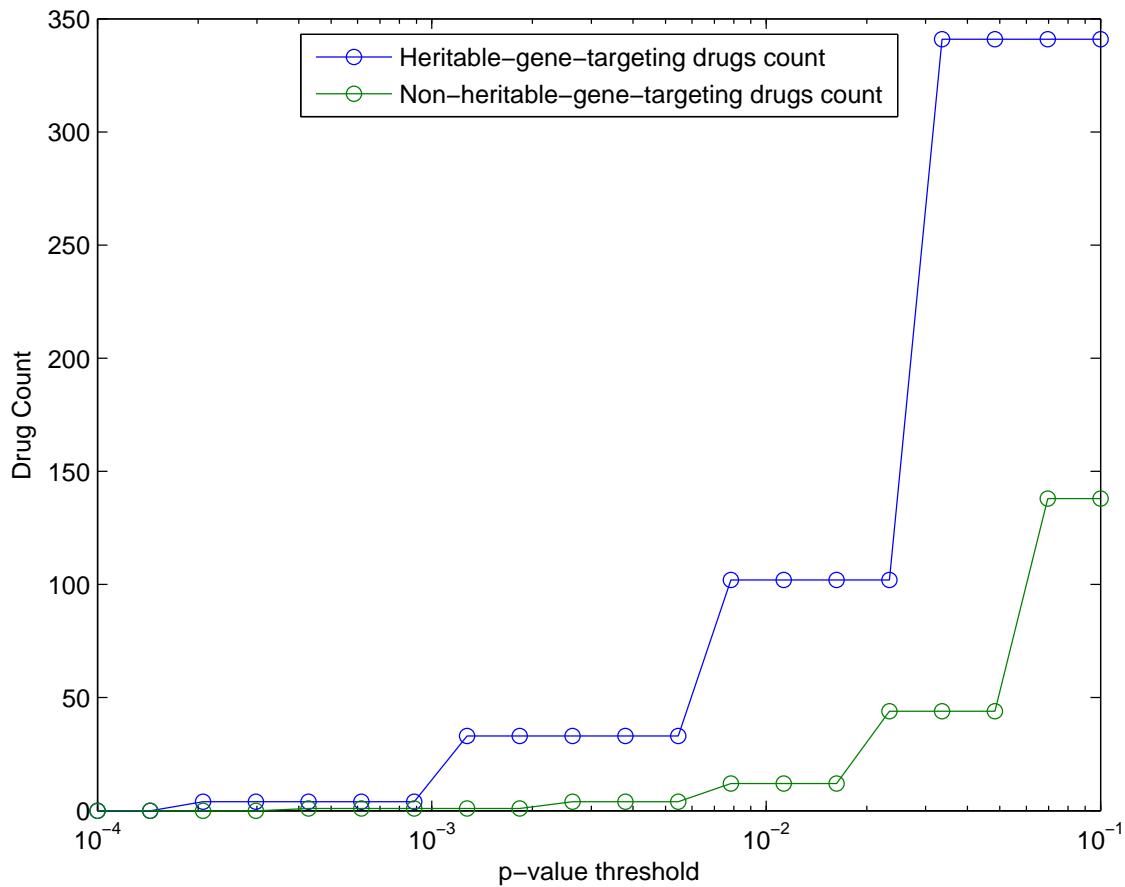


Fig. 4. The blue and green lines indicate the counts of drugs that have significant overlap with heritable or non-heritable genes when selecting different p-value thresholds.

For each feature, we selected a list of value thresholds from minimum to maximum of that feature value. If a known drug influence flow has feature value above the threshold, it is counted as a true-positive. The true-positive rate (TPR) is the ratio between true-positive count and total number of drug influence flows. If a regulation flow above the threshold is not a drug influence flow, it is counted as a false-positive. The false-positive rate (FPR) is the ratio between false-positive count and total number of regulation flows that are not drug influence flows. The TPRs and FPRs across all the thresholds construct receiver operating characteristic (ROC) curve. We use area-under-curve (AUC) as the performance metric. A higher AUC indicates the feature is better for predicting correct drug influence flows. The ROC curves for three features are plotted in Figure 6. It is obvious from the result that using the feature heritable flow strength (AUC = 0.63) is much better than non-heritable flow strength (AUC = 0.38) and regulation flow strength (AUC = 0.44) for predicting drug influence flows.

We also listed the drug influence flows ranked by heritable and non-heritable flow strength in Table 1 and 2.

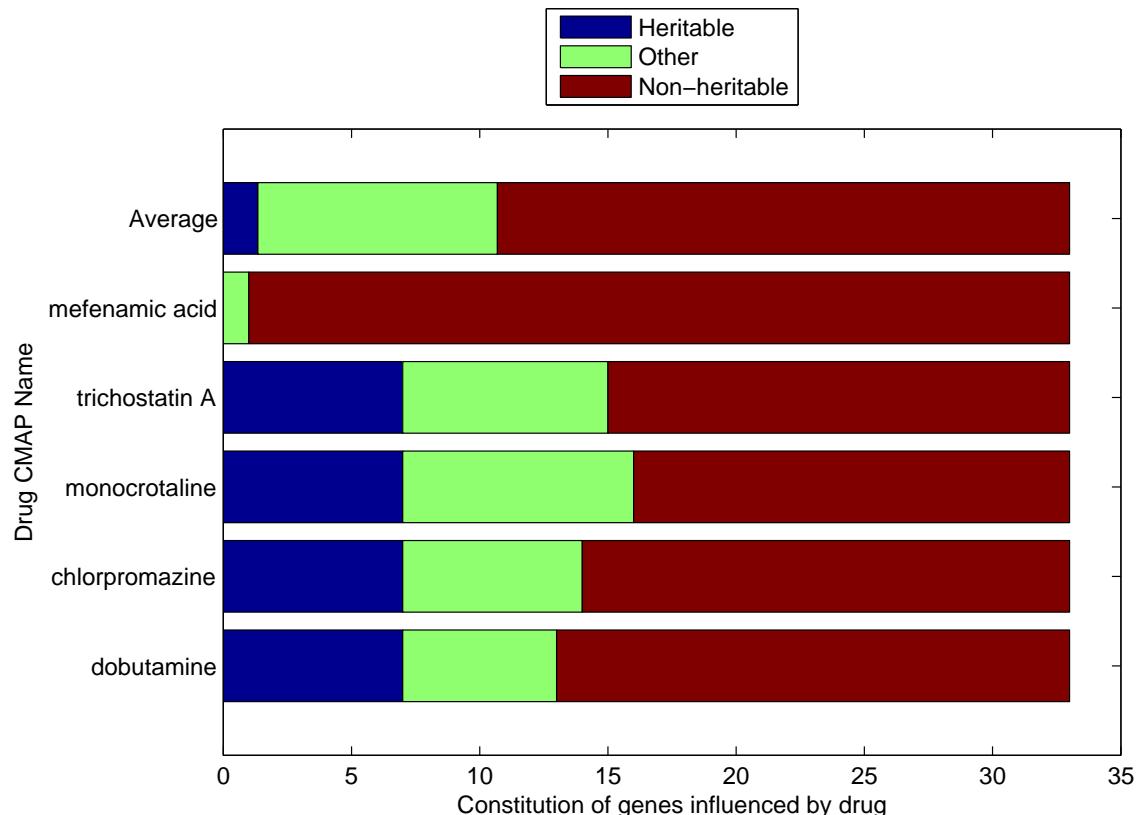


Fig. 5. Non-heritable-gene-targeting drug (“mefenamic acid”) and heritable-gene-targeting drugs “richostatin A,monocrotaline chlorpromazine,dobutamine”). The “Average” bar shows the average constitution of all genes influenced by drugs.

Table 1. Drug influence flows ranked by heritable flow strength (top 10)

Target	Source	G(s,t)	E(s,t)	drug CMap name
ATG12	FOXO3	0.43	0.33	deferoxamine,famotidine,Prestwick-860,azacitidine,bupivacaine
ARPC1B	ACTG1	0.34	0.24	mesoridazine
INADL	CLDN15	0.30	0.29	piroxicam
ARPC4	ACTG1	0.28	0.20	benzylpenicillin,zardaverine
ARPC3	ACTG1	0.25	0.18	phenazopyridine
DGUOK	GUK1	0.24	0.29	benperidol
BCL2L11	DDIT3	0.20	0.21	ChicagoSkyBlue6B
CD22	PTPN6	0.19	0.26	haloperidol,6-bromoindirubin-3'-oxime
TUBB	TUBA1C	0.18	0.65	PF-00539758-00
TK1	DUT	0.18	0.63	tanespimycin

4. Discussion

In this paper, we answered the very first question in our paper: whether a drug aiming to perturb a disease phenotype should target genes whose expression is dominated by heritable or non-heritable factors?

Even though the variance in expression of most genes in the blood sample are driven by non-heritable effects, our first experiment showed that drugs prefer to influence genes driven

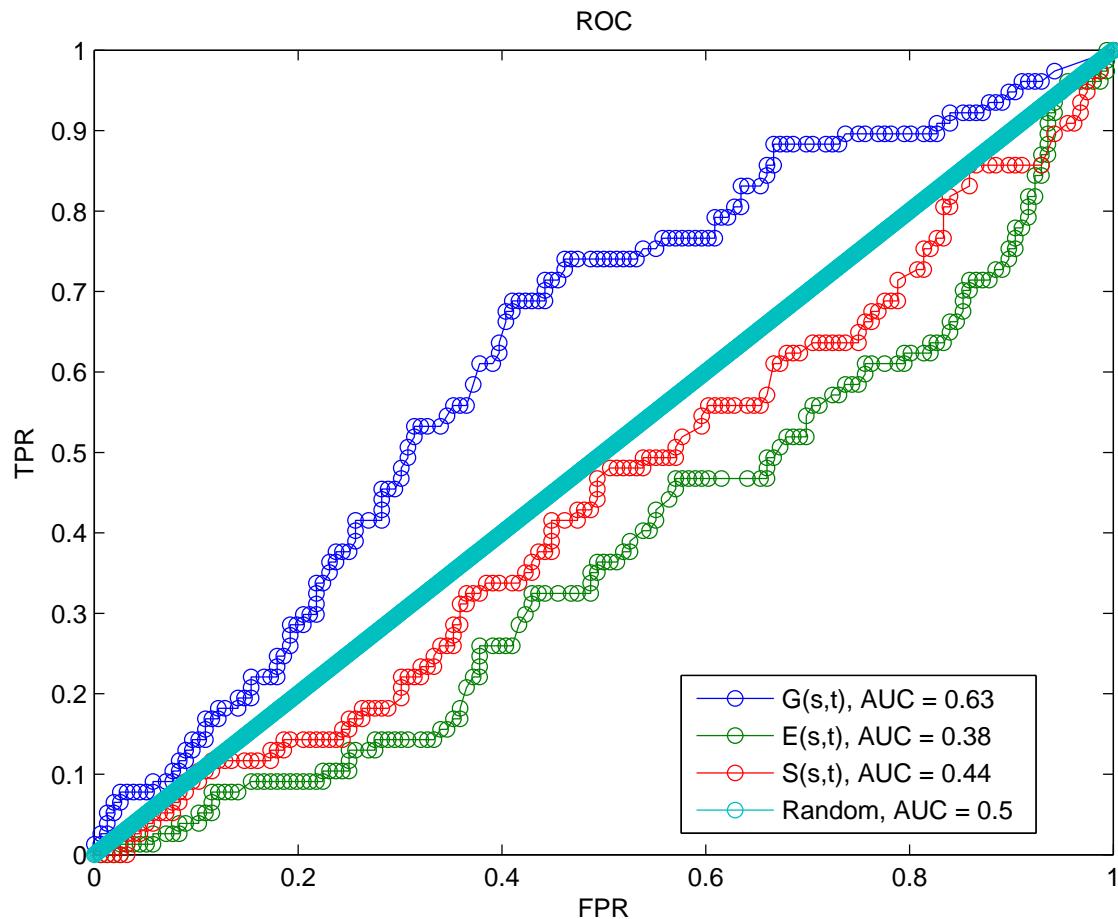


Fig. 6. The receiver operating characteristic curve and AUC for three features as classifiers for drug influence flows.

Table 2. Drug influence flows ranked by non-heritable flow strength (top 10)

Target	Source	$G(s,t)$	$E(s,t)$	drug CMap name
SUCLA2	SDHD	0.09	0.75	epirizole
ANAPC10	MAD2L1	0.04	0.67	nomifensine
TUBB	TUBA1C	0.18	0.65	PF-00539758-00
BTG3	PABPC1	0.03	0.64	deptrropine,phenoxybenzamine,oxprenolol
TK1	DUT	0.18	0.63	tanespimycin
ENOPH1	APIP	0.10	0.62	arcaine
INPP5E	PLCB2	0.12	0.60	furosemide
PFAS	GART	0.13	0.55	S-propranolol
MAP3K4	GADD45B	0.03	0.52	trichostatinA
IVD	HADHA	0.12	0.51	alfuzosin,lomefloxacin,isomethoptene,sulfaquinoxaline

by strong heritable effects rather than non-heritable effects. We then extended this observation to the background of regulatory network, where we found the regulation flow with high heritable flow strength is more likely to be a drug influence flow than flows with high non-heritable strength. The answer to the question is clear from our experiment: a drug aiming to

perturb a disease phenotype should target genes whose expression is dominated by heritable rather than non-heritable factors.

In both experiments, we identified the drugs targeting genes driven by strong heritable or non-heritable effects. These observations and discoveries can help us design drugs targeting more specific and precise regulation flows in the regulatory network to influence the final target gene's expression.

There are plenty of extensions for the current method. We can remove the DAGRN restriction and construct a model the general gene regulatory network with loops and undirected edges, as this is the most common gene regulation pathway. Another possible application is to use the Stimulation-Response matrix for any specific study in gene expression control problem with more constrained goals and resource limitations.

Acknowledgments

The authors want to thank all the reviewers from PSB 2015 for their precious suggestions.

References

1. E. Klechevsky, R. Morita, M. Liu, Y. Cao, S. Coquery, L. Thompson-Snipes, F. Briere, D. Chaussabel, G. Zurawski, A. K. Palucka *et al.*, *Immunity* **29**, 497 (2008).
2. A. C. Nica, L. Parts, D. Glass, J. Nisbet, A. Barrett, M. Sekowska, M. Travers, S. Potter, E. Grundberg, K. Small *et al.*, *PLoS genetics* **7**, p. e1002003 (2011).
3. J. Lamb, E. D. Crawford, D. Peck, J. W. Modell, I. C. Blat, M. J. Wrobel, J. Lerner, J.-P. Brunet, A. Subramanian, K. N. Ross *et al.*, *science* **313**, 1929 (2006).
4. J. Lamb, *Nature Reviews Cancer* **7**, 54 (2007).
5. E. Grundberg, K. S. Small, Å. K. Hedman, A. C. Nica, A. Buil, S. Keildson, J. T. Bell, T.-P. Yang, E. Meduri, A. Barrett *et al.*, *Nature genetics* **44**, 1084 (2012).
6. W. Sadee, *Clinical Pharmacology & Therapeutics* **92**, 428 (2012).
7. C. J. Patel and M. R. Cullen, Genetic variability in molecular responses to chemical exposure, in *Molecular, Clinical and Environmental Toxicology*, (Springer, 2012) pp. 437–457.
8. C. Cotsapas (2008).
9. M. A. Yıldırım, K.-I. Goh, M. E. Cusick, A.-L. Barabási and M. Vidal, *Nature biotechnology* **25**, 1119 (2007).
10. P. Lecca and C. Priami, *Drug discovery today* **18**, 256 (2013).
11. C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)* (Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006).
12. F. V. Rijssdijk and P. C. Sham, *Briefings in Bioinformatics* **3**, 119 (2002).
13. J. Nocedal and S. J. Wright, Quasi-newton methods, in *Numerical Optimization*, (Springer Series in Operations Research and Financial Engineering, 2006).
14. M. Kanehisa, S. Goto, Y. Sato, M. Kawashima, M. Furumichi and M. Tanabe, *Nucleic acids research* **42**, D199 (2014).
15. M. Kanehisa and S. Goto, *Nucleic acids research* **28**, 27 (2000).

AN INTEGRATIVE PIPELINE FOR MULTI-MODAL DISCOVERY OF DISEASE RELATIONSHIPS

BENJAMIN S. GLICKSBERG^{1,2}, LI LI¹, WEI-YI CHENG¹, KHADER SHAMEER¹, JÖRG HAKENBERG¹, RAFAEL CASTELLANOS¹, MENG MA¹, LISONG SHI¹, HARDIK SHAH¹, JOEL T. DUDLEY^{1,2}, RONG CHEN¹

*Department of Genetics and Genomic Sciences*¹

*Department of Neuroscience*²

Icahn School of Medicine at Mount Sinai, 1 Gustave L. Levy Pl.

New York City, NY 10029, USA

Email: Rong.Chen@mssm.edu

In the past decade there has been an explosion in genetic research that has resulted in the generation of enormous quantities of disease-related data. In the current study, we have compiled disease risk gene variant information and Electronic Medical Record (EMR) classification codes from various repositories for 305 diseases. Using such data, we developed a pipeline to test for clinical prevalence, gene-variant overlap, and literature presence for all 46,360 unique diseases pairs. To determine whether disease pairs were enriched we systematically employed both Fishers' Exact (medical and literature) and Term Frequency-Inverse Document Frequency (genetics) methodologies to test for enrichment, defining statistical significance at a Bonferroni adjusted threshold of ($p < 1 \times 10^{-6}$) and weighted $q < 0.05$ accordingly. We hypothesize that disease pairs that are statistically enriched in medical and genetic spheres, but not so in the literature have the potential to reveal non-obvious connections between clinically disparate phenotypes. Using this pipeline, we identified 2,316 disease pairs that were significantly enriched within an EMR and 213 enriched genetically. Of these, 65 disease pairs were statistically enriched in both, 19 of which are believed to be novel. These identified non-obvious relationships between disease pairs are suggestive of a shared underlying etiology with clinical presentation. Further investigation of uncovered disease-pair relationships has the potential to provide insights into the architecture of complex diseases, and update existing knowledge of risk factors.

1. Introduction

With growing genetic and epidemiological knowledge of diseases, it is becoming increasingly vital to develop tools to integrate the bodies of data to better understand clinically relevant connections among diseases. In many instances disease factors are studied independently, but when integrated they have the potential to reveal new disease understanding and accelerate translational findings.¹ Previous studies have created disease networks such as the “human disease network”², which has already been successful in identifying molecular relationships between phenotypically distinct diseases. The researchers formulated disease-specific functional modules by assessing similarity metrics between all disease genes to all genetic disorders, an approach not yet performed at such a large scale. A similar study was extremely successful in associating disease comorbidities within an extensive collection of Electronic Medical Records (EMR) to known genetic variants in complex and Mendelian disorders to infer novel information of disease etiologies³. Disease network methods are also utilized for drug repurposing⁴, in which drugs that are labeled to treat one disease are repurposed to treat another that it is linked to in the network^{5,6,7}. One such analysis utilized a computational network analytical approach to identify that the anticonvulsant topiramate was beneficial in treating irritable bowel disease⁸. Disease networks are also used to discover unidentified components of disease risk⁹.

For the current study, we accumulated a unique database of disease-causing gene variants and performed statistical analysis to determine shared genetic architecture between diseases. We then

overlaid an epidemiological enrichment analysis of co-occurrence rates in an EMR database. We hypothesize that disease pairs that have enrichment within both statistical tests are of highest interest for both known and yet unidentified connections. These unidentified connections will identify new and update existing knowledge of risk factors, elucidate disease mechanisms of action, and provide insight on relative environmental and genetic contribution to disease acquisition.

2. Methods

The workflow of the experiment is displayed in Figure 1. The major components of this analytical pipeline include the gathering, organizing, merging, and analyzing of disease-related data from multiple sources. The following sections will detail the process of each.

2.1. Data Sources

While the data (Figure 1, A) for the current project came from various sources (described below), they can be classified into two groups pertaining to the type of information they contributed: medical or genetic. Due to space constraints, every disease is shown within Figure 4.

2.1.1. Disease Ontology

Disease Ontology¹⁰ (DO) is an open-source repository for integrated information relating to human disease, including but not limited to: OMIM identifiers, International Classification of Diseases (ICD)-9-CM codes, and Unified Medical Language System (UMLS) Concept Unique Identifier (CUI) codes and has been extensively used in large-scale disease analyses¹¹. This repository was especially useful for EMR-related portion of the study, namely the disease to ICD-9 mappings. As there are many criticisms and problems with using the ICD-9-CM classification system¹², particularly when dealing with rare and/or recently discovered diseases, it was necessary to utilize a pre-curated ontology of established and documented mappings for clinical studies¹³. To address this challenge, we filtered from DO all diseases that: 1) had either a direct mapping or an exact synonym match to at least one ICD-9-CM code and 2) present in our genetic information database. At the time of acquisition (June, 2014), DO contained a total of 6,351 unique diseases, with 2,333 of them having at least one ICD-9 map directly or to an exact synonym.

2.1.2. Genetics Repository – VarDi

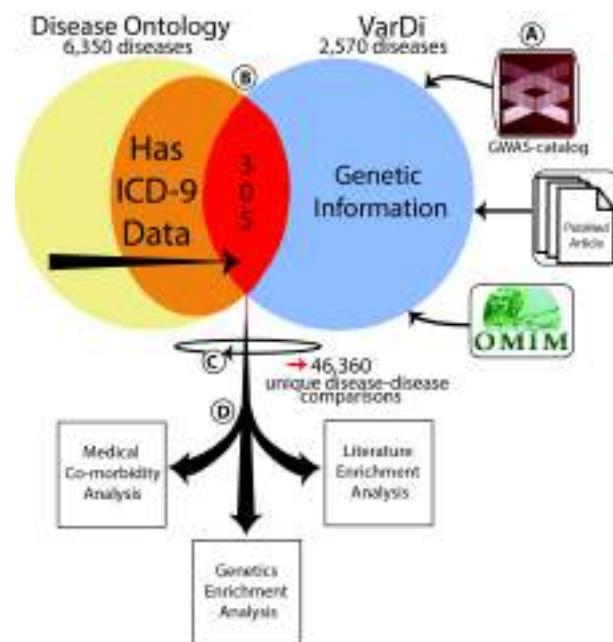


Fig. 1. The overall pipeline for the current experiment. The sources, integration, and analysis of the data are labeled alphabetically and described in detail within the text. A) data sources; B) data organization and filtering; C) unique combinations of disease pairs; D) enrichment analyses.

Our curated genotype-phenotype repository, which we have named VarDi, contains phenotype-gene-mutation mappings for 2,570 diseases. This repository combines information from public online resources, specifically GWAS-catalog¹⁴ and the Online Mendelian Inheritance in Man (OMIM)¹⁵ (acquired June, 2014), as well as a proprietary disease-variant database, built through a combination of a Hadoop-based text mining tool and manual curation. This database is comprised of 24,111 Single Nucleotide Polymorphism (SNP) mutations ($p < 1 \times 10^{-8}$) associated within 3,661 genes in 893 distinct phenotypes with all associations.

Among the online resources, GWAS-catalog was comprised of 4,831 SNPs ($p < 1 \times 10^{-8}$) within 1,838 genes in 776 phenotypes, although some being traits and not diseases. While OMIM contained 5,082 phenotypes, it is difficult to ascertain how many were distinct as subtypes of the same disease were encoded as separate entries. Nonetheless, these phenotypes encompassed 4,211 mutated genes.

2.2. Merging Datasets

With all the disease-gene variant data organized, there were three methods to connect the two repositories, as depicted in Figure 2 and Figure 1, B. From the DO repository, almost all diseases had at least one associated CUI number by which diseases in VarDi were matched to. If there was any discrepancy or more than one possible match due to multiple CUI numbers, the one with the closest name was used. Additionally, most DO entries had at least one associated OMIM code. These diseases were labeled by the DO entry and genetic info taken from each OMIM entry individually. Based on criteria listed above, there were 305 unique disease entries in total that were merged from DO and VarDi that we had both genetic data and at least one associated ICD-9 code. Most diseases were compiled using OMIM matching method (242/305), but CUI matching also added some (63/305).

2.3. Statistical Analyses

In the current study, statistical analyses were performed on every possible disease combination within the selected database to test for enrichment of three different components (Figure 1, C). From the 305 diseases in the database derived at in the previous section, there were 46,306 possible unique combinations. The three types of statistical analyses (Figure 1, D) will be described in detail in the following section.

2.3.1. Genetics Analysis

A primary aim of the current study is to elucidate novel genetic determinants or makers of diseases. Diseases that are genetically linked, in this study defined as having disease causing mutations within the same gene, can facilitate stronger understanding of each disease etiology alone and also add knowledge of new risk factors. The overarching rationale for the notion that disease gene variant association between diseases is indicative of a functional relationship is highlighted in Goh et al.'s (2007) work in generating a human disease network. In our database of 305 diseases, we compiled a total of 1,496 unique genes with disease-causing mutations. The

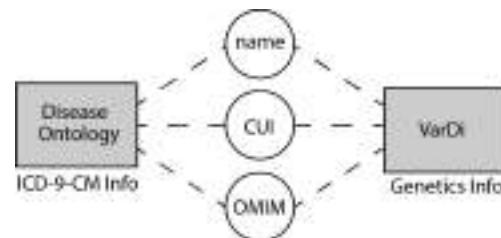


Fig. 2. The three possible components to link Disease Ontology entries to the proprietary database, VarDi.

distribution of the number of genes per disease (6.8 genes per disease on average) was not normalized which was expected due to the unbalanced mix of Mendelian and complex diseases. As expected, broader diseases, such as rheumatoid arthritis and schizophrenia, contained more genes than more specific ones, such as restrictive cardiomyopathy.

To determine this relationship between all unique disease pairs, we directly adopted an extremely informative statistical enrichment methodology from Li Li *et al.* (2014)¹⁶. This study was successful in uncovering risk factors relating to diseases and traits through shared genetic architecture. In their original study¹⁷, they first utilized the Term Frequency-Inverse Document Frequency methodology¹⁸, which weighs the relative frequency of a gene within a disease in proportion to its frequency among all diseases in the database. To test the statistical significance of these scores, we computed a False Discover Rate (FDR) by randomly shuffling (10,000 times) the genes across all diseases. The q-value was calculated as the ratio of the expected number of false positives over the total number of hypotheses tested¹⁹. For this study, the significance threshold of $q < 0.05$ was used.

2.3.2. Medical/Epidemiological Analysis

To determine if any two diseases were phenotypically linked, specifically if they co-occurred in a patient population more than one would expect by chance, we performed a statistical analysis on the patient pool within Mount Sinai Hospital's (MSH) EMR. The MSH is in a uniquely heterogeneous location and receives patients with a variety of phenotypes from diverse ethnicities. The Mount Sinai Data Warehouse (MSDW)²⁰, which houses all the clinical data, currently has 3,691,966 unique patients, over 16 million patient visits recorded, over 1.5 billion patient encounters, and 37,456,873 ICD-9 coded diagnoses documented.

Each of the 305 diseases has at least one associated ICD-9 code obtained from DO. In fact, the distribution of ICD-9 codes per disease is highly skewed towards one single code per disease, but some have multiple. While the average is 1.48 code per disease, this is highly affected by “categorical” diseases that encompass a range of codes, such as “gastrointestinal diseases” which has 60 associated codes respectively. Based on ICD-9 convention, each code could be from three to five numbers long with each proceeding number adding to specificity. As every patient is encoded with the most specific code possible (i.e. full five number code), if a disease had a code that was less than five digits, we automatically assigned that disease every possible five digit code extension.

For every possible unique disease pair combination, we performed a one-sided Fisher's exact statistical test to determine co-morbidity enrichment. The amount of patients that were observed to have at least one ICD-9 code from both diseases at any given time were compared with the amount of unique patients that had at least one ICD-9 code for each disease separately amongst a background group of any possible patient from the disease pool (559,708). A disease combination was deemed statistically significant if the resulting Bonferroni corrected p -value was less than 0.05.

2.3.3. Literature Analysis

To determine how well documented any given disease pair is in the scientific and medical world, a literature enrichment analysis was conducted using a text-mining tool. This tool queries

all abstracts and titles in PubMed for mention of a disease name using Simple Object Access Protocol (SOAP) to access NCBI's Entrez Programming Utilities²¹. A literature enrichment score was determined by performing a one-sided Fisher's exact test on the amount of articles available for each disease pair combination. Each disease was queried in quotes and each disease pair with an 'AND' operator between the two quoted terms to ensure specificity. Specifically, the test compared how many articles returned for the pair to each disease separately (number alone – number together) amongst all unique PubMed articles in the disease space (3,722,357; all diseases queried with an 'OR' operator between quoted terms). A disease combination was deemed statistically significant if the resulting Bonferroni corrected *p*-value was less than 0.05.

3. Results

Figure 3 displays the distribution of significant disease pair connections for EMR and genetics analyses with a literature score filter for pairs that were significant in both tests. As shown in *A*, there were 2,316 pairs of diseases that had a significant amount of co-morbidities in the EMR, while *B* reflects the 213 significant pairs that were enriched in the genetics analysis. *C* shows 2,251 pairs that were enriched in the EMR, but not genetically, while *D* represents the 148 pairs that were enriched genetically but not in the EMR. Key disease pairs of focus were those that had reached significance criteria in both genetic and medical enrichment analysis (*G*), specifically $q < 0.05$ and $p < 0.05/46,360$ respectively.

The two subsections, *E* and *F*, contain disease pairs that while not significant in both analyses, provide noteworthy results. By definition of the analytical procedure, all disease pairs in *G* must have both genes in common as well as co-morbidities in the EMR. The 106 disease pairs in *E* were not significantly enriched in the EMR, but had at least one co-morbidity instance. The 77 disease pairs in *F*, on the other hand, did not achieve significance in the genetics analysis, but had at least one gene in common.

Subsequently, the results in *G* were divided into two sections based on a significant literature enrichment analysis ($p < 0.05/46,360$). Out of the 65 key pairs, 19 are suggestive of novel findings as they are not established in the literature (diagonal line section) while 46 are established in the literature (horizontal line region). The purpose of this distinction is to facilitate the highlighting of putatively novel disease pairs. Accordingly, we hypothesized that pairs that are not represented in the literature, or have an insignificant literature score ($p > 0.05/46,360$) are of more interest and should be further pursued. Conversely, pairs with significant literature score ($p < 0.05/46,360$) theoretically have a recognized link or relation and would be more akin to positive controls. A selection of these results of interest (*G*) is detailed in Table 1, separated by literature significance.

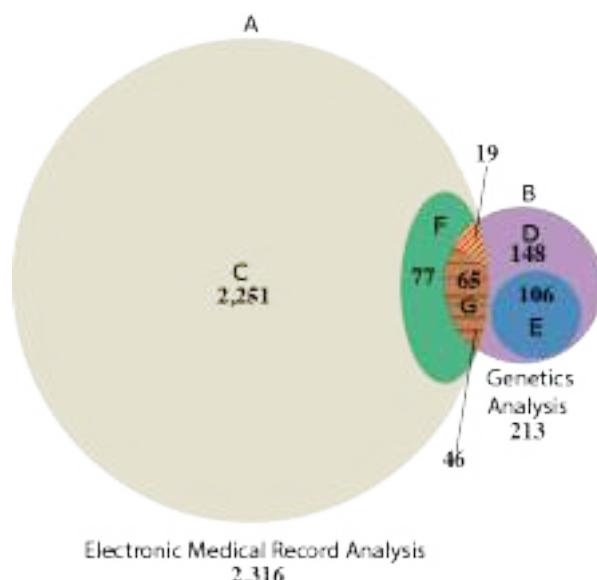


Fig. 3. The distribution of significant disease pair amounts per type of analysis. The letters and shaded regions correspond to subsets of results described in the Results section.

Table 1. Selected Statistically Enriched Disease Pairs. All possible 46,360 disease pairs were generated from the enrichment methodologies described above. The disease pairs listed in Table 1 have reached statistical enrichment in both Genetics and EMR analysis (**Gen.** $q < 0.05$, **EMR** $p < 0.05/46,360$ respectively). The top half is of disease pairs that are not enriched in the literature, while the bottom half are of disease pairs that have an enriched literature score (**Lit.** $p < 0.05/46,360$). All reported p -values pass Bonferroni correction.

Disease Pair		EMR (p)	Gen. (q)	Lit. (p)
Coronary Artery Disease	Hypothyroidism	0	0	1
Lung Cancer	Nasopharynx Carcinoma	6.35×10^{-07}	0.0476	1
Lung Cancer	Hepatocellular Carcinoma	3.16×10^{-11}	0.0168	1
Cerebrovascular Disease	Factor V Deficiency	3.80×10^{-11}	0.0022	1
Hemorrhagic Thrombocythemia	Hypothyroidism	6.67×10^{-15}	0.0092	1
Esophageal Cancer	Hypertension	6.30×10^{-17}	0.0270	1
Esophagitis	Open-angle Glaucoma	3.40×10^{-21}	0.0098	1
Colorectal Cancer	Hepatocellular Carcinoma	1.29×10^{-86}	0.0001	1
Coronary Artery Disease	Alcohol Dependence	1.26×10^{-08}	0.0106	0.9999
Asthma	Sarcoidosis	3.99×10^{-12}	0.0305	0.9999
Myasthenia Gravis	Hypothyroidism	9.78×10^{-11}	0.0099	0.9991
Celiac Disease	Hypothyroidism	1.96×10^{-22}	0.0062	0.9785
Hemochromatosis	Variegate Porphyria	3.01×10^{-12}	0.0078	0.5262
Alcoholic Cirrhosis	Hepatic Steatosis	2.86×10^{-97}	0.0017	0.0021
Hypertrophic Cardiomyopathy	Limb Girdle Muscular Dystrophy	1.04×10^{-12}	0.0365	3.523×10^{-05}
Hemorrhagic Thrombocythemia	Myelofibrosis	1.03×10^{-56}	0.0033	5.99×10^{-09}
Acute Lymphocytic Leukemia	Aplastic Anemia	3.32×10^{-34}	0.0063	9.12×10^{-14}
Velocardiofacial Syndrome	Tetralogy of Fallot	4.33×10^{-32}	0.0056	2.77×10^{-16}
Vitiligo	Hypothyroidism	1.55×10^{-09}	0.0026	2.21×10^{-17}
Systemic Lupus Erythematosus	Membranous Nephropathy	1.42×10^{-13}	0	1.21×10^{-37}
Ankylosing Spondylitis	Ulcerative Colitis	3.13×10^{-15}	0.0468	1.20×10^{-43}
Chronic Obstructive Pulmonary Disease	Lung Cancer	2.87×10^{-262}	0.0064	7.44×10^{-43}
Systemic Lupus Erythematosus	Myasthenia Gravis	5.67×10^{-07}	0.0008	1.22×10^{-57}
Hepatitis B	Primary Biliary Cirrhosis	6.18×10^{-18}	0.0395	2.11×10^{-78}
Ankylosing Spondylitis	Rheumatoid Arthritis	7.41×10^{-38}	0	0
Coronary Artery Disease	Myocardial Infarction	0	0.0125	0
Crohn's Disease	Ulcerative Colitis	0	0	0
Diabetes Mellitus	Hypertension	0	0.0075	0
Double Outlet Right Ventricle	Tetralogy of Fallot	8.24×10^{-38}	0	0
Systemic Lupus Erythematosus	Rheumatoid Arthritis	2.88×10^{-263}	0	0

3.1. Subclass Cluster Interpretation

While all disease pairs resulting in enrichment by either the genetics or EMR analyses (reflected in Figure 3) can be informative, those of highest interest are those that meet criteria in both the genetics and the EMR enrichment tests. The general distribution and amount of connections for each disease in all tests is shown in Figure 4. A disease pair that is enriched in such analyses is not only visible in a clinical population but can be further explored due to known genetic ties. Accordingly, genes that are present in one disease are natural candidates for exploration in the other.

All the 213 genetically enriched pairs (Figure 3, *B*) are interesting as they can inform the molecular mechanisms behind the relationship and the role of the gene variants in the diseases themselves. For instance, Cystic Fibrosis Transmembrane Regulator (CFTR) gene mutations have been implicated in both cystic fibrosis and bronchiectasis, which has facilitated better understanding of bronchiectasis etiopathogenesis²². This disease pair was genetically and clinically enriched in our results (EMR $p = 3.06 \times 10^{-48}$, Gen. $q = 0.0043$). Further analysis of clustering diseases into groups with common genetic overlap can inform both new disease risk approximations based on genetic testing as well as better biological insight into the mechanisms of action²³. Additionally, the 106 pairs that did not reach significance threshold in the EMR analysis but had at least one co-morbidity instance (as seen in Figure 3, *F*) are still worth consideration. If

they were trending towards significance, clinical enrichment might be achieved if a larger data set was used. In fact, 39 of the pairs had significant p -values ($p < 0.05$), but did not pass Bonferroni correction. It is certainly possible that sample size or geography concealed additional co-morbidity instances that otherwise would have made the connection significant.

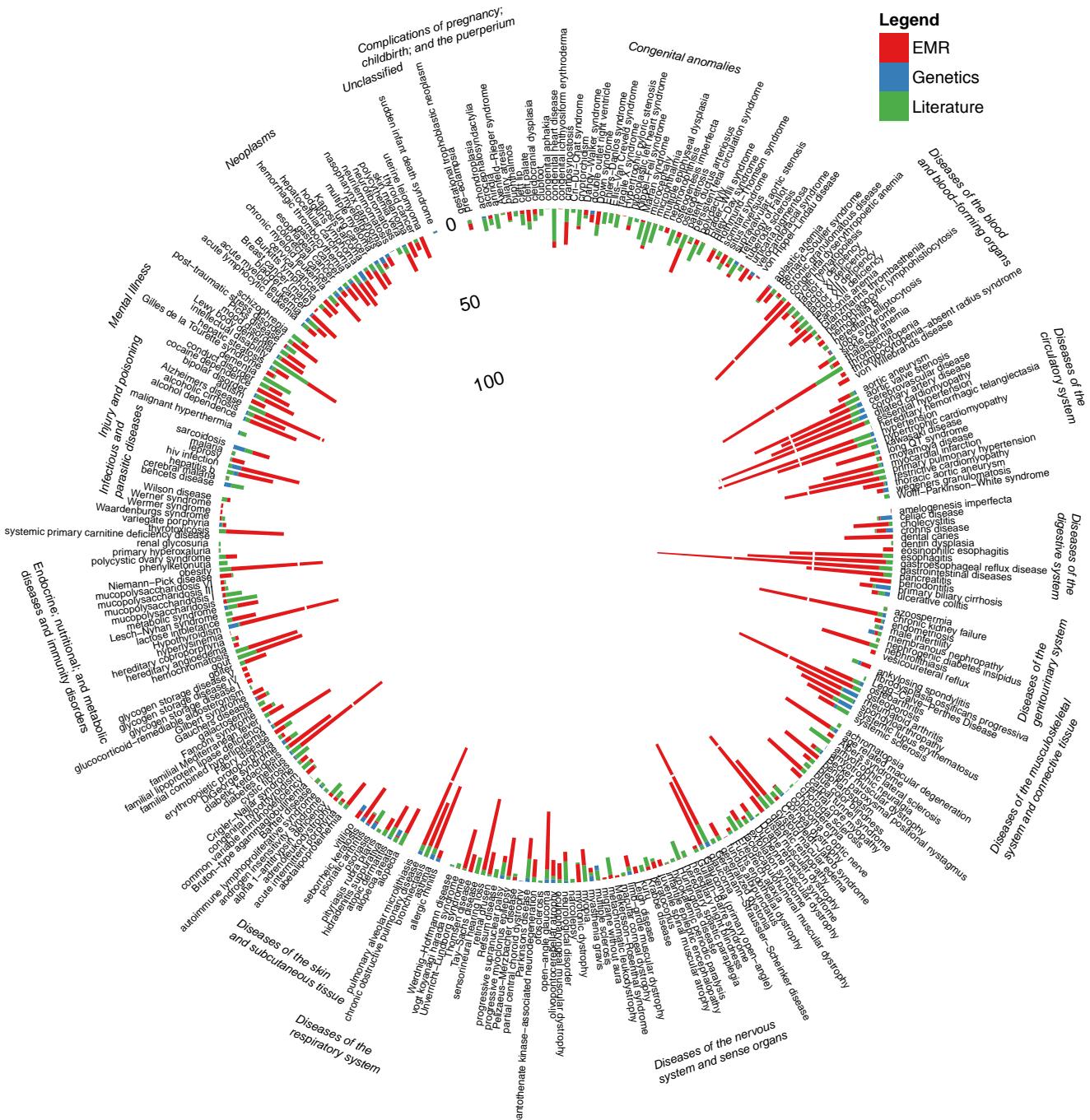


Fig. 4. The overall distribution and amount of significant pairs of each disease for EMR (red), genetics (blue), and literature (green) enrichments. The maximum amount of connections for each disease in each test is 304 (paired with every other disease). The stacked bar is the total number of connections for each test. White demarcations within stacked bar correspond to discrete, annotated y-axis numerical values.

The EMR enrichment analysis produced 2,316 significant disease pairs (Figure 3, A). With over 3.5 million patients and over 1.5 billion encounters in the MSH EMR, it is reasonable to expect the amount of connections. A quick glance at Figure 3 shows that there were over 10x more significant disease pairs in the EMR enrichment compared to genetics enrichment. This is to be expected, as much remains much unknown about the genetic etiology of diseases. All disease pairs that have EMR enrichment (*A*), both with (*G*) and without a genetic connection enrichment (*C*), could have genetic links that have not shown up because the genes themselves have not yet been implicated in the diseases. By keeping our genetic repository as updated as possible will help to expose these “hidden” genetic connections. It is clear that this indicates the need to further identify genetic etiologies of diseases.

Another hypothesis to explain the disease pair enrichment in the EMR alone is that some unrelated factor, namely environment, is responsible for the co-occurrences. This case is certainly possible for all EMR-enriched diseases (*A*), even ones that have shared genetic links (*C*). The manifestation of alcohol dependence and post-traumatic stress disorder (EMR $p=6.17 \times 10^{-66}$, Gen. $q=1$) is clearly dependent on context and situational history.

A hybrid model, which is reflective of the more realistic scenario of gene-by-environment interactions, can be especially used in cases of EMR, but not genetics, enrichment but where the disease pair shares at least one mutated gene, as in the 77 pairs displayed in Figure 3, F. One disease pair that reasonably follows this logic is coronary artery disease and obesity, which we found to be enriched in the EMR ($p=0$), but not genetically ($q=0.6025$), although they did have two mutated gene in common. Although a genetic element is believed to be involved in the co-expression of the disease pair²⁴, more emphasis has been placed on obesity itself being a high risk factor for coronary artery disease.²⁵ While genetic components for aspects of obesity are known²⁶ and suggest high heritability²⁷, environmental factors such as lifestyle and psychological mechanisms are also known to contribute largely to disease acquisition²⁸. Accordingly, components such as these for one disease will invariably affect the co-occurrence for both.

4. Discussion

Our integrative analysis pipeline identified many unidentified disease pairs that suggest novel hypothesis for clinical and biological follow-up. We gathered data for 305 diseases from various repositories and performed three statistical enrichment analyses on genetic variant overlap, clinical presentation, and literature presence on each disease pair for a total of 46,360 unique comparisons. The analyses determined whether shared genes, co-occurrence rates, and common mentions in online articles occurred between a disease pair more often than by chance. Between the three analyses in the pipeline, we produced many interesting pairs including novel associations, which have illuminating implications based on the sub-cluster to which they belong.

The general distribution of the number of significant pairs one disease has for all three analyses can be seen in Figure 4. As expected from the proportion of significant pairs in each test, the EMR enrichment analysis is overrepresented compared to the others. This display is a convenient way to view hub diseases, or diseases that have a large amount of connections. These hub diseases are particularly interesting because the sheer amount of information associated with them can reveal intricate patterns between risk factors. The current study has produced many such hub diseases that require further exploration. A brief analysis on one candidate hub disease result is encouraging of meaningful disease pairs.

Hypothyroidism is an endocrine disorder in which the thyroid gland does not produce enough thyroid hormone. Beyond the typical symptoms, there is converging evidence of some of subsequent effects that are important to understand for proper treatment of the condition. In the current study, hypothyroidism is genetically and epidemiologically linked to many diseases, particularly autoimmune diseases listed in Table 1. Of particular interest is the connection to coronary artery disease (EMR $p=0$, Gen. $q=0$, Lit. $p=1$), where variations in five genes are common to both (*PTPN11*, *ATXN2*, *SH2B3*, *NAA25*, and *C12orf51*). While the literature score implies an unknown connection, there are recent studies that highlight the cardiovascular effects²⁹ of thyroid hormone levels and gland function. Specifically, hypothyroidism has been implicated as a risk factor for coronary artery disease³⁰, impairing cardiac function through mechanisms such as impaired vasodilation³¹, which can lead to atherosclerosis³². Interpretation of findings such as these will be explored in subsequent sections.

4.1. Efficacy of the literature analysis

The literature enrichment portion of the study was used less as an independent measure and more as a way to highlight significant disease pairs that are relatively unknown to direct focus. In this sense, the literature enrichment score was useful and effective. Out of the 65 pairs that were significant in EMR and genetics analysis, 46 were and 19 were not enriched in the literature (p cut-off = 0.05/46,360). Based on our hypothesis, the former group would be already established and can serve as positive controls while the latter would be strong candidates for new findings. In fact, all disease pairs that were of the same characterization root (*i.e.* distal muscular dystrophy and limb-girdle muscular dystrophy) were captured in this filter. Overall the distinction was helpful but there were obvious oversights in each group. After manual review, almost all of the 46 literature enriched pairs had links or were related. Conversely, while many disease pairs in the non-enriched group had known links or relationship, the majority of the unrecognized connections were in this section. These indicated pairs will be discussed in the next section.

One clear limitation to this methodology is the syntax regarding the search criteria. Specifically, we did not filter negative mentions of the diseases, *i.e.* “disease x , but not disease y ,” which is actually the opposite of the intended effect. This issue will be addressed in future iterations of the pipeline. The inherent specificity of a disease name can also confound enrichment. Coronary artery disease and alcohol dependence, for instance, were found to be non-enriched in the literature, suggesting that the connection between them is not well documented. The search term ‘alcohol dependence’ could not be the best representation of the phenotype. Searching ‘Coronary artery disease’ and ‘alcoholism’ in PubMed produced 106 articles, instead of six yielded by the former query. This issue will only affect a small subset of the diseases, however. Further refining of the tool can rectify these issues via search optimization protocols.

4.2. Efficacy of the EMR enrichment analysis

It is important to note, however, that there are some clear factors that might have confounded this EMR analysis and led to spurious correlations. First, while MSH is in a location with a uniquely heterogeneous population, almost all patients are from the New York City area, which can introduce geographic disease bias. Another potential biasing factor has to do with limitations of how diseases are encoded and recorded. Some diseases are so rare or recently discovered that distinct codes do not exist for them. Crigler-Najjar syndrome and Gilbert’s syndrome, for instance,

are similar but different disorders that both are encoded by the ICD-9-CM code 277.4, “Disorders of bilirubin excretion.” Their high level of co-occurrence in the EMR (exact overlap to be specific) is a confounder due to lack of disease specificity and does not have any actual applicability. Unfortunately this issue cannot be readily addressed but will undoubtedly be less problematic when ICD-10 codes can be used instead.

4.3. Notable Significant Disease Pairs

For each of these key 65 pairs (Figure 3, G), extensive further evaluation was conducted to determine impact and relevance. Due to space constraints, a representative subset of notable pairs is displayed in Table 1. As mentioned, these results were split into clusters based on literature enrichment scores that provided a somewhat useful first pass filter to highlight previously unexplored disease pairs.

4.3.1. Disease pairs that are well established act as strong positive controls for the pipeline

Disease pairs across all disease categories that were enriched in the literature (and a few that were not) produced positive control connections that are well established and some that have recently been discovered. The relationship between Coronary artery disease and myocardial infarction is well established³³, and it is encouraging that our pipeline produced strong significance in all three analyses for this pair (EMR $p=0$, Gen. $q=0.0125$, Lit. $p=0$). Similarly, ulcerative colitis and Crohn’s disease are known to be highly related³⁴ and showed up to be strongly significant in all three analyses (EMR $p=0$, Gen. $q=0$, Lit. $p=0$). Diabetes mellitus and Hypertension are also well linked and have been shown to co-occur more often than expected by chance³⁵ (EMR $p=0$, Gen. $q=0.0075$, Lit. $p=0$). Chronic obstructive pulmonary disease (COPD) and lung cancer are both common diseases amongst smokers³⁶ and patients with COPD are at increased risk for developing lung cancer³⁷ and have a robust link in our results (EMR $p=2.87 \times 10^{-262}$, Gen. $q=0.0064$, Lit. $p=7.44 \times 10^{-43}$). Furthermore, recent studies have offered hypotheses to explain the common origins to these “anatomic and functionally disparate diseases.”³⁸ This is a clear example of how the current pipeline can be used to provide context for both genetics and epidemiological prevalence. Tetralogy of Fallot and velocardiofacial syndrome have both genetic links³⁹ and prevalent co-occurrences⁴⁰ and were enriched in our results (EMR $p=4.33 \times 10^{-32}$, Gen. $q=0.0056$, Lit. $p=2.77 \times 10^{-16}$).

4.3.2. Unknown aspects of disease pairs are candidates for further analysis

One of the best possible uses for this pipeline is for identifying candidate gene variants that can explain the link between diseases that have prevalent comorbidities. Esophageal cancer and hypertension are observed in patients⁴¹⁴², but little is known about how they are related. As expected, this pair was not enriched in the literature, but had strong EMR and genetic overlap (EMR $p=6.30 \times 10^{-17}$, Gen. $q=0.027$, Lit. $p=1$). Similarly, esophagitis and open-angle glaucoma do not often show up together in the literature but there is also very little, if any, information documenting any sort of connection between the two, yet our pipeline found a connection (EMR $p=3.40 \times 10^{-21}$ Gen. $q=0.0098$, Lit. $p=1$).

5. Conclusions and Future Directions

The current study combined clinical, genetics, and literature analytical methods to create a pipeline to identify key disease pairs of interest. With this initial iteration of the pipeline, we identified 2,316 and 213 disease pairs that were enriched in the EMR and for shared genetic variants, with 65 in both. Using the analytical approaches listed above, we are able to infer new insights about mechanistic origin, molecular pathways, and risk factors for such pairs. A component to easily adapt to the pipeline is to stratify data based on ethnicity for disease expression to determine if disease pair prevalence is uniform or specific to particular ancestries. We also have categorical data of disease type that we will perform enrichment analysis on to establish if certain connections are more common in certain disease classes. More specific future direction plans are listed briefly in the final section.

5.1 Predicting disease risk through temporal co-morbidity analysis

A large component of Li *et al.*'s work is determining the time course of disease onset in related pairs to identify which is the causal risk factor. For all medically enriched pairs in our pipeline, we plan on incorporating the time line for the disease manifestation to see which is a risk for the other. For instance, we would hypothesize, based on current knowledge,⁴³ that obesity would manifest before diabetes.

5.2 Utilizing connections between diseases for drug related analysis

Another useful type of information that can be derived from this analysis is the case where the treatment for one disease can possibly cause the other. While this pipeline shows clinical presentation relationships, it does not tease out how the disease pairs are related. Systemic lupus erythematosus (SLE) and myasthenia gravis (MG) have an association, albeit a rare one, with a possible common biological explanation along with clinical co-presentation.⁴⁴ A review analyzed the association between these two diseases and offered the possibility that hydroxychloroquine, a drug typically used for the treatment of SLE, could have induced MG, at least in one patient.⁴⁵ Incorporating patient medication information into our pipeline, something that is feasible, in future iterations will be able to uncover these possible scenarios.

6. Acknowledgments

We would like to thank Mount Sinai's Scientific Computing Team for server cluster IT support. This research was supported by Institutional Grants of Rong Chen and Joel Dudley.

References

1. N. H. Shah, *et al.*, *BMC Bioinformatics*, **10** (2009).
2. K. I. Goh, *et al.*, *Proc Natl Acad Sci U S A* **104**, 8685 (2007).
3. D. R. Blair, *et al.*, *Cell*, **155**, 70 (2013)
4. S. Suthram, *et al.*, *PLoS Comput Biol*, **6** (2010).
5. F. Iorio, *et al.*, *Proc Natl Acad Sci U S A*, **107**, 14621 (2010).

6. J. T. Dudley, *et al.*, *J Cardiovasc Transl Res*, **3**, 438 (2010).
7. L. Yang and P. Agarwal, *PLoS One*, **6** (2011).
8. J. T. Dudley, *et al.*, *Sci Transl Med*, **3**, 96 (2011).
9. A. Goris and A. Liston, *Cold Spring Harb Perspect Biol*, **4** (2012).
10. L. M. Schriml, *et al.*, *Nucleic Acids Res*, **40**, 940 (2012).
11. K. Shameer and R. Sowdhamini, *J Clin Bioinforma*, **2**, 8 (2012).
12. A. Hazlewood, *American Health Information Management Association* (2003).
13. K. Shameer, *et al.*, *Hum Genet*, **133**, 95 (2014).
14. L. A. Hindorff, A Catalog of Published Genome-Wide Association Studies [<http://genome.gov/gwastudies>].
15. Online Mendelian Inheritance in Man, OMIM [<http://omim.org>].
16. L. Li, *et al.*, *Sci Transl Med*, **30**, 57 (2014).
17. L. Li, *et al.*, *Pac Symp Biocomput*, 224 (2013).
18. H. C. Wu, *et al.*, *AcM Transactions on Information Systems*, **26**, 13 (2008).
19. J. D. Storey and R. Tibshirani, *Proc Acad Sci U S A*, **100**, 9440 (2003).
20. Mount Sinai Data Warehouse [<https://msdw.mountsinai.org/>].
21. E. Sayers, E-utilities (2009) [<http://eutils.ncbi.nlm.nih.gov>].
22. P. F. Pignatti, *et al.*, *Hum Mol Genet*, **4**, 635 (1995).
23. C. Cotsapas, *et al.*, *PLoS Genet*, **7** (2011).
24. J. D. Brunzell, *Arterioscler Throb Vasc Biol*, **4**, 180 (1984).
25. R. H. Eckel and R. M. Krauss, *Circulation*, **97**, 2099 (1998).
26. C. Bouchard and L. Pérusse, *Annu Rev Neur*, **13**, 337 (1993).
27. J. Wardle, *et al.*, *Am J Clin Nutr*, **87**, 398 (2008).
28. J. O. Hill and J. C. Peters, *Science*, **280**, 1371 (1998).
29. B. Biondi and I. Klein, *Endocrine*, **24**, 1 (2004).
30. O. Mayer Jr., *et al.*, *Vasc Health Risk Manag*, **2**, 499 (2006).
31. J. Lekakis, *et al.*, *Thyroid*, **7**, 411 (1997).
32. Y. Zhang, *et al.*, *Anterioscler Thromb Vasc Biol* (2014).
33. E. G. Nabel and E. Braunwald, *N Engl J Med*, **366**, 54 (2012).
34. J. D. Doecke, *et al.*, *Inflamm Bowel Dis*, **19**, 240 (2013).
35. M. Epstein and J. R. Sowers, *Hypertension*, **19**, 403 (1992).
36. E. Potton, F. McCaughan, and S. James, *Resp Med: COPD Update*, **5**, 34 (2009).
37. S. Raviv, *et al.*, *Am J Respir Crit Care Med*, **183**, 1138 (2011).
38. A. M. Houghton, M. Mouded, and S. D. Shapiro, *Nat Med*, **14**, 1023 (2008).
39. F. Amati, *et al.*, *Hum Genet*, **95**, 479 (1995).
40. D. Young, R. J. Shprintzen, R. B. Goldberg, *Am J Cardiol*, **46**, 643 (1980).
41. L. B. Koppert, *et al.*, *Eur J Gastroenterol Hepatol*, **16**, 681 (2004).
42. T. Kato, *et al.*, *Hepatogastroenterology*, **48**, 1656 (2001).
43. J. M. Chan, *et al.*, *Diabetes Care*, **17**, 961 (1994).
44. G. Vaiopoulos, *et al.*, *Postgrad Med J*, **70**, 741 (1994).
45. M. Jallouli, *et al.*, *J Neurol*, **259**, 1290 (, 2012)

PEAX: INTERACTIVE VISUAL ANALYSIS AND EXPLORATION OF COMPLEX CLINICAL PHENOTYPE AND GENE EXPRESSION ASSOCIATION

MICHAEL A. HINTERBERG, DAVID P. KAO, MICHAEL R. BRISTOW, LAWRENCE E. HUNTER,
J. DAVID PORT, and CARSTEN GÖRG

School of Medicine, University of Colorado, Aurora, CO 80045, USA

E-mail: {michael.hinterberg, david.kao, michael.bristow, larry.hunter, david.port, carsten.goerg}@ucdenver.edu

Increasing availability of high-dimensional clinical data, which improves the ability to define more specific phenotypes, as well as molecular data, which can elucidate disease mechanisms, is a driving force and at the same time a major challenge for translational and personalized medicine. Successful research in this field requires an approach that ties together specific disease and health expertise with understanding of molecular data through statistical methods. We present PEAX (Phenotype-Expression Association eXplorer), built upon open-source software, which integrates visual phenotype model definition with statistical testing of expression data presented concurrently in a web-browser. The integration of data and analysis tasks in a single tool allows clinical domain experts to obtain new insights directly through exploration of relationships between multivariate phenotype models and gene expression data, showing the effects of model definition and modification while also exploiting potential meaningful associations between phenotype and miRNA-mRNA regulatory relationships. We combine the web visualization capabilities of Shiny and D3 with the power and speed of R for backend statistical analysis, in order to abstract the scripting required for repetitive analysis of sub-phenotype association. We describe the motivation for PEAX, demonstrate its utility through a use case involving heart failure research, and discuss computational challenges and observations. We show that our visual web-based representations are well-suited for rapid exploration of phenotype and gene expression association, facilitating insight and discovery by domain experts.

Keywords: personalized medicine, hypothesis testing, visual analytics, gene expression, multidimensional data exploration.

1. Introduction

The diversity of modern multidimensional clinical data enables researchers to define and study subtle, complex sub-phenotypes of patients. For example, rather than defining mere presence or absence of disease, certain types of cancer can be more precisely graded or staged when specific gene sequences and expression characteristics are known [1], and patients with different genotypes may exhibit differential response to drug treatment [2]. Phenotype characterization is a broad challenge in the field of phenomics [3], with much recent effort (e.g, [4]) towards associating phenotype with single nucleotide polymorphisms (SNPs) in genome-wide association studies (PheWAS). But the art of meaningful classification using all available clinical and genetic data still requires significant domain expertise, especially when novel, complex sub-groups are defined. Consequently, clinical domain experts often define phenotypes and generate hypotheses using a certain set of tools and manually curated knowledge sources, whereas statisticians separately perform statistical analyses using an appropriate set of tools, most of which are poorly suited for phenotype definition. This cyclical process involves different users and tools, and therefore tends to be too slow and tedious for efficient collaborative work, presenting a major challenge in translational medical research.

From our collaboration with clinical experts and statisticians we have derived three primary observations regarding this inefficient workflow. First, datasets rich in both clinical and gene expression data may lead to novel insights when a specific, interesting phenotype is defined, and that pheno-

type is associated with the expression data in a biologically plausible manner. Second, data can and should be treated differently based on the domain knowledge and understanding of the audience as well as natural relationships and modeling techniques that are most appropriate for a particular class of data. Finally, an integrative analysis that is both visual and dynamically responsive is more likely to facilitate an iterative and collaborative analytical process that can generate useful insight than text-based scripts and comparisons of static data representations.

Given these observations, we have developed PEAX (Phenotype-Expression Association eXplorer), which allows domain experts to define and explore novel sub-phenotype correlation with gene expression using visual analytics. By integrating phenotype definition with statistical processing in a single tool on a single screen, we seek to inspire novel insight from exploratory analysis, in a more intuitive and collaborative approach than existing tools. We describe the iterative, agile development process driven specifically by use case requirements of cardiology experts; however, we also address general, fundamental challenges identified in personalized medicine involved in clinical phenotype definition, multiple testing, and integrated analysis of heterogenous and missing data [5]. Although issues such as “data dredging” and sparse data cannot be completely avoided, providing clear, comprehensive, and responsive metrics to a domain expert formulating a hypothesis may mitigate some of these effects. With PEAX, we contributed a tool that integrates web visualization with statistical analysis, and its application to a specific biomedical task shows that interactivity and responsiveness can improve existing methods and workflows for data analysis and exploration.

2. Domain Background

The motivating clinical research for PEAX is the analysis of drug efficacy in the clinical trial on the “Effect of β -blockers on Structural Remodeling and Gene Expression in the Failing Human Heart (BORG)” [6]. Heart failure has a devastating and costly impact within the United States, and is responsible for one million hospital visits and 280,000 deaths annually [7]. Patients enrolled in BORG were diagnosed with idiopathic dilated cardiomyopathy (IDCM), a form of heart failure primarily affecting the left ventricle of the heart, and were randomized to one of three different β -blocker treatments. The patients were monitored for up to a year from initial treatment; they exhibited a variable improvement in left ventricular ejection fraction (LVEF) with β -blocker treatment as observed previously [8]. Biopsies from ventricular tissue for each patient were performed prior to β -blocker treatment and at 3 and 12 months; myocardial gene expression of $\sim 34,000$ human mRNAs and ~ 7800 miRNAs was measured, producing longitudinal *in vivo* whole-transcriptome gene expression data in human IDCM patients. The clinical outcome used to measure drug response was improvement in LVEF; the primary aim of the study was to understand molecular mechanisms of LVEF improvement with β -blockers and to identify predictive clinical and/or molecular biomarkers to predict LVEF improvement.

Although the patient size is relatively small, the depth of the data represents a typical clinical trial scenario involving a primary research question of a single outcome tested against thousands of potential biomarkers. Previous analysis of BORG data included standard t-testing of associations between a subset of mRNA and miRNA probes with differential expression changes between responders and non-responders to accomplish the primary aim of the study. A previous collaborative analysis [9] used the machine-learning software package Weka [10] to discover predictive C4.5 trees

for miRNA expression that may be associated as a biomarker for β -blocker drug response. In this prior work, decision trees were seen to be a simpler and accurate predictive model compared to machine learning methods such as support vector machines and random forest, yielding a model of drug responsiveness that was better accepted by cardiologists for description of phenotype. Additional analysis of learning was limited in analyzing the vast potential search space of clinical phenotype and molecular interaction, so that only a fraction of the data were used that could potentially provide knowledge regarding heart failure, as well as the mechanisms of disease and repair (remodeling and reverse remodeling, respectively).

Primary results of the BORG data analysis and new research in the field prompted additional research questions. However, there was a desire to shorten the loop between hypothesis generation and testing, and to move cardiology experts closer to the data analysis phase. In some cases, merely understanding the cohort size and distribution of patients that met specific criteria was important in deciding whether to proceed with further analysis. These research questions, and anticipation of further data exploration, motivated the development of a new workflow.

3. Related Work

Several toolkits can present statistical analysis with **visualization**. Rattle [11] uses a GUI to provide support for data exploration and output of statistical testing. It is useful for general statistical analysis methods but is not particularly enhanced for biomedical data analysis. Tools like JavaStat [12] provide a Java/R interface to support the combined development in Java and R and potentially harness the strengths of both languages. However, development requires an integrated mix of Java and R code; as a result the functional code is not as easily portable and leveragable from existing tools, new tools, and legacy code from user groups that are written natively in either of the individual languages. Several tools address the desire for a GUI by running R as a web server, and providing a web interfacing API and functionality. These include server tools such as RApache [13], and R packages like RServe [14]. RStudio's Shiny [15], on the other hand, provides a framework that supports an R analytical backend that can be tied to browser-based visual displays as well, but also handles the interactivity between visual inputs and outputs, so that extensive coding is not required for this process, while analytical R scripts can generally be leveraged from native R-based projects.

Clinical **hypothesis testing** uses established statistical methodologies, by separating patients into differentiated groups, defined by distinct, measurable features *a priori*; for example, drug responders vs. non-responders. Candidate features, such as mRNA expression are tested for significant differences in expression between patient classes. The popular analysis pipeline Bioconductor [16] contains functions, such as edgeR [17] for differential gene expression. The Gene Expression Omnibus (GEO), a public repository for gene expression experiments, includes online tools such as heatmaps for analysis and display of differentially expressed genes [18]. The actual statistical testing of differential expression in these cases is used to answer specific hypotheses about phenotype-biomarker association.

Testing for associations between defined classes and potential predictors is also generally supported by **machine learning** techniques. Machine learning can be used to build models based on a subset of features, such as gene expression data, that are used to classify or predict membership of an instance, such as a patient, in a defined class group. Weka [10] is a popular tool that wraps sup-

port of several supervised learning techniques, such as C4.5 trees [19], random forests, and support vector machines. Predictive models can be tested and cross-validated. These techniques are based on settled definition of phenotype, so that each model must be individually and iteratively tested for association, requiring an additional step or solution for exploration.

Phenotype discovery can be considered generally as sub-group of latent class analysis, which is a technique applied not just in biological sciences, but also in social and behavioral sciences [20]. Such grouping can be aided by cluster analysis, which is a type of unsupervised learning method that takes all of the known features, and groups patients into distinct clusters based on aggregate similarity using selected features and defined metrics. Clustering has been used for identifying novel disease phenotypes [21]. R contains functions such as hclust and knn, and Weka contains clustering support as well. But unsupervised clustering may result in groupings that are driven by features which are less important or relevant to the discovery process, so feature selection, interpretation and refinement by an expert is still critical, especially in large datasets. Topological data analysis is also a powerful way to explore high-dimensional data, but tools like the Ayasdi Platform [22] are not open source, and such models are not used as widely in clinical medicine as decision trees.

Because of the regulatory relationship between miRNA and gene expression, simultaneous profiling and **integrated analysis** of expression data is useful in further understanding regulatory networks. Experiments that include such profiling look for anti-correlated expression of miRNA and respective mRNA targets, which may be searched in aggregated databases in packages like multiMiR [23] that contain information about observed and predicted miRNA-mRNA interactions. Tools such as mirConnx [24] use this prior knowledge to construct regulatory networks, which can be augmented by expression data from a particular experiment. In general, once a candidate list of potential interacting miRNA/genes is obtained, they can serve as inputs to other methods to look for enrichment that suggests biological interaction. The cBioPortal for Cancer Genomics [25] provides visualization and analysis tools, but this support is tied directly to cancer datasets only.

4. Analysis Task Definition

To define features for our system, as well as prototype, test, and refine, we began with two research questions that supported aims of the BORG study. Even though we use BORG as a driving dataset, these tasks are likely to be supportive of similar research that is rich in clinical, mRNA, and miRNA expression data.

Since BORG utilized three different types of the same class of drug, we want to analyze whether patients with specific drug treatments exhibit different fold changes in gene and/or miRNA expression. Different drugs in the same class may function differently in different patients, potentially exhibiting different side effects and, at a basic level, affect changes in gene expression. First, we sought the ability to compare molecular expression change and association after drug treatment when all patients on β -blocker treatments were analyzed in a pooled fashion, versus a separate analysis of patients grouped by one of three specific β -blocker drugs. Second, we want to test whether drug receptor polymorphisms exhibited different associations with molecular expression data. Even a single nucleotide difference can have drastic effects on individual response to drugs. Therefore, comprehensive understanding of differential drug response should ideally include genotype information of important drug-related SNPs. The BORG dataset includes SNP data of several adrenergic

receptors (e.g. *rs2234888*, *rs1801252*, and *rs1801253*), known or suspected to affect β -blocker response [26, 27]. SNP variants are represented as a categorical variable representing whether a given SNP locus is heterozygous, or either of the homozygous combinations. For the second analysis task, we desired to stratify patients by genotype information for at least one of the SNPs, and compare resulting gene associations.

5. Visual Analytics Approach

The overall goal of our design is to enable informed clinicians who are experts in a certain disease domain to generate and test hypotheses regarding user-specified phenotypes and their associations with gene and miRNA expression data; it was motivated by our previous collaborative work with researchers in the Division of Cardiology in the CU Medical School. In order to bridge the statistical, scientific, and clinical analysis of the BORG dataset, we used a visual analytics approach. The visual analytics principles [28] in PEAX are used to integrate broad and heterogeneous data, and present the analysis results at a variable level of detail in order to facilitate insight from clinical experts. An improved analysis process also allows a group of users to obtain more rapid feedback regarding disease biomarkers and pathological processes, reducing the need for independent work by a dedicated statistician during hypothesis formation and refinement.

We now discuss our design methodology and infrastructure, including some of the visual analytics design choices used in PEAX. (The tool, demo, and code are available at <http://compbio.ucdenver.edu/PEAX>). Based on our initial observations, several major design goals and constraints were immediately apparent for our analysis engine:

- GUI for responsive phenotype definition (*Input*)
- Support for powerful statistical analysis (*Processing*)
- Visual display of processed data (*Output*)

Additional considerations included use of open-source software to provide maximum capability of dissemination and re-use in the research community, and use of web-based applications to reduce installation and platform-dependence overhead. Given these considerations, R is a compelling choice for statistical analysis software, due to popularity and speed of statistical analysis. Using R would also ensure the ability to adapt and integrate ever-increasing library of analysis packages, and R-based analysis results were familiar to our domain experts. However, GUI support is not as well-supported in native use of R.

In order to wrap the power of R statistical analysis in a web-friendly GUI, we employed RStudio's Shiny software as a visual front-end. It provides the ability to design webpages easily with responsive controls and full access and use of Javascript user-interface elements, as well as HTML capability, tied to an R-backend that could be used for data processing. Shiny also provides the capability of "reactive" inputs and outputs, in which changes in inputs can automatically trigger processing events, while newly updated data and UI input elements can be used to trigger updates of data outputs. Finally, we expanded our visual output display by using Data Driven Documents (D3), which is a powerful Javascript-based system for data visualization [29].

Taken together, Shiny allowed us to integrate GUI and data inputs on a webpage. These inputs are visible by R and used to construct a model. The model is processed and tested using statistical algorithms in R, and then displayed as webpage outputs using HTML, Javascript, and D3. Using a

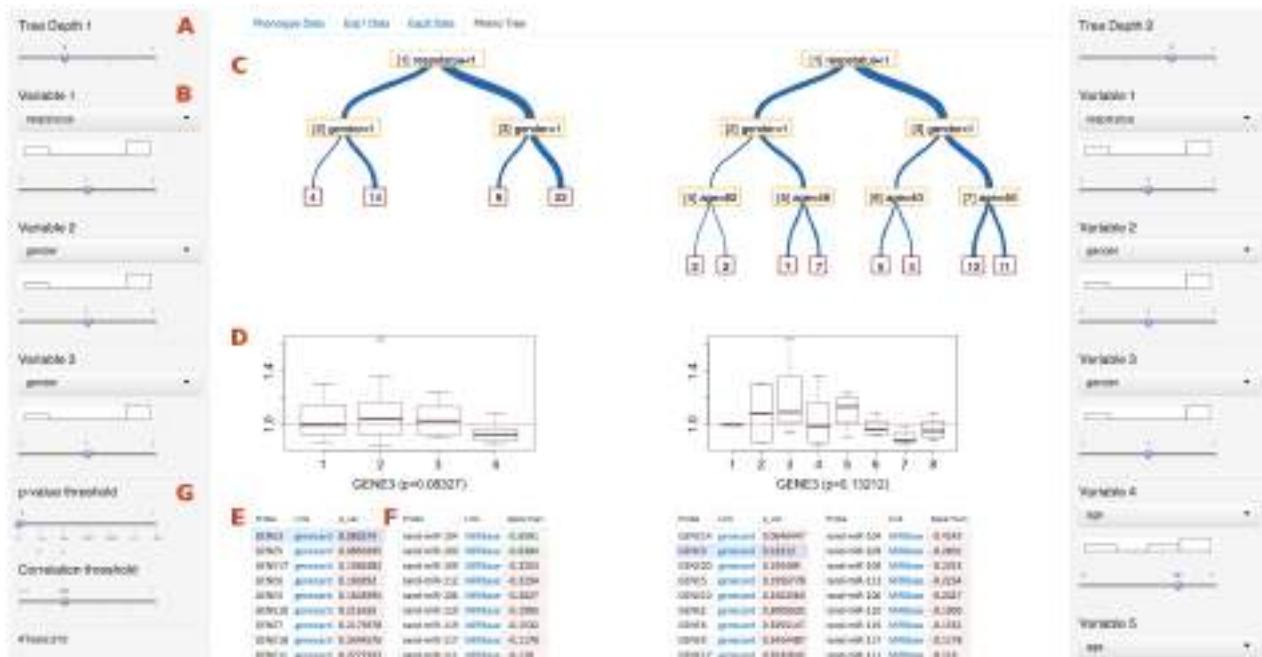


Fig. 1. Sample Screenshot of PEAX. On the left are various controls for defining sub-phenotype, including the depth of the decision tree (A), the decision variable for each node (B), and the decision variable threshold. A histogram of the sub-population of patients is shown below each decision node drop-down box. Together, these controls define the decision tree (C). A boxplot (D) shows distribution of a selected gene from a candidate list of genes (E), while a separate adjacent table (F) shows the top correlating miRNA expression levels with the selected gene. Selectable thresholds (G) allow for highlighting of significant associations in either table. Displayed data are random for illustrative purposes, due to confidentiality of clinical research work described in this paper. The entire set of controls, and resulting tree, boxplot, and associative tables, are replicated on the right side to allow for definition, viewing, and comparing of juxtaposed models.

GUI removes the requirements of being able to script and interpret R code directly, increasing the size of the audience of researchers capable of participating in the analysis process, as well as moving the clinical expert closer to the data analysis. An overview of the PEAX GUI is shown in Figure 1.

We used several visual analytics concepts and design elements to represent data and analysis results. We chose to use binary **decision trees** as visual representations of complex phenotypes. In previous collaborative work, we used C4.5 trees [19] to discover a biomarker associated with drug response [9]. Decision trees are already familiar to clinicians, who work with decision-support systems in diagnosing and defining phenotype, as well as bioinformaticians, who are familiar with common machine-learning techniques that use decision tree analysis. They are even used for crowd-sourced cancer gene expression analysis by users with a variety of levels of research experience and education [30]. A decision tree can cleanly and quickly present the data and relevant cutoffs for decisions in a way that representations through simple tables cannot achieve.

In PEAX, trees are implemented as D3 collapsible trees defined in JSON format. Trees are displayed with the vertical axis representing decision variables, and the horizontal axis representing patient/sample subgrouping. We augmented the decision tree with **visual cues** to show information about the data, primarily by varying line thickness between nodes, so that thicker lines represent a higher percentage of a sub-population of patients that meet a given decision-tree criterion. This becomes more evident when the phenotype definition is changed slightly, providing immediate feed-

back on distribution of phenotype. The leaves of the tree show the size of each group meeting the phenotype definition. Only samples which have their phenotype fully defined by all previous nodes are included in final analysis, so that patients with missing data are automatically excluded from analysis. By noting the number of patients in the final groupings compared to the total population, a researcher can discern how much data are missing.

PEAX also supports the comparison of two decision trees, presented side-by-side. This **juxtaposition** enables several potential use cases and scenarios, in which patients were stratified by drug treatment and/or genotype in different ways. More generally, an analyst can more easily compare two different phenotype definitions, perhaps with slight adjustments, and not have to rely on brain memory to examine differences in phenotype distribution and association. The ability to use and compare multiple, juxtaposed trees provides additional flexibility by providing an extra dimension of exploration and comparison. Since our decision trees are relatively small (they are rarely deeper than three levels) a simple juxtaposition is sufficient for a visual comparison and more advanced approaches for comparing larger trees are not required.

Phenotype variables, shown as decision-tree selection nodes, can be selected with drop-down boxes and adjusted via sliders; a histogram positioned directly above each slider shows the distribution of patients hierarchically and dynamically split by all higher nodes. The histogram quickly shows the **data distribution** based on a particular node and can provide insight into the distribution of the population, thus being useful to determine interesting cutoff values. Augmenting sliders with histograms is an example implementation of a “scented widget,” in which a GUI element is integrated with an embedded visualization, shown to help increase the number of discoveries in data [31].

After a phenotype is defined, a list of mRNAs is displayed, sorted by significance of association with the defined phenotype. Several selectable methods, currently including ANOVA and Kruskal-Wallis, provide **statistical testing** of association. Individual mRNAs can be selected from the list, upon which a boxplot is displayed that shows the difference in distribution of the selected mRNA. A red line indicates a separation between gene expression values that have increased or decreased. The boxplot can provide insight into directionality of upregulation/downregulation of specific genes, or possible dose-response or comparative relationships between more than two phenotypes. If a pattern of expression is found to be interesting based on a selectable threshold for upregulation/downregulation, a researcher can search for other genes that exhibit the same pattern. Once an mRNA is selected, a separate ordered list of the top-correlating miRNA expression values is shown. Each mRNA and miRNA is hyperlinked to online databases, and associations meeting a selectable threshold are highlighted in a different color. The combination of sorted lists, boxplots, and database links provide the capability for the user to achieve “**details on demand**” [28] when exploring a hypothesis.

Although investigating individual gene associations is rapid, we identified the performance bottleneck as the sheer number of analysis-of-variance (aov) tests run when the phenotype definition is modified. In order to improve performance to provide a **responsive interface**, we distributed the aov calculations across the number of available cores. We achieved a noticeable improvement in speed, but the system still lagged by 5-10 minutes for each input adjustment. We further optimized the code by parallelizing several independent steps of the calculation instead of using the built-in

aov function; namely, independent, parallelized computation of column means and sums. Finally, we filtered out low-variant expression data (a common pre-processing step for other forms of analysis [32]), leaving us with a set of 3893 fold-change gene expression values (reduced from the original $\sim 34,000$ values). Using the filtered gene-fold values, PEAX responded to input changes in approximately 30 seconds or less, so as to provide our clinical expert with a system that could meet research needs as well as be sufficiently responsive for data exploration.

6. Case Study

Based on our previous collaborative work with a group of cardiologists, we designed a prototype of PEAX to support the primary research tasks. We then met several times over the course of several weeks with a cardiology expert and refined and tested our system based on initial feedback. Being a researcher as well, this expert was also familiar with script-based analysis in R, and was able to provide comparative feedback with respect to the visual and interactive capabilities of the tool.

To gain trust and experience with PEAX, the domain expert first recreated previous research from his group. He looked into LVEF improvement, the clinical biomarker used for drug response, and its association with gene expression, and was able to verify that the tool was able to provide the same results through a visual interface rather than previous scripting. The domain expert then investigated two additional research aims of the BORG study: examining differences between drug treatment groups (Task I), and examining the effects of β -adrenergic SNPs on drug responsiveness (Task II). Now that the domain expert was more familiar with the tool, he combined these two tasks into a single exploratory analysis. These tasks were completed with a combination of unsupervised, self-directed usage of the tool, succeeded by a follow-up discussion of the results. The researcher stated that he was able to complete the task of initial investigation within half an hour, whereas it would have taken most of a day or more to set up, refine, test, and verify script-based results; he was also able to examine several genes of interest that he may otherwise not have investigated. For these tasks, he designed two trees (Figure 2a and Figure 2c), which are nominally shown side-by-side in PEAX. The first tree simply involves drug responsiveness, whereas the second tree combined drug treatment group with the β 1-adrenergic receptor Arg389Gly SNP, and drug responsiveness. He investigated several potential genes of interest, and quickly identified a gene association ($p < 0.0005$) that had a more pronounced difference when the SNP and drug treatment group interaction was considered. The gene has a moderate differential response between responders and non-responders (Figure 2b), but this difference is mostly due to the drug treatment group (Figure 2d). He commented that the side-by-side comparison gave him additional insight regarding potential interactions when he was able to stratify patients by additional variables and compare to the more general LVEF response vs. non-response gene expression associations to explore possible differences in relative gene expression according to drug treatment and SNP. He suggested that the patterns of differentiation which change when additional variables are added are evident due to the visual side-by-side presentation.

After applying PEAX to Tasks I and II, he then annotated a screen shot (Figure 2c and 2d), which he sent to another member of the clinical research team. This showed us that combined phenotype and association output from the tool could be used quickly for transmitting and discussing results for collaboration within a team of colleagues, and suggested future feature enhancements.

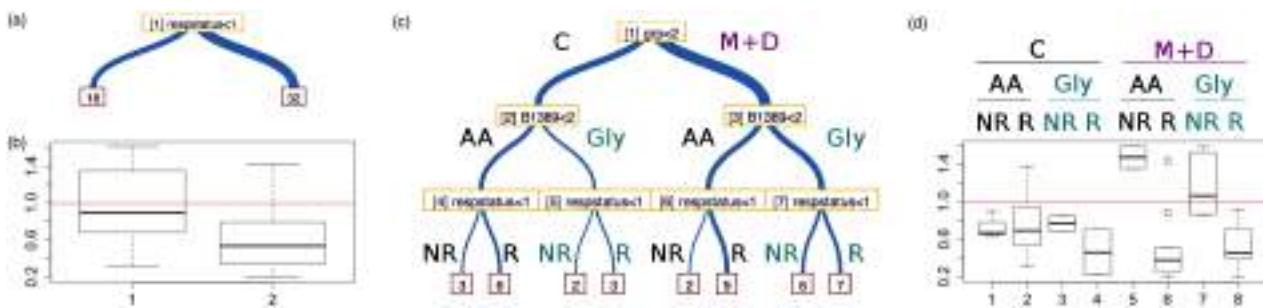


Fig. 2. Sample output from Tasks I and II. The analyst created a 1-node tree for drug responsiveness (a), and examined boxplot distribution for a particular gene of interest (b). He then created a 3-level tree that used drug treatment group *grp*, as well as SNP *B1389* (corresponding to dbSNP *rs1801253*), in addition to drug responsiveness *respstatus* (c). He noticed that the same gene selected previously exhibited a differential response between responders and non-responders in group 3 (d). Tree branch group labels in (c) match those in (d), and are not generated by PEAX, but were added post-analysis by our analyst when showing to a collaborator.

7. Discussion and Lessons Learned

The development of PEAX, including the design phase and preliminary testing, highlighted several insights, challenges, and functional considerations that we used to refine the tool. The use of Shiny to create a rapid, functional, GUI-based prototype with statistical analysis is quite helpful in designing and refining development in an agile fashion. The continuous use of R for analysis allows researchers and developers that already trust R to retain their confidence in underlying methods and support for statistical computing. To our knowledge, PEAX is the first open-source based tool to combine powerful statistical analysis with interactive decision-tree models to enable clinicians to analyze heterogeneous clinical and expression data. Below, we discuss lessons learned during the development and usage of PEAX.

Interactive statistical analysis requires a reactive graphical interface. In order to direct the user in exploratory analysis, an interface must be reactive so as to show the effect of changing aspects of a hypothesis. Implementing our tool in Shiny abstracts the reactive handling of asynchronous user input changes. Comparing different phenotype trees side-by-side, or seeing the difference in data distribution (including the amount of missing data) when making changes to a phenotype definition, for example, provides immediate visual feedback not readily apparent in text-based R-scripting, while still allowing for the power and support of statistical analysis in R.

Interactivity is a driver for optimization of statistical algorithms. The largest performance bottleneck with our test dataset was the number of ANOVA calculations necessary for thousands of candidate genes. While the GUI allowed for easier definition for phenotypes for subsequent testing, a long lag of several minutes caused the GUI to appear unresponsive and slow, and optimization of this step was crucial to achieve acceptable usability. Instead of using R's built-in linear model (*lm*) or analysis of variance functions (*aov*), which provide model results and calculations unnecessary for our interactive display, we decomposed the ANOVA function to optimize the calculation of the F-value. We achieved a considerable speedup by using *colMeans* and *colSums* for intermediate calculations, as well as parallelization using *mclapply* from the *parallel* library to distribute the calculation of column means and the explained group variance for each of the defined phenotype groups. This algorithmic refactoring allowed us to experiment with a reactive and responsive system,

as we were able to demonstrate our prototype on an 8-core system. Parallelization should allow for scalability on larger systems. We also observed that our framework took less than 1GB of memory while running, suggesting that use cases similar to ours are likely to be computationally intensive more than data intensive and therefore potentially more parallelizable.

True “real-time” responsiveness is not always necessary for useful data exploration. Even with performance optimizations, refreshes of the visual display of our data and analysis can take tens of seconds and the tool reactivity cannot be considered “real time.” However, we were surprised to find that our collaborator considered this a vast improvement over more familiar script-based techniques, which were more likely to require tedious and careful setup and checking of data setup before testing an individual hypothesis. In combination with other visual cues that suggested possible interesting patterns, the speed of visual updating encouraged our analyst to continue data exploration and experimentation.

A GUI-based tool allows the domain expert to drive analysis and collaboration. In our iterative design process and case analysis, we found a small, targeted audience of cardiology experts to be responsive and enthusiastic in suggesting features and desire to experiment with the tool, moreso than when similarly presented with an abstract, static list of potential features. With a working prototype, the clinical researcher was able to move closer to the data analysis, and the group was enthusiastic about using the tool to begin analysis tasks that supported prior research grants. Additionally, our collaborator explored data relationships in ways that were surprising to us, by exploration of both previously understood as well as novel gene associations, and by using screenshots of the tool to communicate with colleagues.

Discovery through exploration is based on a comprehensive analysis of all available data. A natural criticism of explorative approaches is the potential for false discovery rates (FDR), especially through “data dredging,” or drawing erroneous conclusions based on excessive analysis of numerous possible relationships. Like any tool, PEAX can and should be used appropriately for the appropriate task. We provide features for adjusting for multiple testing, such as displaying the cumulative number of statistical tests during a session of exploration. The analyst is responsible for appropriate adjustment of resulting significant associations. Another possible use case, not initially envisioned but very manageable with our system, would be to split the dataset into a training and test (holdout) set, whereupon a hypothesis is generated on the training set, and then tested on the holdout set [33]. This approach is especially amenable to our system of allowing for two different decision trees, where the holdout set could be verified side-by-side with the training set.

The evidence for biological plausibility is a comprehensive picture that depends on phenotype input as well as mRNA and miRNA correlation. Phenotype classification is in itself a subjective task performed by expertise, and a motivating factor of the interactivity of phenotype definition and data exploration was to provide expert feedback into the system using a supervised method. The problem of identifying potentially interesting, unknown phenotype definitions has not yet been adequately solved. Therefore, we focused on facilitating domain expert expression of phenotypes and integration of associated molecular data. Correlative mRNA and miRNA biomarkers are more compelling when, taken together, they relate to a plausible biological scenario, which will be determined by the domain expert. PEAX can provide information regarding the data, but interpretation, presentation of results, and experimental follow-up must be done with scientific care.

8. Conclusion

This paper presents a new tool, PEAX, for integrated analysis of complex phenotype definition and association with molecular expression data. This analytic scenario is targeted for clinical experts with mixed datasets involving deep clinical, mRNA, and miRNA expression data, and is designed to abstract statistical analysis scripts while providing useful feedback for exploration of data. We developed a prototype tool, demonstrated its utility through a case study with a domain expert on a cardiology dataset, and report initial observations and lessons learned. We chose decision trees as a simplified model familiar to clinical researchers, and present analysis results on a single interactive screen in order to streamline analysis.

We found the combination of visual interactivity in a web browser, with the statistical analysis capabilities of R, to be a compelling combination to make this type of analysis more rapid, exploratory, and collaborative. We found the generalized functions of defining and testing sub-classes visually to be faster, less error-prone, and more efficient than a process that uses “one-off,” proprietary scripts for very specifically subdividing patient groups based on specific features.

The web-based nature PEAX provides opportunities for scalability in performance and distribution, although clinical data sensitivity issues may require internal application usage. By supporting Javascript and D3 for GUI support, PEAX can continue to leverage new web-based visualizations; and by using R for statistical computation, we can add analysis algorithms based on the extensive and growing library of R functions. In addition to further exploration on BORG and other datasets, planned additional features include highlighting known miRNA-mRNA associations to direct further exploration of plausible miRNA regulation of mRNA related to a hypothesized complex phenotype.

By necessity, PEAX sped up association calculations using analysis of variance by parallelizing several key steps. There is still a lag between user input and redisplay of output, so that the interaction was not considered real-time in our scenario, but to our surprise, our clinical expert was pleased in being able to see new results presented within tens of seconds, as opposed to his previous techniques which involved tedious, text-based scripting. This observation is important for large, multidimensional datasets: interaction may not need to be real-time, if it is tolerably responsive and represents an improvement in the amount of time required to analyze and check through existing techniques, with a reduced cognitive demand for formulating and checking correct syntax of scripts.

Improper use of data-mining techniques to analyze high-dimensional data can lead to spurious, false associations. Therefore, an investigator must draw conclusions based on comprehensive consideration of all of the evidence, and particularly important observations should be validated independently. Nevertheless, a growing number of large datasets are available in which important biological questions have gone unexplored and undiscovered, perhaps because of computational complexity, or proprietary scripting. To address the big data challenges of personalized medicine, integrated statistical and visual analysis tools such as PEAX are needed for rapid data exploration, collaboration, and communication to drive hypothesis generation and testing by clinical experts.

Acknowledgments

This work was supported in part by NIH grants R01 LM008111, 2R01 HL48013, 1R01 HL71118, P20 HL101435-01, and T32 HL007822-12, and by grants from GlaxoSmithKline and AstraZeneca.

References

- [1] L. D. Miller, J. Smeds, J. George, V. B. Vega, L. Vergara, A. Ploner, Y. Pawitan *et al.*, *Proceedings of the National Academy of Sciences of the United States of America* **102**, 13550 (September 2005).
- [2] K. M. Giacomini, C. M. Brett, R. B. Altman, N. L. Benowitz, M. E. Dolan, D. A. Flockhart, J. A. Johnson, D. F. Hayes, T. Klein *et al.*, *Clinical pharmacology and therapeutics* **81**, 328 (March 2007).
- [3] D. Houle, D. R. Govindaraju and S. Omholt, *Nature reviews. Genetics* **11**, 855 (December 2010).
- [4] K. Shameer, J. C. Denny, K. Ding *et al.*, *Human genetics* **133**, 95 (January 2014).
- [5] J. Listgarten, O. Stegle, Q. Morris *et al.*, *Pacific Symposium on Biocomputing* **19**, 247 (2014).
- [6] ClinicalTrials.gov, Effect of Beta-blockers on Structural Remodeling and Gene Expression in the Failing Human Heart (BORG, NCT01798992) (2013), <http://www.clinicaltrials.gov/ct2/show/NCT01798992>.
- [7] V. L. Roger, A. S. Go, D. M. Lloyd-Jones, R. J. Adams, J. D. Berry, T. M. Brown, M. R. Carnethon, S. Dai, G. de Simone, E. S. Ford, C. S. Fox, H. J. Fullerton *et al.*, *Circulation* **123**, e18 (February 2011).
- [8] Investigators, The Beta-Blocker Evaluation of Survival Trial, *The New England journal of medicine* **344**, 1659 (May 2001).
- [9] M. A. Hinterberg, D. Kao, A. Karimpour-Fard, K. Sucharov, L. E. Hunter, D. Port and M. Bristow, *Journal of the American College of Cardiology* **61** (2013).
- [10] M. Hall, H. National, E. Frank *et al.*, *ACM SIGKDD explorations newsletter* **11**, 10 (2009).
- [11] G. J. Williams, *The R Journal* **1**, 45 (2009).
- [12] E. J. Harner, D. Luo and J. Tan, *Computational Statistics* **24**, 295 (September 2008).
- [13] J. Horner, RApache: Web application development with R and Apache (2013), <http://www.rapache.net>.
- [14] S. Urbanek, Rserve: binary R server (2013), <https://rforge.net/Rserve>.
- [15] RStudio, <http://www.rstudio.com/shiny/> (2013).
- [16] R. C. Gentleman, V. J. Carey, D. M. Bates *et al.*, *Genome biology* **5**, p. R80 (January 2004).
- [17] M. D. Robinson, D. J. McCarthy and G. K. Smyth, *Bioinformatics (Oxford, England)* **26**, 139 (January 2010).
- [18] R. Edgar, *Nucleic Acids Research* **30**, 207 (January 2002).
- [19] R. Quinlan, *C4.5: Programs for Machine Learning* (Morgan Kaufmann, 1992).
- [20] L. M. Collins and S. T. Lanza, *Latent Class and Latent Transition Analysis: With Applications in the Social, Behavioral, and Health Sciences(Google eBook)* (John Wiley & Sons, 2010).
- [21] P. Haldar, I. D. Pavord, D. E. Shaw *et al.*, *American journal of respiratory and critical care medicine* **178**, 218 (August 2008).
- [22] Ayasdi, <http://www.ayasdi.com/> (2013).
- [23] Y. Ru, K. J. Keckris, B. Tabakoff, P. Hoffman *et al.*, *Nucleic Acids Research* , 1 (July 2014).
- [24] G. T. Huang, C. Athanassiou and P. V. Benos, *Nucleic acids research* **39**, W416 (July 2011).
- [25] E. Cerami, J. Gao, U. Dogrusoz, B. Gross *et al.*, *Cancer discovery* **2**, 401 (2012).
- [26] M. R. Bristow, G. A. Murphy, H. Krause-Steinrauf, J. L. Anderson, J. F. Carlquist, S. Thaneemit-Chen, V. Krishnan, W. T. Abraham, B. D. Lowes *et al.*, *Circulation. Heart failure* **3**, 21 (January 2010).
- [27] C. M. O'Connor, M. Fiuzat, P. E. Carson *et al.*, *PloS one* **7**, p. e44324 (January 2012).
- [28] D. Keim, G. Andrienko, J.-D. Fekete, C. Görg, J. Kohlhammer and G. Melançon, *Information Visualization: Human-Centered Issues and Perspectives* , 154 (2008).
- [29] M. Bostock, V. Ogievetsky and J. Heer, *IEEE Transactions on Visualization and Computer Graphics* **17**, 2301 (December 2011).
- [30] B. M. Good, S. Loguercio, O. L. Griffith *et al.*, *arXiv preprint arXiv* , 1 (2013).
- [31] W. Willett, J. Heer and M. Agrawala, *IEEE Transactions on Visualization and Computer Graphics* **13**, 1129 (2007).
- [32] S. Bandyopadhyay, S. Mallik and A. Mukhopadhyay, *Computational Biology and Bioinformatics, IEEE/ACM Transactions on* **11**, 95 (Jan 2014).
- [33] S. S. Young and A. Karr, *Significance* **8**, 116 (2011).

T-RECS: STABLE SELECTION OF DYNAMICALLY FORMED GROUPS OF FEATURES WITH APPLICATION TO PREDICTION OF CLINICAL OUTCOMES

GRACE T. HUANG

*Department of Computational and Systems Biology, and
Joint CMU-Pitt PhD Program in computational Biology,
University of Pittsburgh School of Medicine, Pittsburgh, PA 15260, USA
Email: tzh4@pitt.edu*

IOANNIS TSAMARDINOS

*Department of Computer Science, University of Crete, Heraklion, Crete, Greece
Email: tsamard@ics.forth.gr*

VINEET RAGHU

*Department of Computer Science, University of Pittsburgh, Pittsburgh, PA 15213, USA
Email: vkr8@pitt.edu*

NAFTALI KAMINSKI

*School of Medicine, Yale University, New Haven, Connecticut, USA
Email: naftali.kaminski@yale.edu*

PANAYIOTIS V. BENOS

*Department of Computational and Systems Biology
University of Pittsburgh School of Medicine, Pittsburgh, PA 15260, USA
Email: benos@pitt.edu*

Feature selection is used extensively in biomedical research for biomarker identification and patient classification, both of which are essential steps in developing personalized medicine strategies. However, the structured nature of the biological datasets and high correlation of variables frequently yield multiple equally optimal signatures, thus making traditional feature selection methods unstable. Features selected based on one cohort of patients, may not work as well in another cohort. In addition, biologically important features may be missed due to selection of other co-clustered features. We propose a new method, Tree-guided Recursive Cluster Selection (T-ReCS), for efficient selection of grouped features. T-ReCS significantly improves predictive stability while maintains the same level of accuracy. T-ReCS does not require an *a priori* knowledge of the clusters like group-lasso and also can handle “orphan” features (not belonging to a cluster). T-ReCS can be used with categorical or survival target variables. Tested on simulated and real expression data from breast cancer and lung diseases and survival data, T-ReCS selected stable cluster features without significant loss in classification accuracy.

1. introduction

Identifying a minimal gene signature that is maximally predictive of a clinical variable or outcome is of paramount importance for disease diagnosis and prognosis of individual patient outcome and survival. However, biomedical datasets frequently contain highly correlated variables, which generate multiple, equally predictive (and frequently overlapping) signatures. This problem is

particularly evident when sample size is small and distinguishing between necessary and redundant variables becomes hard. This raises the issue of signature stability, which is a measure of a method's sensitivity to variations in the training set. Lack of stability reduces confidence to the selected features. Traditional feature selection algorithms applied on high-dimensional, noisy systems are known to lack stability (2).

In this paper we propose a new feature selection algorithm, named **T**ree-guided **R**ecursive **C**luster **S**election (T-ReCS), which addresses the problem of stability by performing feature selection at the cluster level. Clusters are determined dynamically as part of the predictive signature selection by exploiting a hierarchical tree structure. Formed clusters are of varying sizes depending on user-defined p -value thresholds. Selecting clusters of variables provides an additional potential benefit. *Biologically meaningful biomarkers may not be maximally discriminative, but could be correlated with strongly discriminative features that lack biological interpretation.* T-ReCS was tested on simulated and real data with categorical and survival outcome variables. T-ReCS can efficiently process large datasets with tens of thousands of variables, thus making it ideal for selecting predictive signatures for patient stratification and for development of personalized medicine strategies.

1.1. Related work

To our knowledge, this is the first method for group variable selection with dynamic formation of the groups as part of the feature selection procedure. Group-lasso (3) is the closest method to T-ReCS, but it requires prior knowledge of the groups, while ideally one wants to be able to determine clusters dynamically and the cluster formation to be part of the feature selection process. Localzo and colleagues used subsampling of the training set to identify consensus feature groups, and then perform feature selection on these groups (4). This method is valuable but the determination of clusters precedes the feature selection as well. Ensemble methods have been proposed to address the problem of stability by aggregating the results of different runs of conventional feature selection algorithms. Haury *et al.* (1) conducted a comprehensive comparative study of many of those methods. Jacob *et al.* (5) have presented a method on enforcing clustering structure on multi-task regression problems and these techniques can be adapted to cluster features. Another problem that is somewhat related to feature selection stability (but T-ReCS does not address it) is the selection of multiple signatures (6, 7), because in some cases the members of the signatures may belong to the same clusters.

2. Methods

2.1. Description of T-ReCS

T-ReCS is a modular procedure, which selects group variables in a multi-step process by combining elements of hierarchical clustering with traditional feature selection algorithms. First, the algorithm performs an initial standard feature selection. Suppose the single variables selected are $\{A, B, C\}$. Next, it constructs a hierarchical tree structure from the data that represents the similarity associations between variables. Leafs, at the bottom of the tree, are the single variables.

Each internal node in the tree represents a group of variables (genes). The lower in the tree a node is, the more similar the patterns of its members are. Next, the algorithm climbs the tree up one level at a time per selected variable. If, for example, $\{A,D,E\}$ are clustered together it creates a new feature A' representing the cluster. If the representative of A' is *informationally equivalent* for predicting T , then A is replaced by A' in the set of selected variables which becomes $\{A', B, C\}$. Essentially, the set of selected variables is now $\{\{A,D,E\},B,C\}$. The procedure continues for all selected variables until no informationally equivalent features can be constructed by climbing up the tree. Stability increases since a small perturbation of the data may lead to different initial features to be selected (e.g., $\{D,B,C\}$), but cluster-based selection will still be $\{\{A,D,E\},B,C\}$.

Any appropriate algorithm can in principle be employed for these steps. In our case, for the initial feature selection, we adopt the causal structure finding algorithm Max-Min Parents Children (MMPC) (8). MMPC assumes that the data distribution can be faithfully represented by a Bayesian Network where each variable and the target T serve as nodes. MMPC identifies the parents and children of T (i.e., the adjacencies with T), $PC(T)$, in that network efficiently, without fully reconstructing the network. The output of the MMPC is an approximation (subset) of the Markov Blanket of T , i.e., a minimal subset of variables that renders all other variables conditionally independent and thus can optimally predict T . It was shown that under certain broad conditions, the Markov Blanket is the solution to the variable selection problem (8). Furthermore, Tsamardinos and colleagues have shown that the $PC(T)$ set, in practice, leads to models that are close to optimal for predicting T , while it is significantly less computationally expensive than the full Markov Blanket (9). Therefore, primary feature selection here is equivalent to discovering the $PC(T)$. For generating the tree structure we use ReKS (*Recursive K-means Spectral Clustering*), which was shown to outperform other methods in terms of speed or efficiency and outputs more balanced trees when applied to heterogeneous clinical data (10). Finally, to create the representative features of a cluster we tested the first Principal Component of the cluster, the medoid, and the centroid of the clustered variables.

MatLab was used for implementation of T-ReCS and comparison to other methods. The complexity of T-ReCS is roughly $O(|\varphi|^2)$. Specifically, ReKS is $O(|\varphi|^2)$ (10), MMPC is $O(|\varphi| \cdot |PC(T)| \cdot k)$, and conditional independence tests for ascending the tree is $O(\log |\varphi| \cdot |PC(T)|)$. We note, however, that selection of different methods for single feature selection and tree construction can alter this complexity.

2.2. Deciding informational equivalence

A key innovation of the algorithm is how to determine whether a cluster representative X' at level k of the tree is informationally equivalent to X at a lower level $k+1$. Intuitively, we test whether X should be substituted with X' , a representative of a cluster of variables while maintaining predictive accuracy. We require two conditions to be satisfied: Condition (C1) $\text{Dep}(X'; T | S)$, for every $S \subseteq \{PC(T) \setminus \{X\}\}$, where $\text{Dep}(X ; T / S)$ denotes the conditional dependence of X with T given variables S . This condition needs to be satisfied by MMPC to select a variable in the output. Thus, if (C1) is satisfied MMPC could have selected X' instead of X in the original set of variables if it was available. Intuitively, the test determines that X' carries unique information for predicting

T in any context (subset) of the other selected variables. This is justified by Bayesian Network theory: if the data distribution is faithful to some Bayesian Network, then this condition is satisfied by the parents and children of T . This condition dictates that in the absence of X , a representative X' of a cluster of variables should be selected, which increases stability. *Thus, this condition is responsible for increasing stability.* Condition (C2) is $\text{Ind}(X ; T | X')$, denoting the conditional independence of X with T given X' . This second condition implies that the original variable X is rendered superfluous (redundant) once X' is selected. Thus, X and X' are informationally equivalent for predicting T (at least, when no other variables are considered). *Thus, this condition aims at ensuring that predictive performance is maintained when replacing X with X' .*

2.3. Statistical tests of conditional independence

T-ReCS, like MMPC, uses conditional independence tests to determine inclusion in the final output, based on a corresponding p -value, denoted as $P(X ; T | S)$. If this p -value is below a user-defined threshold (typically, 10^{-2} to 10^{-4} ; see below) we accept dependence, and if it is larger than a threshold (not necessarily the same) we accept independence. The pseudo-code of the algorithm is presented in **Suppl Fig S1**. We emphasize that the procedure constructs new features, corresponding to clusters of variables, *adaptively and dynamically*. It may or may not decide to substitute a variable in the output of MMPC with a representative of a larger cluster. A common framework for constructing hypothesis tests is the framework of a Likelihood Ratio test (11). The Likelihood Ratio computes the deviance $D = -2 \cdot \ln(P_0/P_1)$, where P_0 and P_1 are the null and the alternative model, respectively. D asymptotically follows the chi-square distribution with degrees of freedom equal to the difference in the number of parameters between the two models. From this distribution we can obtain the p -value of the test. For testing the hypothesis $\text{Ind}(X;T | Z)$ the null model is a predictive model for T given Z and the alternative model is a predictive model for T given Z and X . Thus, the ratio tests whether the likelihood for T when X is added is statistically significantly different compared to when X is not given. If yes, then indeed X provides additional information for T given Z and the null hypothesis of independence is rejected. In the following experiments, when T is continuous, we employ linear models (equivalent to testing the partial correlation of X and T given Z); when T is discrete, we employ logistic models; and when T is a right-censored survival variable, we employ the proportional Hazard Cox Regression model) as we did before (12, 13).

2.4. Group variable representation

There are many ways to construct a group variable X' from its members. In this paper, we tested the centroid, medoid and the first component of the principal component analysis (PCA) as cluster representatives since they have been successfully applied on gene expression data before (14-16). Other latent variable representations can be used instead.

2.5. Methods for measuring predictive performance

In this paper, we measured the predictive quality of the sets of selected features either directly (when the network was known, i.e., synthetic data) or indirectly by using the selected features in a regression or classification method.

For binary target variables, we used Support Vector Machine (SVM) (17), which takes continuous predictors as input and outputs a label assignment. SVMs are practical, scalable, and have competitive performance for this type of data (9). We evaluated the performance in terms of (1) classification accuracy rate defined as the sum of number of true positives and true negatives over total number instances and (2) stability as it is defined below.

For time-to-event target variables (a.k.a. survival data) we used the Cox regression (Cox Proportional Hazards Model) (18), which relates the predictor variables to the time that passes before an event of interest occurs. Time-to-event data are typically right-censored, i.e., for some patients we do not know the time of occurrence of the event, but only know that they were event-free up to a time point. Measuring the predictive performance of a survival regression model is also not straightforward as the prediction error can be computed exactly only for the uncensored cases (i.e. when the time to event is known). Several measures have been proposed to measure performance for survival analysis (19-22). We select the Concordance Index (CI) (19) as it is one of the most commonly used measures for survival models. Intuitively, the CI measures the fraction of all pairs of patients, whose predicted survival time is correctly ordered by the regression model. Scenarios in which the order of observed survival cannot be determined due to censorship are excluded from the calculation.

2.6. Methods for measuring stability

For this paper, stability is a measure of how consistently the same variables are selected across different cross validation runs. Typically, a measure like Tanimoto set-similarity (23) is used to characterize the agreement or percentage of overlap between two sets of features. In our case, however, each set of features can contain single- or group-variables, and the Tanimoto set-similarity alone would not suffice. For example, $F_i = \{\{A,B,C\}, \{D,E\}\}$ may be selected in cross-validation fold i and $F_j = \{\{A,C\}, \{B,D\}\}$ may be the selection at fold j . Before computing set-similarities between elements of F_i and F_j , the elements of these sets need to be matched. In this example, one question is whether $\{A,B,C\}$ in F_i should be matched with $\{B,D\}$ or $\{A,C\}$ in F_j ? To find the best matching, we use maximum weight matching (24) to build a bipartite graph between elements of F_i and F_j . The weights on the edges correspond to the Tanimoto set-similarity $S(s, w) = |s \cap w| / |s \cup w|$, where $s \in F_i$ and $w \in F_j$. In this example, $\{A,B,C\}_i$ is matched to $\{A,C\}_j$, and $\{D,E\}_i$ to $\{B,D\}_j$ where the indexes denote membership in F_i and F_j respectively. After the optimal matching is found, the weights normalized to a total sum of 1, and we take the sum of normalized weights of the selected edges of this matching as a metric of stability between the selected variables in each pair of folds. The overall stability is the average pair-wise stability over all pairs of cross-validation folds and ranges between zero (no stability) and one (absolute stability). Note that, when only single variables are selected, this definition of stability reduces to the average Tanimoto similarity of the selected variables over each pair of folds.

2.7. Datasets used in this paper

2.7.1. Synthetic data

Simulated gene expression data were created using the linear Gaussian Bayesian network structure shown in **Suppl Fig S2**. The network includes the target variable T , a set of 25 variables that are ancestors of T , a set of 25 variables that are descendants of T , and 44 variables that do not have a path to T . Parents are nodes 26-28, children are nodes 29-31, connected variables are nodes 1-25 and 32-56 and unconnected variables are nodes 57-100. The non-immediate relatives to T have an average out-degree of 2. Each node has continuous values analogous to that of gene expression data. The target variable we observe is binary; this is akin to case/control studies or observing two disease subtypes in patients. To generate the data, we model the value of each variable as a linear function of its parents with equal weights, with a Gaussian noise. In order to simulate the effects of co-linearity between variables, for every variable in the dataset we created a total of 10 datasets \times 1,000 samples each. Each dataset had an increasing amount of Gaussian noise, ranging from $N(0,0.05)$ to $N(0,2.5)$. In addition, we similarly created one test set with 5000 samples.

2.7.2. Biological data

Large scale biomedical datasets. We used three large-scale biomedical datasets. Haury *et al.* (1) study used gene expression data from four metastatic breast cancer cohorts (GEO numbers GSE1456, GSE2034, GSE2990, GSE4922), each with >125 patient samples to a total of 819 samples. The second is a breast cancer cell line dataset (25), which contains mRNA expression in 60 breast cancer cell lines (24 basal and 36 luminal). The third dataset is miRNA expression data from the *Lung Genomics Research Consortium* (LGRC) (26), which includes samples from patients with chronic obstructive pulmonary disease (COPD; 210 patients) and idiopathic pulmonary fibrosis (IPF; 249 patients). In all these datasets, T-ReCS was used to identify gene signatures predictive of the particular target variable (relapse or not, breast cancer type and COPD or IPF, respectively).

Censored dataset. We used the censored benchmarking datasets from (12) which consisted of six publicly available gene expression datasets (27-32). The six sets of censored survival data range in size from 86 to 295 cases with 70 to 8,810 variables, and the events of interests are either metastasis or survival.

3. Results

We tested T-ReCS on (1) a set of simulated data, (2) a set of six benchmarking gene expression datasets, and (3) one set of biological (cell lines) and two of biomedical (clinical) data. These datasets were selected to cover cases with either binary or survival target variables. We compare T-ReCS performance to a baseline produced by single variable selection. We also compare it against ensembles constructed from features selected from different folds of cross validation data. We perform 10-fold cross validations when the sample size allows. For datasets with sample size less than 200, we perform two repetitions of 5-fold cross validation. For a fair comparison, on all

instances, the single variable MMPC component was run with the same significance threshold $\alpha=0.05$ and size of maximum conditioning set $k=5$.

3.1. Evaluation of T-ReCS

3.1.1. T-ReCS evaluation on synthetic data and comparison to other methods

On the synthetic dataset MMPC recovered on average 5.5 out of 6 PC(T) members, with 0.8 false positives that are almost always the least significant selected variables. This confirms that the single variable MMPC is successfully recovering the planted variables. We also note that the spectral clustering method (ReKS) clusters together most of the noisy copies of the variables. Non-singleton clusters are often connected by an edge, indicating that there is high correlation between them and the clustering is justified. In fact, 75.7% of all the unique clusters that were selected under the most lenient parameter combination contain copies of a single “seed” variable;

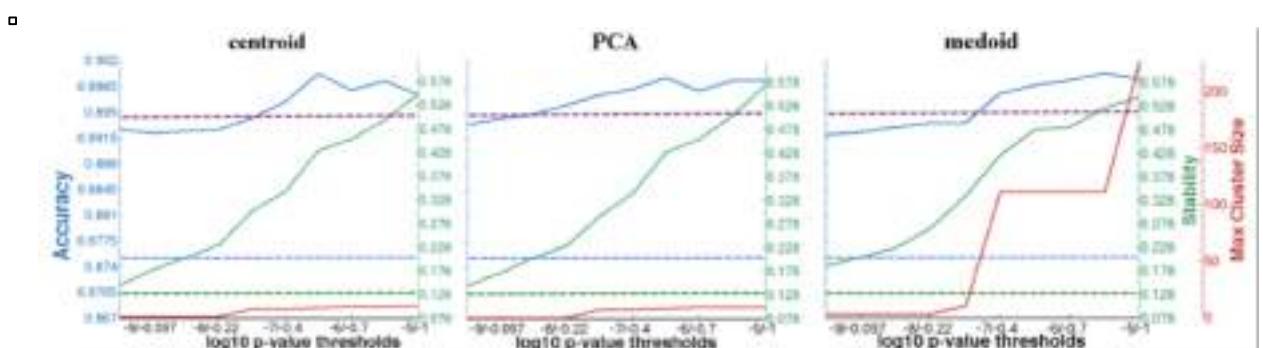


Fig. 1. T-ReCS performance on synthetic data. Accuracy (blue), Stability (green) and Cluster size (red) across 10 parameter combinations (left to right, most stringent to most permissive) for three different cluster representation methods applied on simulated data. Plotted in dotted lines in corresponding colors are single variable selection baseline results. The purple dotted line is the ensemble baseline accuracy.

while another 18.6% contain “foreign” variables seeded from a different variable; and a mere 5.7% of the selected clusters have copies of variables from more than one foreign seed variables. This result confirms that ReKS is indeed creating valid data partitions for T-ReCS.

The average cross validation accuracy, stability, and cluster size for different thresholds are plotted in **Fig 1** for three methods for cluster definition. As expected, we see a trend of increasing stability and cluster size toward the top of the tree (left to right), with the accuracy displaying more subtle variations with a slight spike in the middle region. We plot the baseline stability and accuracy in dotted lines in corresponding colors. These are the average performance of single variable MMPC across the cross validation runs. Both T-ReCS accuracy and stability are improved over the corresponding baselines. For comparison purposes, we plot the baseline of the ensemble consisting of the union of single variables selected from all the 10 cross validation runs, which we use to train SVM models across the 10 training sets. The average test accuracy is plotted in purple dotted line, and we can see that in the more permissive half of the parameter range, T-ReCS performs the same or better than simple ensemble average. Lastly, we observe that centroid and PCA methods produce very similar results, while medoid allows for larger clusters to be

formed, possibly because the same member continues to be the “medoid” of the cluster as it advances up the tree, masking the “noise” that other cluster members may otherwise introduce.

T-ReCS run on 10 subsets of the synthetic dataset and it identified a total of 66 group features, 58 of which contained representatives of at least one of the six members of the $PC(T)$ (nodes 26-31). Three others contained only distantly connected nodes and five contained unconnected nodes. Out of the ten testing sets, T-ReCS recovered all 6 $PC(T)$ nodes in six, 5 $PC(T)$ nodes in three and 4 $PC(T)$ nodes in one. The results were the same regardless of the method used for representing the cluster. We compare T-RECS to SVM Recursive Feature Elimination (RFE), lasso and Elastic Net (E-Net) methods on 10 subsets. For comparison, we retained the top seven features of each run (total: 70 features for each method). SVM RFE recovered instances of nodes 27, 29, 30, 31 only two times and representatives of nodes 29, 30, 31 eight times. Lasso recovered instances of nodes 27, 29, 30, 31 four times and instances of nodes 27, 29, 31 six times. E-Net only recovered instances of nodes 29, 31 on all runs. The detailed results are presented in **Suppl Table S1**.

3.1.2. Comparison of T-ReCS to other feature selection methods on biological datasets

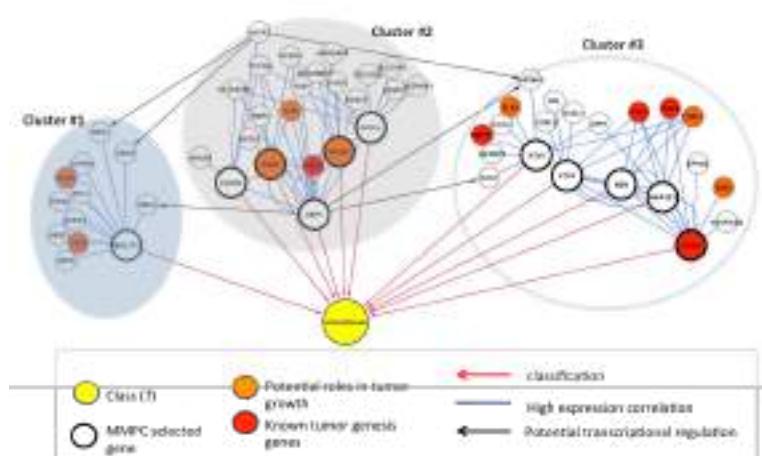


Fig. 3. T-ReCS applied on breast cancer cell line expression data.

Selected gene groups predictive of Basal vs. Luminal subtypes are presented.

balance between accuracy and stability compared to these methods (**Fig 2**). Perhaps more importantly, the figure shows that T-ReCS is on the pareto frontier, i.e., it is never simultaneously

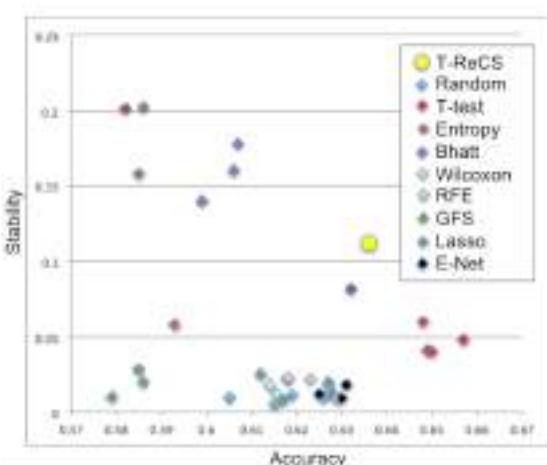


Fig. 2. Comparison of T-ReCS with other feature selection methods. Accuracy and stability tradeoff is calculated as in Haury *et al.* (1) from which the other results are obtained. *RFE*: recursive feature elimination; *GFS*: greedy forward selection; *E-Net*: elastic network.

Haury *et al.* (1) performed a comprehensive study on the influence of feature selection on accuracy and stability of molecular signatures. They compared eight filter and wrapper feature selection methods, including E-Net, RFE and Lasso in single run or ensemble runs. We also run T-ReCS on the same four public datasets (33-36) and we calculated the same measure of accuracy and stability. We found that T-ReCS strikes a very good

dominated in both stability and accuracy; thus, T-ReCS offers a new trade-off point of stability *vs* accuracy not found in any other method.

3.2. Application of T-ReCS to biomedical datasets

3.2.1. T-ReCS on breast cancer cell line data

T-ReCS run on a set of 60 Basal and Luminal breast cancer cell lines (25) and identified three groups of genes that differentiate the two subtypes (**Fig 3**). MMPC identified several genes as predictive of basal *vs* luminal in one or more rounds of cross-validation (**Fig 3**, “MMPC selected genes”, bold cycles), but most are not known to be involved with breast cancer. However, their clusters contain members such as FOXA1 (known to be involved in ESR-mediated transcription in breast cancer cells) and GATA3 (a known marker of luminal breast cancer). Additionally, when we examined the local potential regulatory relationships between the selected genes and their group variables, we found potential XBP1 and GATA3 transcription factor binding sites in other members as evidenced by the fact that these two genes are the two regulatory hubs in this network (**Fig 3**). This observation suggests that a method like T-ReCS that performs variable selection on groups of variables and additionally provides contextual information around the selected groups could provide more biologically robust and meaningful biomarkers.

3.2.2. T-ReCS on lung patient data

miRNA expression data from LGRC were analyzed with respect to disease diagnosis (COPD *vs* ILD). MMPC, ran on the 210 COPD and 249 ILD samples, returned seven miRNAs as maximally predictive of diagnosis, none of which was reported as associated with COPD or IPF on a recent comprehensive review article (37). When T-ReCS performed cluster selection starting from these seven miRNAs, it identified 33 additional miRNAs, (**Suppl Fig S3**; each miRNA label marks the cluster that includes it). Four of the 33 had distinct role in these diseases according to the Sessa *et al.* review (*p*-value=10⁻⁴). miR-1274a, is the most highly induced miRNA in smoker COPD patients compared to normal smoker individuals (38). miR-146a is believed to participate in a feedback loop with its target, COX2, that limits prostaglandin E₂ production and thus controls inflammation. In fibroblasts from COPD patients, miR-146a is induced at lower levels than in normal fibroblasts (39). miR-21 has been strongly associated to IPF *via* the TGF- β signaling pathway (40). miR-154 is a SMAD3 regulated miRNA, whose expression is increased in IPF lung fibroblasts leading to increases in cell proliferation and migration (41). Transfection with miR-154 leads to the activation of the WNT pathway in NHLF cells. WNT and TGF- β are the two most important pathways involved in IPF (41). Further literature search showed that other MMPC/T-ReCS selected miRNAs that have also been reportedly associated with COPD or IPF are miR-24 (COPD) (38), miR-135b (COPD) (42) and miR-376a (IPF) (43). Interestingly, T-ReCS cluster #7 includes seven of the nine miRNAs with confirmed high expression in both IPF lungs and embryonic lungs (41): miR-127, miR-299-5p, miR-382, miR-409-3p, miR-410, as well as miR-154, and miR-487b. Overall, twelve of the 40 T-ReCS selected miRNAs as diagnostic to COPD or IPF are known to be associated with these diseases. The targets of 25 of the 40 miRNAs, as defined by the mirConnX (44) prior network, are presented in **Suppl Fig S4**.

3.2.3. T-ReCS on survival (censored) data

We evaluated T-ReCS on survival data by comparing it with the Survival MMPC (SMMPC) algorithm (13) on the same six clinical data sets that were in the 2010 publication (27-32). In general, stability improves from the baseline by a substantial margin, while accuracy (in terms of CI) hovers around baseline, with small increase or decrease across parameter combinations (**Table 1**). The size of the chosen group variables largely stays within the range of 10 members. This indicates that our method gains in stability without severe loss of accuracy, compared to the single variable selection baseline.

Table 1. Benchmarking censored datasets used in T-ReCS evaluation.

Dataset	#Cases (Cens)	#Vars	Event	% Improvement Stability	CI
Vijver	295 (207)	70	metastasis	+19~28%	-1.2~1.7%
Veer	78 (44)	4751	metastasis	+0~9%	-1.8~11%
Ros02	240 (102)	7399	survival	+23.5~41%	-2.7~0.6%
Ros03	92 (28)	8810	survival	+100~194%	-5.3~1.2%
Bullinger	116 (49)	6283	survival	+6.7~19%	-1.3~3%
Beer	86 (62)	7129	survival	+39~80%	-7.3~8.5%

3.3. Discussion

T-ReCS main novelty is on the dynamic nature of cluster formation, by statistically evaluating their predictive equivalence. Compared to other methods it was able to recover more of the true parents and children of the target variable. We believe this is because T-ReCS will cluster together most of the instances of a node. In addition, it was able to uncover more biological information than single feature selection methods. A body of work has been accumulating on structured sparsity using sparsity-inducing regularizers. Such approaches impose a hierarchy of group structure on the variables and penalties apply on the groups (45). Typically, the group structure stems from prior knowledge, while in T-ReCS it is learned dynamically. But, the main difference between T-ReCS and structured regularization methods is that the former is based on statistical tests of independence, while the latter on regularization and optimization theory. The former has the advantage of theoretically guaranteeing an optimal (and minimal) solution under certain conditions. On the other hand, T-ReCS only includes a subset of the variables in each independence test, which may lead to sub-optimal solutions if the conditions are not met. Overall, we believe T-ReCS addresses an important problem in biomedicine in a robust way. Below we explain some details of the algorithm, which we feel require further clarification.

3.3.1. Group vs single variable selection

We demonstrated the stability improvement of the algorithm over single variable selection and ensemble baseline on simulated data. Significant improvement of stability was achieved with minimum change in accuracy. This is somewhat expected, but this is the first time that the cluster structure is determined dynamically as part of the search process. T-ReCS uses two conditions to achieve this. One condition is designed to enhance stability by substituting single or group variables with larger group variables. The other condition is designed to maintain the predictive accuracy of the initial variables as they are substituted by group variables. Besides improved stability T-ReCS selected group variables contained more biological information than the single ones as we showed in the breast cancer and the biomedical datasets.

3.3.2. Selection of *p*-value threshold

Varying the thresholds of the two conditions affects the output cluster sizes and subsequently the accuracy and stability of the algorithm. A stringent set of thresholds would prevent the procedure from advancing far beyond the initial set of single variables (reducing stability), while moderate thresholds allow larger group variables to be selected (possibly, at the expense of accuracy). As we relax the parameters the expected gain in stability was observed, but the loss in accuracy was minimal at the *p*-value threshold range of 10^{-2} to 10^{-4} , suggesting that this may be a parameter region that is more suitable for biological data. The accuracy reflects a tradeoff between overfitting (from the more stringent range of the parameters) and loss of predictive signals (in the more relaxed range of the parameters). A closer look in the distribution of *p*-values of these two tests also confirms that this parameter range is most effective in thresholding the clusters in the bottom portion of the tree. Alternatively, cross validation can be performed on all input datasets and the parameters selected based on best combined accuracy and stability.

3.3.3. Group variable representation methods

We also investigated its performance over a range of parameter combinations using three distinctive cluster representation methods. Similar performance was observed between centroid and PCA, while medoid tends to produce slightly more dissimilar behaviors. We suspect that this is because a medoid does not represent an “average” behavior of a cluster; it is merely a member of the cluster that is most similar to everyone else. As the cluster size increases, the identity of this member could remain unchanged, in which case the cluster may be allowed to grow very large without affecting the predictive performance, and too many noisy members could be erroneously recruited. On the other hand, medoid could also be susceptible to fluctuations of the member composition in the scenario that a current cluster joins with a larger, dissimilar cluster and the identity of medoid switches all of a sudden. For this reason, we recommend centroid as the preferred collapsing method since it produces more gradual change in stability across many parameter ranges, but unlike PCA it has also a straightforward interpretation.

3.3.4. Future work

We have also begun systematically applying our method on a number of large-scale studies (e.g., TCGA datasets, METABRIC (46), LGRC (26)). While our method was tested only on gene expression datasets in this study, it can be easily adapted to other high-dimensional systems such as methylation and SNP data to provide predictive models as well as biological intuition. Additionally, the modular structure of the algorithms paves the way for a novel group feature selection framework in which alternative clustering step, hypothesis tests, and different variants of the causal discovery algorithm can be employed. The results presented here are promising both in terms of computational performance as well as biological implications.

Acknowledgements. Supplementary material for this work can be found on our web site (<http://www.benoslab.pitt.edu/huangPSB2015.html>). This work was supported by NIH grant U54HG008540 to PVB. IT was partially supported by the EPILOGEAS GSRT ARISTEIA II project, No 3446.

References

1. Haury AC, Gestraud P, Vert JP. 2011. *PLoS ONE* 6: e28210

2. He Z, Yu W. 2010. *Comput Biol Chem* 34: 215-25
3. Yuan M, Lin Y. 2007. *J R Stat Soc B* 68: 49-67
4. Loscalzo S, Yu L, Ding C. 2009. *Consensus group based stable feature selection*. Presented at 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD--09), Paris, France
5. Jacob L, Bach F, Vert JP. 2009. Clustered Multi-Task Learning: A Convex Formulation. In *Twenty-Second Annual Conference on Neural Information Processing Systems (NIPS 2008)*, ed. D Koller, D Schuurmans, Y Bengio, L Bottou, pp. 745-52. Vancouver, BC, Canada: Curran
6. Ein-Dor L, Kela I, Getz G, Givol D, Domany E. 2005. *Bioinformatics* 21: 171-8
7. Statnikov A, Aliferis CF. 2010. *PLoS Comput Biol* 6: e1000790
8. Tsamardinos I, Aliferis CF. 2003. *Towards Principled Feature Selection: Relevancy, Filters and Wrappers*. Presented at 8th International Workshop on Artificial Intelligence and Statistics
9. Aliferis CF, Statnikov A, Tsamardinos I, Mani S, Koutsoukos XD. 2010. *Journal of Machine Learning Research* 11: 171-234
10. Huang GT, Cunningham KI, Benos PV, Chennubhotla CS. 2013. *Pac Symp Biocomput* accepted
11. Neyman J, Pearson ES. 1992. *On the problem of the most efficient tests of statistical hypotheses*. New York: Springer
12. Lagani V, Tsamardinos I. 2010. *Bioinformatics* 26: 1887-94
13. Lagani V, Tsamardinos I. 2013. *Computational and Structural Biotechnology Journal* 6
14. Gasch AP, Eisen MB. 2002. *Genome Biol* 3: RESEARCH0059
15. Kashef R, Kamel MS. 2008. In *Image Analysis and Recognition*, pp. 423-34. Berlin Heidelberg: Springer-Verlag
16. Langfelder P, Horvath S. 2007. *BMC Syst Biol* 1: 54
17. Vapnik V, Chapelle O. 2000. *Neural Comput* 12: 2013-36
18. Cox DR. 1972. *J R Stat Soc B* 34: 187-220
19. Heagerty PJ, Lumley T, Pepe MS. 2000. *Biometrics* 56: 337-44
20. Dybowski R. 2000. *Neural computation in medicine: perspectives and prospects*. Presented at Artificial Neural Networks in Medicine and Biology (ANNMB-00)
21. Harrell FE, Jr. 2002. *Regression modeling strategies*: Springer
22. Graf E, Schmoor C, Sauerbrei W, Schumacher M. 1999. *Stat Med* 18: 2529-45
23. Rogers DJ, Tanimoto TT. 1960. *Science* 132: 1115-8
24. Gibbons A. 1985. *Algorithmic graph theory*: Cambridge University Press
25. Enerly E, Steinfeld I, Kleivi K, Leivonen SK, Aure MR, et al. 2011. *PLoS ONE* 6: e16915
26. LGRC. Lung Genomics Research Consortium.
27. van de Vijver MJ, He YD, van't Veer LJ, Dai H, Hart AA, et al. 2002. *N Engl J Med* 347: 1999-2009
28. van 't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, et al. 2002. *Nature* 415: 530-6
29. Rosenwald A, Wright G, Chan WC, Connors JM, Campo E, et al. 2002. *N Engl J Med* 346: 1937-47
30. Rosenwald A, Wright G, Wiestner A, Chan WC, Connors JM, et al. 2003. *Cancer Cell* 3: 185-97
31. Bullinger L, Dohner K, Bair E, Frohling S, Schlenk RF, et al. 2004. *N Engl J Med* 350: 1605-16
32. Bair E, Tibshirani R. 2004. *PLoS Biol* 2: E108
33. Pawitan Y, Bjohle J, Ammer L, Borg AL, Eghazi S, et al. 2005. *Breast Cancer Res* 7: R953-64
34. Wang Y, Klijn JG, Zhang Y, Sieuwerts AM, Look MP, et al. 2005. *Lancet* 365: 671-9
35. Sotiriou C, Wirapati P, Loi S, Harris A, Fox S, et al. 2006. *J Natl Cancer Inst* 98: 262-72
36. Ivshina AV, George J, Senko O, Mow B, Putti TC, et al. 2006. *Cancer Res* 66: 10292-301
37. Sessa R, Hata A. 2013. *Palm Circ* 3: 315-28
38. Ezzie ME, Crawford M, Cho JH, Orellana R, Zhang S, et al. 2012. *Thorax* 67: 122-31
39. Sato T, Liu X, Nelson A, Nakanishi M, Kanaji N, et al. 2010. *Am J Respir Crit Care Med* 182: 1020-9
40. Liu G, Frigeri A, Yang Y, Milosevic J, Ding Q, et al. 2010. *J Exp Med* 207: 1589-97
41. Milosevic J, Pandit K, Magister M, Rabinovich E, Ellwanger DC, et al. 2012. *Am J Respir Cell Mol Biol* 47: 879-87
42. Halappanavar S, Nikota J, Wu D, Williams A, Yauk CL, Stampfli M. 2013. *J Immunol* 190: 3679-86
43. Pandit KV, Corcoran D, Yousef H, Yarlagadda M, Tzouvelekis A, et al. 2010. *Am J Respir Crit Care Med* 182: 220-9
44. Huang GT, Athanassiou C, Benos PV. 2011. *Nucleic Acids Res* 39: W416-23
45. Jenatton R, Mairal J, Obozinski G, Bach F. 2011. *J Mach Learn Res* 12: 2681-720
46. Curtis C, Shah SP, Chin SF, Turashvili G, Rueda OM, et al. 2012. *Nature* 486: 346-52

META-ANALYSIS OF DIFFERENTIAL GENE CO-EXPRESSION: APPLICATION TO LUPUS

SUMIT B. MAKASHIR

*Lindner College of Business, University of Cincinnati, 2925 Campus Green Dr.,
Cincinnati, OH 45221, USA
Email: makashsb@mail.uc.edu*

LEAH C. KOTTYAN

*Center for Autoimmune Genomics and Etiology, Cincinnati Children's Hospital, 3333 Burnet Ave.,
Mail Location 15021, Cincinnati, OH 45229-3039, USA
Email: Leah.Kottyan@cchmc.org*

MATTHEW T. WEIRAUCH*

*Center for Autoimmune Genomics and Etiology, and Division of Biomedical Informatics
Cincinnati Children's Hospital, 3333 Burnet Ave.,
Mail Location 15021, Cincinnati, OH 45229-3039, USA
Email: Matthew.Weirauch@cchmc.org*

We present a novel statistical framework for meta-analysis of differential gene co-expression. In contrast to standard methods, which identify genes that are over or under expressed in disease vs controls, differential co-expression identifies gene pairs with correlated expression profiles specific to one state. We apply our differential co-expression meta-analysis method to identify genes specifically mis-expressed in blood-derived cells of systemic lupus erythematosus (SLE) patients. The resulting network is strongly enriched for genes genetically associated with SLE, and effectively identifies gene modules known to play important roles in SLE etiology, such as increased type 1 interferon response and response to wounding. Our results also strongly support previous preliminary studies suggesting a role for dysregulation of neutrophil extracellular trap formation in SLE. Strikingly, two of the gene modules we identify contain SLE-associated transcription factors that have binding sites significantly enriched in the promoter regions of their respective gene modules, suggesting a possible mechanism underlying the mis-expression of the modules. Thus, our general method is capable of identifying specific dysregulated gene expression programs, as opposed to large global responses. We anticipate that methods such as ours will be more and more useful as gene expression monitoring becomes increasingly common in clinical settings.

1. Introduction

Proper control of gene expression is vital to cellular function, and improper control is thought to play a role in many diseases. There are >50,000 datasets currently available in the Gene Expression Omnibus database [1], covering a wide variety of disease states, tissue types, and conditions. Development of methods for extracting relevant information from these rich data remains a limiting step in the longstanding goal to translate them into useful clinical information.

Differential expression analysis is by far the most popular method for identifying genes that are mis-expressed in a disease state, having been utilized in hundreds of studies to date (reviewed in [2-9]). Such analyses apply statistical methods such as t-tests to cohorts of individuals in order to identify genes with high expression levels in the disease state and low expression levels in normal conditions (or *vice versa*) (reviewed in [10]). However, these analyses often result in large lists of genes that must subsequently be further decomposed into specific dysregulated pathways and expression programs. Furthermore, differences in gene expression levels between individuals due to factors such as genotype, disease severity or stage, and age/gender/race may confound results from standard differential expression analysis methods.

In contrast, differential co-expression analysis is a relatively underutilized, complementary procedure capable of revealing subtle but important relationships missed by standard differential expression methods [11-20]. In differential co-expression, pairs of genes are identified that have correlated expression patterns in one state (e.g., disease), but not the other (e.g., normal conditions). Such methods are therefore capable of capturing relationships that might be influenced by differences between individuals within each cohort. Further, groups of genes that are co-expressed only in one state might represent regulatory programs that are specific to that state, such as pathways that are over-active in a disease.

Despite its clear utility, gene expression data contains a large amount of noise, and findings are thus not always reproducible. Moreover, differences in cohorts, treatments, and experimental methodologies can lead to disparate findings between studies. Meta-analysis (statistically combining results from multiple studies) is a particularly useful tool to address these issues and extract relevant signals from multiple related datasets. Indeed, several studies have employed meta-analysis techniques to identify genes that are consistently co-expressed across studies (e.g., [21-24]), or to identify gene pairs that are specifically co-expressed under certain conditions [17]. Notably, disease signatures are largely consistent across studies [25], emphasizing the potential benefits of applying meta-analysis-based techniques to gene expression compendia.

To our knowledge, there is no standard statistical framework for performing meta-analysis of differential co-expression. In this study, we propose a novel statistical framework, and apply it to identify gene regulatory programs that are mis-expressed in systemic lupus erythematosus (SLE), an incurable, debilitating, and potentially fatal disease affecting an estimated 3.5 million people worldwide [26]. Our results offer new insights into the etiology of SLE, highlighting the potential benefits of differential co-expression in the analysis of disease gene expression datasets.

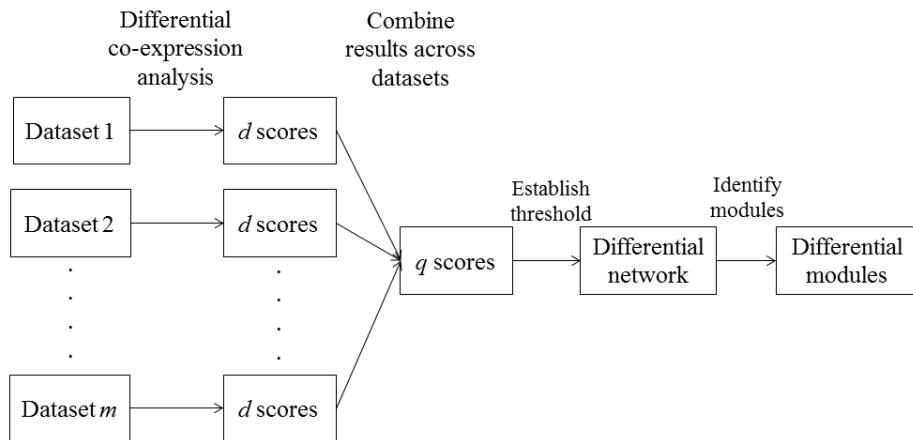


Fig 1. Schematic overview of the method. There are three major components: 1) differential co-expression analysis (e.g. disease vs. control) within each study; 2) combination of results across studies; and 3) gene network and module identification. Differential co-expression analysis of each dataset provides a d score for each gene pair. The d scores are then combined across datasets, resulting in q scores. The q scores are used to construct a differential network, which is decomposed into gene modules representing mis-expressed regulatory programs.

2. Methods

2.1. Statistical framework for meta-analysis of differential co-expression

Let X, Y denote two random variables (e.g., the expression levels of two genes across a collection of people) that follow a bivariate normal distribution under each of the two conditions A, B (e.g., diseased, healthy). Let ρ_A and ρ_B be the correlations between X and Y under conditions A and B, respectively. Let there be m independent studies measuring the two random variables X, Y under both conditions. Under these assumptions, we developed a statistical framework to test the hypothesis that there does not exist a differential relationship between X and Y . The framework takes into account data (e.g., a matrix of gene expression values) from m different studies, along with the sample size of each study. A schematic overview of the method is provided in Fig 1.

2.1.1. s score definition

Within each of the m studies, the sample Pearson correlation coefficient between the two random variables X and Y is calculated as given in Eq. (1):

$$r_{X,Y} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} \quad (1)$$

Note that this computation needs to be performed under each condition (A & B) separately.

The Pearson correlation coefficient, r , is then transformed to Fisher's z score using Fisher's transform as given in Eq. (2):

$$z = \frac{1}{2} \ln \left(\frac{1+r}{1-r} \right) \quad (2)$$

Note: $-\infty < z < \infty$.

Thus, within each study, two z scores are calculated for each pair X,Y by transforming the two sample Pearson correlation coefficients, one under each condition.

Within each study, the difference between the z scores is next calculated as given in Eq. (3):

$$d = z_A - z_B \quad (3)$$

Note: d can also be defined as: $d = z_B - z_A$.

Finally, we define the s score as the sum of the d scores from the m studies, and compute it as:

$$s = \sum_{k=1}^m d_k \quad (4)$$

2.1.2. Probability distribution of s scores

Under the assumption that two random variables X,Y follow a bivariate normal distribution with a correlation of ρ , the random variable Z is approximately normally distributed [27, 28]:

$$Z \sim N \left(\frac{1}{2} \ln \left(\frac{1+\rho}{1-\rho} \right), \frac{1}{n-3} \right) \quad (5)$$

Note: n is the sample size (e.g., the number of conditions/individuals in the expression dataset).

The random variable D, by the d score definition given in Eq. (3), is then the difference of two random variables Z_A and Z_B . Since Z_A and Z_B are normally distributed random variables and are independent by definition, the random variable D will also follow a normal distribution, with the expected value equal to the difference of the expected values of Z_A and Z_B , and variance equal to the sum of the variances of Z_A and Z_B :

$$D \sim N \left(\frac{1}{2} \ln \left(\frac{1+\rho_A}{1-\rho_A} \right) - \frac{1}{2} \ln \left(\frac{1+\rho_B}{1-\rho_B} \right), \left(\frac{1}{n_A-3} \right) + \left(\frac{1}{n_B-3} \right) \right) \quad (6)$$

We next define a random variable S that is a linear combination of the normal random variables D_1, D_2, \dots, D_m , which are all independent by definition. Thus, using the result stated in Eq. (5), the random variable S will follow a normal distribution as given in Eq. (7).

$$S \sim N\left(m \left(\frac{1}{2} \ln \left(\frac{1+\rho_A}{1-\rho_A} \right) - \frac{1}{2} \ln \left(\frac{1+\rho_B}{1-\rho_B} \right) \right), \sum_{k=1}^m \left(\frac{1}{n_{A_k}-3} + \frac{1}{n_{B_k}-3} \right) \right) \quad (7)$$

Further, under the assumption that there does not exist a differential relationship between X and Y , ρ_A will be equal to ρ_B , and the expected value of the random variable S will be zero.

The random variable S can then be converted to a standard normal random variable, denoted Q :

$$Q = \frac{S-0}{\sqrt{\sum_{k=1}^m \left(\frac{1}{n_{A_k}-3} + \frac{1}{n_{B_k}-3} \right)}} \quad (8)$$

Since Q is a standard normal random variable, for any observed s score, the corresponding q score can be calculated using the result stated in Eq. (8). The P-value ($P > |q|$) can then be calculated easily using the standard normal distribution function. We note that this P-value provides a way to test the hypothesis that there does not exist a differential relationship between X and Y .

2.2. SLE dataset selection for meta-analysis

We selected a total of five gene expression studies after screening several based on cell type, sample size, patient treatment, etc. (Table 1). All five studies measure gene expression in Peripheral Blood Mononuclear Cells (PBMC) of SLE patients and healthy individuals. We excluded any samples from individuals that received intensive immunosuppressive therapy (e.g., greater than 20 mg/day of steroids or interferon alpha Kinoid immunization).

2.3. Data pre-processing

Data pre-processing consisted of two operations: normalization and probe ID mapping. For three datasets, only pre-normalized data were available, which we used as-is. We quantile normalized the remaining two datasets, using the *normalize.quantile* function from the *preprocessCore* R library. We mapped gene probe IDs from each study to Ensembl gene IDs. In the cases where multiple probes mapped to one Ensembl ID (e.g. for genes with multiple splice variants), we used the mean expression value of the probes. These mappings resulted in a total of 18,663 unique genes across the five studies.

Table 1. Gene expression datasets used in this study

Database	ID	Study Title	N (SLE)	N (CTRL)
GEO	GSE22098	Whole blood transcriptional profiles of patients with active tuberculosis (TB) and other inflammatory and infectious diseases	110	81
GEO	GSE20864	Human peripheral blood cells: systemic lupus erythematosus vs healthy individual	21	45
GEO	GSE39088	Down-regulation of Interferon signature in systemic lupus erythematosus patients by active immunization with Interferon alpha-Kinoid	26	46
GEO	GSE8650	Blood Leukocyte Microarrays to Diagnose Systemic Onset Juvenile Idiopathic Arthritis and Follow IL-1 blockade	13	21
Array Express	E-MTAB-145	Transcription profiling of human separated leukocyte subsets in SLE and vasculitis	13	25

2.4. *Meta-analysis of differential gene co-expression in SLE*

We performed differential gene co-expression meta-analysis using the framework described in section 2.1, and outlined in Fig 1. Here, the random variables X, Y represent the expression levels of two genes across a collection of individuals, and conditions A and B represent diseased (SLE) and healthy (control) conditions, respectively. Note that in applying the meta-analysis framework to SLE, we converted any negative z scores to a z score of zero, since negative correlations are non-transitive, and thus might have adverse effects on subsequent clustering analyses. This is a conservative modification, in the sense that it may lower the d score, but can never increase it; this modification thus will not result in any false positives in our network. The 18,663 unique genes translate into 174,144,453 pairwise relationships. However not all genes are present in all studies, and not all genes present in a study have the minimum number (3) of non-missing observations we required to compute correlation coefficients. In this study, we only considered gene pairs for which we could compute correlation coefficients in at least four studies, a total of 85,276,272 pairs (~49%). We used this value to adjust our final P-values for multiple testing with the Benjamini-Hochberg procedure, using *p.adjust* function in R, requiring a significance level of 0.05.

2.5. Comparison of Type I error rates and statistical power

We empirically estimated Type I error rate and statistical power using simulation-based approaches. We performed 10,000 iterations wherein we created between 2 and 10 studies, with each study consisting of between 25 and 100 samples/observations (all values were chosen uniformly). To create expression levels for two genes across the conditions, we picked random values independently from two standard normal distributions for the Type I error rate analysis, since its null hypothesis assumes no correlation. For the power analysis, which has the null hypothesis that there is some underlying correlation, we simulated expression levels by picking from a bivariate normal distribution with expected values [0, 0], variances [1, 1], and correlation (rho) values drawn from a uniform (0,1) distribution. We calculated the sample Pearson correlation between the two genes across conditions, and performed meta-analysis using one of the three methods (ours, Fisher's, or Stouffer's). Type I error rate and power were then calculated as the fraction of meta-analysis results across the 10,000 iterations with P-value < 0.05 (i.e., $\alpha=0.05$). We repeated this process 100 times and report the mean.

2.6. Differential expression analysis

We also performed standard differential gene expression analysis. For each dataset, we used a two sample independent t-test to estimate the significance of the difference of the expression levels of each gene between SLE and control conditions, and calculated the mean across the five datasets.,

2.7. Identification and annotation of gene modules

We identified densely connected regions of the final network (gene modules) using the fastgreedy algorithm in the R *igraph* package. We annotated modules by calculating their overlap with known biological processes using Fisher's exact test in ToppGene [29]. We identified enriched transcription factor (TF) binding sites in 1,000 base upstream regions of genes in each module using a large collection of human TF binding motifs [30] and the Pscan algorithm [31]. We identified TFs with enriched ChIP-seq binding peaks in the promoters of module genes using data obtained from UC Santa Cruz [32] and PAZAR [33], and a novel statistical method. Briefly, for each ChIP-seq dataset, we calculated its observed overlap with the promoter regions of the genes of the module. We then built a distribution of expected overlap values from 1,000 randomly selected promoters. The distribution of the overlap scores from the randomized data resembles a normal distribution (not shown), which we used to generate a Z-score and P-value for the observed number of peaks that overlap each module. P-values were corrected using Bonferroni's method.

2.7. Availability

Source code is available at <https://tf.cchmc.org/pubs/makashir2014/>. Full results are available from MTW upon request.

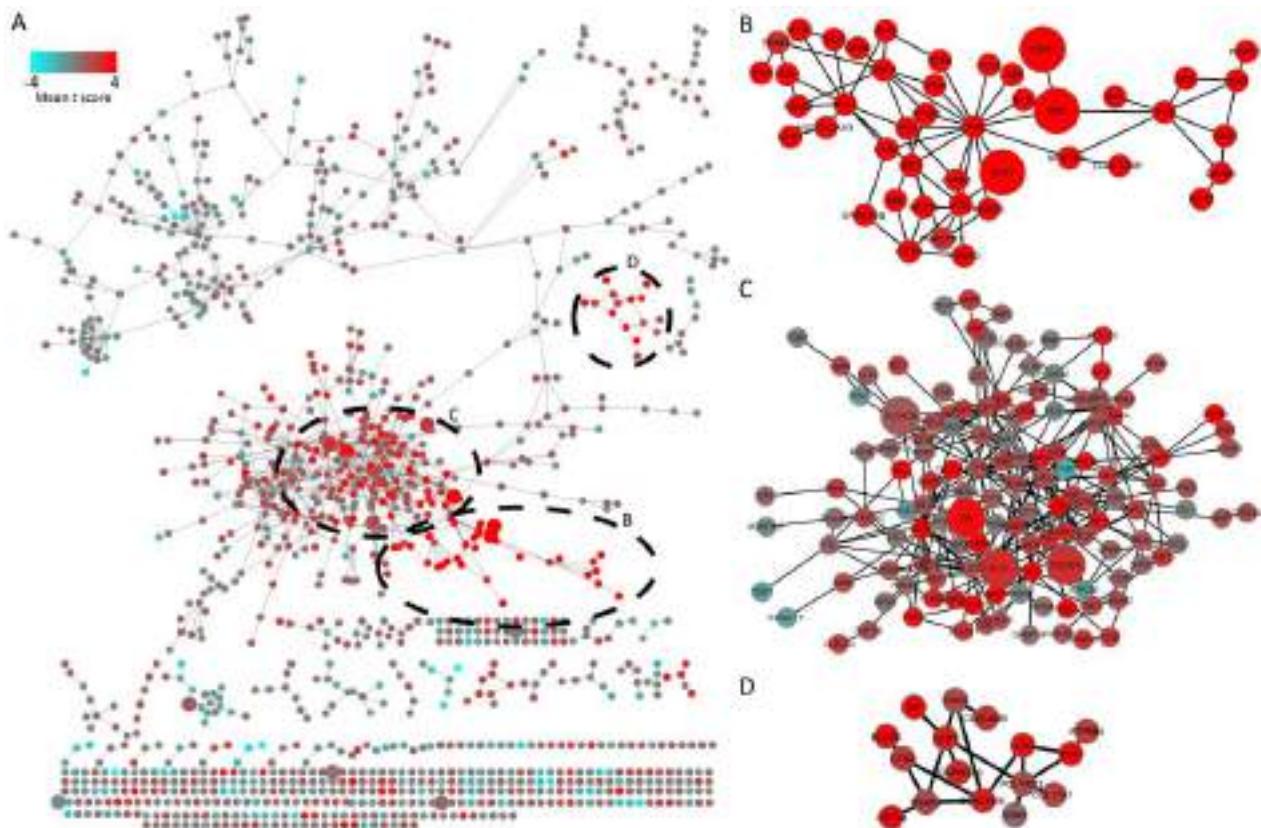


Fig.2. Global view of the SLE-specific gene co-expression network. Each node represents a gene, with edges between genes indicating co-expression relationships specific to SLE. Nodes are colored based on standard differential expression analysis; red indicates over-expression, while blue indicates under-expression in SLE, compared to healthy controls (see key, upper left). Large nodes indicate genes genetically associated with SLE. The network has been laid out such that highly connected groups of nodes are located near each other in 2D space. A. Full network. Dashed circles indicate gene modules highlighted in panels B-D and discussed in the text.

3. Results

3.1. A network of genes specifically co-expressed in SLE

Using simulations, we empirically estimated the Type I error rate of our method to be 0.0496, which is comparable to that of two standard statistical meta-analysis procedures, Fisher's method (0.0494), and Stouffer's method (0.0496) (see Methods). Further, our method has significantly greater statistical power than these methods (0.8798 vs. 0.8527 and 0.8534 for Fisher and Stouffer, respectively; $P < 10^{-16}$, paired t-test with Bonferroni multiple testing correction; see Methods).

We applied our procedure to identify pairs of genes specifically co-expressed in blood samples from SLE patients, relative to controls. The resulting network, which contains 1,250 nodes and 1,310 links, is depicted in Fig 2A. Edges in the network indicate pairs of genes that are strongly co-expressed in SLE patients, but not in normal individuals, and thus might indicate regulatory relationships specific to the disease state.

Twelve genes contained in our network are genetically associated with SLE, a 2.8-fold enrichment over random expectation ($P \sim 0.001$), and a near identical number to that obtained from a list of the top 1,250 differentially expressed genes (14). Interestingly, only seven of these genes are identified by both methods, highlighting the complementary nature of differential expression and co-expression based approaches. We note that we used a strict threshold to generate the network, resulting in a largely sparse network. Notwithstanding, there is a clear modularity to the network, with one large component containing mostly down-regulated genes (top), a second large component containing mostly up-regulated genes (middle), and several smaller components. We therefore used a graph-theoretic algorithm to extract gene modules from the network (see Methods), which might correspond to specific regulatory programs that are mis-expressed in SLE.

3.2. Gene co-expression modules specific to SLE

3.2.1. Type I interferon response

The most strongly up-regulated genes in the network comprise a single gene module enriched for genes involved in the Type I interferon response ($P < 10^{-29}$), a process with a well-established history of elevated levels in the sera of SLE patients [34, 35]. Specifically, the module contains many interferon inducible genes, including OAS family members, and *IFIT1*, *IFIT5*, *IFIH1*, and *MX1* [36, 37] (Fig 2B). Intriguingly, this module also contains three SLE-associated genes (*STAT1*, *IRF7*, and *IFIH1*), all of which have risk variants associated with an increase in Type I IFN activity [37-39]. Given that *STAT1* and *IRF7* are transcription factors (TFs) with well-known roles in the interferon response [40, 41] and established elevated expression and activity in SLE [42, 43], we hypothesized that their heightened activity in SLE might play a role in the elevated expression of Type I interferon genes in SLE. Indeed, two lines of evidence strongly support this possibility. First, the most strongly enriched TF binding motifs in the promoter regions of the genes of this module are for *IRF3* (highly related to *IRF7*) and *STAT1* ($P < 10^{-15}$ and 10^{-11} , respectively, see Methods). Second, *STAT1* and *IRF1* ChIP-seq binding peaks are over 200-fold enriched in the promoters of these genes ($P < 10^{-186}$ and 10^{-177} , respectively), covering 25 and 17 of the 44 genes in the module, respectively (see Methods) (we could not locate ChIP-seq data for *IRF7*; *IRF1* is highly related, and recognizes a highly related DNA binding motif). In summary, our data suggest that this Type I interferon response gene module is specifically over-expressed in SLE patients as a consequence of the elevated activity of *STAT1* and *IRF7*.

3.2.2. Cell movement and response to wounding

Our method also identified a module strongly enriched for genes involved in response to wounding ($P < 10^{-7}$) and leukocyte cell migration ($P < 10^{-6}$) (Fig 2C). Four of the genes contained in this module are encoded in regions genetically associated with SLE (*ELF1*, *LYN*, *FCGR2A*, and *FCGR3B*). *ELF1* is a TF, and we again found that *ELF1*'s DNA binding motif is significantly enriched in the promoters of the genes in this module ($P < 10^{-3}$). Strikingly, *ELF1* ChIP-seq binding events in B cells are present in the promoters of 64 of the 134 genes in this module (>12-fold enrichment, $P < 10^{-80}$). Although the precise role of cell migration in SLE has yet to be fully elucidated, recent studies in mice have demonstrated that blocking chemotaxis of immune effector cells into organs can inhibit immune-mediated damage such as nephritis [44]. Further analysis of the genes in this module will likely reveal the role played by *ELF1* in the dysregulated co-expression of mediators of cell migration, and the contribution of this pathway to the pathoetiology of SLE.

3.2.3. Immune defense against extracellular organisms

Finally, we identified a gene module containing genes with elevated expression levels involved in immune defense against extracellular organisms such as response to bacterium, fungus, and symbiotic cells ($P < 10^{-4}$) (Fig 2D). In particular, twelve out of sixteen (75%) of these genes have established roles in neutrophil maturation and extracellular trap formation (including *DEFA4*, *CTSG*, *ELANE*, *PGYRP1*, *OLR1*, and *CD24*) [45-47]. Among these, *DEFA4* (a defensin involved in neutrophil extracellular traps) is particularly intriguing, since it is known to impact some of the immune-mediated tissue damage that occurs in response to autoantibody deposition in SLE [48]. Further, both *DEF4A* and *CTSG* (a serine protease that is also released by neutrophils during neutrophil extracellular trap formation) are expressed at higher levels in SLE patients with active disease [49]. Together, these results support recent studies suggesting a role for neutrophil extracellular traps in SLE etiology [50] and, importantly, identify specific members of this pathway whose regulatory programs might be compromised in SLE patients.

4. Discussion

Despite numerous advantages to differential co-expression-based approaches, methods such as ours remain largely underutilized. Here, we present a novel general differential co-expression meta-analysis framework, and show how it can be used in conjunction with differential expression techniques, transcription factor binding analysis, and genetic association information to identify dysregulated gene expression programs in SLE. Notably, our method identifies several regulatory programs not captured in a previous meta-analysis study of pathway differential expression in SLE

(both capture the interferon response) [51]. We anticipate that integrated methods such as ours will have increasing utility as data from genome-scale technologies become more and more prevalent. Notably, results from our method offered considerable biological insights, despite data limitations relating to differing study designs, differing array platforms, and heterogeneous cell populations. It is likely that applications to well-controlled datasets (e.g., single population or single cell RNA-seq data) will afford even higher resolution views into regulatory programs that are disrupted in human diseases.

5. Conclusions

We developed a novel statistical methodology for differential gene co-expression meta-analysis. Our method is complementary to traditional differential expression methods, and is capable of identifying specific dysregulated expression programs, as opposed to large global responses. Our framework is flexible and general, and thus can be applied to other diseases (e.g, cancer), or other genome-wide data types (e.g. RNA-seq). Here, application to SLE captured a significant number of genetically associated genes, and provided new insights into SLE etiology. Specifically, our data implicate the STAT1 and IRF7 TFs in the elevated Type I Interferon response in SLE, and suggest a role for the ELF1 TF in a novel dysregulated cell movement pathway. Importantly, these predictions are strongly supported by enrichment of binding sites and ChIP-seq binding events for these TFs in the promoters of the associated gene modules. Further, we have revealed a new potential role for the dysregulation of specific neutrophil extracellular trap genes in SLE. As technologies for measuring gene expression continue to mature, we anticipate that meta-analysis methods such as ours will play a vital role in obtaining high-confidence biological insights into the gene expression programs of many human diseases.

5. Acknowledgments

We thank Xiaoting Chen and Siddharth Dixit for computational support, and Josh Stuart for helpful discussions.

References

1. Barrett, T., et al. *Nucleic Acids Res.* **41**(Database issue):D991-5 (2013).
2. Jakobs, T.C. *Cold Spring Harb Perspect Med.* **4**(7) (2014).
3. Ruppert, V. and B. Maisch. *Herz.* **37**(6):619-26 (2012).
4. Bacher, U., A. Kohlmann, and T. Haferlach. *Cancer Treat Rev.* **36**(8):637-46 (2010).

- 5.Kinter, J., T. Zeis, and N. Schaeren-Wiemers. *Int MS J.* **15**(2):51-8 (2008).
- 6.Fehrmann, R.S., et al. *Oncologist.* **12**(8):960-6 (2007).
- 7.Nanni, L., et al. *J Mol Cell Cardiol.* **41**(6):934-48 (2006).
- 8.Lemmer, E.R., S.L. Friedman, and J.M. Llovet. *Semin Liver Dis.* **26**(4):373-84 (2006).
- 9.Mandel, M. and A. Achiron. *Lupus.* **15**(7):451-6 (2006).
- 10.Cui, X. and G.A. Churchill. *Genome Biol.* **4**(4):210 (2003).
- 11.Kostka, D. and R. Spang. *Bioinformatics.* **20 Suppl 1**:i194-9 (2004).
- 12.Choi, J.K., et al. *Bioinformatics.* **21**(24):4348-55 (2005).
- 13.Wang, B.D., et al. *Mol Cancer.* **9**:98 (2010).
- 14.Fu, S., X. Pan, and W. Fang. *Mol Med Rep.* **10**(2):713-8 (2014).
- 15.Lai, Y. *Bioinformatics.* **24**(5):666-73 (2008).
- 16.de Jong, S., et al. *PLoS One.* **7**(6):e39498 (2012).
- 17.Gillis, J. and P. Pavlidis. *BMC Bioinformatics.* **10**:306 (2009).
- 18.Fang, G., et al. *Pac Symp Biocomput:*145-56 (2010).
- 19.Li, K.C. *Proc Natl Acad Sci U S A.* **99**(26):16875-80 (2002).
- 20.Oldham, M.C., S. Horvath, and D.H. Geschwind. *Proc Natl Acad Sci U S A.* **103**(47):17973-8 (2006).
- 21.Kim, S.K., et al. *Science.* **293**(5537):2087-92 (2001).
- 22.Lee, H.K., et al. *Genome Res.* **14**(6):1085-94 (2004).
- 23.Guan, Y., et al. *PLoS Comput Biol.* **4**(9):e1000165 (2008).
- 24.Stuart, J.M., et al. *Science.* **302**(5643):249-55 (2003).
- 25.Dudley, J.T., et al. *Mol Syst Biol.* **5**:307 (2009).
- 26.Danchenko, N., J.A. Satia, and M.S. Anthony. *Lupus.* **15**(5):308-18 (2006).
- 27.Fisher, R.A. *Biometrika.* **10**:507-521 (1915).
- 28.Fisher, R.A. *Metron* **1**:3-32 (1921).
- 29.Chen, J., et al. *Nucleic Acids Res.* **37**(Web Server issue):W305-11 (2009).
- 30.Weirauch, M.T., et al. *Cell.* **158**(6):1431-43 (2014).
- 31.Zambelli, F., G. Pesole, and G. Pavesi. *Nucleic Acids Res.* **37**(Web Server issue):W247-52 (2009).
- 32.Rosenbloom, K.R., et al. *Nucleic Acids Res.* **41**(Database issue):D56-63 (2013).
- 33.Portales-Casamar, E., et al. *Nucleic Acids Res.* **37**(Database issue):D54-60 (2009).
- 34.Hooks, J.J., B. Detrick-Hooks, and A.I. Levinson. *J Am Vet Med Assoc.* **181**(10):1111-4 (1982).
- 35.Hooks, J.J., et al. *N Engl J Med.* **301**(1):5-8 (1979).
- 36.Ghodke-Puranik, Y. and T.B. Niewold. *Int J Clin Rheumatol.* **8**(6) (2013).
- 37.Kariuki, S.N., et al. *J Immunol.* **182**(1):34-8 (2009).
- 38.Mavragani, C.P., et al. *Front Immunol.* **4**:238 (2013).
- 39.Niewold, T.B. *J Interferon Cytokine Res.* **31**(12):887-92 (2011).
- 40.Darnell, J.E., Jr., I.M. Kerr, and G.R. Stark. *Science.* **264**(5164):1415-21 (1994).
- 41.Marie, I., J.E. Durbin, and D.E. Levy. *EMBO J.* **17**(22):6660-9 (1998).
- 42.Dong, J., et al. *Lupus.* **16**(2):101-9 (2007).
- 43.Lin, L.H., P. Ling, and M.F. Liu. *J Rheumatol.* **38**(9):1914-9 (2011).
- 44.Bignon, A., et al. *J Immunol.* **192**(3):886-96 (2014).
- 45.Parlato, M., et al. *J Immunol.* **192**(5):2449-59 (2014).
- 46.Ghosh, A., et al. *Invest Ophthalmol Vis Sci.* **50**(9):4185-91 (2009).
- 47.Martinelli, S., et al. *J Biol Chem.* **279**(42):44123-32 (2004).
- 48.Villanueva, E., et al. *J Immunol.* **187**(1):538-52 (2011).
- 49.Tamiya, H., et al. *Rheumatol Int.* **27**(2):147-52 (2006).
- 50.Bouts, Y.M., et al. *Autoimmunity.* **45**(8):597-601 (2012).
- 51.Arasappan, D., et al. *BMC Med.* **9**:65 (2011).

MELANCHOLIC DEPRESSION PREDICTION BY IDENTIFYING REPRESENTATIVE FEATURES IN METABOLIC AND MICROARRAY PROFILES WITH MISSING VALUES

ZHI NIE[†], TAO YANG[†], YASHU LIU[†], BINBIN LIN[†], QINGYANG LI[†], VAIBHAV A NARAYAN[‡], GAYLE WITTENBERG[‡], JIEPING YE[†]

[†]*Department of Computer Science and Engineering,
Center for Evolutionary Medicine and Informatics, The Biodesign Institute,
Arizona State University, Tempe, AZ 85287, USA*

[‡]*Johnson & Johnson Pharmaceutical Research & Development, LLC,
Titusville, NJ, USA*

E-mail: [†]{ Zhi.Nie, T.Yang, Yashu.Liu, Binbin.Lin, Qingyang.Li, Jieping.Ye }@asu.edu,
[‡]{ VNaray16, GWittenb }@its.jnj.com

Recent studies have revealed that melancholic depression, one major subtype of depression, is closely associated with the concentration of some metabolites and biological functions of certain genes and pathways. Meanwhile, recent advances in biotechnologies have allowed us to collect a large amount of genomic data, e.g., metabolites and microarray gene expression. With such a huge amount of information available, one approach that can give us new insights into the understanding of the fundamental biology underlying melancholic depression is to build disease status prediction models using classification or regression methods. However, the existence of strong empirical correlations, e.g., those exhibited by genes sharing the same biological pathway in microarray profiles, tremendously limits the performance of these methods. Furthermore, the occurrence of missing values which are ubiquitous in biomedical applications further complicates the problem. In this paper, we hypothesize that the problem of missing values might in some way benefit from the correlation between the variables and propose a method to learn a compressed set of representative features through an adapted version of sparse coding which is capable of identifying correlated variables and addressing the issue of missing values simultaneously. An efficient algorithm is also developed to solve the proposed formulation. We apply the proposed method on metabolic and microarray profiles collected from a group of subjects consisting of both patients with melancholic depression and healthy controls. Results show that the proposed method can not only produce meaningful clusters of variables but also generate a set of representative features that achieve superior classification performance over those generated by traditional clustering and data imputation techniques. In particular, on both datasets, we found that in comparison with the competing algorithms, the representative features learned by the proposed method give rise to significantly improved sensitivity scores, suggesting that the learned features allow prediction with high accuracy of disease status in those who are diagnosed with melancholic depression. To our best knowledge, this is the first work that applies sparse coding to deal with high feature correlations and missing values, which are common challenges in many biomedical applications. The proposed method can be readily adapted to other biomedical applications involving incomplete and high-dimensional data.

Keywords: melancholic depression, sparse coding, missing value, clustering, disease prediction, biomarker identification, feature learning.

1. Introduction

Understanding the fundamental biology underlying melancholic depression is a very challenging problem of great clinical importance for researchers from medical and psychiatric research communities. Unlike some other subtypes of depression, melancholic depression is described as “mainly biologically based rather than determined by personality or life circumstances”,¹

which motivates researchers to discover biological evidence of the disease. Research with regard to this aspect has made progress in recent years. For instance, it has been shown that an elevated level of concentration of certain metabolites in plasma is found among the depressive patients with melancholia.² More recently, Gabbay *et al.*³ pointed out the significance of kynurenine pathway in adolescent depression with melancholic features through comparing adolescents with melancholic depression with non-melancholic depression and healthy adolescents. Also, recently through gene ontology and pathway analyses, certain biological functions of differentially expressed mRNAs were identified as related to fundamental metabolic processes and brain disorders.⁴

On the other hand, recent advances in biotechnologies have made it possible to detect a large number of metabolites from human tissue extract.⁵ Meanwhile, the microarray technology has taken us from being able to analyze the biological functions of only a few related genes or proteins at one time to the place where global investigation of cellular activities is possible.⁶ With data on such a large scale available, one promising approach that can potentially offer us a deeper understanding of collective impact of numerous factors involved in the pathogenesis of melancholic depression and its prospective treatments is to build predictive models based on all of the information available using machine learning approaches. However, the “curse of dimensionality” due to the fact that the number of variables of interest far exceeds the number of samples available renders most of traditional classification/regression algorithms less effective in this setting. Furthermore, strong empirical correlations between the variables, especially in the case of microarray data where there is high degree of linear dependence between expression measures of a group of genes sharing the same biological pathways,⁷ tremendously limit the prediction performance of traditional machine learning algorithms. Another major issue with data collected on a large scale is the presence of missing values, which is ubiquitous in biomedical applications.

Most of the existing methods were designed to deal with either the problem of strong empirical correlations between the variables or the problem of missing values. For instance, Bühlmann *et al.*⁸ recently proposed a bottom-up agglomerative clustering algorithm to deal with correlations between the variables, but their method cannot be readily used in the context of missing values. As for the issue of missing values, there are two basic approaches to dealing with missing data. We can either discard the samples with missing values or impute the missing data. The shortcoming of the first approach is obvious. It does not make full use of available information. The second approach, i.e., imputation of missing data, generally involves certain assumptions about the missing pattern of the data which may not be satisfied in applications. The most commonly used imputation technique, EM, for example, assumes that data is sampled from a Gaussian distribution and the missing-at-random (MAR) assumption is satisfied.

We hypothesize that the problem of missing values might potentially benefit from the correlations between the variables; for example, a variable with missing values could borrow information from its correlated variables. However, simply imputing the missing values of a variable by exploiting information from its correlated variables still leaves the problem of empirical correlations between the variables unsolved. Therefore, instead of discarding the

incomplete samples or imputing the missing values, we attempt to generate a compressed set of representative features for all the samples from the data with a group of correlated variables represented by one or a few features. We demonstrate in this paper that sparse coding, which has been shown to be very effective in object recognition and image denoising applications,^{9,10} is desirable for such a task. Specifically, we apply sparse coding in such a way that the learned dictionary corresponds to a set of representative features and each variable is represented as a sparse combination of these features. Furthermore, we develop an efficient algorithm to solve the proposed sparse coding formulation to deal with missing values.

We apply the proposed algorithm to datasets of metabolic and microarray profiles collected from a group of subjects consisting of both patients with melancholic depression and healthy controls. Results from our experiments revealed that features obtained from our method significantly outperform those generated from several baseline methods based on traditional clustering methods and standard data imputation techniques. In particular, in comparison with our baseline methods, the representative features learned by the proposed method achieve much better performance in predicting the disease status of the subjects with melancholic depression on both datasets. In addition, on the dataset of metabolic profiles, we found that most of the known metabolites within each cluster are biologically relevant. These results demonstrate the promise of the proposed method for learning from incomplete and high-dimensional biomedical data.

The rest of the paper is organized as follows. In section 2, we formulate the sparse coding problem in the presence of missing values. In section 3, we describe the datasets used in the analysis and present experimental results. Section 4 concludes the paper.

2. Learning Representative Features via Sparse Coding

In this section, we present our sparse coding formulation to learn a compressed representative set of features such that the observations from all the samples on each variable can be represented as a sparse linear combination of these learned features. The proposed formulation can naturally deal with missing values.

Suppose we are given a dataset of m samples and their observations on n variables with missing values which we denote as $\mathcal{X} = \{(x_1, \Omega_1), \dots, (x_n, \Omega_n)\}$. Each x_i ($1 \leq i \leq n$) is an m dimensional column vector representing measurements of all the samples on the i -th variable (e.g., concentrations of the i -th metabolite or measurements of the i -th gene expression), and Ω_i is an ordered set of integers ranging from 1 to m including the indices of samples whose measurements on the i -th variable are observed. If there is no missing value in x_i , then Ω_i includes all integers between 1 and m . Our goal is to use sparse coding to learn a set of k representative features such that each variable x_i can be well represented by a sparse combination of these k features. In the presence of missing values, the sparse coding problem can be formulated as the following optimization problem:

$$\begin{aligned} & \min_{\mathcal{D}, z_1, \dots, z_n} \sum_{i=1}^n \frac{1}{2} \|\mathcal{P}_{\Omega_i}(\mathcal{D}z_i - x_i)\|_2^2 + \lambda \|z_i\|_1 \\ & \text{s.t.} \quad \|\mathcal{D}_j\|_2 \leq 1; 1 \leq j \leq k, \end{aligned} \tag{1}$$

Algorithm 2.1 Stochastic Coordinate Coding with Missing Values**Initialization:**

Samples $X = \{x_1, x_2, \dots, x_n\}$, missing indices $\Omega = \{\Omega_1, \Omega_2, \dots, \Omega_n\}$, $\lambda \in \mathbb{R}$, initial dictionary $\mathcal{D}^0 \in \mathbb{R}^{m \times k}$, initial combination coefficients $Z = \{z_1, z_2, \dots, z_n\}$, number of iterations T .

```

1:  $\mathcal{H} \in \mathbb{R}^{k \times k} \leftarrow 0$ 
2: for  $t = 1$  to  $T$  do
3:   for  $i = 1$  to  $n$  do
4:     Update coefficients via a few iterations of coordinate descent according to (3)
5:      $z_i \leftarrow \arg \min_{z_i} f_{\mathcal{D}}(z_i) \equiv \frac{1}{2} \|\mathcal{P}_{\Omega_i}(\mathcal{D}z_i - x_i)\|_2^2 + \lambda \|z_i\|_1$ .
6:     Update Hessian matrix
7:      $\mathcal{H} = \mathcal{H} + z_i z_i^T$ ,
8:     Update the dictionary  $\mathcal{D}^{i-1}$  column by column
9:     for  $j \in \{t | 1 \leq t \leq k, t \in \mathbb{N}, z_i(t) \neq 0\}$  do
10:       $u_j = \mathcal{P}_{\Omega_i}(\mathcal{D}^{i-1}) - \frac{1}{\mathcal{H}[j,j]} z_i(j) * \mathcal{P}_{\Omega_i}(\mathcal{D}^{i-1} * z_i - x_i)$ .
11:       $\mathcal{P}_{\Omega_i}(\mathcal{D}_{:,j}^i) \leftarrow u_j$ .
12:       $\mathcal{D}_{:,j} \leftarrow \frac{1}{\max\{\|\mathcal{D}_{:,j}\|_2, 1\}} \mathcal{D}_{:,j}$ .
13:    end for
14:  end for
15:   $\mathcal{D}^0 \leftarrow \mathcal{D}^n$ 
16: end for
Output:  $\mathcal{D}^n$  .

```

where z_i represents sparse combination coefficients (also called sparse code) for x_i , and $\mathcal{D} \in \mathbb{R}^{m \times k}$, the dictionary or codebook, represents the learned set of features with its j -th column denoted as $\mathcal{D}_{:,j}$. $\mathcal{P}_{\Omega_i}(\cdot)$ projects a matrix into its submatrix consisting of rows indexed by Ω_i . The cardinality of Ω_i is denoted by m_i . Minimization of the first term in (1) leads to a feature set \mathcal{D} such that the observed entries of each variable can be well represented by the features in \mathcal{D} . Note that variables with similar combination coefficients are likely to be correlated. Minimization of the second term in (1) induces sparsity on combination coefficients of each variable, enforcing each variable to be represented by only a small subset of features in \mathcal{D} . λ controls the sparsity of each z_i . The larger the λ is, the sparser each z_i will be. With a proper λ , minimization of these two terms combined will yield a feature set \mathcal{D} such that the observed part of each variable can be well represented by a small subset of features from \mathcal{D} .

Although the problem in (1) is convex with respect to either z_i or \mathcal{D} , it is not jointly convex. Thus, it is difficult to obtain a globally optimal solution. Most algorithms solving the sparse coding problem alternate the step of optimizing over z_i with a fixed \mathcal{D} and the step of optimizing over \mathcal{D} with a fixed z_i .¹¹ In this paper, we extend the framework proposed by Lin *et al.*,¹² which applies to data without missing values, to solve sparse coding with missing values in the data matrix. The detailed description of the algorithm to solve the above problem is presented in Algorithm 2.1.

With a fixed \mathcal{D} , updating z_i amounts to solving a Lasso¹³ problem which can be formulated

as follows:

$$\min_{z_i} f_{\mathcal{D}}(z_i) \equiv \frac{1}{2} \|\mathcal{P}_{\Omega_i}(\mathcal{D}z_i - x_i)\|_2^2 + \lambda \|z_i\|_1. \quad (2)$$

Suppose that we iteratively update the sparse code of each variable for T epochs. The total number of Lasso problems involved is Tn . Even with state-of-the-art solvers, the total cost of solving so many Lasso problems is prohibitive, particularly in the case of the microarray data where there are usually at least tens of thousands of genes involved. In our algorithm, we adopt the strategy of solving the Lasso problem incrementally by updating only the support of z_i for a few times via coordinate descent with a warm start. This strategy has proved to be computationally efficient in practice while still yielding competitive performance.

In the algorithm, each time we pick one element, say the j -th element z_{ij} ($1 \leq j \leq k$), to update with all the other coordinates fixed. Under this circumstance, (2) can be converted to a problem with a closed form solution. Let $z_i = z_i^s$ before z_{ij} is updated. Let $z_i = z_i^{s+1}$ after z_{ij} is updated, and $\bar{z}_{ij} = [z_{i,1}, \dots, z_{i,j-1}, z_{i,j+1}, \dots, z_{i,k}]^T$, $\bar{\mathcal{D}}_{\cdot j} = [\mathcal{D}_{\cdot 1}, \dots, \mathcal{D}_{\cdot j-1}, \mathcal{D}_{\cdot j+1}, \dots, \mathcal{D}_{\cdot k}]$. Apparently, $\bar{z}_{ij}^{s+1} = \bar{z}_{ij}^s$, and z_{ij}^{s+1} is the only unknown variable. Plugging z_i^{s+1} into $f_{\mathcal{D}}$, we have

$$\begin{aligned} f_{\mathcal{D}}(z_i^{s+1}) &= \frac{1}{2} \|\mathcal{P}_{\Omega_i}(x_i - \bar{\mathcal{D}}_{\cdot j} \bar{z}_{ij}^{s+1})\|_2^2 - (\mathcal{P}_{\Omega_i}(x_i - \bar{\mathcal{D}}_{\cdot j} \bar{z}_{ij}^{s+1}))^T \mathcal{P}_{\Omega_i}(\bar{\mathcal{D}}_{\cdot j}) z_{ij}^{s+1} \\ &\quad + \frac{1}{2} \|\mathcal{P}_{\Omega_i}(\bar{\mathcal{D}}_{\cdot j})\|_2^2 (z_{ij}^{s+1})^2 + \lambda |z_{ij}^{s+1}| + \lambda \|\bar{z}_{ij}^{s+1}\|_1 \\ &= \frac{1}{2} \|\mathcal{P}_{\Omega_i}(x_i - \bar{\mathcal{D}}_{\cdot j} \bar{z}_{ij}^s)\|_2^2 - (\mathcal{P}_{\Omega_i}(x_i - \bar{\mathcal{D}}_{\cdot j} \bar{z}_{ij}^s))^T \mathcal{P}_{\Omega_i}(\bar{\mathcal{D}}_{\cdot j}) z_{ij}^{s+1} \\ &\quad + \frac{1}{2} \|\mathcal{P}_{\Omega_i}(\bar{\mathcal{D}}_{\cdot j})\|_2^2 (z_{ij}^{s+1})^2 + \lambda |z_{ij}^{s+1}| + \lambda \|\bar{z}_{ij}^s\|_1. \end{aligned}$$

By setting $\partial f_{\mathcal{D}}(z_i^{s+1}) / \partial z_{ij}^{s+1} = 0$, we have

$$-(\mathcal{P}_{\Omega_i}(x_i - \bar{\mathcal{D}}_{\cdot j} \bar{z}_{ij}^s))^T \mathcal{P}_{\Omega_i}(\bar{\mathcal{D}}_{\cdot j}) + \|\mathcal{P}_{\Omega_i}(\bar{\mathcal{D}}_{\cdot j})\|_2^2 (z_{ij}^{s+1}) + \lambda \text{sign}(z_{ij}^{s+1}) = 0.$$

Adding and subtracting $\|\mathcal{P}_{\Omega_i}(\bar{\mathcal{D}}_{\cdot j})\|_2^2 z_{ij}^s$, we have

$$-(\mathcal{P}_{\Omega_i}(x_i - \mathcal{D}z_i^s))^T \mathcal{P}_{\Omega_i}(\bar{\mathcal{D}}_{\cdot j}) + \|\mathcal{P}_{\Omega_i}(\bar{\mathcal{D}}_{\cdot j})\|_2^2 (z_{ij}^{s+1}) - \|\mathcal{P}_{\Omega_i}(\bar{\mathcal{D}}_{\cdot j})\|_2^2 (z_{ij}^s) + \lambda \text{sign}(z_{ij}^{s+1}) = 0.$$

This equation has a closed form solution which is given by

$$z_{ij}^{s+1} = S_a \left(\frac{(\mathcal{P}_{\Omega_i}(x_i - \mathcal{D}z_i^s))^T \mathcal{P}_{\Omega_i}(\bar{\mathcal{D}}_{\cdot j})}{\|\mathcal{P}_{\Omega_i}(\bar{\mathcal{D}}_{\cdot j})\|_2^2} + z_{ij}^s \right). \quad (3)$$

where S is the shrinkage operator defined by $S_\alpha(x) = (|x| - \alpha)_+ \text{sign}(x)$, $x, \alpha \in \mathbb{R}$ and $a = \lambda / \|\mathcal{P}_{\Omega_i}(\bar{\mathcal{D}}_{\cdot j})\|_2^2$. Note that although x_i may have missing values, z_i does not contain missing values. Only the rows of \mathcal{D} corresponding to the rows of x_i where values are observed are used to update z_i . It is worth emphasizing that we only update all the coordinates of z_i in the first iteration due to the fact that the dictionary has changed since z_i was updated last time. For iterations afterwards, only the support of z_i is updated.

With a fixed z_i , we only use the newly updated z_i and the corresponding x_i to update \mathcal{D} using gradient descent. The problem can be formulated as follows:

$$\min_{\mathcal{D}} g_{z_i}(\mathcal{D}) \equiv \frac{1}{2} \|\mathcal{P}_{\Omega_i}(x_i - \mathcal{D}z_i)\|_2^2 \quad \text{s.t. } \|\mathcal{D}_{\cdot j}\|_2 \leq 1, \quad 1 \leq j \leq k. \quad (4)$$

The gradient of g_{z_i} with respect to $\mathcal{P}_{\Omega_i}(\mathcal{D}_{\cdot j})$ is $\nabla_{\mathcal{P}_{\Omega_i}(\mathcal{D}_{\cdot j})} g_{z_i} = \mathcal{P}_{\Omega_i}(\mathcal{D}z_i - x_i)z_{ij}$. Note that only the columns of \mathcal{D} corresponding to the support of z_i need to be updated. As to the learning rate, following the practice of Mairal *et al.*¹¹ and Lin *et al.*,¹² we set the learning rate to be $1/\mathcal{H}[j, j]$ where \mathcal{H} is initialized to be zero and accumulates $z_i z_i^T$. Finally, $\mathcal{D}_{\cdot j}$ is normalized to be within the unit ball.

Note that most existing works apply sparse coding on signals, e.g., images to learn a representation of each signal. One novel aspect of the proposed framework is that we apply sparse coding to learn a sparse representation for each variable and use the dictionary \mathcal{D} as features where each row represents a sample and each column represents a feature. In addition, most sparse coding formulations assume that the data is complete, while our proposed framework can naturally deal with missing values in the data. From the perspective of clustering, different from those traditional clustering algorithms such as Kmeans which assign each data point to one cluster, sparse coding can be considered as a soft version of clustering in that it allows a point to belong to different clusters at the same time, depending on the number of non-zero elements in its sparse representation vector. Such flexibility is desirable in many applications, since some data points may be close to multiple clusters. The sparse representation of a data point may be a zero vector, especially when regularization parameter λ in the algorithm is set to be a large value. In this case, the data point can be regarded as an outlier or a noisy point.

3. Data and Experiments

The datasets used in our analysis were collected from a study initiated by Brain Resource Company (BRC) and Johnson & Johnson Pharmaceutical Research & Development, L.L.C(J&J PRD). The overall objective of the study is to identify the best molecular profiles, cognitive and psychophysiological biomarkers in people with depression. In the study, about 100 depressive subjects evenly distributed in gender and age as well as an equal number of matched healthy controls are recruited nationwide by BRC in Australia. All the subjects that are included in the study have been screened to satisfy certain criteria on Hamilton Depression Rating Scale (HAM-D) score, CORE score,¹⁴ toxicology tests, and so on. Part of the study is dedicated mainly to collecting the following information from all the participants : a) Personal medical history; b) Cognition; c) Electrical brain-body function (EBBF); d) Brain structure (e.g. structural MRI, functional MRI); e) Molecular profiles (which includes metabolite, microarray, protein and transcripts profiles). However, not all the subjects have all five blocks of information or all sub-categories of one type of information recorded due to a variety of reasons such as participant dropout, failure of quality control, long storage time, etc. In our analysis, we use metabolite and microarray data from the molecular profiles to demonstrate the effectiveness of the proposed algorithm in dealing with correlated variables as well as the significant discriminative power of the resulting compressed set of features. As for the target, we are interested in melancholic depression. The decision of whether or not a subject is diagnosed with melancholic depression is made on the basis of pyschomotor findings in the CORE scale which consists of 18 items measuring a subject's interactiveness, motor agitation, etc. The score of each item ranges from 0 (no symptom) to 3 (severe symptom). A subject will be labeled as melancholically depressive if he or she has a total score on CORE over 8.

3.1. Analysis on Metabolic Profiles

3.1.1. Data Preprocessing

During the stage of medical screening, a sample of 20ml of plasma was obtained from each of the participants by BRC and was later on sent to J&J PRD where the molecular profiling analysis was carried out. Based on Gas chromatography-mass spectrometry (GS-MS) and Liquid chromatography-mass spectrometry (LC-MS/MS), 272 peaks were acquired with 160 of them being known metabolites and the rest unknown. Considering the fact that concentrations of metabolites change with the increase of storage time, we removed all the samples stored for over 200 days and performed a linear regression of concentration with storage time at the temperature of -20 degrees centigrade on the remaining samples to control for the confounding effects caused by storage time. Also, over 40 metabolites whose concentrations were detected to be highly correlated with storage time were excluded from our analysis. After the pre-processing, we are left with 118 samples and 228 metabolites in total. Among all the 118 samples, 21 were diagnosed with melancholic depression and 97 were healthy controls. About 1.27% of all the entries in the data matrix are missing.

The method of sparse coding proposed in this paper can deal with missing entries in the data matrix. To demonstrate the capability of our method to generate a compressed discriminative set of features even under the presence of missing values, we also include several baseline methods for comparison, which impute the missing entries using some standard missing value imputation techniques including: 1) HalfMin: Impute the missing entries on each variable by filling in half of the minimum of the observed values on that variable; 2) KNN: Find the k nearest neighbors of the variable with missing values based on observed part and assign the missing values to be a weighted combination of its nearest neighbors with the weight determined by the inverse of the Euclidean distance between the variable concerned and the neighbor; 3) Expectation Maximization (EM): Assuming that the underlying distribution of the samples follows a mixture of Gaussian distribution, it iterates between updating the posterior probability of each of the data points and updating the mean, covariance matrix and mixing coefficient of each Gaussian component and filling in the missing entries with conditional expectation given the observed part; 4) Singular value decomposition (SVD): Assuming that there is an inherent low-rank structure in the data, it fills in the missing entries with the values obtained from the low-rank approximation of the data.

Before further data analysis, each variable was normalized to have zero mean and unit standard deviation. In the case of variables with missing values, we simply omitted the missing values when computing the mean and standard deviation.

3.1.2. Classification

With the ratio of the number melancholic depressive subjects to the number of healthy controls being almost 1 to 5, the dataset is extremely imbalanced. Direct application of traditional classification methods like Support Vector Machine (SVM) in this situation would severely biased the classifier toward majority class. Drawing on the experience from Dubey *et al.*(2014),¹⁵ we implemented a scheme which combines the techniques of data undersampling and model ensemble methods to deal with the issue of data imbalance.

Table 1. Classification performance on metabolic profiles

Method	RF Vote				RF Weighted Vote			
	Acc	Sen	Spe	AUC	Acc	Sen	Spe	AUC
HalfMin	0.7624	0.6333	0.7922	0.7128	0.7624	0.6333	0.7922	0.7128
EM	0.7367	0.5833	0.7711	0.6772	0.7290	0.5833	0.7611	0.6722
KNN	0.7624	0.6833	0.7822	0.7328	0.7624	0.6333	0.7922	0.7128
SVD	0.7450	0.5833	0.7811	0.6822	0.7547	0.6333	0.7822	0.7078
HC	0.7540	0.8000	0.7489	0.7744	0.7214	0.6500	0.7411	0.6956
Kmeans	0.7778	0.7167	0.7922	0.7544	0.7861	0.7167	0.8022	0.7594
SC	0.8315	0.8333	0.8344	0.8339	0.8315	0.8333	0.8344	0.8339

In this scheme, samples from each of the two classes were randomly partitioned into 10 folds of (approximately) equal size. One fold from both classes were set aside for testing and the rest were used as training set. During the training stage, we used all the samples from the minority class and randomly subsampled with replacement the same amount of samples from the majority class to build a classifier. The process of subsampling was repeated p times (in our experiments, we choose $p = 30$) so that p different classifiers will be built on the same training set. Each of the p classifiers will give a prediction of the label of each testing sample. In the ensemble stage, predictions from different classifiers were combined in different ways. In our experiments we adopted two strategies to combine the predictions from different classifiers. The first strategy counts the number of times that a given testing sample is predicted positive and the number of times the sample is predicted negative. The final label of the sample is given by the majority of the votes. If there is a tie, then we randomly assign the testing sample to one of the two classes. The second strategy weights the prediction of each of the classifier with its confidence accompanying the prediction. The final predicted label is determined by the sign of the confidence weighted sum of prediction from each of the p classifiers. Each of the 10 folds from both classes is used as the testing fold once so that each of the samples is used as testing sample exactly once. We regard it as a convention throughout the paper that the positive class consists of subjects with melancholic depression and the negative class consists of healthy controls. The basic classifiers we used in the paper include SVM with linear kernel and Random Forest (RF). We used the following four measures to evaluate the performance of ensemble of classifiers: accuracy, sensitivity, specificity and area under curve (AUC).

In our experiments, we compared classification performance on features yielded from our method (SC, in abbreviation) with those generated by different data imputation and clustering methods. We tried different initializations, different values of K (number of keywords in the dictionary or number of clusters) and different values of λ (regularization parameter) on our method, and tried different initializations and different values of K on Kmeans and hierarchical clustering.

The classification performance is reported in Table 1. Due to space limit and the fact that SVM generally performed worse than RF on this dataset, we only report the classification performance by RF. In using KNN for data imputation, we tried a range of values for k and report the results from $k = 3$ since it gives the best performance. Also for Kmeans

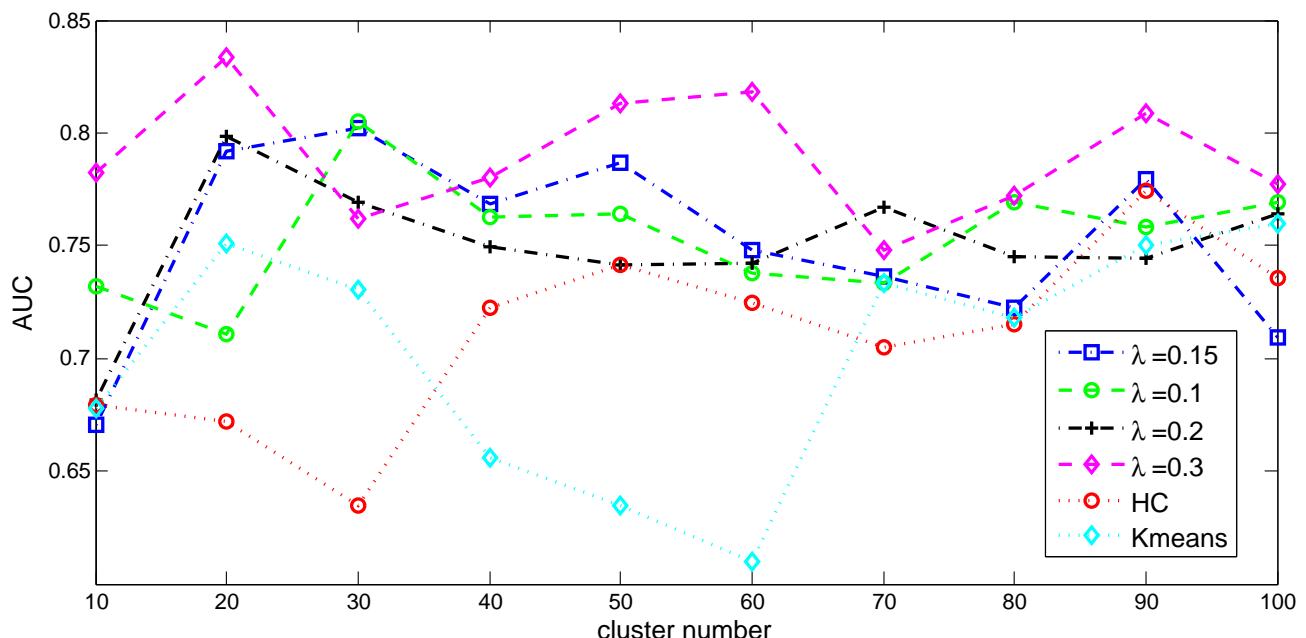


Fig. 1. Changes in AUC score with varying number of clusters for sparse coding with different regularization parameters, Kmeans and hierarchical clustering.

and hierarchical clustering, we imputed the raw design matrix using KNN before we applied these clustering techniques as KNN gave the best classification performance among all data imputation techniques. All the three clustering methods share one parameter, K , which we varied between 10 and 100 with a step size of 10. For sparse coding, there is an extra parameter λ which we set to be 0.1, 0.15, 0.2 and 0.3 in our experiments.

Fig. 1 shows how the AUC score changes when we ran classification on features generated by different clustering algorithms for different values of K . We can see that although the classification performance fluctuates with a growing number of clusters for all the clustering algorithms, sparse coding generally yields feature sets with stronger discriminative power when λ is set to be 0.3. This implies that there are indeed several clusters of metabolites in our dataset since a larger λ tends to drive the combination coefficient for each metabolite to be sparse and several metabolites are potentially outliers.

It is also of great interest to explore the groups of metabolites that are clustered together by looking into the coefficient matrix Z . The metabolites clustered into the i th group, which is represented by the i th column of \mathcal{D} , are those metabolites corresponding to the nonzero entries of the i th row of matrix Z . We looked into the most discriminative features (measured in terms of their p values) in the feature set \mathcal{D} on which the best classification performance is achieved and their corresponding rows in the coefficient matrix Z . The most discriminative feature, which has a p-value of 2.93×10^{-16} , corresponds to six metabolites with four of them being unknown, one of them belonging to the general category of “Complex lipids, fatty acids and related” and one of them belonging to the general category of “Amino acids and related”. The second most discriminative feature which has a p-value of 6.64×10^{-16} corresponds to twenty-four metabolites, with sixteen of them falling in the category “Amino acids and related”, two of them belonging to the category “Nucleobases and related”, five of them being unknown

and one of them belonging to the category “Hormones, signal substances and related”. The third most discriminative feature which has a p-value of 7.92×10^{-16} corresponds to a group of nineteen metabolites, with nine of them falling into the category “Complex lipids, fatty acids and related”, six of them unknown, two of them being “unknown lipid” and one of them being “Vitamins, cofactors and related”. From these, we can see that sparse coding does produce meaningful clusters in that most of the known metabolites assigned into the same cluster belong to related categories.

3.2. Analysis on Microarray Profiles

Microarray is another modality of data collected on the subjects. This dataset included information on 54675 transcripts from 123 subjects with 28 of them being labeled as positive (with melancholic depression) and 95 of them being labeled as negative (normal control). The same framework introduced in 3.1.2 to deal with extreme imbalance between two classes was also used on this dataset. Among the 54675 transcripts, a list of 2261 transcripts was identified as related to depression and a list of 3297 transcripts was identified as related to immune system. We ran classification on all three sets of data. There is no missing value in this dataset.

We also compared the classification performance on features generated by sparse coding, Kmeans and hierarchical clustering. For all the clustering methods, we set the number of clusters to be 100, 200, 500, 1000. For sparse coding, we used the same set of values for λ that were used on the metabolic profile. We report the best classification performance on features generated by sparse coding on each λ over all the K values and best classification performance on features generated by Kmeans and hierarchical clustering over all the K s. we only report the classification performance by SVM since SVM generally outperforms Random Forest on this dataset.

It is evident from Table 2, Table 3, Table 4 that the feature sets yielded by sparse coding have superior discriminative power than those generated by traditional clustering methods and raw data across all these three sets of data. In particular, feature sets obtained through sparse coding give rise to significantly improved sensitivity in classification performance, implying that it allows prediction with high accuracy of disease status in those who are diagnosed with melancholic depression. Overall, the feature sets given by sparse coding produce the best performance when $K = 500$. However, as shown in the results, a proper choice of λ is important as well.

4. Conclusion and Future Work

In this paper, we propose a method to learn a compressed set of representative features through an adapted version of sparse coding which is capable of simultaneously clustering variables with strong empirical correlation and dealing with the missing values in the design matrix. We apply the proposed method on datasets of metabolic and microarray profiles collected from a group of subjects consisting of patients with melancholic depression and healthy controls. Results show that our method can not only produce meaningful clusters of variables, but also generate a set of representative features which demonstrate superior discriminative power than those generated by traditional clustering and data imputation techniques. In particular,

Table 2. Classification performance on all genes

Method	Acc	SVM Vote			AUC	SVM Weighted Vote			AUC
		Sen	Spe			Acc	Sen	Spe	
Raw data	0.6411	0.6167	0.6400	0.6283	0.6090	0.6167	0.5978	0.6072	
KM	0.6333	0.6833	0.6200	0.6517	0.6172	0.7167	0.5878	0.6522	
HC	0.6007	0.6333	0.5878	0.6106	0.6167	0.6333	0.6089	0.6211	
SC($\lambda=0.1$)	0.6172	0.7833	0.5656	0.6744	0.6019	0.7833	0.5456	0.6644	
SC($\lambda=0.15$)	0.6578	0.7000	0.6389	0.6694	0.6578	0.7500	0.6278	0.6889	
SC($\lambda=0.2$)	0.6668	0.7667	0.6400	0.7033	0.6411	0.7667	0.6067	0.6867	
SC($\lambda=0.3$)	0.6578	0.7500	0.6289	0.6894	0.6744	0.7500	0.6511	0.7006	

Table 3. Classification performance on genes related to depression

Method	Acc	SVM Vote			AUC	SVM Weighted Vote			AUC
		Sen	Spe			Acc	Sen	Spe	
Raw data	0.6822	0.5833	0.7056	0.6444	0.6822	0.5833	0.7056	0.6444	
KM	0.6597	0.6333	0.6633	0.6483	0.6284	0.6667	0.6122	0.6394	
HC	0.6681	0.6167	0.6756	0.6461	0.6394	0.6283	0.6386	0.6334	
SC($\lambda=0.1$)	0.7245	0.7833	0.7044	0.7439	0.7091	0.7833	0.6844	0.7339	
SC($\lambda=0.15$)	0.7573	0.7000	0.7689	0.7344	0.7559	0.7000	0.7678	0.7339	
SC($\lambda=0.2$)	0.7245	0.7500	0.7167	0.7333	0.7176	0.7000	0.7178	0.7089	
SC($\lambda=0.3$)	0.6912	0.7500	0.6733	0.7117	0.6906	0.7500	0.6722	0.7111	

Table 4. Classification performance on genes related to immune system

Method	Acc	SVM Vote			AUC	SVM Weighted Vote			AUC
		Sen	Spe			Acc	Sen	Spe	
Raw data	0.6981	0.6167	0.7156	0.6661	0.6828	0.6167	0.6956	0.6561	
KM	0.7079	0.7167	0.7067	0.7117	0.7079	0.7167	0.7067	0.7117	
HC	0.6975	0.7500	0.6822	0.7161	0.6975	0.7500	0.6822	0.7161	
SC($\lambda=0.1$)	0.6911	0.7833	0.6622	0.7228	0.6911	0.7833	0.6622	0.7228	
SC($\lambda=0.15$)	0.7149	0.8333	0.6822	0.7578	0.7309	0.8667	0.6933	0.7800	
SC($\lambda=0.2$)	0.7065	0.7333	0.6933	0.7133	0.7225	0.7833	0.7033	0.7433	
SC($\lambda=0.3$)	0.7399	0.8333	0.7144	0.7739	0.7476	0.8333	0.7244	0.7789	

on both datasets, we found that in comparison with those traditional clustering algorithms, feature sets yielded by sparse coding give rise to significantly improved sensitivity scores, suggesting that learned features allow prediction with high accuracy of disease status in those who are diagnosed with melancholic depression.

One interesting future direction is to extend the current method to deal with data with multiple modalities and block-wise missing patterns (i.e., one sample may lack observations on one or more modalities). Simply concatenating different types of data is not appropriate in this situation since there is a high risk that pseudo-correlation may be detected between variables belonging to different data types which are not really related possibly due to a limited number of observations available on these variables. One direction is to use sparse

coding to simultaneously learn a group of features shared by all data types and individual features specific to each data type.

5. Acknowledgement

This work is supported in part by grants from NIH (R01 LM010730) and NSF (IIS-0953662, IIS-1421057, and IIS-1421100).

References

1. P. J. McGrath, A. Y. Khan, M. H. Trivedi, J. W. Stewart, D. W. Morris, S. R. Wisniewski, S. Miyahara, A. A. Nierenberg, M. Fava and A. J. Rush, *Journal of Clinical Psychiatry* **69**, 1847 (2008).
2. C. M. Mazure, M. B. B. Jr., F. H. Jr., K. B. Miller and J. Nelson, *Biological Psychiatry* **22**, 1469 (1987).
3. V. Gabbay, R. G. Klein, Y. Katz, S. Mendoza, L. E. Guttman, C. M. Alonso, J. S. Babb, G. S. Hirsch, and L. Liebes, *J Child Psychol Psychiatry* **51**, 935 (2010).
4. Z. Liu, X. Li, N. Sun, Y. Xu, Y. Meng, C. Yang, Y. Wang and K. Zhang, *PLOS One* **9**, p. e93388 (2014).
5. D. S. Wishart, T. Jewison, A. C. C. Guo, M. Wilson, C. Knox, Y. Liu, Y. Djoumbou, R. Mandal, F. Aziat, E. Dong, S. Bouatra, I. Sinelnikov, D. Arndt, J. Xia, P. Liu, F. Yallou, T. Bjorndahl, R. Perez-Pineiro, R. Eisner, F. Allen, V. Neveu, R. Greiner and A. Scalbert, *Nucleic acids research* **41**, D801 (January 2013).
6. J. D. Hoheisel, *Nat Rev Genet* **7**, 200 (March 2006).
7. M. R. Segal, K. D. Dahlquist and B. R. Conklin, *Journal of Computational Biology* **10**, 961 (2003).
8. P. Bühlmann, P. Rütimann, S. van de Geer and C.-H. Zhang, *Journal of Statistical Planning and Inference* **143**, 1835 (2013).
9. J. Yang, K. Yu, Y. Gong and T. Huang, Linear spatial pyramid matching using sparse coding for image classification, in *IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, 2009.
10. X. Lu, H. Y. Yan, P. Yan, L. Li and X. Li, Image denoising via improved sparse coding, in *Proceedings of the British Machine Vision Conference*, (BMVA Press, 2011). <http://dx.doi.org/10.5244/C.25.74>.
11. J. Mairal, F. Bach, J. Ponce and G. Sapiro, *J. Mach. Learn. Res.* **11**, 19 (March 2010).
12. B. Lin, Q. Li, Q. Sun, M.-J. Lai, I. Davidson, W. Fan and J. Ye, *CoRR abs/1407.8147* (2014).
13. R. Tibshirani, *Journal of the Royal Statistical Society, Series B* **58**, 267 (1994).
14. G. Parker and D. Hadzi-Pavlovic, *Melancholia: A Disorder of Movement and Mood* (Cambridge University Press, New York, 1996).
15. R. Dubey, J. Zhou, Y. Wang, P. M. Thompson and J. Ye, *NeuroImage* **87**, 220 (2014).

BAYCLONE: BAYESIAN NONPARAMETRIC INFERENCE OF TUMOR SUBCLONES USING NGS DATA

SUBHAJIT SENGUPTA¹, JIN WANG², JUHEE LEE³, PETER MÜLLER⁴,
KAMALAKAR GULUKOTA⁵, ARUNAVA BANERJEE⁶, YUAN JI^{1,7,*}

¹*Center for Biomedical Research Informatics, NorthShore University HealthSystem*

²*Department of Statistics, University of Illinois at Urbana-Champaign*

³*Department of Applied Mathematics and Statistics, University of California Santa Cruz*

⁴*Department of Mathematics, University of Texas Austin*

⁵*Center for Molecular Medicine, NorthShore University HealthSystem*

⁶*Department of Computer & Information Science & Engineering, University Of Florida*

⁷*Department of Health Studies, The University Of Chicago*

In this paper, we present a novel feature allocation model to describe tumor heterogeneity (TH) using next-generation sequencing (NGS) data. Taking a Bayesian approach, we extend the Indian buffet process (IBP) to define a class of nonparametric models, the categorical IBP (cIBP). A cIBP takes categorical values to denote homozygous or heterozygous genotypes at each SNV. We define a subclone as a vector of these categorical values, each corresponding to an SNV. Instead of partitioning somatic mutations into non-overlapping clusters with similar cellular prevalences, we took a different approach using feature allocation. Importantly, we do not assume somatic mutations with similar cellular prevalence must be from the same subclone and allow overlapping mutations shared across subclones. We argue that this is closer to the underlying theory of phylogenetic clonal expansion, as somatic mutations occurred in parent subclones should be shared across the parent and child subclones. Bayesian inference yields posterior probabilities of the number, genotypes, and proportions of subclones in a tumor sample, thereby providing point estimates as well as variabilities of the estimates for each subclone. We report results on both simulated and real data. BayClone is available at <http://health.bsd.uchicago.edu/yji/soft.html>.

Keywords: Categorical Indian buffet process; Heterozygosity; Latent feature model; NGS data; Random categorical matrices; Subclones; Tumor heterogeneity.

1. Introduction

1.1. Background

Tumorigenesis is a complex process.^{1,2} A wide variety of genetic features that promotes tumors are involved in this process, including the acquisition of somatic mutations that allow tumor cells to gain advantages over time compared to normal cells. As such, a tumor is oftentimes heterogeneous consisting of multiple subclones with unique genomes, a phenomenon called tumor heterogeneity (TH). Multiple recent reviews^{3–8} support the existence of subclones within tumors. Specifically, cancer cells undergo Darwinian-like clonal somatic evolution and tumor formation is dependent on acquisition of oncogenic mutations. In fact it has been found that individual tumors have a unique clonal architecture that is spatially and temporally evolving, which poses challenges as well as opportunities on individualized cancer treatment. We consider the differences in subclones arising from single nucleotide variations (SNVs), although

*Address for Correspondence: Research Institute, NorthShore University HealthSystem, 1001 University Place, Evanston, IL 60201, USA. Email: koaeraser@gmail.com

there can be other differences such as copy number variations. An SNV represents modification to a single DNA sequence. A scaffold of SNVs along the same haploid genome constitutes a *haplotype*. A pair of haplotypes gives rise to a subclonal genome.

Next-generation sequencing (NGS) experiments use massively parallel sequenced short reads to study long genomes. The short reads are mapped to the reference genome based on sequence similarities. Mapped reads are used to produce estimates of SNVs, small indels and copy number (CN) variations along the genome. In this paper we use the whole-genome sequencing (WGS) or whole-exome sequencing (WES) data to model the variant allele fraction (VAF) at an SNV, defined as the fraction of short reads that bear a variant sequence (compared to the reference genome). Innovatively, we infer subclones using scaffolds of SNVs, or haplotypes.

1.2. Main idea

Most multicellular organisms have two sets of chromosomes – they are called diploids. Diploid organisms have one copy of each gene (and therefore one allele) on each chromosome. At each locus, two alleles can be *homozygous* if they share the same genotypes, or *heterozygous* if they do not. In a recent paper⁹ the authors use an Indian buffet process (IBP)¹⁰ that assumes that SNVs are homozygous, where both alleles are either mutated or wild-type. However, biologically there are three possible allelic genotypes at an SNV: homozygous wild-type (no mutation on both alleles), heterozygous mutant (mutation on only one allele), or homozygous mutant (mutation on both alleles). Therefore, the IBP model is not sufficient to fully describe the subclonal genomes.

Our main idea is to extend IBP to categorical IBP that allows three values, 0, 0.5, and 1, to describe the corresponding genotypes at each SNV. Such an extension is mathematically non-trivial as we show later. More importantly, it allows for a principled and powerful statistical inference on TH. Different from existing methods based on Dirichlet processes,^{11,12} IBP and cIBP allow one SNV to appear in multiple subclones. We argue that this is more realistic and agrees with the fundamental evolutionary theory of clonal expansion. In particular, somatic mutations occurred in early tumor development should be shared by child subclones.

To start, note that each SNV can be associated with a non-negative number of subpopulations. Consider a finite number of S SNV loci and assume that an unknown number of C subclones are present. We introduce an $S \times C$ ternary matrix, $\mathbf{Z} = [z_{sc}]$ where each z_{sc} denotes the allelic variation at SNV site s for subclone c , $s = 1, 2, \dots, S$; $c = 1, 2, \dots, C$. Specifically, we let $z_{sc} \in \{0, 0.5, 1\}$ be a ternary random variable to denote three possible genotypes at the locus, homozygous wild-type ($z_{sc} = 0$), heterozygous variant ($z_{sc} = 0.5$), and homozygous variant ($z_{sc} = 1$); see Figure 1. Each sample is potentially an admixture of the subclones (columns of \mathbf{Z}), mixed in different proportions. Given \mathbf{Z} , we can denote the proportions of the C subclones by $\mathbf{w}_t = (w_{t0}, w_{t1}, \dots, w_{tC})$ for sample t , where $0 < w_{tc} < 1$ for all c and $\sum_{c=0}^C w_{tc} = 1$. Therefore, the contribution of a subclone to the VAF at an SNV is $0 \times w_{tc}$, $0.5 \times w_{tc}$ or $1 \times w_{tc}$, if the subclone is homozygous wild-type, heterozygous or homozygous mutant at the SNV, respectively. We develop a latent feature model (Section 2.3) for the entire matrix \mathbf{Z} to uncover the unknown subclones that constitute the tumor cells and given the data, we aim to

infer two quantities, \mathbf{Z} and \mathbf{w} , by a Bayesian inference scheme.

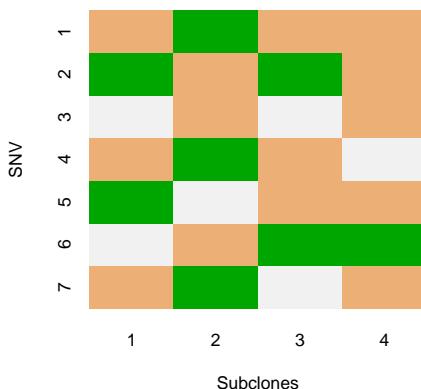


Fig. 1. Illustration of cIBP matrix \mathbf{Z} for subclones in a tumor sample. Colored cells in green=1, brown=0.5, and white=0 represent homozygous variants, heterozygous variants, and homozygous wild-type, respectively.

As shown in Figure 1, a subclone is defined by a vector of categorical values in $\{0, 0.5, 1\}$ representing the genotypes at specific SNV location. For example, in Figure 1 there are seven different SNV locations and four subclones. SNV 5 takes values 1 in subclone one, 0 in subclone two, and 0.5 in subclones three and four. Therefore, the same mutation is shared by two subclones (three and four).

1.3. Existing Methods

Recent rapid development has generated useful tools for subclonal inference, notably represented by SciClone,¹³ TrAp,¹⁴ PhyloSub,¹⁵ and Clomial,¹⁶ among others. TrAp nicely described the issue of solution identifiability stating the need to have sufficient sample size for unique mathematical solutions. SciClone and Clomial assume a binary matrix, focusing SNVs at copy neutral regions with heterozygous mutations. PhyloSub carefully considered possible genotypes at SNVs accounting for potential copy number changes. BayClone differs from these methods in that it novelly proposes a categorical IBP model as a nonparametric approach, accounting for the three potential genotypes at each SNV. BayClone does not take the indirect route of estimating nested clusters in a tree structure, as in PhyloSub. Instead, it directly outputs subclones with overlapping SNVs. As a model choice, BayClone does not assume a phylogenetic tree structure for the inferred subclones since in any given tumor sample not all the subclones on the phylogenetic tree may be present. Instead, existing subclones may only represent nodes on a subset of branches of the phylogenetic tree. With this notion, BayClone does not assume a phylogenetic tree but rather use data to infer subclones with overlapping SNVs.

The remainder of the paper is organized as follows. In section 2, we elaborate on the proposed probability model. Section 3 describes model selection and posterior inference. In the following section, we report experimental results, one with simulated data and another by real-life data from an NGS experiment. In the final section we conclude with discussion and future work.

2. Probability Model

2.1. Latent feature model with IBP

In latent feature model, each data point is generated by a vector of latent feature values. In our case, each subclone (one column of \mathbf{Z}) is a latent feature vector and a data point is the observed VAF. The IBP model is used to define a prior on the space of binary matrices that indicate the presence of a particular feature for an object, with the number of columns in the matrix (corresponding to features) being potentially unbounded. The detailed construction of IBP can be found in Ref. [10]. We consider a constructive definition of IBP as follows. For each component z_{sc} in the binary matrix \mathbf{Z} , assume

$$\begin{aligned} z_{sc} | \pi_c &\sim Bern(\pi_c), \\ \pi_c | \alpha &\sim beta(\alpha/C, 1), \quad c = 1, \dots, C, \end{aligned} \tag{1}$$

where $Bern(\pi_c)$ is the Bernoulli distribution and $\pi_c \in (0, 1)$ is the probability $Pr(z_{sc} = 1)$ *a priori*. Also, the marginal $p(\mathbf{Z}) = \prod_{c=1}^C p(\mathbf{Z}_c) = \int p(z_{sc} | \pi_c)p(\pi_c)d\pi_c$ factors assuming conditional independence, where \mathbf{Z}_c is the c -th column vector. When $C \rightarrow \infty$, the marginal distribution of \mathbf{Z} (as an equivalence class) exists and is called IBP. We extend the IBP model to a categorical setting, where each entry of the matrix is not necessarily 0 or 1, but a set of integers in $\{0, 1, \dots, Q\}$ where Q is fixed *a priori*. We call the extended model categorical IBP (cIBP) and use it as a prior in exploring subclones of tumor samples. In upcoming discussion, SNVs correspond to objects (rows) and subclones correspond to feature (columns) in the \mathbf{Z} matrix.

2.2. Development of cIBP

We discuss the development of the cIBP for a general case with an arbitrary Q . A straightforward extension of IBP in (1) would be to replace the underlying beta distribution of π_c with a Dirichlet distribution, and replace the Bernoulli distribution of z_{sc} with a multinomial distribution. However, as $C \rightarrow \infty$, Ref. [17] showed that the limiting distribution is degenerate. Instead, utilizing a Beta-Dirichlet distribution defined in Ref. [18] we propose a construction given C and Q : let $\{1, \dots, Q\}$ be the possible values z_{sc} takes. Then we assume

$$\pi_c \sim \text{Beta-Dirichlet } (\alpha/C, 1, \underbrace{\beta, \dots, \beta}_{Q \text{ of them}}); \quad z_{sc} | \pi_c \sim \text{Multi}(1, \pi_c). \tag{2}$$

Integrating out π_c in (2), the probability of a $(Q + 1)$ -nary matrix, \mathbf{Z} is

$$p(\mathbf{Z}) = \left(\frac{1}{\prod_{s=1}^S (s + \alpha/C)} \right)^C \prod_{c=1}^{C_+} \left(\frac{\alpha}{C} \cdot \frac{1}{Q} \right) \frac{(S - m_c)!}{S!} \times \prod_{j=1}^{m_c-1} \left[\frac{(j + \alpha/C)}{(j + Q\beta)} \right] \frac{1}{\beta} \prod_{q=1}^Q \frac{\Gamma(\beta + m_{cq})}{\Gamma(\beta)},$$

where m_{cq} denotes the number of rows possessing value $q \in \{1, \dots, Q\}$ in column c , i.e., $m_{cq} = \sum_{s=1}^S \mathbb{I}(z_{sc} = q)$ and $m_c = \sum_{q=1}^Q m_{cq}$. This gives birth to a random matrix with C columns, each entry taking a discrete value in a set of $(Q + 1)$ values. It can be shown that the limiting distribution of \mathbf{Z} (as an equivalent class) exists and is called the cIBP.¹⁷

2.3. Sampling model

Suppose there are T tumor samples in the data in which S SNVs are measured for each sample. Let N_{st} be the total number of reads mapped to SNV s in sample t , $s = 1, 2, \dots, S$ and $t = 1, 2, \dots, T$. Among N_{st} reads, assume n_{st} possess a variant sequence at the locus. We assume a binomial sampling model

$$n_{st} \stackrel{\text{indep.}}{\sim} \text{Binomial}(N_{st}, p_{st}), \quad (3)$$

where p_{st} is the expected proportion of variant reads.

We assume that the matrix \mathbf{Z} follows a finite version of cIBP in (2), $\mathbf{Z} \sim \text{cIBP}_C(Q = 2, \alpha, [\beta_1, \beta_2])$. Recall that $\mathbf{w}_t = (w_{t0}, w_{t1}, \dots, w_{tC})$ denotes the vector of subclonal weights. We assume \mathbf{w}_t follows a Dirichlet prior given by,

$$\mathbf{w}_t \stackrel{\text{indep.}}{\sim} \text{Dirichlet}(a_0, a_1, \dots, a_C).$$

As we have mentioned earlier that, each sample t potentially consists of several subclones with different proportions. Thus the variant reads must come from those subclones possessing variant alleles. In other words, parameters p_{st} can be modeled as a linear combination of variant alleles $z_{sc} \in \{0, 0.5, 1\}$ weighted by the proportions of subclones bearing the alleles. Remember that, $z_{sc} = 0, 0.5$ and 1 means that there is no mutation, heterozygous mutation and homozygous mutation at SNV position s for subclone c , respectively. Apparently, when a subclone bears no variant alleles, i.e., $z_{sc} = 0$, the contribution from that subclone to p_{st} should be zero. We assume the expected p_{st} is a result of mixing subclones with different proportions. Mathematically, given \mathbf{Z} and \mathbf{w} we assume

$$p_{st} = \sum_{c=1}^C w_{tc} z_{sc} + \epsilon_{t0}. \quad (4)$$

Equation (4) is a key model assumption. It allows us to back out the unknown subclones from a decomposition of the expected VAF p_{st} as a weighted sum of latent genotype calls z_{sc} with weights w_{tc} being the proportions of subclones. Importantly, we assume these weights to be the same across all SNV's, $s = 1, \dots, S$. In other words, the expected VAF is contributed by those subclones with variant genotypes, weighted by the subclone prevalences. Subclones without variant genotype on SNV s do not contribute to the VAF for s since all the short reads generated from those subclones are normal reads.

In (4) ϵ_{t0} is an error term defined as $\epsilon_{t0} = p_0 w_{t0}$, where $p_0 \sim \text{Beta}(\alpha_0, \beta_0)$. Importantly ϵ_{t0} is devised to capture experimental and data processing noise. Specifically, p_0 is the relative frequency of variant reads produced as error from upstream data processing and takes a small value close to zero; w_{t0} absorbs the noise left unaccounted for by $\{w_{t1}, \dots, w_{tC}\}$. Ignoring the error term ϵ_{t0} , model (4) can also be considered as a non-negative matrix factorization (NMF) if it is written as a matrix format, in which p_{st} could be replaced by observed VAFs. Our proposed feature allocation models differ from (NMF) in that we take a probabilistic approach effectively accounting for the noise in the data and providing variabilities measures for the model parameters using probability inference. We discuss the inference below.

3. Model Selection and Posterior Inference

3.1. MCMC simulation

In order to infer the sampling parameters from the posterior distribution, we use Markov chain Monte Carlo (MCMC) simulations. The Gibbs sampling method is used to update z_{sc} , whereas the Metropolis-Hastings (MH) sampling is used to get the samples of w_{tc} and p_0 . We omit detail except the one for sampling z_{sc} . Due to exchangeability, we let SNV s be the last customer. Let $\mathbf{z}_{-s,c}$ be the set of assignment of all other SNVs but SNV s for subclone c , m_{cq}^- the number of SNVs with level q , not including SNV s and $m_c^- = \sum_{q=1}^Q m_{cq}^-$. We obtain,

$$p(z_{sc} = q | \mathbf{z}_{-s,c}, rest) \propto \left(\frac{m_c^-}{s} \right) \times \left(\frac{\beta_q + m_{cq}^-}{\beta^* + m_c^-} \right) \prod_{t=1}^T \binom{N_{st}}{n_{st}} (p'_{st})^{n_{st}} (1 - p'_{st})^{(N_{st} - n_{st})}$$

for any c such that $m_c^- > 0$, where $rest$ includes the data and current MCMC values for all the other parameters. Also, p'_{st} is value of p_{st} by plugging the current MCMC values and setting $z_{sc} = q$.

3.2. Choice of C

The number of subclones C in cIBP is unknown and must be estimated. We discuss a model selection to select the correct value for C . We use predictive densities as a selection criterion. Let \mathbf{n}_{-st} denote the data removing n_{st} . Also denote the set of parameters for a given C by $\boldsymbol{\eta}^C$. The conditional predictive ordinate (CPO)¹⁹ of n_{st} given \mathbf{n}_{-st} is given by

$$CPO_{st} = p(n_{st} | \mathbf{n}_{-st}) = \int p(n_{st} | \boldsymbol{\eta}^C, \mathbf{n}_{-st}) p(\boldsymbol{\eta}^C | \mathbf{n}_{-st}) d\boldsymbol{\eta}^C. \quad (5)$$

The Monte-Carlo estimate of (5) is the harmonic mean of the likelihood values²⁰ $p(n_{st} | \boldsymbol{\eta}_l^C)$,

$$\hat{p}(n_{st} | \mathbf{n}_{-st}) \approx \frac{1}{L^{-1} \sum_{l=1}^L p(n_{st} | \boldsymbol{\eta}_l^C)^{-1}} \quad (6)$$

where $\boldsymbol{\eta}_l^C$'s are MCMC draw's and L is the number of iterations. We take each data point out from \mathbf{n} and compute average *log-pseudo-marginal likelihood* (LPML) over this set as $L^C = \sum_{n_{st} \in \mathbf{n}} \log[\hat{p}(n_{st} | \mathbf{n}_{-st})]$. For different values of C , we compare the values of L^C and choose that \hat{C} which maximizes L^C .

3.3. Estimate of \mathbf{Z}

The MCMC simulations generate posterior samples of the categorical matrix \mathbf{Z} and other parameters. Directly taking sample average is not desirable since it will result in an estimated matrix with entries taking values outside the set $\{0, 0.5, 1\}$. Instead, we define a posterior point estimate of \mathbf{Z} similar to that in Ref. [9], i.e.,

$$\hat{\mathbf{Z}} = \arg \min_{\mathbf{Z}'} \frac{1}{L} \sum_{l=1}^L d(\mathbf{Z}^{(l)}, \mathbf{Z}') \quad (7)$$

where $\mathbf{Z}^{(l)}, l = 1, \dots, L$ are MCMC samples. The term $d(\mathbf{Z}^{(l)}, \mathbf{Z}')$ is a distance with the following definition. Note that the MCMC samples $\mathbf{Z}^{(l)}$ may have different labels for Z across iterations.

Therefore we introduce a permutation for comparing any two matrices. For two matrices Z and Z' , let $D_{cc'}(\mathbf{Z}, \mathbf{Z}') = \sum_{s=1}^S |z_{sc} - z'_{sc}|$ for two columns c and c' . We define a distance $d(\mathbf{Z}, \mathbf{Z}') = \min_{\zeta} \sum_{c=1}^C D_{c,\zeta_c}(\mathbf{Z}, \mathbf{Z}')$ where $\zeta_c, c = 1, \dots, C$ is a permutation of $\{1, \dots, C\}$. Having the permutation ζ_c resolves the potential label-switching issue in the MCMC samples.

4. Results

4.1. Simulated Data

We demonstrate the performance of BayClone with two sets of simulated data. First, we take a set of $S = 100$ SNV locations and consider $T = 30$ samples. The true number of latent subclones is $C = 4$ in this experiment. The true Z values are given in the left most panel of Figure 2. We generate the true proportion matrix w by setting $w_{t0} = 0.05$ to account for the

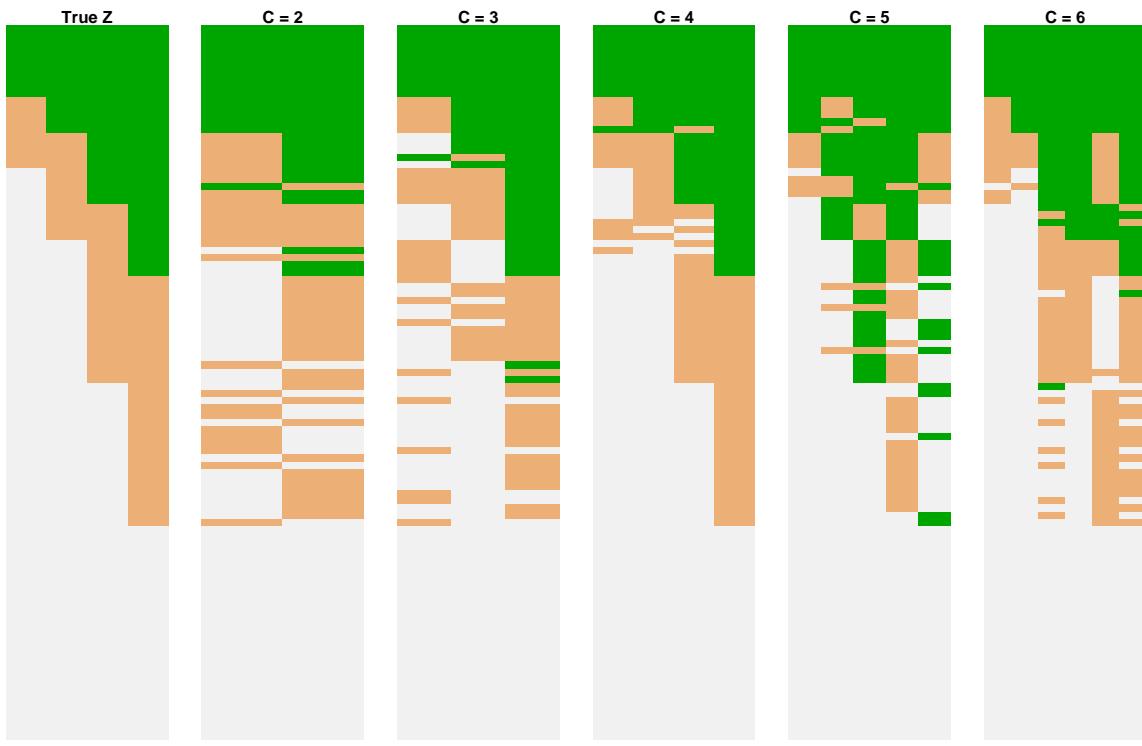


Fig. 2. True Z and estimate \hat{Z} in (7) with green standing for homozygous mutation i.e. $z_{sc} = 1$, brown for heterozygous mutation i.e. $z_{sc} = 0.5$ and white for homozygous wild type i.e. $z_{sc} = 0$. The model with $C = 4$ fits the data the best.

background noise in sample t , and the rest w_{tc} 's from the permutations of $(0.5, 0.3, 0.1, 0.05)$ (where $c = 1, 2, 3, 4$). We take the true p_0 as 0.01 and fix $N_{st} = 50$ for all $s = 1, 2, \dots, 100$ and $t = 1, 2, \dots, 30$. Finally we generate n_{st} from $\text{Binomial}(N_{st}, p_{st})$. Hyperparameters are set up as follows: for w_t : $a_0 = a_1 = a_2 = \dots = a_C = 1$, for π_c : $\alpha = 1$, $\beta_1 = \beta_2 = 2$, and for p_0 : $\alpha_0 = 1$, $\beta_0 = 100$. Given C , we randomly initialize the binary matrix Z and draw the initial p_0 from the specified prior. The initial w_t are generated by drawing gamma random variables from the prior $\theta_t \sim \text{Gamma}(a_0, a_1, \dots, a_C)$, and then normalizing them. That is, $w_{tc} = \theta_{tc}/(\sum_{k=0}^C \theta_{tk})$.

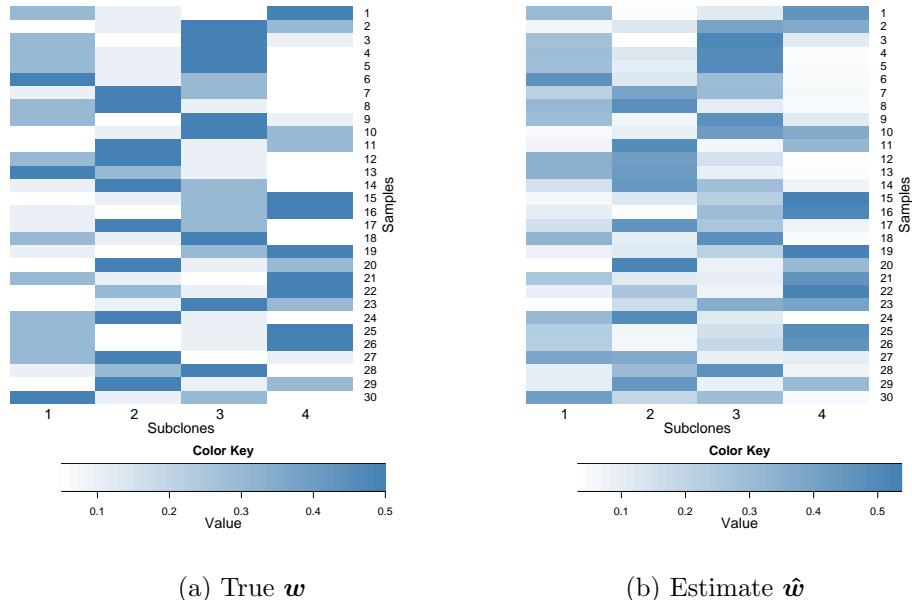


Fig. 3. True \mathbf{w} and estimated proportions $\hat{\mathbf{w}}$ for $\hat{C} = 4$ with simulated data.

For MCMC simulations, we run 4,000 iterations, discard the initial 2,000 as burn in, and take one sample every 5th iteration. The Markov chain converges quickly. Table 1 presents the average LPML for various values of C . As we can see, L^C is maximized at $\hat{C} = 4$, which is

Table 1. LPML L^C for C values. The simulation truth is 4.

C	2	3	4	5	6
L^C	-9144.4	-6664.1	-4992.869	-5218.707	-5034.129

the true C . We find the estimate $\hat{\mathbf{Z}}$ in (7) based on the posterior samples drawn from MCMC simulations. In Figure 2, we compare the truth with estimates $\hat{\mathbf{Z}}$ for different values of C . Also we plot the true \mathbf{w} and estimate $\hat{\mathbf{w}}$ across all the samples using $\hat{C} = 4$ in Figure 3. They have almost identical values. As model checking we computed the difference between the true p_{st} and the posterior mean \hat{p}_{st} , for different model C . When $\hat{C} = 4$, the difference of \hat{p}_{st} and true p_{st} is the smallest (results not shown). Also the posterior mean of p_0 is 0.0107 for the correct value of C , which is very close to the simulation truth $p_0 = 0.01$. All the other parameters in the model were closely estimated under the Bayesian model as well.

Lastly, we compare the simulation results with PyClone,¹² which uses Dirichlet process to partition SNVs into mutation clusters. In Figure 4, we plot the true $\mathbf{p} = [p_{st}]$, estimate $\hat{\mathbf{p}}$ by our model and cellular prevalences inferred by PyClone, which is equivalent to $\hat{\mathbf{p}}$ in our models. PyClone estimates six SNV clusters. Also, the L1-norm, $\sum_{s,t} |p_{st} - \hat{p}_{st}|$ equals 35.24 for our method, compared to 132.04 for PyClone. The fitting is worse than our model when $C = 4$. We note that the comparison to PyClone is not to choose a superior method as PyClone does not directly provide estimates on subclones and their cellularities. Instead, it only aims to infer subclonal frequencies of SNVs.

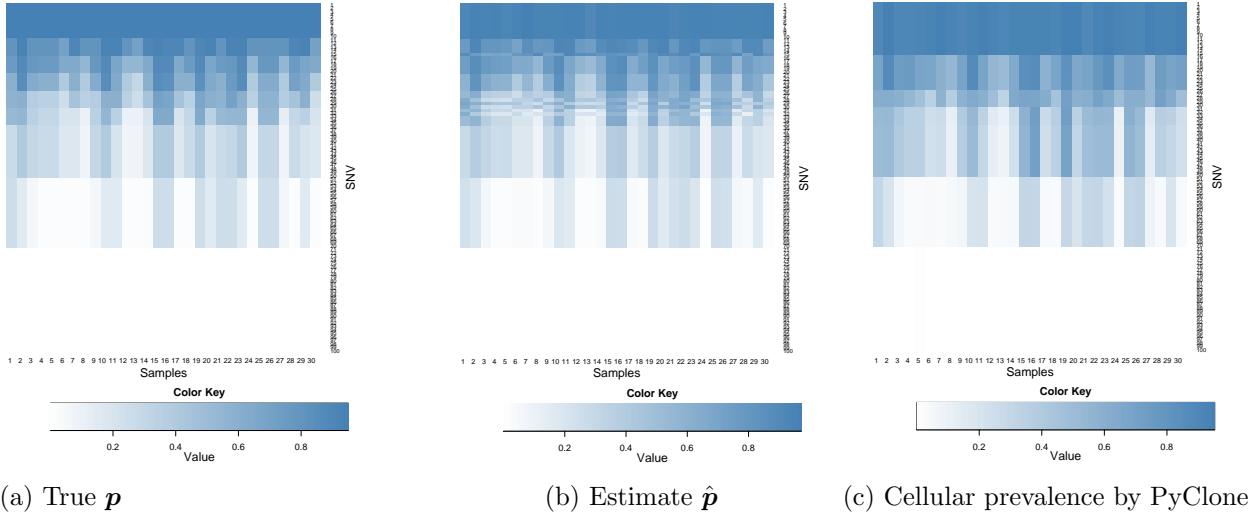


Fig. 4. True and estimated (by our model) expected VAF and cellular prevalence inferred by PyClone

As a second simulation, we use the same setting but reduce the numbers of SNVs and samples to $S = 70$ and $T = 7$. We find that again $\hat{C} = 4$ according to the LPML criterion and in Figure 5 we find that $\hat{\mathbf{Z}}$ is closest to the true \mathbf{Z} when $\hat{C} = 4$. Estimate of \mathbf{w} when $\hat{C} = 4$ is close to the truth (results not shown) and so are the other model parameters. In summary, the model performs well with only $T = 7$ samples.

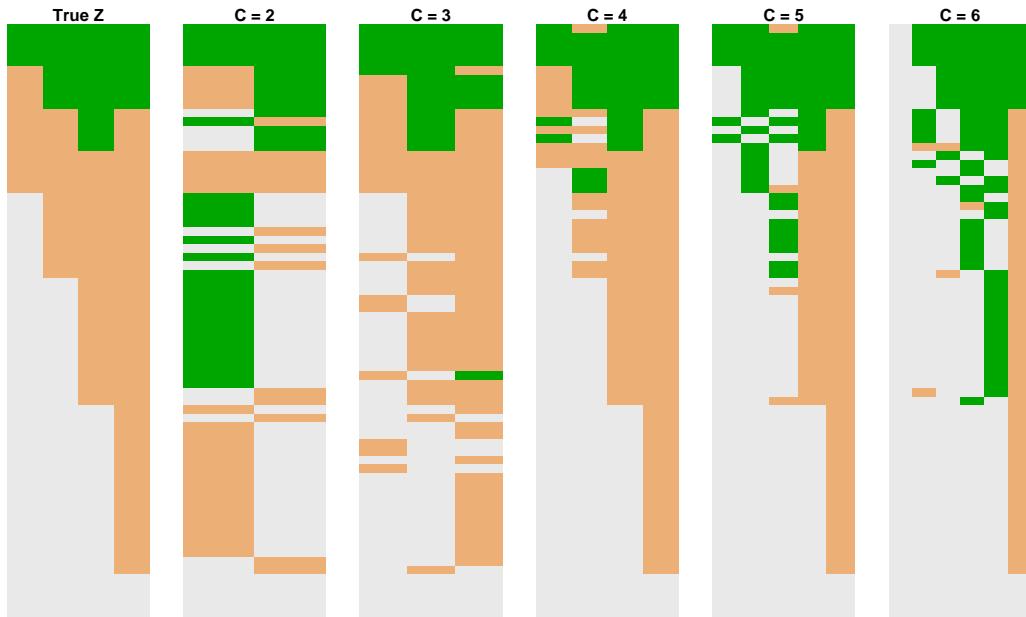


Fig. 5. True \mathbf{Z} and estimate $\hat{\mathbf{Z}}$ in (7) with green standing for homozygous mutation i.e. $z_{sc} = 1$, brown for heterozygous mutation i.e. $z_{sc} = 0.5$ and white for homozygous wild type i.e. $z_{sc} = 0$. The model with $C = 4$ fits the data with $S = 70$ SNVs and $T = 7$ samples the best.

4.2. Intra-Tumor Lung Cancer Samples

We record whole-exome sequencing for four surgically dissected tumor samples taken from a single patient diagnosed with lung adenocarcinoma. A portion of the resected tumor is flash frozen and another portion is formalin fixed and paraffin embedded (FFPE). Two different specimens are taken from the frozen portion of the resected tumor and another two from the FFPE portion. Genomic DNA is extracted from all four specimens and an exome capture is done using Agilent SureSelect v5+UTR probe kit. The exome library is then sequenced in paired-end fashion on an Illumina HiSeq 2000 platform. Only two specimens are sequenced on each to ensure a high depth of coverage. We map the reads to the human genome (version HG19)²¹ using BWA²² and called variants using GATK.²³ Post-mapping, the mean coverage of the samples is around 100 fold.

We restrict our attention to the SNVs that (i) exhibit significant coverage in all our samples (total number of mapped reads N_{st} are ranged in [100, 240]) and (ii) have reasonable chance of mutation (the empirical fractions n_{st}/N_{st} in [0.25, 0.75]). This filtering left us with 12,387 SNV's. We then randomly select $S = 150$ for computational purposes. In summary, using the above notations, the data record the read counts (N_{st}) and mutant allele read counts (n_{st}) of $S = 150$ SNVs from $T = 4$ tumor samples.

The large values for N_{st} make the binomial likelihood very informative. For the prior specification, we adopt the same hyperparameters in the simulation study. We ran MCMC for 6,000 iterations, discarding the first 3,000 iterations as initial burn-in and thinning by 3. We consider $C = 2, 3, 4, 5, 6$ and use LPML to select the best C . LPML values for $C = 2, 3, 4, 5$ and 6 are $-1991.87, -1991.82, -1992.64, -1993.57$ and -1994.67 , respectively. So the LPML is maximized at $\hat{C} = 3$ implying that three distinct subclones are present. Conditioning on $\hat{C} = 3$, the estimate $\hat{\mathbf{Z}}$ is shown in Figure 6(a). The proportions of the three subclones in each of the four samples are plotted in Figure 6(b). A phylogenetic tree is hypothesized in Figure 6(c). In particular, subclone 1 appears to be the parent giving birth to two branching child subclones 2 and 3. Comparing columns in Figure 6(a), we hypothesize that subclones 2 and 3 arise by acquiring additional somatic mutations in the top portion of the SNV regions where subclone 1 shows “white” color, i.e., homozygous wild type. The three subclones share the same genotype in the middle and lower half of the SNVs (the large chunk of “brown” bars in Figure 6(a)), suggesting that these could be either somatic mutations acquired in the parent subclone 1, or germline mutations. All four tumor samples have similar proportions of the subclones, showing lack of geographical heterogeneity although each sample is mosaic. This is expected since the four tumor samples were dissected from regions that were close by on the original lung tumor.

Clinically, our analysis provides valuable information for treatment considerations. Since each tumor sample is mosaic consisting of three subclones, detailed mutational annotation could be conducted to seek potential biomarker mutations for targeted therapy. Also, combinational drugs could be considered if possible to specifically target each subclone. Since the four tumor samples possess similar proportions of subclones, the tumor appears to be homogeneous spatially. The results from our subclonal analysis could be used as a future reference should the disease progress or relapse. For example, future subclonal analysis could be

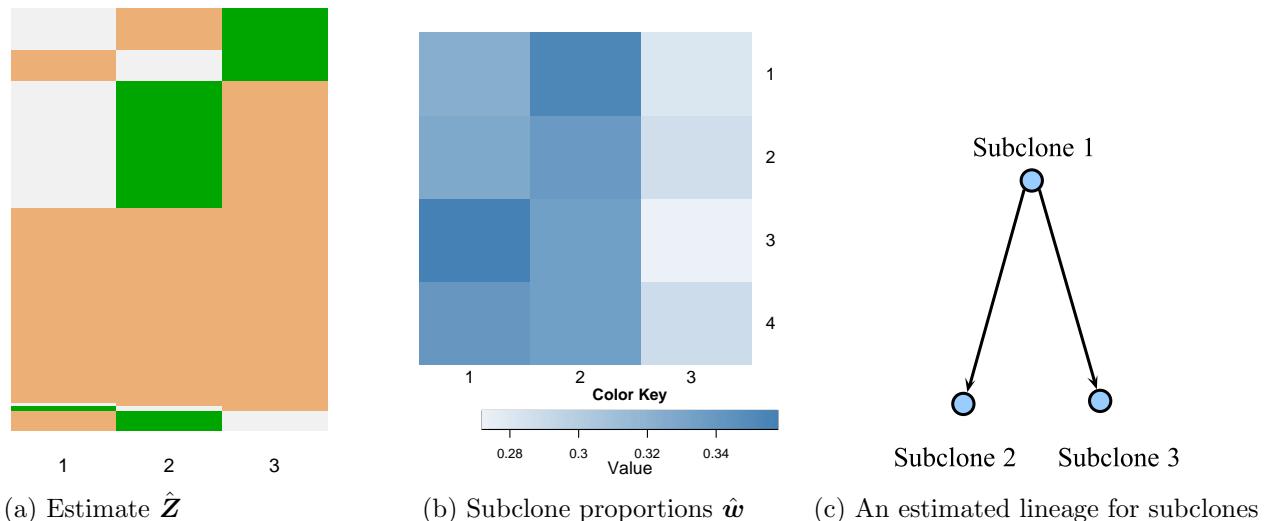


Fig. 6. Subclone structures, proportions and a possible lineage for the lung cancer data.

compared to the existing one to understand the temporal genetic changes.

5. Discussion and future work

One of the major motivations to detect the heterogeneity in tumors is personalized medicines.²⁴ Measure of heterogeneity can be useful as a prognosis marker.²⁵ Using NGS data to study the co-existence of genetically different subpopulations across tumors and within a tumor can shed light on cancer development. The main feature of BayClone is the model-based and principled inference on subclonal genomes for a set of SNVs, which directly genotypes subclones and the associated variabilities. Although not shown, posterior variances are easily obtained using MCMC samples for the Z and w matrices in our examples. More importantly, the feature allocation model, cIBP, reflects the underlying evolutionary biology of clonal expansion, such as the population “infinite sites assumption” in which mutations occurred in parental subclones are passed on to all offsprings. However, we do not explicitly enforce this assumption in modeling the SNVs, fearing that some subclones might not be present in the tumor samples as they lose to other subclones in the fitness selection. Instead, we explicitly model overlapping SNVs across subclones and let the data fit to dictates the subclonal genotypes. This is a distinction from clustering-based approaches in the existing literature.

There can be a number of possible extensions to the current model. First, the number of SNVs examined in this paper was relatively limited (about 150). Other than computational complexity, there is no limitation on extending the current model to analyze a large set of SNVs. We have begun to investigate efficient computational algorithms to take on a large number of SNVs, see Ref. [26].

As another important extension, we are considering joint modeling SNVs and copy number variations (CNVs) using linked feature allocation models. Briefly, we could consider a sampling model for the total read counts N_{st} to estimate the sample copy numbers, conditional on which a couple of feature allocation models can be linked for estimating subclonal copy numbers and

DNA sequences.

References

1. R. A. Weinberg, *The biology of cancer* (Garland Science New York, 2007).
2. N. Navin, A. Krasnitz, L. Rodgers, K. Cook, J. Meth, J. Kendall, M. Riggs, Y. Eberling, J. Troge, V. Grubor *et al.*, *Genome research* **20**, 68 (2010).
3. N. D. Marjanovic, R. A. Weinberg and C. L. Chaffer, *Clinical chemistry* **59**, 168 (2013).
4. V. Almendro, A. Marusyk and K. Polyak, *Annual Review of Pathology: Mechanisms of Disease* **8**, 277 (2013).
5. K. Polyak, *The Journal of clinical investigation* **121**, p. 3786 (2011).
6. J. Stingl and C. Caldas, *Nature Reviews Cancer* **7**, 791 (2007).
7. M. Shackleton, E. Quintana, E. R. Fearon and S. J. Morrison, *Cell* **138**, 822 (2009).
8. D. L. Dexter, H. M. Kowalski, B. A. Blazar, Z. Fligiel, R. Vogel and G. H. Heppner, *Cancer Research* **38**, 3174 (1978).
9. J. Lee, P. Müller, Y. Ji and K. Gulukota, *A Bayesian Feature Allocation Model for Tumor Heterogeneity*, tech. rep., UC Santa Cruz (2013).
10. T. L. Griffiths and Z. Ghahramani, *Journal of Machine Learning Research* **12**, 1185 (2011).
11. S. Nik-Zainal, P. Van Loo, D. C. Wedge, L. B. Alexandrov, C. D. Greenman, K. W. Lau, K. Raine, D. Jones, J. Marshall, M. Ramakrishna *et al.*, *Cell* **149**, 994 (2012).
12. A. Roth, J. Khattra, D. Yap, A. Wan, E. Laks, J. Biele, G. Ha, S. Aparicio, A. Bouchard-Côté and S. P. Shah, *Nature methods* (2014).
13. C. A. Miller, B. S. White, N. D. Dees, M. Griffith, J. S. Welch, O. L. Griffith, R. Vij, M. H. Tomasson, T. A. Graubert, M. J. Walter *et al.*, *PLoS computational biology* **10**, p. e1003665 (2014).
14. F. Strino, F. Parisi, M. Micsinai and Y. Kluger, *Nucleic acids research* **41**, e165 (2013).
15. W. Jiao, S. Vembu, A. G. Deshwar, L. Stein and Q. Morris, *BMC bioinformatics* **15**, p. 35 (2014).
16. H. Zare, J. Wang, A. Hu, K. Weber, J. Smith, D. Nickerson, C. Song, D. Witten, C. A. Blau and W. S. Noble, *PLoS computational biology* **10**, p. e1003703 (2014).
17. S. Sengupta, J. Ho and A. Banerjee, *Two Models Involving Bayesian Nonparametric Techniques*, tech. rep., University Of Florida (2013).
18. Y. Kim, L. James and R. Weissbach, *Biometrika* **99**, 127 (2012).
19. A. E. Gelfand, Model determination using sampling-based methods, in *Markov chain Monte Carlo in practice*, (Springer, 1996) pp. 145–161.
20. L. Held, B. Schrödle and H. Rue, Posterior and cross-validatory predictive checks: a comparison of mcmc and inla, in *Statistical Modelling and Regression Structures*, (Springer, 2010) pp. 91–110.
21. D. M. Church, V. A. Schneider, T. Graves, K. Auger, F. Cunningham, N. Bouk, H.-C. Chen, R. Agarwala, W. M. McLaren, G. R. Ritchie *et al.*, *PLoS Biology* **9**, p. e1001091 (2011).
22. H. Li and R. Durbin, *Bioinformatics* **25**, 1754 (2009).
23. A. McKenna, M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly *et al.*, *Genome research* **20**, 1297 (2010).
24. D. L. Longo, *N Engl J Med* **366**, 956 (2012).
25. A. Marusyk, V. Almendro and K. Polyak, *Nature Reviews Cancer* **12**, 323 (2012).
26. Y. Xu, P. Muller, Y. Yuan, Y. Ji and K. Gulukota, *arXiv preprint arXiv:1402.5090* (2014).

HUMAN EVOLUTIONARY GENOMICS AND THE SEARCH FOR THE GENES THAT MADE US HUMAN

JAMES M. SIKELA

*Department of Biochemistry and Molecular Genetics
Human Medical Genetics and Neuroscience Programs
University of Colorado School of Medicine
Aurora CO 80045, USA
Email: james.sikela@ucdenver.edu*

1. Introduction:

Due to advances in genome sequencing and analysis we can now carry out comparisons of genomic data with unprecedented scope and detail, both within and between species. As a result gene and genomic changes that have occurred specifically in the human lineage can now be identified. This field is one of the most interesting areas of human biology, and there has now emerged, due in large part to our ability to carry out computational analyses of large genome datasets, a growing list of significant discoveries. This workshop will survey the latest findings in this rapidly advancing field, including both key discoveries and remaining challenges (e.g. complexity of some genomic regions, difficulties in moving from identification of genomic differences to linking them with human-specific traits, etc.).

The types of human lineage-specific genomic changes, and the methods used to detect them, cover a wide range, and speakers have been chosen with this in mind. Speakers include individuals focused on 1) computational analysis at the transcriptional level particularly in brain (D. Geschwind; E. Lein), 2) computational genomics to identify rapidly evolving genomic sequences including non-coding sequences (D. Kostka), 3) computational analysis of protein coding sequences under selection in the human lineage (E. Vallender), 4) identification of gene coding regions that have undergone extreme human lineage-specific copy number expansions (J. Sikela), and 5) computational approaches to identifying genomic changes that distinguish “modern” humans from archaic hominins (J. Wall).

2. Speakers and Abstracts:

Transcriptional Features of Human Brain Development

Ed S. Lein, PhD

Allen Institute for Brain Science

How the genome guides brain formation is still poorly understood, as is our understanding of which features of brain development are conserved across all mammals versus specific to human. To address these questions we have created a series of high anatomical resolution transcriptional atlases of the developing human, non-human primate and mouse brain, and performed systematic analyses of transcriptional dynamics during brain development and between species. We find the transcriptome is tremendously dynamic across brain development, particularly in prenatal and early postnatal stages, with both similarities and many differences between rodents and primates. These differences in gene regulation likely underlie the unique

features of human brain structure and function, and provide clues about the locus of action for genes associated with neurodevelopmental diseases.

Transcriptional Networks in Human Brain

Daniel Geschwind, MD/PhD

University of California, Los Angeles School of Medicine

Understanding the evolution of human cognitive phenotypes has benefited greatly from comparison with non-human primates. Studies of genetic and transcriptional changes in brain genes has identified many candidates, but studies of transcription are often limited by the notion that gene expression differences between species are mostly neutral. We have developed network approaches to analysis of transcription data that allow us to put changes in gene expression in a functional context, permitting us to identify changes that are likely to be relevant to brain evolution on the human lineage. We have also used this framework to connect gene expression networks to other types of networks in human brain to connect different levels of brain function.

Lineage-specific Accelerated Evolution in Five Primates

Dennis Kostka, PhD

University of Pittsburgh School of Medicine

Regulatory evolution has been proposed to explain diversity of closely related species, and there is great interest in identifying the genetic basis of lineage-specific traits. We developed a statistical phylogenetic approach to identify genomic regions with lineage-specific accelerated substitution rates (linARs) and applied it to human and four non-human primates. We find an enrichment of human-specific linARs in non-coding regions with epigenetic marks of regulatory sequences, particularly nearby neurodevelopmental genes. Comparing across primates, similar loci and pathways harbor distinct linARs from multiple species. Thus, shared biological processes may have been independently targeted by adaptive events in multiple primate lineages.

Detecting Signatures of Selection through Interspecific Comparisons

Eric J. Vallender, PhD

Harvard Medical School

Selection on proteins can be identified through the ratio of fixation of nonsynonymous changes to synonymous changes (dN/dS). With the proliferation of genomic sequence, the ability to broadly and without bias survey protein evolution is now easily accessible. Using genomes from forty-five mammalian species, we recently interrogated 2,350 genes associated with neuropsychiatric disease in humans. In doing so, we identified strong signatures of purifying selection across mammals with a moderate elevation, indicative of an apparent relaxation of constraint, in catarrhine primates and a large, pervasive, apparent acceleration in cetaceans. This research not only demonstrates a general lack of selection on coding changes in genes associated with these diseases, but also highlights some of the challenges associated with molecular

evolution in the post-genomic era including ortholog identification, genome error and misannotation, and difficulties in functional attribution of genes and proteins.

DUF1220 Protein Domains: Gene Sequences Showing Extreme Human-specific Copy Number Expansion

James Sikela, PhD

University of Colorado School of Medicine

Gene duplication is thought to be the primary means by which evolution creates new genes and biochemical processes that have facilitated the evolution of complex organisms from primitive ones. Gene duplications (gene dosage increases) that are specific to the human lineage have been identified in well over 130 genes some of which are excellent candidates to underlie human lineage-specific traits. Among these sequences are those encoding DUF1220 protein domains, which show the largest human lineage-specific copy number increase of any protein coding region in the genome and have been linked to human brain expansion.

Admixture at Different Timescales in Human Genomes

Jeff Wall, PhD

University of California, San Francisco

Admixture between diverged populations is a common phenomenon in human history. We review some of the methods for detecting this admixture, and show how it provides evidence for admixture between modern humans and 'archaic' human groups across a wide range of contemporary human populations. The strongest signal of this admixture occurs in sub-Saharan African populations, which is consistent with the extent of presumed opportunities for admixture that have been proposed from archeological research.

DISCOVERY INFORMATICS IN BIOLOGICAL AND BIOMEDICAL SCIENCES: RESEARCH CHALLENGES AND OPPORTUNITIES

VASANT HONAVAR

*College of Information Sciences and Technology, Pennsylvania State University
University Park, PA 16802, USA
Email: vhonavar@ist.psu.edu*

New discoveries in biological, biomedical and health sciences are increasingly being driven by our ability to acquire, share, integrate and analyze, and construct and simulate predictive models of biological systems. While much attention has focused on automating routine aspects of management and analysis of "big data", realizing the full potential of "big data" to accelerate discovery calls for automating many other aspects of the scientific process that have so far largely resisted automation: identifying gaps in the current state of knowledge; generating and prioritizing questions; designing studies; designing, prioritizing, planning, and executing experiments; interpreting results; forming hypotheses; drawing conclusions; replicating studies; validating claims; documenting studies; communicating results; reviewing results; and integrating results into the larger body of knowledge in a discipline. Against this background, the PSB workshop on Discovery Informatics in Biological and Biomedical Sciences explores the opportunities and challenges of automating discovery or assisting humans in discovery through advances (i) Understanding, formalization, and information processing accounts of, the entire scientific process; (ii) Design, development, and evaluation of the computational artifacts (representations, processes) that embody such understanding; and (iii) Application of the resulting artifacts and systems to advance science (by augmenting individual or collective human efforts, or by fully automating science).

The workshop, which is especially timely in the context of the NIH Big Data to Knowledge (BD2K) initiative, brings together a group of scientists with complementary expertise to explore research challenges and opportunities in the informatics of discovery in biomedical sciences including, but not limited to:

- Representation and modeling languages with precise formal semantics, for describing, sharing, and communicating scientific models, theories, and hypotheses in biomedical sciences.
- Novel approaches to interactive visualization and exploration of complex biomedical data.
- Sophisticated approaches to construction of comprehensible and communicable predictive models and discovery of causal mechanisms from disparate types of observational and experimental data, literature, images, spatial, temporal, richly structured e.g., network data in biomedical sciences.
- Effective approaches for acquiring and making effective use of background assumptions, hypotheses, knowledge, beliefs and conjectures, arguments, domain expertise, and process descriptions in biomedical sciences.
- Algorithms and tools for automating various aspects of discovery.

Several individuals have contributed to the workshop: Tim Clark, Harvard University; Michel Dumontier, Stanford University; Yolanda Gil, University of Southern California; Lawrence Hunter, University of Colorado at Denver; Marylyn Ritchie, Pennsylvania State University; Andrey Rzhetsky, University of Chicago; Alan Ruttenberg, University at Buffalo; Neil Smalheiser, University of Illinois at Chicago; Nigam Shah, Stanford University.

INVITING THE PUBLIC: THE IMPACT ON INFORMATICS ARISING FROM EMERGING GLOBAL HEALTH RESEARCH PARADIGMS

RICHARD GAYLE

SpreadingScience

Woodinville, WA 98072

Email: gayler@spreadingscience.com

MARK MINIE

University of Washington

Seattle, WA

Email: mark@meminie.com

ERIK NILSSON

inSilicos

Seattle, WA

Email: erik.nilsson@insilicos.com

This workshop will focus on disruptive processes impacting research arising from the increasing ability of individuals to create, curate and share data with scientists. Encompassing processes from funding research to providing samples to creating algorithms, including the public will require new approaches even as it opens up new possibilities. We will hear from a few researchers at the forefront of these disruptive processes, followed by a moderated discussion with the audience about these topics.

1. Why invite the public?

1.1. Two trends

Two accelerating trends in biomedical informatics are producing disruptive effects on research projects. They are also creating new opportunities to apply novel approaches in the analysis of data and the creation of knowledge.

Due to the increasing computational power of smartphones, medical apps and powerful handheld medical devices, we are seeing a radical change in how data are generated and who controls access. The onrushing onslaught of medical data that will arise from the quantified-self^a and the personalized medicine^b revolutions holds the promise of generating enormous storehouses of data, waiting to be examined using novel informatics approaches. However, this also can create barriers for analysis of the data as access is often dispersed to patients.

Perhaps, the second trend may provide a novel approach for solving the difficulties of the first. The formerly separate, once distinct boundaries between patient, researcher, funder, student and entrepreneur are overlapping. In some instances, they are totally disappearing. This allows collaborations and data sharing to happen in unique ways. It also permits the creation of new research and analytical approaches for informatics – if we are creative enough to use them.

Tremendous stores of data are being produced. Yet, knowledge can only be created by examining information in a social setting. We need to understand and utilize new approaches for generating, examining or analyzing the data – by expanding the idea of who collaborators are.

1.2. Some Problems and Some Questions

Research into human health is becoming more global every day. Health records are increasingly digitized, decentralized and personalized. This process can disrupt the ability of researchers to ask new questions and investigate new solutions. It is also impacting the lay population in ways to add greater obstacles as well as opportunities.

Electronic health records (EHR) are rapidly being adopted in developed countries. The number of US hospitals that adopted EHR^c tripled between 2010 and 2012, with over 40% of all hospitals using such systems. This rapid adoption is also apparent in many developing countries. For example, in Malawi, touchscreen digital devices^d are being used for the healthcare needs of over 40,000 HIV patients. Can researchers get ethical access to such widely dispersed or incompatible data? What new informatics approaches will be required to acquire and analyze this data?

While digitization of medical records is being driven from the top-down, the quantified-self and personalized medicine movements are being driven from the bottom-up, with individuals controlling their own health data and creating large personal databases of increasingly granular information. Can community-building approaches help gain access to this data? How will this affect analysis?

^a http://en.wikipedia.org/wiki/Quantified_Self

^b http://en.wikipedia.org/wiki/Personalized_medicine

^c <http://health.usnews.com/health-news/news/articles/4013/07/08/us-hospitals-triple-use-of-electronic-health-records-report>

^d <http://www.plosmedicine.org/article/info:doi/10.1371/journal.pmed.1000319%23s5>

Companies such as Theranos^e are working to fragment the collection of human medical data further, putting devices into pharmacies for customer use that can produce large amounts of personal health data for a few dollars. The rapidly dropping costs for testing of more “-omes” means that the data are becoming not only decentralized, under the control of individuals, but also potentially accessible to a much larger groups, including researchers. What approaches for dealing with the torrents of data that will satisfy all participants will need to be developed?

The continuing digitization of worldwide health data will provide important information simply unattainable before. Although this presents opportunities for examination, it also raises novel barriers for accessing the data as control of the data becomes dispersed amongst individuals. How will the increasing lack of boundaries between individuals controlling the data and those wishing to examine it affect the analysis of the data?

2. The Solutions?

Highly collaborative, open access research networks often destroy the barriers between doctors and patients, between the researcher and the entrepreneur, between the lab bench and the clinic. While disruptive to some processes, these efforts can make it easier for the researcher to gain access to an individual’s data, simply because in many cases the people collecting the data on themselves become a part of the research collaboration.

The rapidly dropping costs for many informatics investigations remove many obstacles formerly seen, allowing the public to fund research they find useful, without asking or waiting for permission^f. For example, the people of Puerto Rico used bake sales and art auctions to raise money to produce the genomic sequence of the Puerto Rican parrot. The resulting data were then assembled and annotated by the local college.^{1,2} This full circle connection between community and academia bootstrapped a much larger effort to examine tropical parrot genomes, produced multiple informatics tools and created the Caribbean Genome Center^g. Research continues to appear derived from this example of public-academia cooperation.³

The Pacific Symposium on Biocomputing has always been at the forefront of new technologies, algorithms and research processes. One paradigm already highlighted has been increasingly collaborative approaches between different researchers to leverage network effects while examining large amounts of data. These collaborations are just one subset of the "public", which can extend from a handful of scientists to millions of people.

The tremendous pressure for reducing the cost/time of drug development is driving new approaches. The overlap of what had been separate roles along the drug development pipeline (i.e. nonprofit biomedical research centers now working on drug development and manufacturing) is just one example. The very idea of who is a collaborator, as well as who is the public, is changing in ways that can actually enhance the ability to examine and analyze the torrent of health data beginning to appear. In all cases, scalable informatics approaches need to be created to deal with the workflow and to help produce results in novel ways.

^e http://news.walgreens.com/article_display.cfm?article_id=5820

^f <http://blogs.biomedcentral.com/gigablog/2014/09/30/community-genomes-from-the-peoples-parrot-to-crowdfiernding/>

^g <https://www.facebook.com/Caribbean.Genome.Center>

Informatics are dealing with the collision of these formerly separate spheres in ways that hold enormous potential, but only if we can become more comfortable with the new paradigms these endeavors require. Coupled with the emerging and exponentially increasing ability of individuals to collect or control huge amounts of personal medical data, informatics requires appreciation of the unique approaches needed to take advantage of this important emerging resource of raw data.

There are many efforts attempting to leverage the two emerging trends of widespread digital health records and the disintegrating boundaries between stakeholders along the drug development pipeline. Inviting the public to fund research has spawned a dedicated site – Experiment^h. New databases are being created by connecting with the publicⁱ. Researchers have mimicked fantasy football^j to help invite the public into the lab. Organizations such as Ingenuity are asking people to become involved in informatics with the chance of winning a prize^k.

Deeper entanglements of research with the public are also happening. For example, several projects have used open challenges that permit the public, along with researchers, to examine data in order to crowdsource better algorithms. These have involved breast cancer models^l, cancer survival rates^m, Alzheimer'sⁿ and Parkinson's disease^o. Open competitions for predicting drug responses to arthritis drugs by examining genetic data have also been announced.⁵

In many cases, the creation and support of open science communities may permit greater collaborations, as well as decreased barriers between the public and academia. Sites such as 23andme^p or patientslikeme^q are forcing new approaches for communicating science as well as analyzing data. Understanding how communities deal with information and innovation will be critical for many future research projects.

We hope to hear from researchers who are embracing new approaches for creating informatics databases, for crowd sourcing the analysis of the databases, for disintermediating the examination of data and for leveraging low cost, low barrier-to-entry, open research approaches to more rapidly understand the complex systems being studied today in health investigations.

3. For Discussion

This workshop has invited researchers working to discover the best practices for combining the dual trends of increasing digital health data with increasingly overlapping roles of scientific collaboration between researchers and interested communities that may control the data.

Richard Gayle, President of SpreadingScience, will be moderating this workshop. Using his commercial experience (as a researcher at a biotechnology company and as Vice-President of a

^h <https://experiment.com>

ⁱ <http://www.northeastern.edu/pollastri/collaborate/>

^j <http://www.kplu.org/post/seattle-scientists-look-make-drug-research-more-fantasy-football>

^k <http://www.ingenuity.com/blog/customer-stories/causal-variant-challenge-update-our-ipad-winner-and-next-challenge>

^l <http://www.ploscompbiol.org/article/info%3Adoi%2F10.1371%2Fjournal.pcbi.1003047>

^m <http://www.ceoalzheimersinitiative.org/global-ceo-initiative-alzheimer's-disease-announces-big-data-challenge-find-new-predictors-cogniti-0>

ⁿ <http://news.patientslikeme.com/press-release/patientslikeme-and-sage-bionetworks-launch-open-science-study-people-parkinsons-diseas>

^o <https://www.23andme.com>

^p <http://www.patientslikeme.com>

small startup), he has recently been examining how information transits communities. Enlarging the communities of practice that can examine data and produce knowledge is a key aspect. He has helped organize successful crowdfunding projects for research, including the Arkyd project^q, which raised over \$1.5 million for a satellite for the people. He has also raised money for his own research using these approaches^r.

Daniel McDonald is currently at the University of Colorado at Boulder, working in the lab of Rob Knight. He is involved with the American Gut Project,^s examining the microbiomes found in humans. This project not only provides a service to the public but also collects data from the same public for research, requiring new informatics tools to be developed. It has so far raised almost \$600,000 through crowdfunding approaches, with over 6000 people submitting the samples used in the study^t. Some preliminary data are already openly available^u. He has been directly involved in designing and implementing bioinformatics tools for reducing data bottlenecks when dealing with a variety of “-omics.”⁶

Jonathan Eisen, at the University of California, Davis, has long been focused on open science approaches for engaging the public. These have also involved gamifying^v the research. He has extended high throughput sequencing approaches from the human gut to the built environment^w. He has taken a lead on examining the communication barriers that these processes can produce^x.

Following short presentations from the three speakers, we will have a moderated discussion. We hope to create a vibrant conversation including all participants regarding current and future approaches for including a wider public in research processes.

The entire healthcare field has finally found Moore’s Law, seeing enormous and rapid change as patient records become fully digitized. Quickly understanding the processes involved to effectively manage successful human health research projects using these unique datasets will be as important as developing new analytical tools to examine the data.

References

1. S. J. O’Brien, *GigaScience*, **1**, 13 (2012)
2. T. Oleksyk, *GigaScience*, **1** 14 (2012)
3. Y. Afanador, *Conservation Genetics Resources*, doi 10.1007/s12686-014-0232-6 (2014)
4. A. Margolin et al., *Sci Transl Med*, **5**, 381 (2013)
5. R. Plenge et al., *Nature Genetics*, **45**, 468 (2013)
6. D. McDonald et al., *Gigascience*. **1**, 7 (2012).

^q <https://www.kickstarter.com/projects/arkydforeveryone/arkyd-a-space-telescope-for-everyone-0>

^r <http://www.rockethub.com/projects/28741-consider-the-facts-moving-people-to-deliberative-thinking>

^s <http://american-gut.org>

^t <https://fundrazr.com/campaigns/4Tqx5>

^u http://american-gut.org/wordpress/wp-content/uploads/2013/09/module1_Sept_16_small.pdf

^v <http://phylogenomics.wordpress.com/gut-check-the-microbiome-game/>

^w <http://microbe.net>

^x <http://icis.ucdavis.edu>

TRAINING THE NEXT GENERATION OF QUANTITATIVE BIOLOGISTS IN THE ERA OF BIG DATA

KRISTINE A. PATTIN AND ANNA C. GREENE

*Institute for Quantitative Biomedical Sciences, Dartmouth College
Hanover, NH 03755, USA*

Email: Kristine.A.Pattin@Dartmouth.edu, Anna.C.Greene@Dartmouth.edu

RUSS B. ALTMAN

*Department of Genetics, Stanford University
Stanford, CA 94305, USA
Email: russ.altman@stanford.edu*

KEVIN B. COHEN, ELIZABETH WETHINGTON, CARSTEN GÖRG

*Computational Bioscience Program, University of Colorado School of Medicine
Aurora, CO 80045, USA
Email: kevin.cohen@ucdenver.edu*

LAWRENCE E. HUNTER

*Computational Bioscience Program, University of Colorado School of Medicine
Aurora, CO 80045, USA
Email: larry.hunter@ucdenver.edu*

SPENCER V. MUSE

*Department of Statistics, North Carolina State University
Raleigh, NC 27695, USA
Email: muse@ncsu.edu*

PREDRAG RADIVOJAC

*Department of Computer Science and Informatics, Indiana University
Bloomington, IN 47405, USA
Email: predrag@indiana.edu*

JASON H. MOORE

*Institute for Quantitative Biomedical Sciences, Dartmouth College
Hanover, NH 03755, USA
Email: Jason.H.Moore@Dartmouth.edu*

1. Workshop Focus

Francis Collins recently stated that, “the era of ‘Big Data’ has arrived, and it is vital that the NIH play a major role in coordinating access to and analysis of many different data types that make up this revolution in biological information.”¹ With this, Philip E. Bourne was named as the Associate Director for Data Science at the NIH, the first permanent appointment of this position. Additionally, through the Big Data Initiative started in 2012, the Obama Administration invested \$200 million dollars in “big data” research that promises “to greatly improve the tools and techniques needed to access, organize, and glean discoveries from huge volumes of digital data.”² The term “big data” extends beyond the research arena into the popular press. CNN Money has named the data scientist job as one of the best jobs in America (#32/100).³ Harvard Business Review Magazine has named the data scientist as “the sexiest job” of the 21st Century.⁴ Yet, what exactly is a data scientist? The focus of this workshop is to discuss key skill sets for biomedical data scientists to determine if they differ from a standard bioinformatics curriculum.

- Is there any substantive difference between a biomedical data scientist and a biomedical informaticist? If so, how do we train one versus the other?
- Are current bioinformatics curricula evolving to encompass the realm of data science?

- Are there obsolete lessons in coursework that could be replaced with more modern technical information?

We will have 6 scientists from various quantitative fields describe how their program's curriculum is structured and what changes they have made or anticipate they could make to strengthen and align the program with current practices in data science and bioinformatics. They will also speak to where future training in bioinformatics should go in the growing era of big data.

2. Workshop Contributions

The speakers were asked to respond to the following question: What is a data scientist? Do the key skill sets for biomedical data scientists differ from a standard bioinformatics curriculum?

Russ B. Altman. I think that the concept of a “Data scientist” has emerged within industries where large amounts of information are collected and managed by individuals with skills in statistics, computer science, information science and related disciplines. For many of these industries, there is no tradition of employees with this skill set—they were used to hiring engineers from the traditional engineering disciplines or (in some cases) natural scientists from biology, chemistry, and physics. Sometimes they may have hired a statistician, but this was usually for study design or analysis of relatively orderly “controlled” data. The phenomenon of an employee with a firm grounding in statistics, but also with ability to write and run programs to handle relatively large amounts of data[Footnote1], and apply the principles of data mining and machine learning is new in many fields. In addition, there are skills from informatics that are also critical including understanding the use and maintenance of controlled terminologies and ontologies.

Within biomedical research, the field of biomedical informatics has existed (arguably) since the early 1960's when Ledley & Lusted outlined in *Science*⁵ some of the major challenges to information sciences in biomedicine. In the early 1980's programs emerged to train professionals in biomedical informatics. The curricula that emerged were, in many cases, very similar to the curricula created today for data scientists; they included a strong background in computer science, statistics, probability, decision theory and (importantly) courses in the domain of application. The final element is quite important so that the individual understands the major questions and challenges in the domain, and knows when certain problems have been solved, and when they are unsolved. The main concern about undifferentiated data scientists who lack domain knowledge is whether they will be as efficient and effective as practitioners with an understanding of the underlying application area. For biomedicine, there is little doubt that the best data scientists will be those who understand the special features and challenges in biology or medicine, and thus make assumptions and approximations that are valid and not fatal.

[Footnote1] In this context, “Big Data” can be defined as any data set that is mission critical to the organization and bigger than what their current infrastructure can handle. As soon as the infrastructure and staff adjust to “Big Data” it becomes regular data.

Kevin B. Cohen, Elizabeth Wethington, Carsten Görg. Examining the advertisements for open positions for data scientists on the popular Monster.com job-seeking website shows that biomedical science is well-represented in the data science job market (at least on the date of search). The search [data scientist] returns 180 job openings, and [data scientist biomedical] returns 8. (The search [data scientist health] returns 30, but many of these simply mention that they provide health insurance to employees.) Examining the advertisements themselves is revealing. The sample returned by the [data scientist biomedical] search is small—8 positions in total—but some trends emerge.

The first thing of note is that the list of required skills for most of these positions is short. This might be somewhat surprising. Data scientists are typically thought of as some sort of engineer or statistician on the one hand, or as a sort of jack-of-all-trades on the other. These advertisements suggest that a jack-of-all-trades is not needed, but rather that a relatively small set of skills will suffice in most (although not all) cases. (This is apparently true of getting the job—whether or not a limited skill set would be sufficient to keep the job is less obvious.) In the small skill sets mentioned in most of these advertisements, two specific skills predominate. Databases are mentioned in three of them, and statistics are mentioned in three of them—not the same three.

How does this compare to a standard bioinformatics curriculum? A recent highly unscientific (and unpublished) survey of bioinformatics doctoral programs showed that databases were not part of any them, and statistics was not covered in an independent class in any of them. It is, however, unlikely that students typically leave a bioinformatics doctoral program without any background in statistics or databases—it is likely that they enter their doctoral program already having a background in these areas. If not, they are likely to pick it up in the course of their education, although our survey suggests that they are not doing so in their coursework. The key skill sets for biomedical data scientists do seem to differ from a standard bioinformatics curriculum.

Lawrence E. Hunter. Taking “Data Science” as defined in Vasant Dhar’s CACM article⁶ summarized as “the generalizable extraction of knowledge from data” and requiring “an integrated skill set spanning mathematics, machine learning, artificial intelligence, statistics, databases and optimization, along with a deep understanding of the craft of problem formulation to engineer effective solutions”; while there are clear overlaps between bioinformatics and biomedical data science, there are also important differences. Significant aspects of bioinformatics fall outside of this definition. For example, many of the key methods for dealing with protein structural data (e.g. molecular dynamics simulation or structural visualization) are not subsumed by Data Science, but are clearly of fundamental importance in bioinformatics. Likewise, many of the critical techniques for handling massive short read sequencing data would not be included in a reasonable definition of data science. A well-trained bioinformatician knows about computational techniques that are important in contemporary molecular biology, but that are not clearly part of the Data Science toolkit. Furthermore, an effective bioinformatics researcher will have deeper domain knowledge than is typically assumed for a data scientist. Many important innovations in bioinformatics have come from a deep familiarity with the underlying biology or even more frequently with the experimental methodologies that generate the data to be analyzed. Insights into the idiosyncrasies of instruments such as mass spectrometers and hybridization arrays have led to dramatic improvements in informatics methods not available to those who treat data as a “given”. Perhaps the most important difference, however, is not about the computational methods or domain knowledge of the practitioners, but about the goals of the scientific work. Philosophers of science Carl Craver and Lindley Darden have eloquently described the central role of elucidation of mechanism in biology.⁷ Data science is largely concerned with finding patterns in data. While such patterns have the potential to be extremely helpful to understanding living systems, their identification is the beginning of biology, not the end. Biologists insist on mechanistic understandings of the phenomena they observe, not merely predictive ones. Bioinformatics must always be acutely sensitive to the needs of biologists to hypothesize and test mechanisms, not just to find predictive patterns.

Spencer V. Muse. The Harvard Business Review recently dubbed data science “the sexiest job of the 21st century.” Acknowledging the challenge of defining the field, they suggested that “data scientists’ most basic, universal skill is the ability to write code”. This claim implies that data

science is grounded in computer science. Reflecting a different perspective, celebrity statistician Nate Silver remarked that data science is a “sexed up term for statistician.” The truth likely falling somewhere in the middle, most data scientists would likely agree that they work at the intersection of computer science and statistics, with a heavy dose of discipline-specific knowledge thrown into the mix. The demand for these individuals has presented a workforce and training challenge. The fundamental difficulty is one shared by most emerging interdisciplinary areas: the traditional educational paths in the constituent fields lack the breadth or flexibility to allow students to easily become data scientists. A core set of skills for data science training is beginning to emerge, though. Students must be proficient programmers able to work with large heterogeneous data sets, often distributed across multiple locations. (Note that few traditionally-trained statisticians have those skills.) Students must also be fluent with a wide range of statistical techniques, have a strong knowledge modeling complex data, and be able to combine those skills to build advanced statistical analysis tools. (Abilities rarely found in traditionally-trained computer scientists.)

It is no surprise that the influx of data has created tremendous demand for data scientists. Under the umbrella of “biomedical informatics” we now have specialization in areas including medical informatics, clinical informatics, and bioinformatics. In the same way that one would not expect an endocrinologist to perform well as a cardiologist simply because they are both physicians, one should not expect to place someone trained in, say, bioinformatics into a medical informatics position and get satisfactory performance. While there is certainly a high degree of overlap (e.g. complex, high-order database searches; network construction; data mining and predictive modeling), the details of the needs and tools in each specialty are driven by the fundamentals specific to each, and one can neither be an effective developer or user of the tools without being firmly grounded in the underlying discipline.

Predrag Radivojac. Data science may best be described as a discipline whose intellectual core derives from the interplay between statistics and computer science. Statistics generally provides frameworks for modeling and inference from data. Numerous such approaches have been proposed in both predictive and descriptive scenarios as well as for characterizing inference methods. Computer science, on the other hand, studies computing paradigms for implementing such approaches. It generally provides algorithmic framework for solving statistically formulated problems, given the resources such as a particular computer architecture, clock time and memory. In addition, computer science provides a framework to formally address data management, software engineering and visualization issues. Various concepts from other disciplines also contribute to data science; for example, those from physics, biology, psychology, logic, information theory and others.

A biomedical data scientist must possess core competencies in statistics and computer science, but must also understand the biomedical side of the equation. Biomedical expertise may come from a diverse set of sub-disciplines, including molecular biology, developmental biology, evolutionary biology, biochemistry, analytical chemistry, genetics, pharmacology and neuroscience, but also a combination thereof. Overall, a biomedical data scientist must not only have deep domain expertise and the ability to identify important biomedical problems, but also the ability to formally pose such problems within statistical and computer science frameworks and then properly solve them.

I believe that the core skills of a biomedical data scientist significantly overlap with those of a bioinformatics scientist, but the main difference may come from the emphasis on particular problems rather than the ability of such scientists to be able to tackle them. A data scientist may have a larger and deeper focus on data modeling problems, perhaps of the systems biology or functional genomics flavor, whereas a narrowly defined bioinformatician may be more focused on

algorithmic issues such as sequence analysis. However, a large number of traditional bioinformaticians regularly handle data modeling issues typically by developing and applying machine learning methodologies as well as by creating tools for biologists and medical scientists.

In my view, biomedical data science is a suitable umbrella term for a host of other disciplines that rely on biomedical data to ultimately produce knowledge. At this time, however, it is too early to tell whether it will have transformative impact on biomedical research beyond what other (overlapping) disciplines have already initiated. Consequently, the development of biomedical data science curricula may not require significant restructuring of the more traditional but broadly defined bioinformatics, biomedical informatics, or systems biology curricula on many campuses.

Jason H. Moore. Data science is a rapidly emerging discipline that combines pieces of computer science and statistics to manage and analyze big data across different domains. At face value, this definition is not much different than some definitions of bioinformatics where the big data is coming from biological or biomedical sources. One possible difference between the two disciplines is with regard to the integration of statistics. Bioinformatics has traditionally focused much more on the computational sciences including algorithms, databases, high-performance computing, machine learning and software engineering, for example. This is likely due, in part, to the lack of formal statistics training in computer science curricula. Fully exploiting the potential of big data requires an equal mix of computational and statistical sciences. For example, a working knowledge of statistical inference can significantly complement machine learning approaches to big data where false-positives (type I errors) and false-negatives (type II errors) are common. Similarly, the ability to complement computational methods such as support vector machines with statistical methods such as logistic regression expand the analytical toolbox in useful ways. The demand is there and data scientists are few and far between given the rarity of in depth training in both computational and statistical sciences. Given the need for this unique blend of skills and expertise it might be time for bioinformatics training programs to consider adding additional courses in statistics to the curriculum. These courses would not replace those in algorithms and databases but rather extend the requirements. Additional courses will take additional time to complete. This is likely unappealing to some but may be necessary to fully prepare our graduate students for a world of big data.

References

1. “NIH Names Dr. Philip E. Bourne First Associate Director for Data Science”. 2013. <http://www.nih.gov/news/health/dec2013/od-09.htm> [Feb 13 2014].
2. “Obama Administration Unveils ‘Big Data’ Initiative: Announces \$200 Million In New R&D Investments”. 2012. http://www.whitehouse.gov/sites/default/files/microsites/ostp/big_data_press_release_final_2.pdf [Feb 13 2014].
3. “IT Data Scientist - Best Jobs in America 2013”. 2013. http://money.cnn.com/pf/best-jobs/2013/snapshots/32.html?iid=BestJobs_fl_list [Feb 13 2014].
4. “Data Scientist: The Sexiest Job of the 21st Century - Harvard Business Review”. 2012. <http://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century/> [Feb 13 2014].
5. R.S. Ledley, L.B. Lusted. “Reasoning Foundations of Medical Diagnosis: Symbolic logic, probability, and value theory aid our understanding of how physicians reason”. *Science*, 1959. 130(3366): 9–21. 10.1126/science.130.3366.9.
6. V. Dhar. “Data science and prediction”. *Commun. ACM*. ACM, 2013. 56(12): 64–73. 10.1145/2500499.
7. L. Darden. book: Craver, Carl F. and Lindley Darden (2013), *In Search of Biological Mechanisms: Discoveries across the Life Sciences*. Chicago, IL: University of Chicago Press. 2013.

ERRATUM

[SARAH A. PENDERGRASS, SHEFALI S. VERMA, EMILY R. HOLZINGER, CARRIE B. MOORE, JOHN WALLACE, SCOTT M. DUDEK, WAYNE HUGGINS, TERRIE KITCHNER, CAROL WAUDBY, RICHARD BERG, CATHERINE A. MCCARTY, and MARYLYN D. RITCHIE (2012) NEXT-GENERATION ANALYSIS OF CATARACTS: DETERMINING KNOWLEDGE DRIVEN GENE-GENE INTERACTIONS USING BIOFILTER, AND GENE-ENVIRONMENT INTERACTIONS USING THE PHENX TOOLKIT. *Biocomputing 2013*: pp. 147-158, doi: 10.1142/9789814447973_0015]

]

"This corrects the above-titled article. There was an error in the case-control label for a subset of samples. This was corrected and analyses were re-run. The thrust of the results and discussion did not change, but these results are more precise and corrected."

**NEXT-GENERATION ANALYSIS OF CATARACTS: DETERMINING
KNOWLEDGE DRIVEN GENE-GENE INTERACTIONS USING BIOFILTER, AND
GENE-ENVIRONMENT INTERACTIONS USING THE PHENX TOOLKIT***

SARAH A. PENDERGRASS

*Center for Systems Genomics, Department of Biochemistry and Molecular Biology, The Pennsylvania State University, 503 Wartik Lab
University Park, PA 16802, USA
Email: sap29@psu.edu*

SHEFALI S. VERMA

*Center for Systems Genomics, Department of Biochemistry and Molecular Biology, The Pennsylvania State University, 512 Wartik Lab
University Park, PA 16802, USA
Email: szs14@psu.edu*

MOLLY A. HALL

*Center for Systems Genomics, Department of Biochemistry and Molecular Biology, The Pennsylvania State University, 512 Wartik Lab
University Park, PA 16802, USA
Email: mah546@psu.edu*

EMILY R. HOLZINGER

*Center for Systems Genomics, Department of Biochemistry and Molecular Biology, The Pennsylvania State University, 512 Wartik Lab
University Park, PA 16802, USA
Email: Emily.R.Holzinger@vanderbilt.edu*

CARRIE B. MOORE

*Center for Systems Genomics, Department of Biochemistry and Molecular Biology, The Pennsylvania State University, 512 Wartik Lab
University Park, PA 16802, USA
Email: ccb12@psu.edu*

JOHN R. WALLACE

*Center for Systems Genomics, Department of Biochemistry and Molecular Biology, The Pennsylvania State University, 512 Wartik Lab
University Park, PA 16802, USA
Email: jrw32@psu.edu*

SCOTT M. DUDEK

*Center for Systems Genomics, Department of Biochemistry and Molecular Biology, The Pennsylvania State University, 512 Wartik Lab
University Park, PA 16802, USA
Email: sud23@psu.edu*

WAYNE HUGGINS

*RTI International
Research Triangle Park, NC, USA
Email: whuggins@rti.org*

TERRIE KITCHNER

Marshfield Clinic

Marshfield, WI, USA

Email: Kitchner.Terrie@mcrf.mfldclin.edu

CAROL WAUDBY

Marshfield Clinic

Marshfield, WI, USA

Email: WAUDBY.CAROL@mcrf.mfldclin.edu

RICHARD BERG

Marshfield Clinic

Marshfield, WI, USA

Email: Berg.Richard@mcrf.mfldclin.edu

CATHERINE A. MCCARTY

Essential Rural Health

Duluth, MN, USA

Email: CMcCarty@eirh.org

MARYLYN D. RITCHIE

Center for Systems Genomics, Department of Biochemistry and Molecular Biology, The Pennsylvania State University, 512 Wartik Lab, University Park, PA 16802, USA

Email: Marylyn.ritchie@psu.edu

Investigating the association between biobank derived genomic data and the information of linked electronic health records (EHRs) is an emerging area of research for dissecting the architecture of complex human traits, where cases and controls for study are defined through the use of electronic phenotyping algorithms deployed in large EHR systems. For our study, cataract cases and controls were identified within the Marshfield Personalized Medicine Research Project (PMRP) biobank and linked EHR, which is a member of the NHGRI-funded electronic Medical Records and Genomics (eMERGE) Network. Our goal was to explore potential gene-gene and gene-environment interactions within these data for 527,953 and 527,936 single nucleotide polymorphisms (SNPs) for gene-gene and gene-environment analyses, respectively, with minor allele frequency > 1%, in order to explore higher level associations with cataract risk beyond investigations of single SNP-phenotype associations. To build our SNP-SNP interaction models we utilized a prior-knowledge driven filtering method called Biofilter to minimize the multiple testing burden of exploring the vast array of interaction models possible from our extensive number of SNPs. Using Biofilter, we developed 57,376 prior-knowledge directed SNP-SNP models to test for association with cataract status. We selected models that required 6 sources of external domain knowledge. We identified 13 statistically significant SNP-SNP models with an interaction with p-value < 1×10^{-4} , as well as an overall model with p-value < 0.01 associated with cataract status. We also conducted gene-environment interaction analyses for all GWAS SNPs and a set of environmental factors from the PhenX Toolkit: smoking, UV exposure, and alcohol use; these environmental factors have been previously associated with the formation of cataracts. We found a total of 782 gene-environment models that exhibit an interaction with a p-value < 1×10^{-4} associated with cataract status. Our results show these approaches enable advanced searches for epistasis and gene-environment interactions beyond GWAS, and that the EHR based approach provides an additional source of data for seeking these advanced explanatory models of the etiology of complex disease/outcome such as cataracts.

* This work supported by the following grants: U19 HL0659625, R01 LM010040, U01 HG006389

1. Introduction

DNA biobanks coupled to electronic health records (EHR) have become a valuable resource for investigating the genetic architecture of complex traits, as EHR contain a wide array of medical information including billing codes and clinical laboratory measurements, often yielding a large sample size. Through carefully defining phenotypes, and using deployable algorithms that combine multiple sources of information in the EHR, cases and controls can be defined for association studies, such as defining age-related cataract cases and controls [1,2]. The Marshfield Personalized Medicine Research Project biobank (Marshfield PMRP) and linked EHR, used for the study described herein, is one such resource [3]. The Marshfield PMRP is a member of the NHGRI-funded electronic Medical Records and Genomics (eMERGE) Network, a network of similar biobanks coupled with EHR based data [4].

Cataracts are a leading cause of blindness globally [5], and are believed to arise from a combination of age, environmental factors, and heritable factors [6]. Thus, understanding the genetic etiology of cataracts, coupled with the effect of environment as a modifier, could have a profound impact on human health. For our study, algorithms proven for age-related cataract case identification [2] were deployed in the Marshfield PMRP EHR to identify 2580 cataract cases and 1367 controls, with further study details presented in Table 1. A total of 527,953 (gene-gene interactions) and 527,936 (gene-environment interactions) single nucleotide polymorphisms (SNPs) were available after PMRP genotyping coupled with quality control filtering and selection for SNPs with a minor allele frequency > 1%.

Table 1. Marshfield Cataract Study Description

	Gene-Environment Analysis	Gene-Gene Analysis
Age	> 50	> 50
Ancestry	European American	European American
Total Samples	2,033	3,377
Cases	1,242	2,192
Controls	791	1,185
Males	821	1,408
Females	1,212	1,969
SNPs	527,936	527,953

Single SNP-phenotype associations are a dominant study design used in most genome-wide association studies (GWAS), however, more complex models that include interactions may more accurately describe the relationship between genetic variation and complex outcomes. Investigating all gene by gene (GxG), and in extension, all SNP by SNP (SNPxSNP) pairwise models is possible depending on the number of SNPs that have been genotyped. Unfortunately, the multiple hypothesis testing burden and risk of Type I error is inflated when investigating all pairwise models. A different approach can be used, utilizing prior biological knowledge methods directing model development. Thus, to investigate more complex models beyond single SNP-Phenotype associations for the Marshfield PMRP cataract dataset, we used the prior knowledge accessible through Biofilter 1.0 (a new implementation of Biofilter after the original description in [7]) to direct the investigation of pairwise GxG interaction models based on the following resources: the Kyoto Encyclopedia of Genes and Genomes (KEGG) [8], Reactome [9], Gene Ontology (GO) [10], the Protein families database (Pfam) [11], NetPath [12], Biological General Repository for Interaction Datasets (BioGrid) [13], and the Molecular INTeraction Database (MINT) [14]. Using the Biofilter, we developed 57,376 prior-knowledge directed SNPxSNP models to test for association with cataract status.

In addition, for this study we investigated gene-environment interactions (GxE), as there are clearly known environmental exposures that increase cataract risk, and when incorporated into analyses, may provide new models for the contribution of both environment and genetic architecture to cataracts. The Marshfield PMRP collected standardized Phenotypes and eXposures (PhenX) measures as a member of the PhenX Real-world, Implementation, SharingING (PhenX RISING) project. PhenX has the goal of defining standard phenotypic measures through a framework of measurement protocols via a web-based toolkit [15]. Environmental exposures such as smoking, sun exposure, and alcohol use, have been associated with increased cataract rates [16]. Thus we used 12 PhenX defined environmental exposures to investigate GxE interactions for the Marshfield PMRP cataract data focused on smoking, UV exposure, and alcohol use measures.

Through integrating EHR data, advanced bioinformatics tools, and PhenX, we can pursue advanced searches for epistasis and gene-environment interactions in genome-wide studies of common disease.

2. Methods

2.1. *Marshfield EHR and Age-Related Cataract Case Identification*

The Marshfield PMRP is a population based biobank with ~20,000 subjects, aged 18 years and older, enrolled in the Marshfield Clinic healthcare system in central Wisconsin [3]. DNA, plasma, and serum samples are collected at the time the enrollee completes a written informed consent document, with allowance for ongoing access to the linked medical records. PMRP participants also complete questionnaires, including responses regarding smoking history, occupation, and diet.

To identify cataract surgery cases aged 50 years and older within the PMRP, Current Procedural Terminology (CPT) codes in the Marshfield Clinic EHR were used. A research coordinator manually abstracted additional information to identify the eye affected, the type and severity of the cataract, and the level of visual acuity prior to the cataract surgery. This was also done to remove any cases with non-age related cataracts.

To identify individuals with diagnosed cataracts but without surgery, and to identify the type of cataract, International classification of diseases, 9th revision (ICD-9) codes and CPT codes were used, coupled with Natural Language Processing (NLP) and Intelligent Character Recognition (ICR) of free-text in the EHR. NLP and exclusion criteria were used to identify individuals with congenital and traumatic cataracts for omission from the study. Further details of the identification of cataract cases and controls and the efficacy of the EHR defined phenotyping can be found in Waudby et al., 2011 [2]. All total, the procedures used on the EHR identified 2,192 cases and 1,185 controls for gene-gene analysis and 1,242 cases and 791 controls for gene-environment analysis.

2.2. *Genotyping*

The eMERGE network and the Center for Inherited Disease Research (CIDR) at Johns Hopkins university performed the genotyping of the Marshfield PMRP samples, using the Illumina Human660W-Quadv1 A platform with total of 560,635 SNPs and 96,731 intensity-only probes. Bead Studio version 3.3.7 was used by CIDR for the genotyping calls. The total

cohort genotyped included 3947 samples from the Marshfield PRMP, 21 blind duplicates, and 85 HapMap controls. The HapMap concordance rate was 99.8% and the blind duplicate reproducibility rate was 99.99%. For quality control and data cleaning the eMERGE quality control (QC) pipeline developed by the eMERGE Genomics Working Group [17] was used. Any SNPs with a minor allele frequency > 1%, SNP call rate > 99%, Sample Call Rate > 99% were used in further analysis. After QC and allele frequency filtering using PLINK [18], a total of 527,953 and 527,936 SNPs were used for further gene-gene and gene-environment analyses, respectively.

2.3. PhenX

The standardized phenotypic and environmental consensus measures for Phenotypes and eXposures (PhenX) [15] were used to capture the environmental variables used in this study. The PhenX Toolkit (<https://www.phenxtoolkit.org/>) offers high-quality, well-established, standard measures of phenotypes and exposures for use in epidemiological studies.

The Marshfield PRMP is part of the PhenX RISING consortium, which is comprised of seven groups funded by the National Human Genome Research Institute (NHGRI) and the Office of Behavioral and Social Sciences Research (OBSSR) to incorporate PhenX (<https://www.phenxtoolkit.org/>) measures into existing population-based genomic studies.

For this initiative, Marshfield PRMP subjects with GWAS data who were alive with known, non-institutionalized addresses and who had given consent for re-contact were mailed a 32-page self-administered questionnaire that contained 35 PhenX measures across a range of phenotypic domains including alcohol and tobacco use questions (McCarty et al. 2012, *in preparation*). For this study, we considered 12 of these measures, shown in Table 2.

2.4. BioFilter 1.0

For the SNPxSNP analysis, Biofilter 1.0 was used. Biofilter has been upgraded from the initial Biofilter 0.5 [7], with the addition of more data sources, improved the handling of data, and the development of an eternal database for prior knowledge called the Library of Knowledge Integration (LOKI). Biofilter 1.0 and LOKI are freely available for non-commercial research institutions. For full details see: <http://ritchielab.psu.edu/ritchielab/software>.

Biofilter 1.0 utilizes prior biological knowledge through accessing the data of several publicly available biological information databases, all compiled within the LOKI database developed specifically for Biofilter. The data sources selected for Biofilter contain information on networks, connections, and/or pathways that establish relationships between genes and gene products. Biofilter is a “gene based” approach, thus all the region information (such as genes) and position information (such as SNPs) are mapped to genes within LOKI.

The following sources that are compiled within LOKI were used for the Biofilter model building: the Kyoto Encyclopedia of Genes and Genomes (KEGG) [8], Reactome [9], Gene Ontology (GO) [10], the Protein families database (Pfam) [11], and NetPath [12], Biological General Repository for Interaction Datasets (BioGrid) [13], and the Molecular INTeraction Database (MINT) [14]. The database source used in LOKI solely for the purpose of mapping SNPs to genes is the National Center for Biotechnology (NCBI) dbSNP [19] database.

Table 2. The PhenX measures used for this study

PhenX Measure	Survey Question
PX030301 Alcohol 30Day Frequency	During the past 30 days, on how many days did you drink one or more drinks of an alcoholic beverage?
PX030301 Alcohol 30Day Quantity	During the past 30 days, how many drinks did you usually have each day?
PX030602 Cigarette Smoking 100	Have you smoked at least 100 cigarettes in your entire life?
PX030602 Cigarette Smoking Current	Do you now smoke cigarettes every day, some days, or not at all?
PX030602 Cigarette Smoking Everyday 6Month	Have you EVER smoked cigarettes EVERY DAY for at least 6 months?
PX030802 Everyday Smoker Quantity 1Day	On the average, about how many cigarettes do you now smoke each day?
PX030802 Someday Smoker Days 1Month	On how many of the past 30 days did you smoke cigarettes?
PX030802 Someday Smoker Quantity 1Day	On the average, on those days, how many cigarettes did you usually smoke each day?
PX030802 Former Smoker Smoking 6Month	Have you EVER smoked cigarettes EVERY DAY for at least 6 months?
PX030802 Former Smoker Quantity 1DayA	When you last smoked every day, on average how many cigarettes did you smoke each day?
PX030802 Former Smoker Quantity 1DayB	When you last smoked fairly regularly, on average how many cigarettes did you smoke each day?
PX061301 Weekend Sun Hours Last Decade	On a typical weekend day in the summer, about how many hours did you generally spend in the mid-day sun in the past ten years?

The following process was used within Biofilter 1.0 to develop the SNPxSNP models used in prior knowledge directed association testing. Figure 1 shows a simplified example of how the Biofilter 1.0 model generation process works. First, the input list of SNPs are mapped to genes within Biofilter. Next, comprehensive pairs of genes that are all terminal leaves of the graph for Pathway 1 in Source 1, and Pathway 2 in Source 1 are generated, only for genes that contain SNPs in the input list of SNPs.

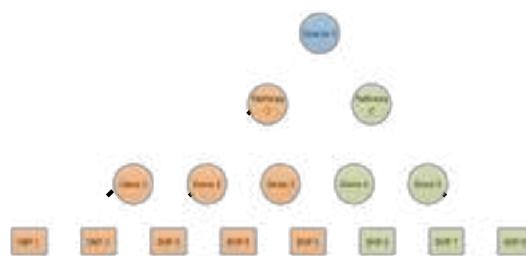


Figure 1. Simplified model for one Biofilter 1.0 database source with 2 pathways, 5 genes, and 8 SNPs

Implication scores are used in Biofilter to give each pairwise model a “score” indicating how many sources have that connected pair of genes represented, the higher the implication score, the more sources have indicated a connection between a pair of genes. The implication index is a measure of the number of data sources providing evidence of an interaction, a sum of the number of data sources supporting each of the two genes and the connection between them. In the case of

our simplified example, for Genes 1-5, that all contain SNPs within the input list, the following pairwise Gene-Gene models would result, each with an implication score of 1:

Gene1 – Gene 2
Gene1 – Gene 3
Gene 2 – Gene 3
Gene 4 – Gene 5

This process continues through all other sources used for Biofilter. Each time a pairwise combination of genes is found in another source (such as the pair Gene1-Gene2), the implication score for that pairwise model will be increased by 1. Lastly, the G-G models are broken into all pairwise combinations of SNPs across the genes, *within P1 or P2*. The SNP-SNP models would look like the following:

SNP1-SNP3
SNP1-SNP4
SNP1-SNP5
SNP2-SNP3
SNP2-SNP4
SNP2-SNP5
SNP3-SNP5
SNP3-SNP4
SNP6-SNP7
SNP6-SNP8

This same process was used within Biofilter 1.0 to develop the SNPxSNP models used for our prior knowledge directed association testing. First, the 527,953 SNPs were mapped to their corresponding genes. Next, the genes corresponding to the SNPs of the dataset were mapped to the gene-relationship graphs for each LOKI source used. After this mapping process, gene pairs were exhaustively generated for each occurrence of two genes within a single pathway and single source. Implication scores were calculated for the pairwise models. After the gene-gene models were developed in Biofilter, the models were divided into exhaustive SNP-SNP pairs for association testing.

Table 3 indicates the number of models that were found at each implication score cutoff. An implication index cutoff of 4 actually incorporates all possible pairwise models for all SNPs we had for this study, a total of 603,032 models. We found an implication score cutoff of 6 resulted in a balance between a large group of models for exploration (57,376 models), but still maintained a very computationally feasible set of associations to investigate, limiting our type 1 error rate more than using all exhaustive pairs of SNP-SNPs or some of the less stringent implication score cutoffs. With a requirement for an implication index of 6, as we had in this study, the gene-gene relationship or known interaction had to be found in nearly all of the data sources we used within LOKI.

Table 3. Number of Resulting Models for Each Implication Score Cutoff

Implication Index Cutoff	Number of Models
4	603,032
5	337,113
6	57,376
7	2479

2.5. Statistical Analysis

For the SNPxSNP models generated through the use of Biofilter, PLATO [20] was used to determine the significance of the interaction via likelihood ratio test (LRT) of the full versus

reduced models, using logistic regression, where the full model was: SNP1 + SNP2 + SNP1*SNP2 and the reduced model was: SNP1 + SNP2 for all of the pairwise sets of SNPs generated by Biofilter with an implication index of 6. For the GxE (SNPxE) models, the same methods were employed using PLATO; however the full model was: SNP1 + ENV1 + SNP1*ENV1 and the reduced model was: SNP1 + ENV1 for all the possible unique SNPxE pairs, from the set of 527,936 SNPs and the PhenX variables described earlier in methods. Again, the outcome was case control status for cataracts. The GGPlot2 [21] package in R was used for Figure 2.

3. Results

3.1. GxE Results

Figure 2 shows a Manhattan plot of the association results for the PhenX GxE models that had interaction with LRT p-values $\leq 1 \times 10^{-4}$, a total of 782 models exhibited an interaction with a p-value $\leq 1 \times 10^{-4}$ associated with cataract status. The top five GxE interaction results for each PhenX measure are also presented in Table 4, sorted by chromosome to highlight results similar across SNPs and regions for multiple PhenX measures. The measurement “Weekend Sun Hours Last Decade” a survey question asking “On a typical weekend day in the summer, about how many hours did you generally spend in the mid-day sun in the past ten years?” with the SNP rs6447541, located in an intron of *GABR1* on chromosome 4, with an association LRT p-value of 2.35×10^{-8} , was the most significant interaction found when compared to the other 12 PhenX measurements we used in our GxE analysis.

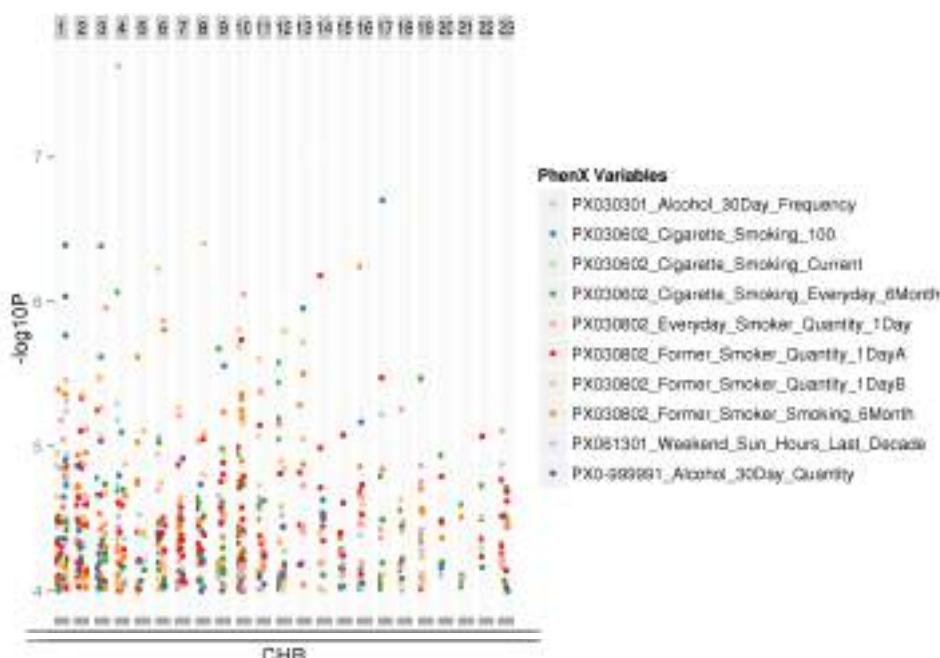


Figure 2. Manhattan plot of the association results for the GxE interaction models. Displayed are the results for the 12 PhenX measures that had interaction p-values $< 1 \times 10^{-4}$. Two PhenX variables did not have an interaction p-value less than 1×10^{-4} .

Table 4. Five most significant results for each PhenX measurement, sorted by chromosome and position

RSID	PhenX Variable	Chr:BP	P-value	Gene
rs7529518	PX030602_Cigarette_Smoking_100	1:200718422	1.71x10 ⁻⁶	CAMSAP2
rs2292097	PX030602_Cigarette_Smoking_100	1:200843768	9.23x10 ⁻⁷	GPR25*
rs10800745	PX030602_Cigarette_Smoking_100	1:200849676	4.06x10 ⁻⁷	GPR25*
rs11117581	PX061301_Weekend_Sun_Hours_Last_Decade	1:216613761	5.70x10 ⁻⁶	USH2A*
rs11117582	PX061301_Weekend_Sun_Hours_Last_Decade	1:216622208	3.48x10 ⁻⁶	USH2A*
rs10495409	PX061301_Weekend_Sun_Hours_Last_Decade	1:238255679	4.84x10 ⁻⁶	MTND5P18
rs607949	PX030602_Cigarette_Smoking_Current	1:43695708	1.12x10 ⁻⁵	WDR65
rs581503	PX030802_Former_Smoker_Smoking_6Month	1:61329593	4.03x10 ⁻⁶	NFIA*
rs11803470	PX030301_Alcohol_30Day_Frequency	1:95117783	1.42x10 ⁻⁵	ABCD3*
rs2587695	PX030802_Former_Smoker_Quantity_1DayA	2:120321817	4.62x10 ⁻⁶	PCDPI
rs262302	PX030301_Alcohol_30Day_Frequency	2:180931461	1.34x10 ⁻⁵	CWC22*
rs5994737	PX030602_Cigarette_Smoking_Current	22:33804804	3.07x10 ⁻⁵	LARGE
rs3846094	PX030602_Cigarette_Smoking_Everyday_6Month	3:101601159	2.40x10 ⁻⁶	NFKBIZ *
rs11720478	PX030602_Cigarette_Smoking_Everyday_6Month	3:101637806	4.10x10 ⁻⁷	NFKBIZ*
rs12495970	PX030802_Everyday_Smoker_Quantity_1Day	3:194928138	1.10x10 ⁻⁶	XXYLT1
rs11735349	PX030602_Cigarette_Smoking_Everyday_6Month	4:16506826	8.61x10 ⁻⁷	LDB2
rs157606	PX030301_Alcohol_30Day_Frequency	4:16699795	5.05x10 ⁻⁶	LDB2
rs283018	PX030301_Alcohol_30Day_Frequency	4:16740168	6.61x10 ⁻⁶	LDB2
rs6447541	PX061301_Weekend_Sun_Hours_Last_Decade	4:47215939	2.35x10 ⁻⁸	GABRB1
rs16888770	PX030802_Former_Smoker_Smoking_6Month	5:21586180	2.41x10 ⁻⁶	GUSBP1
rs13183503	PX030602_Cigarette_Smoking_Current	5:81515885	4.04x10 ⁻⁵	ATG10
rs9376419	PX030802_Everyday_Smoker_Quantity_1Day	6:139801295	1.36x10 ⁻⁶	TXLNB
rs3798756	PX030802_Former_Smoker_Smoking_6Month	6:152529260	1.56x10 ⁻⁶	SYNE1
rs3094549	PX030802_Former_Smoker_Quantity_1DayB	6:29355148	5.91x10 ⁻⁷	ORI2D2*
rs4712006	PX061301_Weekend_Sun_Hours_Last_Decade	6:52245415	8.61x10 ⁻⁶	PAQR8
rs3889488	PX030802_Former_Smoker_Quantity_1DayB	8:141544748	3.96x10 ⁻⁷	AGO2
rs6987670	PX030602_Cigarette_Smoking_Current	8:9883177	3.18x10 ⁻⁵	MSRA*
rs10968388	PX030602_Cigarette_Smoking_Everyday_6Month	9:28210699	2.11x10 ⁻⁶	LINGO2
rs9783135	PX030802_Everyday_Smoker_Quantity_1Day	10:129937722	8.90x10 ⁻⁷	MKI67*
rs12360020	PX030802_Former_Smoker_Quantity_1DayB	10:15264322	1.57x10 ⁻⁶	FAM171A1
rs2820100	PX030802_Former_Smoker_Quantity_1DayA	10:84491173	1.84x10 ⁻⁶	NRG3
rs6592528	PX030802_Everyday_Smoker_Quantity_1Day	11:73377350	4.22x10 ⁻⁶	PLEKHBI*
rs4944859	PX030802_Everyday_Smoker_Quantity_1Day	11:73424135	4.22x10 ⁻⁶	RAB6A
rs7977795	PX030802_Former_Smoker_Quantity_1DayB	12:132096632	1.59x10 ⁻⁶	SFSWAP*
rs7972947	PX030602_Cigarette_Smoking_Everyday_6Month	12:2170433	2.64x10 ⁻⁶	CACNA1C
rs775474	PX030602_Cigarette_Smoking_Current	12:70075933	2.60x10 ⁻⁵	BEST3
rs680711	PX030602_Cigarette_Smoking_100	13:101814804	1.11x10 ⁻⁶	NALCN
rs4772995	PX030802_Former_Smoker_Smoking_6Month	13:109410933	3.15x10 ⁻⁶	MYO16
rs7983958	PX030802_Former_Smoker_Quantity_1DayB	13:96473682	1.90x10 ⁻⁶	UGGT2
rs1957480	PX030802_Former_Smoker_Quantity_1DayA	14:44397890	6.59x10 ⁻⁷	X10IF4BP1*

rs11644531	PX030802_Formal_Smoker_Smoking_6Month	16:6008824	5.72×10^{-7}	<i>RBFOX1</i> *
rs8075882	PX030301_Alcohol_30Day_Frequency	17:55469362	6.01×10^{-6}	<i>MSI2</i>
rs1443269	PX030802_Formal_Smoker_Quantity_1DayA	17:55894564	3.35×10^{-6}	<i>MRPS23</i> *
rs9911607	PX030802_Formal_Smoker_Quantity_1DayA	17:55895539	3.35×10^{-6}	<i>MRPS23</i> *
rs7210514	PX030602_Cigarette_Smoking_100	17:67793814	1.99×10^{-7}	<i>KCNJ16</i> *

Table Abbreviations: Chr = Chromosome; BP = Base pair location of SNP; RSID = SNP ID; P-value = P-value of the interaction; Gene = Gene symbol of gene is within or nearest to (*indicates nearest gene is listed)

3.2. GxG Results

The top Biofilter 1.0 derived GxG models are presented in Table 5. A total of 13 models had an LRT p-value $< 1 \times 10^{-4}$ and full model p-value < 0.01 . A total of 9 genes were in the thirteen models. Of these models, the most significant was for a model with *SOS1*, which encodes a guanine nucleotide exchange factor for RAS proteins, and *FYN*, which is a member of the protein-tyrosine kinase oncogene family.

Table 5. The 13 SNPxSNP models with an interaction p-value $< 1 \times 10^{-4}$ after association testing of the Biofilter derived pairwise models.

SNP1	Gene 1	SNP2	Gene 2	Interaction P-value
rs2888586	<i>SOS1</i>	rs706885	<i>FYN</i>	1.29×10^{-6}
rs2888586	<i>SOS1</i>	rs17072912	<i>FYN</i>	2.14×10^{-6}
rs2888586	<i>SOS1</i>	rs11964650	<i>FYN</i>	2.97×10^{-6}
rs2888586	<i>SOS1</i>	rs9372313	<i>FYN</i>	6.32×10^{-6}
rs17446875	<i>CDH2</i>	rs6121791	<i>CDH4</i>	2.64×10^{-5}
rs9384805	<i>FYN</i>	rs11017910	<i>DOCK1</i>	2.67×10^{-5}
rs11083252	<i>CDH2</i>	rs6121791	<i>CDH4</i>	4.39×10^{-5}
rs13135792	<i>KIT</i>	rs10515074	<i>PIK3RI</i>	4.74×10^{-5}
rs631428	<i>COL4A1</i>	rs3803231	<i>COL4A2</i>	6.67×10^{-5}
rs613116	<i>COL4A1</i>	rs3803231	<i>COL4A2</i>	6.99×10^{-5}
rs17704348	<i>FYN</i>	rs4751282	<i>DOCK1</i>	8.85×10^{-5}
rs17446875	<i>CDH2</i>	rs1110359	<i>CDH4</i>	8.85×10^{-5}
rs809193	<i>FYN</i>	rs11594969	<i>DOCK1</i>	9.64×10^{-5}

4. Discussion

The results presented herein are an exploration of the use of multiple novel approaches for investigating gene and phenotype associations within EHR based data. We performed an analysis with PhenX derived measures, seeking GxE interaction models for the Marshfield Cataract data set. The majority of the significant interactions were found for smoking related measures. We did find some highly correlated PhenX measures with significant interactions for SNPs within similar regions, such as the results on chromosome 1 for SNPs rs2292097 and rs7529518, for smoking related phenotypes. Through searches in the NCBI catalog [22], as well as the National Center for Biotechnology (NCBI) dbSNP [19], these two SNPs, as well the SNP in our most significant GxE model, did not show previous GWA level significant associations for any phenotypes.

We also performed an exploratory analysis with Biofilter 1.0, an updated and improved implementation of the originally published Biofilter. The results are intriguing, and provide the basis for hypotheses that can be investigated further, highlighting how Biofilter results have a biological context that provide additional information for resulting models. Interestingly, three of the models that passed our significance cutoff contained two of the same genes, *FYN*, a member of

the protein-tyrosine kinase oncogene family implicated in cell growth, and *DOCK1*, dedicator of cytokinesis 1. These models as a whole implicate genes related to cell growth, the cell cycle, and epidermal growth.

We are currently developing Biofilter 2.0 which will include additional database sources and allow for the use of other position and region based information beyond SNPs and genes, such as copy number variation (CNV) data, evolutionary conserved regions, and regulatory regions, allowing users to incorporate additional sources of prior knowledge as well as utilize other sources of genetic variation measurement data, with a more user-friendly interface.

Our results provide more complex models for an association between genetic variation and cataract outcome, moving beyond the more standard SNP-phenotype associations. The models found we intend to investigate further and warrant additional investigation of the environment and genetic variables contributing to these more complex models. These bioinformatics approaches can be used with other datasets, to expand the investigation of the relationship between genetic architecture and phenotypic outcome. With these approaches that consider the complexity of the data and harness the power of novel bioinformatics tools, we will elucidate the missing heritability of complex traits.

Acknowledgments

This work was supported by the following grants: U19 HL0659625, R01 LM010040, U01 HG006389

References

1. Peissig PL, Rasmussen LV, Berg RL, Linneman JG, McCarty CA, et al. (2012) Importance of multi-modal approaches to effectively identify cataract cases from electronic health records. Journal of the American Medical Informatics Association : JAMIA 19: 225-234.
2. Waudby CJ, Berg RL, Linneman JG, Rasmussen LV, Peissig PL, et al. (2011) Cataract research using electronic health records. BMC Ophthalmol 11: 32.
3. McCarty CA, Wilke RA, Giampietro PF, Wesbrook SD, Caldwell MD (2005) Marshfield Clinic Personalized Medicine Research Project (PMRP): design, methods and recruitment for a large population-based biobank. Personalized Medicine 2: 49-79.
4. Pathak J, Pan H, Wang J, Kashyap S, Schad PA, et al. (2011) Evaluating Phenotypic Data Elements for Genetics and Epidemiological Research: Experiences from the eMERGE and PhenX Network Projects. AMIA Summits Transl Sci Proc 2011: 41-45.
5. Michael R, Bron AJ (2011) The ageing lens and cataract: a model of normal and pathological ageing. Philos Trans R Soc Lond B Biol Sci 366: 1278-1292.
6. Hammond CJ, Duncan DD, Snieder H, de Lange M, West SK, et al. (2001) The heritability of age-related cortical cataract: the twin eye study. Invest Ophthalmol Vis Sci 42: 601-605.

7. Bush WS, Dudek SM, Ritchie MD (2009) Biofilter: a knowledge-integration system for the multi-locus analysis of genome-wide association studies. *Pac Symp Biocomput*: 368-379.
8. Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, et al. (1999) KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res* 27: 29-34.
9. Matthews L, Gopinath G, Gillespie M, Caudy M, Croft D, et al. (2009) Reactome knowledgebase of human biological pathways and processes. *Nucleic Acids Res* 37: D619-622.
10. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, et al. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature genetics* 25: 25-29.
11. Sonnhammer EL, Eddy SR, Durbin R (1997) Pfam: a comprehensive database of protein domain families based on seed alignments. *Proteins* 28: 405-420.
12. Kandasamy K, Mohan SS, Raju R, Keerthikumar S, Kumar GS, et al. (2010) NetPath: a public resource of curated signal transduction pathways. *Genome Biol* 11: R3.
13. Stark C, Breitkreutz BJ, Chatr-Aryamontri A, Boucher L, Oughtred R, et al. (2011) The BioGRID Interaction Database: 2011 update. *Nucleic Acids Res* 39: D698-704.
14. Licata L, Brigand L, Peluso D, Perfetto L, Iannuccelli M, et al. (2012) MINT, the molecular interaction database: 2012 update. *Nucleic Acids Res* 40: D857-861.
15. Stover PJ, Harlan WR, Hammond JA, Hendershot T, Hamilton CM (2010) PhenX: a toolkit for interdisciplinary genetics research. *Curr Opin Lipidol* 21: 136-140.
16. Abraham AG, Condon NG, West Gower E (2006) The new epidemiology of cataract. *Ophthalmol Clin North Am* 19: 415-425.
17. Turner S, Armstrong LL, Bradford Y, Carlson CS, Crawford DC, et al. (2011) Quality control procedures for genome-wide association studies. *Curr Protoc Hum Genet Chapter 1: Unit1* 19.
18. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, et al. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *American journal of human genetics* 81: 559-575.
19. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, et al. (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res* 29: 308-311.
20. Grady BJ, Torstenson E, Dudek SM, Giles J, Sexton D, et al. (2010) Finding unique filter sets in plato: a precursor to efficient interaction analysis in gwas data. *Pac Symp Biocomput*: 315-326.
21. Wickham H (2009) *ggplot2: elegant graphics for data analysis*: Springer New York.
22. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, et al. (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences of the United States of America* 106: 9362-9367.