

BIOS 516 Introduction to Large-Scale Biomedical Data Analysis

Lecture 2

Steve Qin

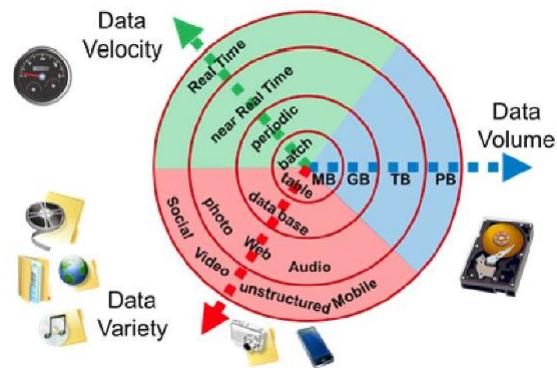
September 7, 2021

Background

- What is Big Data?
- What is the big deal?
- Where can I find biomedical Big Data?
- How can we take advantage of it?

BigData

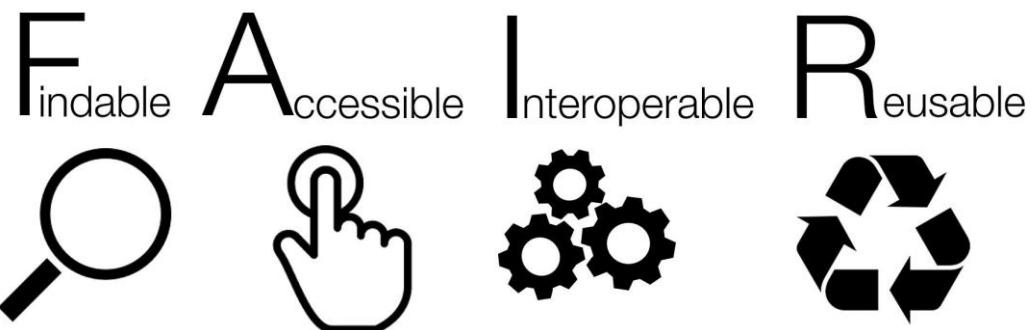
- Volume
- Variety
- Velocity



https://en.wikipedia.org/wiki/Big_data

By Ender005 - Own work, CC BY-SA 4.0, <https://commons.wikimedia.org/w/index.php?curid=49888192>

FAIR principals



By SangyaPundir - Own work, CC BY-SA 4.0, <https://commons.wikimedia.org/w/index.php?curid=53414062>

Where is biomedical BigData?

- GEO, ArrayExpress
- mSigDB
- ENCODE
- TCGA,
- GTEx
- PheGenI, GWAS catalog
- 1000 Genomes,
- UKBB
- ...

GEO, SRA, ArrayExpress and GSA

- Repositories
- For sharing high-throughput experimental data, often required by publishers.
 - Originally designed to share microarray data
- Data uploaded by members of the whole research community
- Capture and display rich metadata, and enables query of the metadata.
 - e.g., mouse brain, K562 cell lines.
- No quality control, honor system
- Totally free. No registration is required.
- Operate like a collection of supplemental data of papers.

Gene Expression Omnibus



GEO is a public functional genomics data repository supporting MIAME-compliant data submissions. Array- and sequence-based data are accepted. Tools are provided to help users query and download experiments and curated gene expression profiles.

Getting Started	Tools	Browse Content
Overview	Search for Studies at GEO DataSets	Repository Browser
FAQ	Search for Gene Expression at GEO Profiles	DataSets: 4348
About GEO DataSets	Search GEO Documentation	Series: 159662
About GEO Profiles	Analyze a Study with GEO2R	Platforms: 22535
About GEO2R Analysis	Studies with Genome Data Viewer Tracks	Samples: 4599547

MIAME and MINSEQE guidelines

[MIAME](#) (Minimum Information About a Microarray Experiment)

[MINSEQE](#) (Minimum Information About a Next-generation Sequencing Experiment)

- Raw data for each assay (e.g., CEL or FASTQ files)
- Final processed (normalized) data for the set of assays in the study (e.g., the gene expression data count matrix used to draw the conclusions in the study)
- Essential sample annotation (e.g., tissue, sex and age) and the experimental factors and their values (e.g., compound and dose in a dose response study)
- Experimental design including sample data relationships (e.g., which raw data file relates to which sample, which assays are technical, which are biological replicates)
- Sufficient annotation of the array or sequence features examines (e.g., gene identifiers, genomic coordinates)
- Essential laboratory and data processing protocols (e.g., what normalization method has been used to obtain the final processed data)

An example: <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM5366693>

Web portal of large consortia

- ENCODE, TCGA, GTEx, 1000 Genomes ...
- Only data produced by members of the consortia, follow a set of protocols.
- High quality data, often with extensive quality control.
- Publicly available, popular in the research community.
- Often used as benchmark data.

ENCODE Data Encyclopedia Materials & Methods Help Search... Sign in / Create account

ENCODE: Encyclopedia of DNA Elements

About ENCODE Project | **Getting Started** | **Experiments**

Search ENCODE portal ⓘ

ENCODE Q | **Functional Characterization Experiments**

About ENCODE Encyclopedia

candidate Cis-Regulatory Elements

Search for candidate Cis-Regulatory Elements ⓘ

Hosted by [SCREEN](#)

	TF ChIP-Seq	Histone ChIP-Seq	Dnase-seq	polyA+ RNA-seq	total RNA-seq	miRNA-seq	ATAC-seq	DNase array	eCLIP	small RNA-seq	WGS	RAMPAGE	long-read RNA-seq	RNA microarray	genotyping array	CAGE	microRNA counts	RepL-seq	RNA Bind-n-Seq	RRBS	ChIP-PET	Hi-C	
Human GRO18 Q	379	1934	772	397	283	181	198	121	2	67	162	104	88	2	7	17	101	17	8				
Mouse mm10 Q	42	91	9	20	3	7	7	1		1	9	2					7	1					
immune system areas 46	30	103	56																				
liver	5	79	25	16	17	7	9			1	10	2	4				8						
heart	8	38	24	11	20	8	21	5	2	2	5	4	19				2						
adrenal gland	17	72	22	15	5	6	4	3		4	10	5					4						
stomach																							
cell line	2452	665	171	143	109	24	120	26	87	223	110	18	29	33	67	73	50	8	92	50	81	39	
K562	644	19	4	15	13	2	2	3	120	7	1	1	3	2	9	2	9	1	6	1	11	1	
HepG2	667	15	2	11	6	2	2	3	103	3	2	2	6	2	6	1	6	1	6	2	4	1	
GM12878	187	15	1	13	5	2	2	3		6	1	1	3	7	2	6	1	6	2	2	7		
MCF-7	147	18	4	4	1	1	2	2		7								3	6				
HEK293	196	6					2										1	2		2	2		
primary cell	66	506	257	82	188	354	23	38	38	24	7	16	4	57	37	30	1	12	27	10	14		
T-cell	11	51	1	3	12	4											1						
macrophage	4		1	78																			
CD4-positive, alpha-beta memory T cell	19	1		2	44				2														
CD14-positive monocyte	1	21	7	2	28													1					
endothelial cell of umbilical vein	13	16	2	5	1			1									1	1	5	1	6	2	2
whole organisms	996		41	68													19						
whole organism	996		53	64													15						
carcass			8	4													4						
in vitro differentiated cells	46	229	23	22	18	71	10	15	5	11	14	6	10	2	4	1	4	1	2	2	9		
motor neuron	18					10											2						

The screenshot shows the GTEx Portal interface. At the top, there's a navigation bar with links for Home, Databases, Expression, Cells & Tissues, Design Data, and Documentation. Below the navigation is a banner with a blue background and white text. The main area has several sections:

- Resource Overview:** Shows "Current Release (v8)" with "View 4 Sample Identifiers" and "View Sample ID (microarray)".
- Browse:** Options include "By gene ID", "By variant or ID", "By Tissue", and "Morphology image viewer".
- Explore GTEx:** Options include "Browse and search data by gene", "Browse and search data by variant", "Browse and search data by tissue", and "Browse and search GTEx morphology images".
- Data Overview:** Includes "Learn more about available single cell data" and "Browse and search single cell expression by gene and tissue".
- Annotations:** A table with columns for Description, Name, and Size. It lists several files:
 - GTEX_Analysis_v8_Annotations_SampleAttributesDS.txt (16K)
 - GTEX_Analysis_v8_Annotations_SubjectPhenotypesDS.txt (11K)
 - GTEX_Analysis_v8_Annotations_SampleAttributesDS.txt (11M)
 - GTEX_Analysis_v8_Annotations_SubjectPhenotypesDS.txt (20K)
- RNA-Seq Data:** A table with columns for Description, Name, and Size. It lists several files:
 - Gene read counts (876M)
 - Gene TPMs (1.6G)
 - Median gene-level TPM by tissue. Median expression was calculated from the file GTEX_Analysis_2017-06-05_v8_RNASeQCv1.1.9_gene_median_tpm.gct.gz (6.7M)

On the right side of the interface, there is a large scatter plot titled "n >= 70". The x-axis is labeled "Number of Samples" and ranges from 100 to 800. The y-axis represents the number of genes. The plot contains numerous colored dots representing different genes, with a legend indicating "eGenes" (blue circles) and "sGenes" (black diamonds). Most points are clustered between 100 and 600 samples, with a few outliers extending towards 700 samples.

Specialty databases

- mSigDB, GWAS catalog
- Data often not produced by the database owner.
- Designed for secondary or tertiary analyses.
- Data parsed, collected, and often processed and QCed.
- The database just serve as an access point for the collection.
- Often easy to query and free to download.

GSEA Gene Set Enrichment Analysis

- Home
- Downloads
- Molecular Signatures Database
- Documentation
- Contact
- Terms

MSigDB Molecular Signatures Database

Molecular Signatures Database v7.4

Overview

The Molecular Signatures Database (MSigDB) is a collection of annotated gene sets for use with GSEA software. From this web site, you can:

- Search for gene sets by keyword.
- Search for gene sets by name or collection.
- Examine a gene set and its annotations. See, for example, the HALLMARK_APOPTOSIS gene set page.
- Download gene sets.
- Investigate gene sets:
 - Compute overlaps between your gene set and gene sets in MSigDB.
 - Categorize members of a gene set by gene families.
 - View the expression profile of a gene set in a provided public expression compendia.
 - Investigate the gene set in the online biological network repository NCDE.

Licence Terms

GSEA and MSigDB are available for use under these license terms.

Please register to download the GSEA software and the MSigDB gene sets, and to use our web tools. After registering, you can log in at any time using your email address. Registration is free. Its only purpose is to help us track usage for reports to our funding agencies.

Current Version

MSigDB database v7.4 updated April 2021. Release notes.

Citing the MSigDB

To cite your use of the Molecular Signatures Database

GWAS Catalog

The NHGRI-EBI Catalog of human genome-wide association studies

Search the catalog Examples: breast carcinoma, rs7329174, Yair, Zg37.1, HBS1L, E:16000000-25000000

Collections

Download

Download a full copy of the GWAS Catalog in spreadsheet format as well as current and older versions of the GWAS diagram in PDF format.

Summary statistics

Documentation and access to full summary statistics for GWAS Catalog studies where available.

Submit

Submit summary statistics to GWAS Catalog.

Documentation

Including FAQs, our curation process, training materials, related resources, a list of abbreviations and API documentation

Diagram

Explore an interactive visualization of all SNP-trait associations with genome-wide significance (p < 5e-8)

Ancestry

An introduction to our ancestry curation process



Various ways to use these data

- Construct informative prior for Bayesian inference
- Build null distributions
- Features to be used in ML algorithms
- Develop supervised ML models
 - As positive training data
 - As negative training data
- Mine novel biological knowledge

Use Big Data to construct informative prior and null distribution

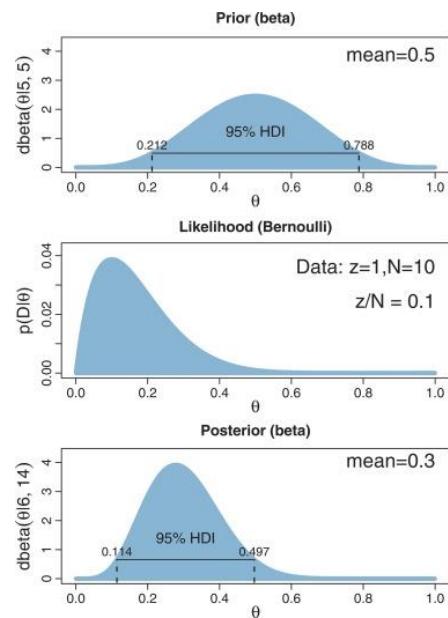
Bayesian inference

Bayes' Theorem

- Bayes' theorem is most commonly used to estimate the state of a hidden, causal variable H based on the measured state of an observable variable D :

$$p(H|D) = \frac{p(D|H)p(H)}{p(D)}$$

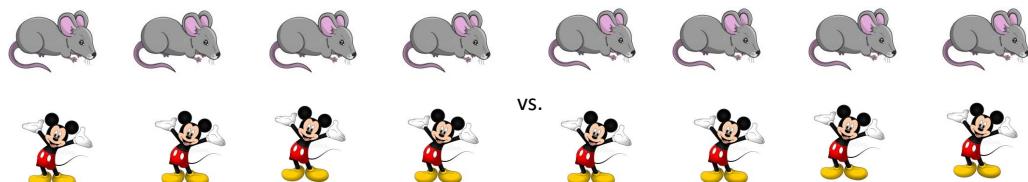
Likelihood Prior
Posterior Evidence



Detection of DE genes

- A classical problem in gene expression microarray study: detect differentially expressed (DE) genes.
- DE genes: genes from various samples are expressed differentially in different cell types, tissues, developmental stages or diseases.
- Typically the number of replicates is rather low.

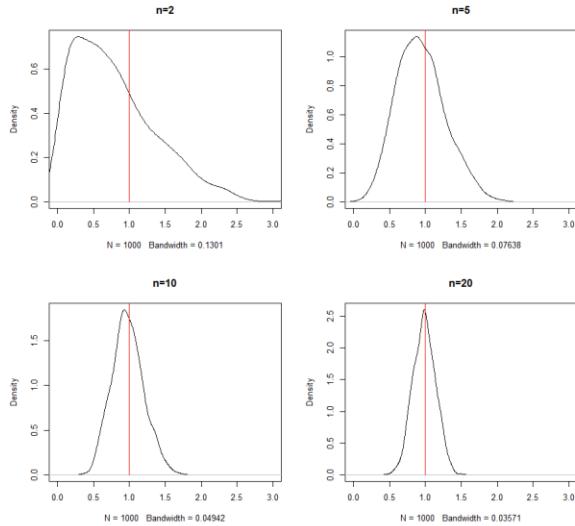
We wish:



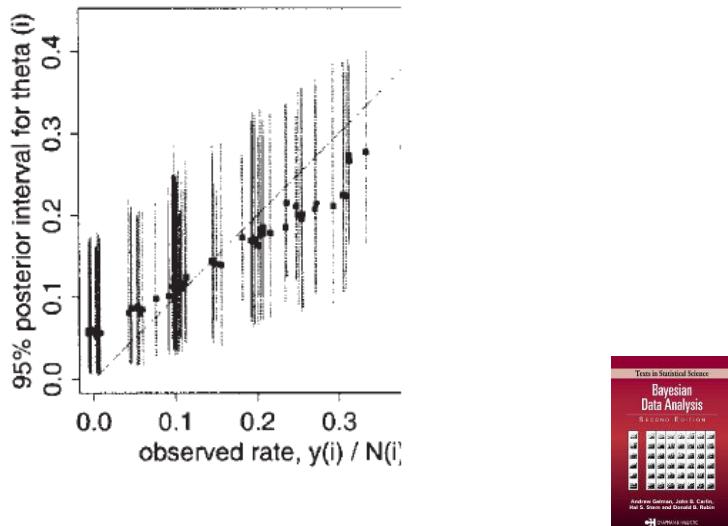
But in reality, we often have:



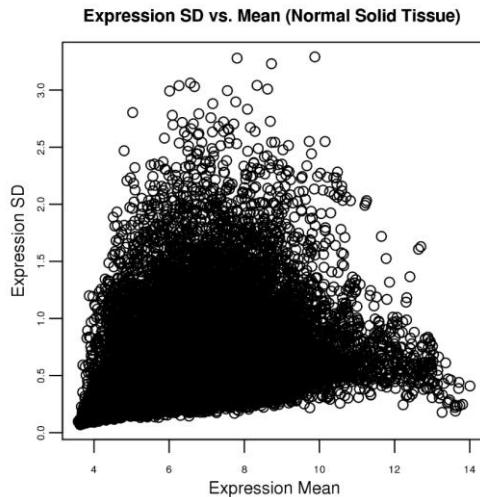
The problem



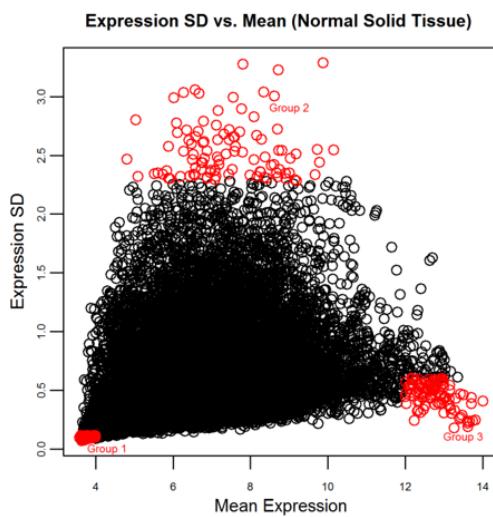
How does hierarchical model work



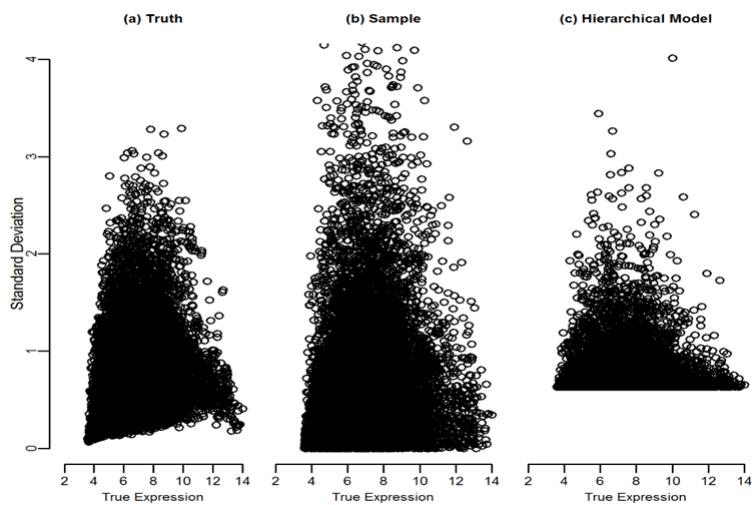
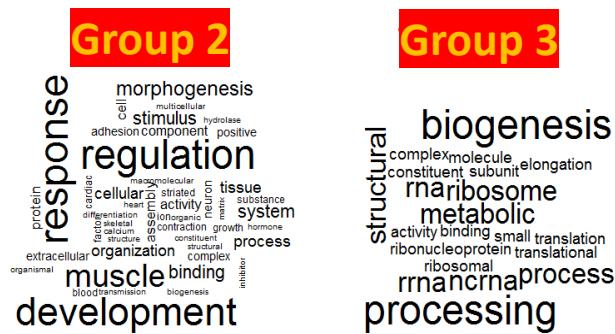
Std dev vs mean



Std dev vs mean



Diverse functions



Drawbacks of hierarchical models

- Restrict to current dataset.
- May overcorrect, especially at the low end.
- Inflated variance means much less discovery power—conservative.

Public databases



- 4,600,533 samples
- 159,703 series



- 74,706 experiments
- 2,557,032 assays

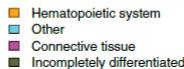
A microarray compendium

CORRESPONDENCE

A global map of human gene expression

To the Editor:

Although there is only one human genome sequence, different genes are expressed





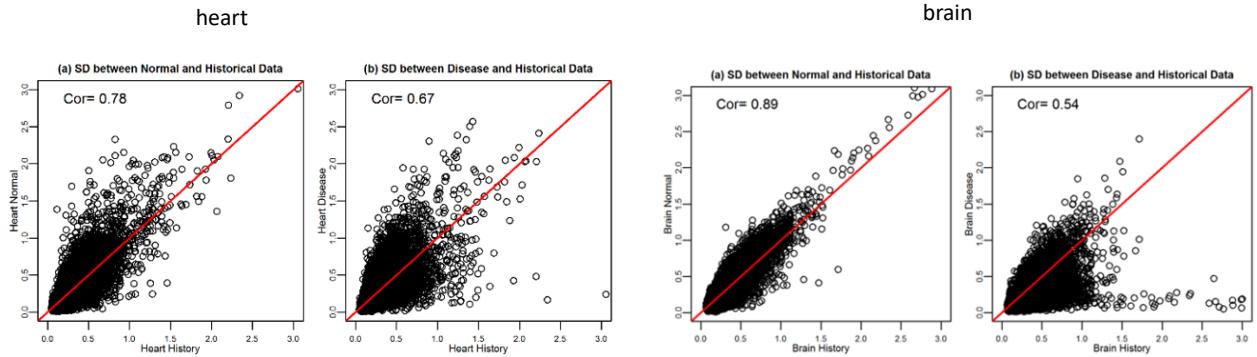
- 5,372 samples
- 206 different studies
- From 163 different labs
- Normalized

Lukk et al. 2010.

The global gene expression map

4 meta groups		15 groups	
Group	# of samples	Group	# of samples
cell line	1259	blood neoplasm cell line	166
		non neoplastic cell line	262
		solid tissue neoplasm cell line	831
disease	765	blood non neoplastic disease	388
		solid tissue non neoplastic disease	377
neoplasm	2315	breast cancer	672
		germ cell neoplasm	71
		leukemia	567
		nervous system neoplasm	112
		non breast carcinoma	288
		non leukemic blood neoplasm	334
		other neoplasm	167
		sarcoma	104
normal	1033	normal blood	467
		normal solid tissue	566

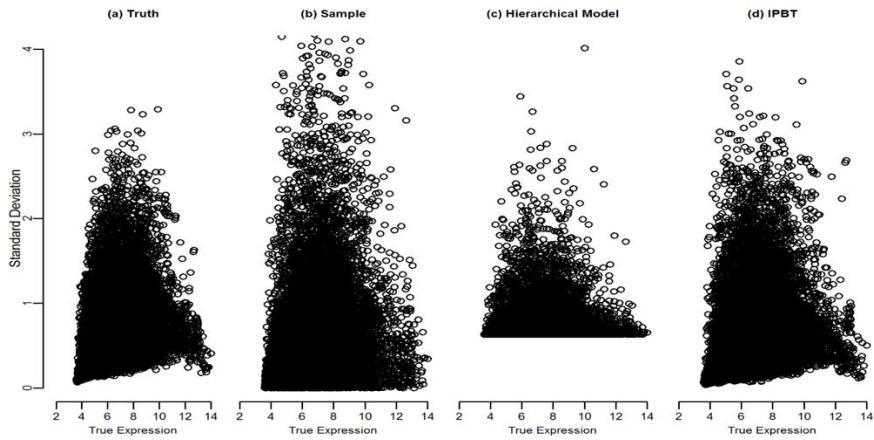
Standard deviations from different studies



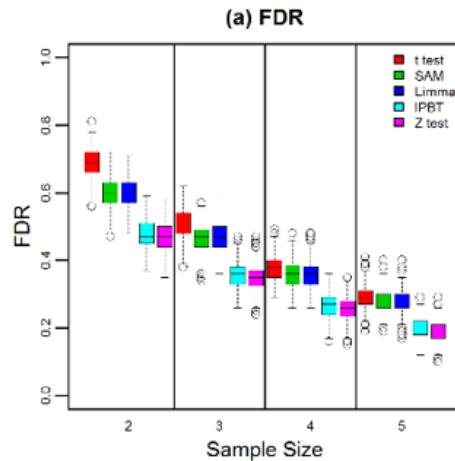
Informative Prior Bayesian Test (IPBT)

- Use historical data to build gene-specific, informative priors.
- Conduct Bayesian inference on σ_i , the standard deviation of gene i .
- Either calculate a Bayes factor or test statistics of an adjusted t -test and rank genes based on that.

Compare variance estimates



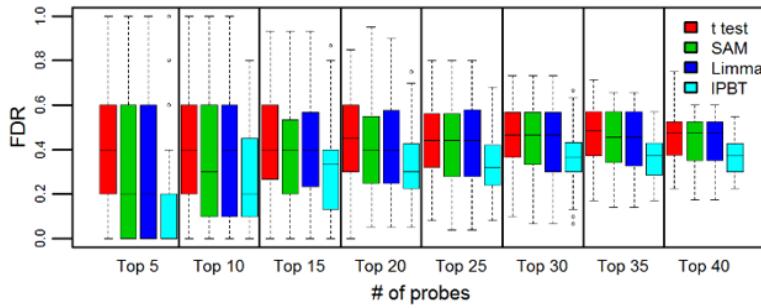
FDR boxplot



Spike-in experiment

- FDR when first k declare significant

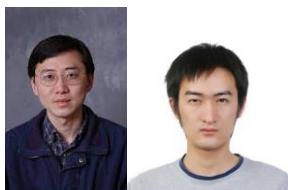
FDR for Spike-in Experiment



Summary

- Gene-specific properties such as variance can be captured by exploiting existing data that are public-available.
- Utilizing historical data in detecting differentially expressed genes is a better alternative than classical hierarchical model.
- Using informative prior can overcome difficulties faced in low-sample size inference problems.
- It is possible to reduce the number of replicates.

Bioinformatics, 2015, 1–4
doi:10.1093/bioinformatics/btv316
Advance Access Publication Date: 30 October 2015
Original Paper



Gene expression

Bayesian inference with historical data-based informative priors improves detection of differentially expressed genes

Ben Li¹, Zhaonan Sun², Qing He¹, Yu Zhu^{2,*} and Zhaohui S. Qin^{1,3,*}

¹Department of Biostatistics and Bioinformatics, Rollins School of Public Health, Emory University, Atlanta, GA 30322, USA; ²Department of Statistics, Purdue University, West Lafayette, IN 47906, USA and ³Department of Biomedical Informatics, Emory University School of Medicine, Atlanta, GA 30322, USA

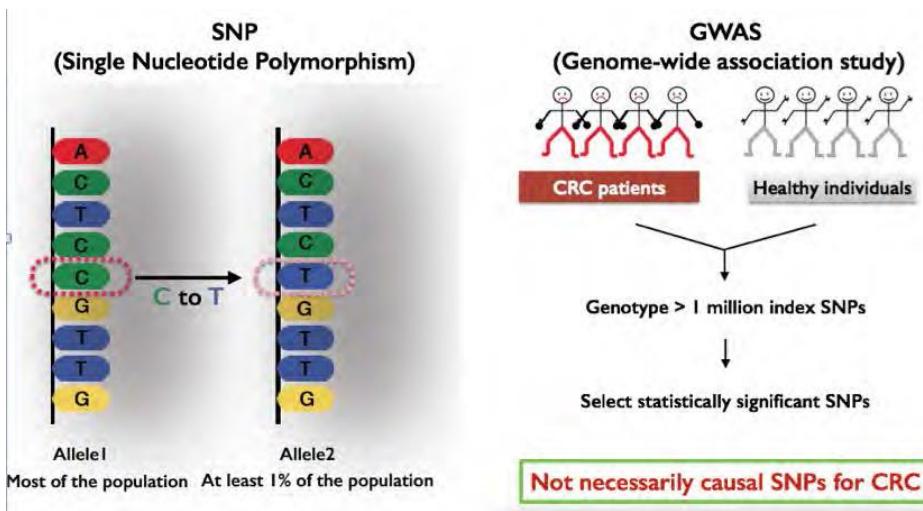
*To whom correspondence should be addressed.

Associate Editor: Jonathan Wren

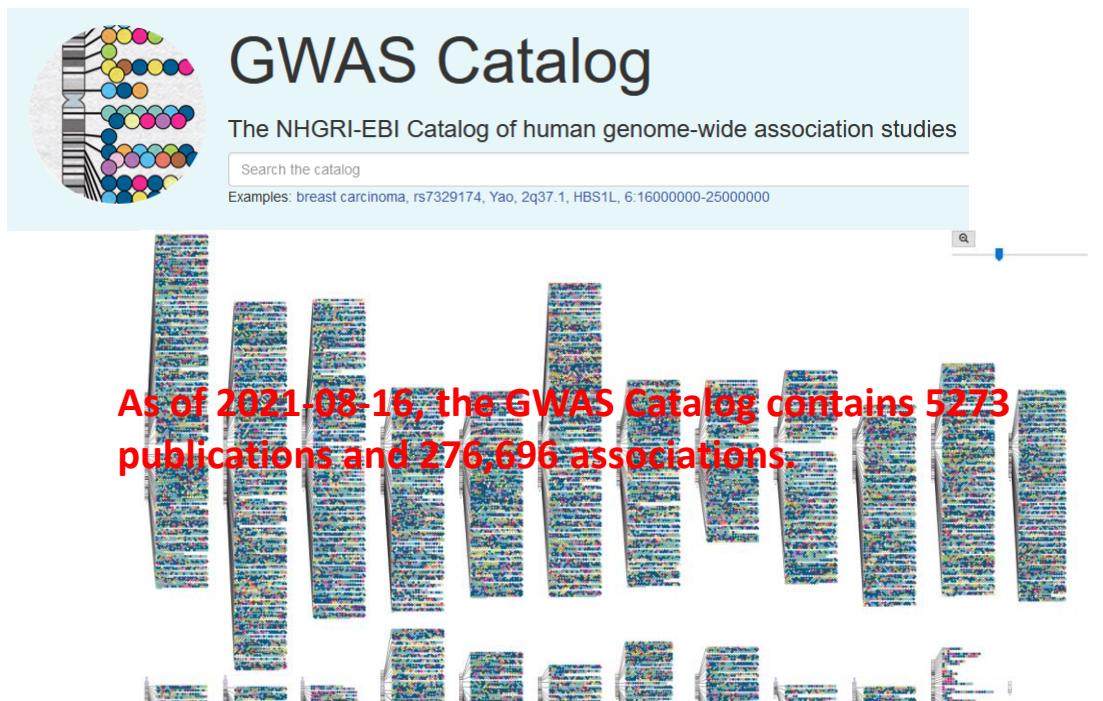
Received on April 15, 2015; revised on September 28, 2015; accepted on October 26, 2015

Use Big Data to annotate different parts of the genome

Genome-wide Association Studies

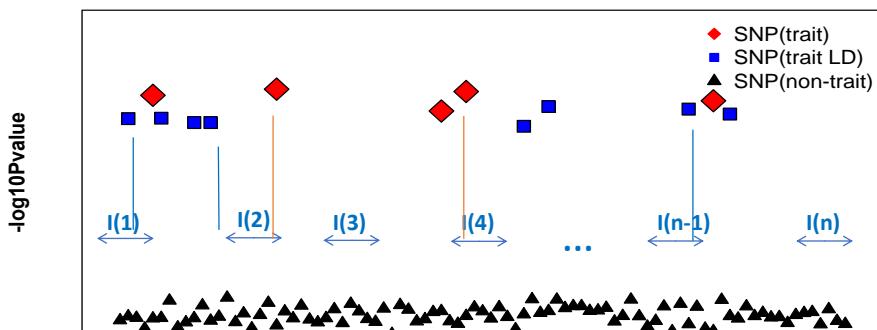


From Peggy J. Farnham.



traseR: trait-associated SNP enrichment analysis

- The goal is to link GWAS SNPs to genomic loci to uncover potential connections



Example query result (H3K4me1 peaks in T cell)

Table 1. Top-ranked traits for peripheral T cell H3K4me1 peaks

Trait	P value	OR	#taSNP hits	#taSNP
All	2.7e-48	1.5	1846	30 553
Behcet syndrome	4.4e-23	6.3	59	274
Diabetes mellitus, type 1	1.7e-11	5.0	33	185
Lupus erythematosus	6.2e-09	3.9	32	223
Arthritis, rheumatoid	1.4e-07	5.1	20	112
Multiple sclerosis	1.6e-05	2.9	26	236
Autoimmune diseases	5.2e-05	15.9	6	15

	inside	outside
#SNP(trait LD)	87	326
#SNP(non-trait)	165,441	3,812,459

traseR: trait-associated SNPs

- Easy-to-use bioinformatics tools that is capable of uncovering potential connections between genomic loci and complex diseases through known GWAS variants.
- Provide annotation to interesting genomic loci found by experiments.

Bioinformatics, 2016, 1–3

doi: 10.1093/bioinformatics/btv741

Advance Access Publication Date: 18 December 2015

Applications Note



Genome analysis

traseR: an R package for performing trait-associated SNP enrichment analysis in genomic intervals

Li Chen¹ and Zhaohui S. Qin^{2,3*}

¹Department of Mathematics and Computer Science, ²Department of Biostatistics and Bioinformatics, Rollins School of Public Health, Emory University, Atlanta, GA 30322 USA and ³Department of Biomedical Informatics, Emory University School of Medicine, Atlanta, GA 30322, USA

*To whom correspondence should be addressed.

Associate editor: John Hancock

Received on 16 October 2015; revised on 27 November 2015; accepted on 12 December 2015

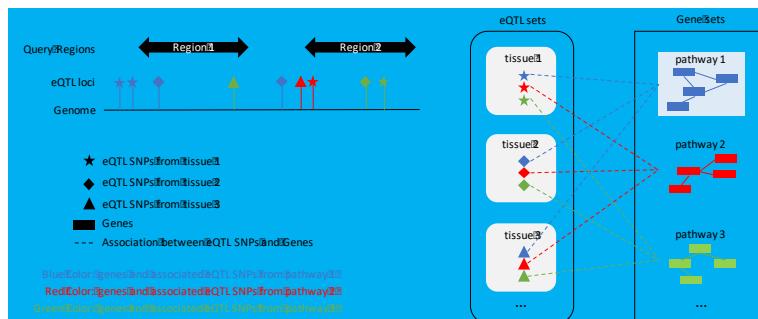


Loci2path

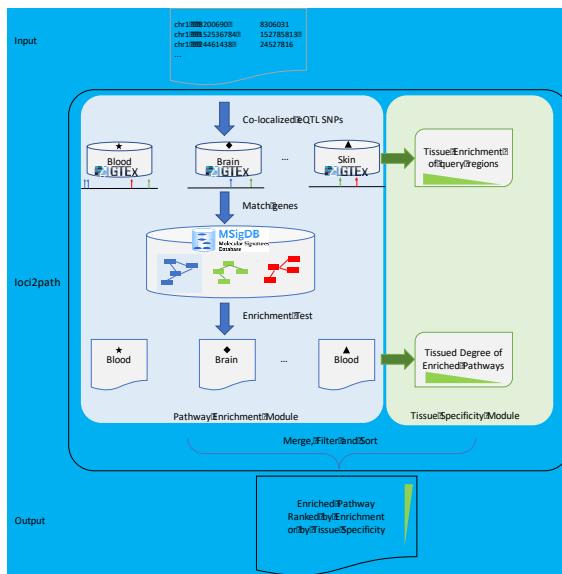
- Annotating a given genomic locus or a set of genomic loci for the non-coding part of the genome.
- Takes advantage of the newly emerged, genome-wide and tissue-specific expression quantitative trait loci (eQTL) information to help annotate a set of genomic intervals in terms of transcription regulation.
- key advantages
 - no longer rely on proximity to link a locus to a gene which has shown to be unreliable;
 - provide the regulatory annotation under the context of specific tissue types which is important.



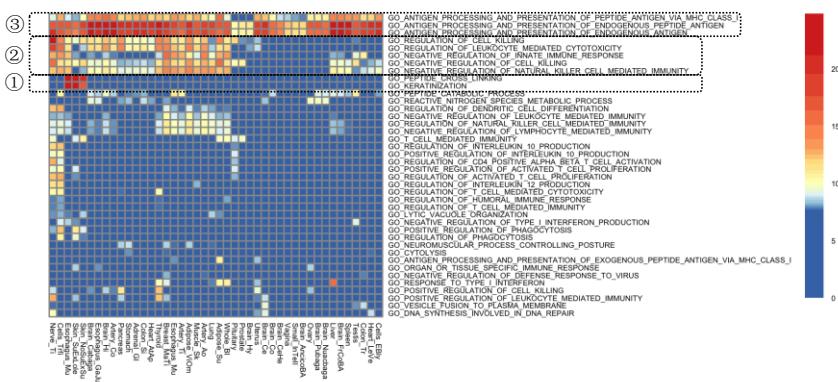
How does it work?



Workflow



Result - Psoriasis



Summary

- Annotate non-coding variants using eQTL resources
- Tissue specificity provide additional information
- Tissue degree suggest different categories of pathways involved in the pathogenesis of psoriasis
- Enrichment in pathways across immune diseases reveals common gene sets of shared disease risks

Bioinformatics, 36(3), 2020, 690–697
doi: 10.1093/bioinformatics/btz669
Advance Access Publication Date: 27 August 2019
Original Paper



Genome analysis
Regulatory annotation of genomic intervals based on tissue-specific expression QTLs

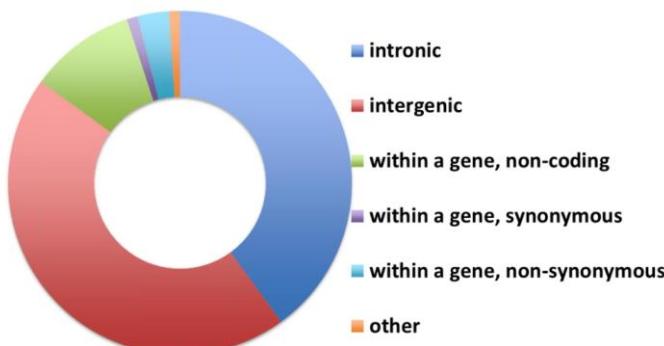
Tianlei Xu ¹, Peng Jin ² and Zhaojun S. Qin ^{3,*}

Use Big Data as features and training data in ML

Goal: annotate non-coding variants

- Many computational tools developed for coding variants
 - SIFT
 - PolyPhen
- Method is needed to annotate the majority (90%) of GWAS-identified variants of complex diseases which are non-coding

GWAS SNP catalog: Genomic location of SNPs



Adapted from Freedman et al. nature genetics, 2011

ational Human Genome Research Institute

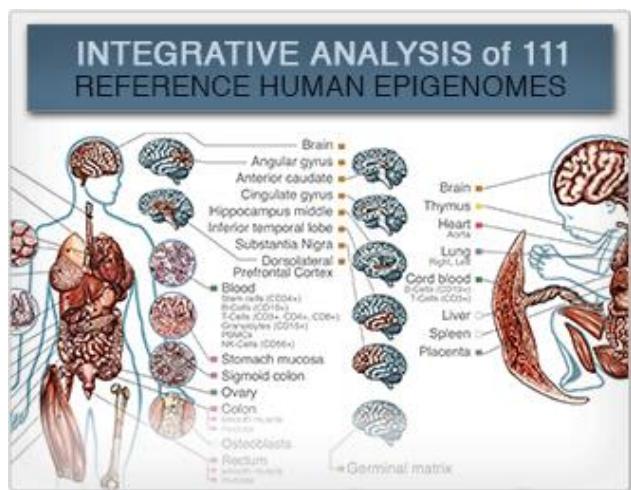
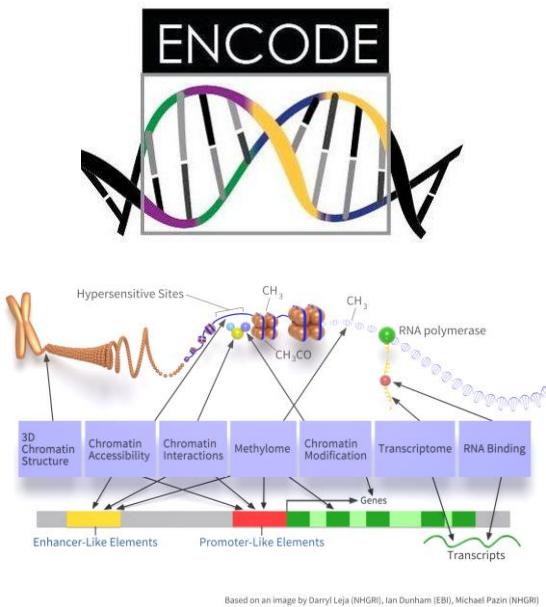
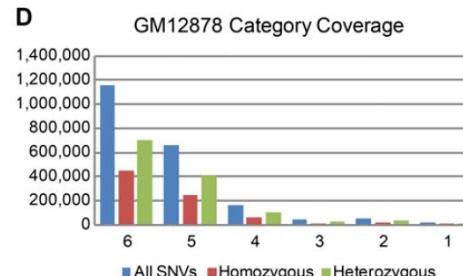
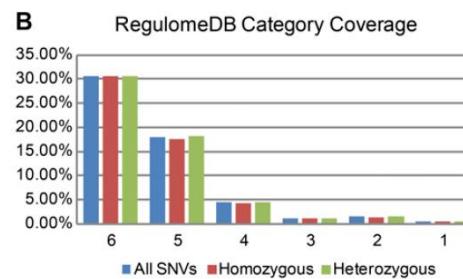


Table 2. RegulomeDB variant classification scheme

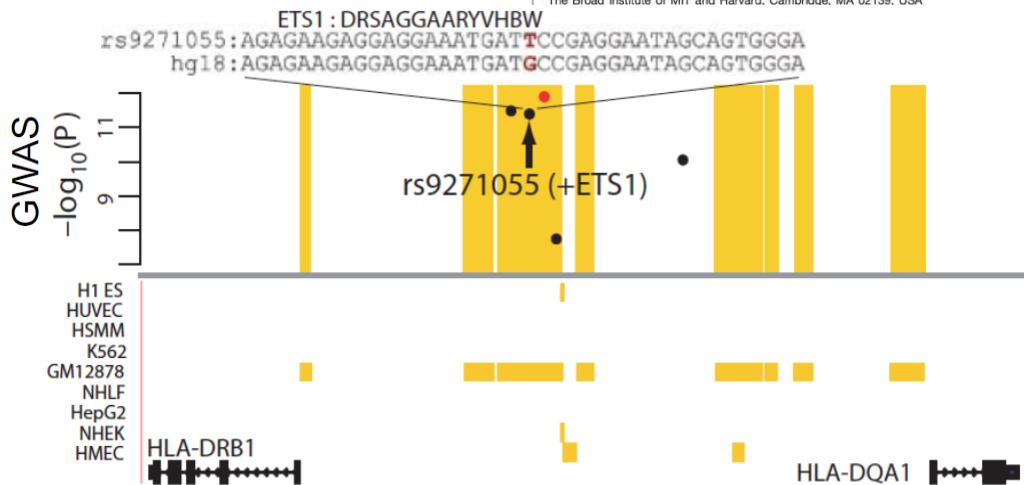
Category scheme	
Category	Description
1a	Likely to affect binding and linked to expression of a gene target
1b	eQTL + TF binding + matched TF motif + matched DNase footprint + DNase peak
1c	eQTL + TF binding + any motif + DNase footprint + DNase peak
1d	eQTL + TF binding + matched TF motif + DNase peak
1e	eQTL + TF binding + any motif + DNase peak
1f	eQTL + TF binding + matched TF motif
	0.55%
2a	Likely to affect binding
2b	TF binding + matched TF motif + matched DNase footprint + DNase peak
2c	TF binding + any motif + DNase footprint + DNase peak
	1.48%
3a	TF binding + matched TF motif
3b	Less likely to affect binding
	1.16%
4	TF binding + any motif + DNase peak
5	TF binding + DNase peak
6	Motif hit



Boyle et al. 2012



HaploReg v4.1



Existing methods for annotating noncoding variants



CADD
nature genetics
TECHNICAL REPORTS

A general framework for estimating the relative pathogenicity of human genetic variants

Martin Kircher^{1,3}, Daniela M Witten^{1,3}, Preethi Jain^{1,4}, Brian J O’Roak^{1,5}, Gregory M Cooper¹ & Jay Shendre¹



Eigen/EigenPC
nature genetics
TECHNICAL REPORTS

A spectral approach integrating functional genomic annotations for coding and noncoding variants

Juliana Ionita-Laza^{1,4}, Kenneth McCallum^{1,4}, Bin Xu² & Joseph D Bushman^{3,5,7}



PAFA
Genome Medicine

Priority and functional assessment of noncoding variants associated with complex diseases

Lin Zhou^{1,2} and Fangting Zhao^{1,2,3*}

GWAVA
nature genetics
BRIEF COMMUNICATIONS

Functional annotation of noncoding sequence variants

Graham R S Ritchie^{1,2}, Ian Dunham³, Delphine Lefèvre⁴ & Paul Black^{1,2}



GenoCanyon
SCIENTIFIC REPORTS

A Statistical Framework to Predict Functional Non-Coding Regions in the Human Genome Through Integrated Analysis of Annotation Data

Giongkittit Lai,¹ Vining He,¹ Zhihuan Sun,¹ Yousel Chengi,¹ Kai-Hao Cheung,^{1,2,3} & Hung-jen Zhuang¹



PINES
Genome Biology

PINES: phenotype-informed tissue weighting improves prediction of pathogenic noncoding variants

Comellu A, Bodda^{1,2,4}, Adele A, Mitchell¹, Alex Blomendrala⁴, Aaron G, Day Williams^{1,5}, Helio Rung^{1,6} and Sharmin R, Sunayav^{1,3,4*}



Information sources for identifying non-coding variants?

- Phylogenetic conservation
 - PhastCon scores
 - GERP scores
- Genomic profile
 - Whether it overlap with any known transcription factor binding motif?
- Epigenomic profile
 - TF binding
 - Histone modification
 - DNA methylation
 - ...

DIVAN: DIsease-specific Variant ANnotation

- A unique model for each disease/phenotype
 - 45 diseases from 12 categories.
- Using trait-associated SNPs identified by GWAS as training data
- Using genomic and epigenomic data as features
- Use machine learning techniques to distinguish risk variants from benign variants.

GWAS SNP collection (1)

The screenshot shows the NCBI GWAS Catalog search interface. The search query is set to "DIABETES MELLITUS". The results table displays 34 associations across various chromosomes, with the top hit being rs11585396 on chromosome 1 with a p-value of 6.19×10^{-6} . The columns include Source, Trait, SNP, p-value, Chr, Position, Gene Region, Context, and PubMed link.

Source	Trait	SNP	p-value	Chr	Position	Gene Region	Context	PubMed
dbGaP	Diabetes Mellitus	rs11585396	6.19×10^{-6}	1	50817360	C1orf87 NFIA	Intergenic	
dbGaP	Diabetes Mellitus	rs4915919	2.09×10^{-6}	1	64195043	PGM1 ROR1	Intergenic	
dbGaP	Diabetes Mellitus	rs17016501	6.03×10^{-6}	1	106438336	CDK4PS PRMT6	Intergenic	
dbGaP	Diabetes Mellitus	rs10495124	2.78×10^{-6}	1	219502193	LYPLAL1 ZC3H11B	Intergenic	
dbGaP	Diabetes Mellitus	rs17591522	3.42×10^{-6}	1	219533768	LYPLAL1 ZC3H11B	Intergenic	
dbGaP	Diabetes Mellitus	rs6704463	3.42×10^{-6}	1	219547825	LYPLAL1 ZC3H11B	Intergenic	
dbGaP	Diabetes Mellitus	rs1915260	8.64×10^{-6}	1	239109813	KRT18P32 RPL39P10	Intergenic	
dbGaP	Diabetes Mellitus	rs10495875	1.66×10^{-6}	2	38055097	CDC42EP3 FAM82A1	Intergenic	17903298
dbGaP	Diabetes Mellitus	rs10495875	5.68×10^{-6}	2	38055097	CDC42EP3 FAM82A1	Intergenic	17903298

GWAS SNP collection (2)

The screenshot shows the National Heart, Lung, and Blood Institute (NHLBI) GRASP (Genome-Wide Repository of Associations Between SNPs and Phenotypes) website. The main page features a search bar and navigation links for Public, Health Professionals, Researchers, Clinical Trials, News & Resources, and About NHLBI. A sidebar on the right is titled "Now Accepting GWAS Results Submissions" and provides instructions for submitting results to the GRASP mailing list.

GRASP: Genome-Wide Repository of Associations Between SNPs and Phenotypes

Overview

GRASP includes all available genetic association results from papers, their supplements and web-based content meeting the following guidelines:

- All associations with $P < 0.05$ from GWAS defined as $> 25,000$ markers tested for 1 or more traits.
- Study exclusion criteria: CNV-only studies, replication/follow-up studies testing $< 25k$ markers, non-human only studies, article not in English, gene-environment or gene-gene GWAS where single SNP main effects are not given, linkage only studies, aCGH/LOH only studies, heterozygous homozygosity (genome-wide or long runs), studies that only present genome-wide association results, studies that only present linkage analysis results, studies which we judge as redundant with prior studies since they do not provide significant inclusion of new samples or exposure of new results (e.g., many methodological papers on the WTCCC and FHS GWAS).
- More detailed methods and resources used in constructing the catalog are described at the "Methods & Resources" page.

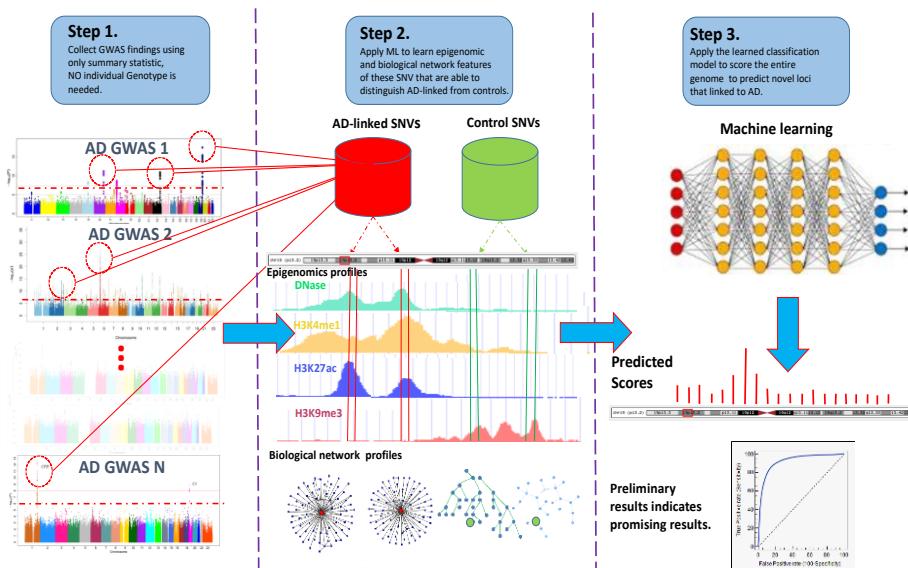
[Search the catalog](#)

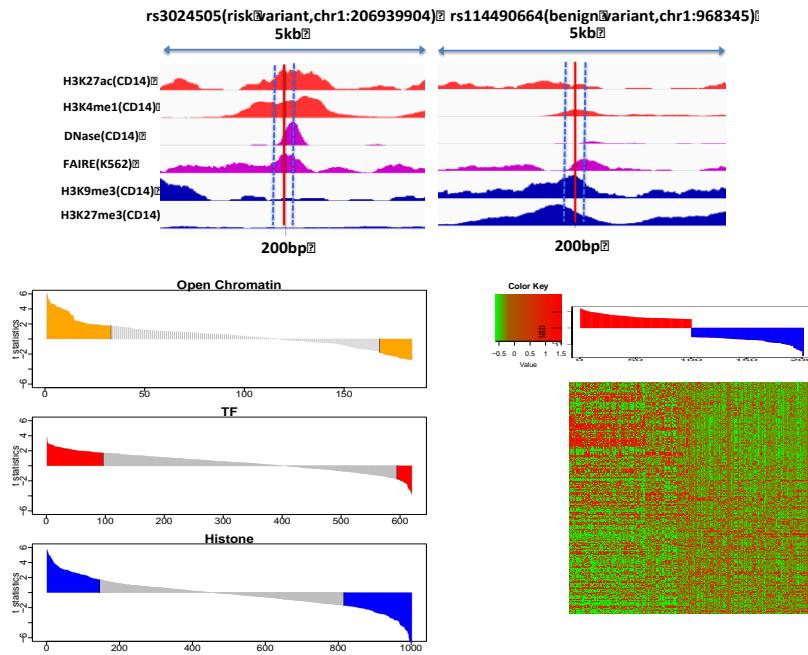
Epigenomic features utilized

Data Source	cell line	TF/Histone	feature
REMC DNase	73	-	73
REMC Histone	109	31	735
ENCODE DNase	80	-	80
ENCODE FAIRE	31	-	31
ENCODE TF(HAIB)	19	76	293
ENCODE TF(SYDH)	31	100	279
ENCODE Histone	18	42	267
ENCODE RNA Polymerase	31	2	49
Total	261	217	1806



DIVAN

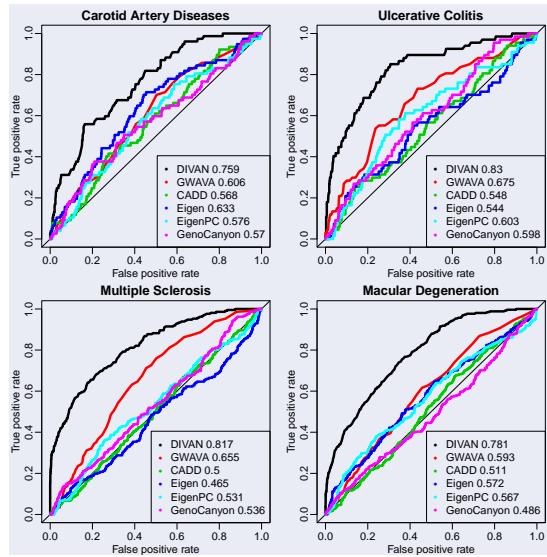




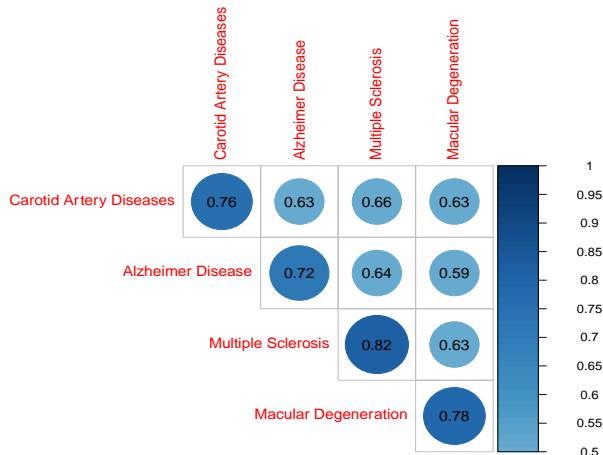
Choose GWAS variants

- We choose 45 diseases/phenotypes spanning 12 disease/phenotype classes, with at least 50 disease-SNP associations from Association Results Browser
- Benign SNPs are sampled from the 1000 Genomes (Phase I) with same distance (SNP to nearest TSS) distribution as risk SNPs

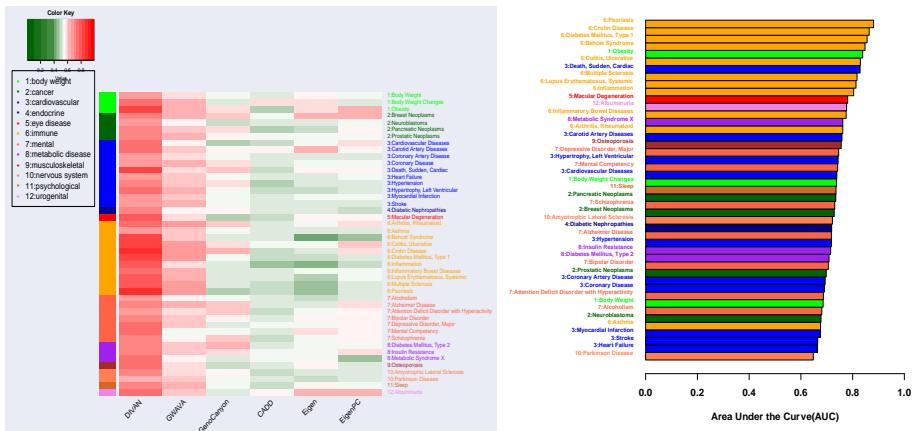
Performance comparison on four diseases



The importance of disease-specificity



AUC values of the 45 diseases tested



AUC from 0.66 to 0.88 with median 0.74

Immune diseases are best predicted

DIVAN website

<https://sites.google.com/site/emorydivan/>

DIVAN Updated Nov 9, 2016, 4:46 PM

DIVAN

DIVAN

Welcome to the home page of DIVAN !

DIVAN (Disease-specific Variant ANnotation), is a feature selection, ensemble-learning framework for disease-specific noncoding variant annotation and prioritization. DIVAN considers thousands of epigenomic annotations and is able to handle the class imbalance and "large p, small n" problem. Unlike most existing computational tools, DIVAN is able to provide scores that gauge a variant's impact in a disease-specific manner. Currently, DIVAN has been trained to evaluate variants for 45 different diseases/trait.

We have pre-computed the scores for each of the 45 disease/trait for each base of the whole genome (hg19). We also provide different software to score known variants and arbitrary genomic regions.

For questions or concerns, please contact Steve Qin (zhaohui.qin@emory.edu) or Li Chen (lchen@emory.edu).

DIVAN is free and open-source software, released under the [GNU General Public License v3](#).

EMORY

Summary

- Disease-specific risk variant identification is feasible
- Training data obtained from GWAS results and 1000 Genomes databases.
- Features are collected from genomics profiling data stored in ENCODE, REMC.

Chen et al. *Genome Biology* (2016) 17:252
DOI 10.1186/s13059-016-1112-z

Genome Biology



METHOD

Open Access

DIVAN: accurate identification of non-coding disease-specific risk variants using multi-omics profiles



CrossMark

Li Chen¹, Peng Jin² and Zhaohui S. Qin^{3,4*}

Extract insights from Big Data

Omicseq:

An omics data search engine



OMICSEQ

An information hub for genomic data

Gene Pathway

Human Enter keywords here... Settings

Genes:
EGFR, KRAS, ERBB2, POU5F1,
FOXA1...

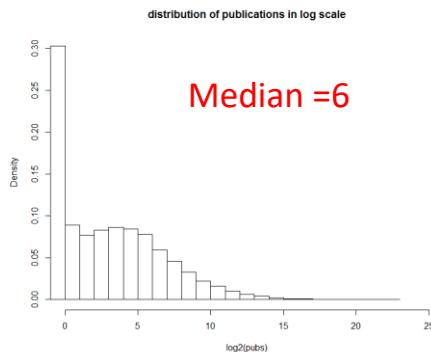
Pathway:
Apoptosis-GO, RNA elongation...

Literature is the major source of biomedical knowledge



- Accurate (with quality control)
- Specific and definitive
- With established infrastructure and technology to conduct effective literature mining

How much do we know?



- For an obscure gene, little information is known in the literature.

69

NCBI Resources How To

PubMed.gov
US National Library of Medicine
National Institutes of Health

PubMed sumo1p1 | Create RSS Create alert Advanced

Article types Format: Summary ▾ Sort by: Most Recent ▾ Send to
Clinical Trial
Review
Customize ...

Text availability
Abstract
Free full text
Full text

PubMed Commons Reader comments Trending articles

Publication dates
5 years
10 years
Custom range...

Species Humans Other Animals

[Clear all](#) [Show additional filters](#)

See [SUMO1P1 SUMO1 pseudogene 1](#) in the Gene database

Search results
Items: 3

1. [Phenotype and Tissue Expression as a Function of Genetic Risk in Polycystic Ovary Syndrome.](#)
Pau CT, Mosbruger T, Saxena R, Welt CK.
PLoS One. 2017 Jan 9;12(1):e0168870. doi: 10.1371/journal.pone.0168870. eCollection 2017.
PMID: 28068351 [Free PMC Article](#)
[Similar articles](#)

2. [Cross-ethnic meta-analysis of genetic variants for polycystic ovary syndrome.](#)
Louwers YV, Stolk L, Uitterlinden AG, Laven JS.
J Clin Endocrinol Metab. 2013 Dec;98(12):E2006-12. doi: 10.1210/jc.2013-2495. Epub 2013 Oct 8.
PMID: 24106282
[Similar articles](#)

3. [Genome-wide study identifies PTPRO and WDR72 and FOXQ1-SUMO1P1 interaction associated with neurocognitive function](#)

Literature is limited

- Only interesting (for authors and the journal, not necessarily for all audience) and significant findings were reported
- Mundane events, like most TF binding, gene expression changes do not make it to the papers
- Polished yet subjective and selective

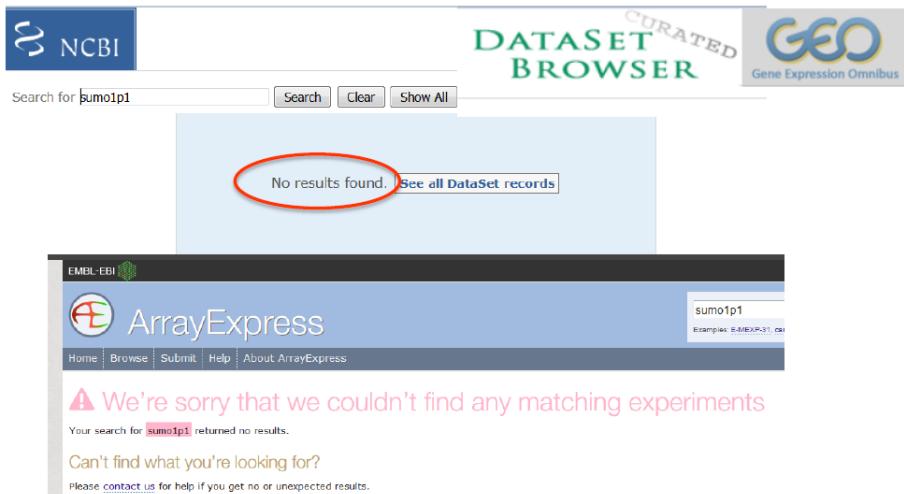


71



72

Perform a search in public data repositories



Our goals

- To develop a website that links to **ALL** the biomedical data that ever produced
- The database only stores data that are **processed** and **ready-to-use**
- To build a search engine from which one can get information on any **gene**.
- Do not rely on the metadata.



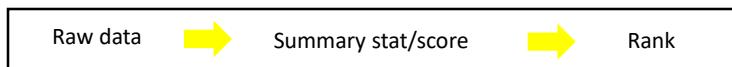
- Collect different variety of –omics data.
- Develop standardized protocols to process different types of data.
- Store these processed data in databases.
- Collect metadata.
- Develop a query engine for dataset searching.
- Develop a ranking algorithm “TrackRank”.
- Facilitating easy downloading of the processed, ready-to-analyze data.

Data types to be included

- From experimental assays
 - FPKM values from RNA-seq,
 - Read counts at promoters from ChIP-seq,
 - P-values of detecting DE genes using microarray,
 - Pausing index from GRO-seq,
 - Average methylation level at the promoters from BS-seq

Compare across data types (I)

- Our hypothesis is that a gene “plays an important role in a dataset” if its score ranks at the top among all genes in the genome.
- We developed trackRank algorithm to rank datasets using this idea.



77

A

ID	type	geneA	geneB	geneC	geneD	geneE	geneF	geneG	geneH
001	Ch	5.8	6.2	0.1	18.5	2.3	8.10	0.9	7.5
002	RN	0.001	103.0	25.3	1.2	0.5	28.0	10.5	2.1
003	BS	0.02	0.85	0.03	0.18	0.99	0.08	0.91	0.15
004	SO	0	5	3	10	2	1	0	0
005	CNV	0.3	1.5	0.2	-0.1	1.8	2.3	-0.3	0.5

Gene-based scores

B

ID	type	geneA	geneB	geneC	geneD	geneE	geneF	geneG	geneH
001	Ch	5/8	4/8	8/8	1/8	6/8	2/8	7/8	3/8
002	RN	8/8	1/8	3/8	6/8	7/8	2/8	4/8	5/8
003	BS	8/8	3/8	7/8	4/8	1/8	6/8	2/8	5/8
004	SO	8/8	2/8	3/8	1/8	4/8	5/8	8/8	8/8
005	CNV	6/8	3/8	7/8	8/8	2/8	1/8	6/8	4/8

Percentiles

C

ID	type	geneA	geneB	geneC	geneD	geneE	geneF	geneG	geneH
003	BS	8/8	3/8	7/8	4/8	1/8	6/8	2/8	5/8
005	CNV	6/8	3/8	7/8	8/8	2/8	1/8	6/8	4/8
004	SO	8/8	2/8	3/8	1/8	4/8	5/8	8/8	8/8
001	Ch	5/8	4/8	8/8	1/8	6/8	2/8	7/8	3/8
002	RN	8/8	1/8	3/8	6/8	6/8	7/8	4/8	5/8

Query gene

OMICSEQ
An information hub for genomic data

Gene miRNA PathWay MultiGene Genomic Region datasetSearch DiseasesRank

hg19

Search
Setting

Genes:
EGFR,KRAS, ERBB2, POU5F1, FOXA1...

miRNA:
has-let-7b,has-mir-100...

Pathway:
Apoptosis-GO, RNA elongation...

Multigene:
HOXA1,HOXA2, HOXA3...

Genomics regions:
Chr2 start: 33805280 end: 33808250...

Omicseq result page

Recent Search

Gene miRNA Pathway Multigene Genomic Region Dataset Search Diseases Rank

hg19
ERG
Search Setting

Advanced

Search "ERG" * 98 (top 1% of total 34896) results (2.405155 seconds) about NM_001138154
(hg19) Chr: 21, Start: 39751950, End: 40033704, Strand: +
You can also search NM_001136155, NM_001243429, NM_001243429, NM_004449, NM_182918
Maybe you want to know more in PubMed, Wikipedia, Google, WikiGenes, GeneCards, HGNC, BioGPS, ClinBase,
GENATLAS, GOPubMed, H-InvDB, QuickGO, Reactome.

Rank	DataSetID	DataType	sample	tissue/status/factor	Order/Total	Percentile(%)	Study	Lab	More info
1	41198	CNV	TCGA-laml-tumor	Bone Marrow tumor	12/22039	0.054	TCGA		MetaData PubMed GEO Download more
2	41250	CNV	TCGA-laml-tumor	Bone Marrow tumor	19/22039	0.086	TCGA		MetaData PubMed GEO Download more
3	1200203	Summary Track	TCGA-cesc	Cervical Tumor	15/15397	0.097	TCGA Firehose	BROAD GDAC	MetaData PubMed GEO Download more
4	1200571	Summary Track	TCGA-brca	Breast Tumor	19/16080	0.118	TCGA Firehose	BROAD GDAC	MetaData PubMed GEO Download more
5	41389	CNV	TCGA-laml-tumor	Bone Marrow tumor	26/22039	0.118	TCGA		MetaData PubMed GEO Download more
6	47239	CNV	TCGA-blca-tumor	Bladder tumor	35/22039	0.159	TCGA		MetaData PubMed GEO Download more
7	39473	CNV	TCGA-luad-tumor	Lung tumor	36/22039	0.163	TCGA		MetaData PubMed GEO Download more
8	200329	ChIP-seq(P)	bone marrow derive...	mesenchymal Normal H...	51/30792	0.165	Epigenome Ro...	Broad	MetaData PubMed GEO Download more
9	34685	CNV	TCGA-stad-tumor	Stomach tumor	46/22039	0.209	TCGA		MetaData PubMed GEO Download more
10	44892	CNV	TCGA-lihc-tumor	Liver tumor	52/22039	0.236	TCGA		MetaData PubMed GEO Download more
11	101095	ChIP-seq(P)	HSMM	Skeletal Normal H3K27...	77/30792	0.250	ENCODE	Broad	MetaData PubMed GEO Download more
12	201806	ChIP-seq(P)	lung_fetal day8 F	Lung Normal input	77/30792	0.250	Epigenome Ro...	Broad	MetaData PubMed GEO Download more

Summary

- Developed Omicseq: a omics data search engine, and a biological knowledge discovery tool.
- Does not rely on metadata
- Powered by trackRank algorithm
- Powerful resource for data mining
- Try it <http://www.omicseq.org>



OMICSEQ
An information hub for genomic data

Gene Pathway

Human Enter keywords here...

Advanced

Genes:
EGFR, KRAS, ERBB2, POU5F1,
FOXA1...

Pathway:
Apoptosis-GO, RNA elongation...

Nucleic Acids Research, 2017 1
doi: 10.1093/nar/gkx258

Omicseq: a web-based search engine for exploring omics datasets

Xiaobo Sun¹, William S. Pittard², Tianlei Xu¹, Li Chen¹, Michael E. Zwick³, Xiaoqian Jiang⁴, Fusheng Wang^{5,6} and Zhaohui S. Qin^{2,7,*}

¹Department of Mathematics and Computer Science, Emory University, 400 Dowman Drive, Atlanta, GA 30322, USA, ²Department of Biostatistics and Bioinformatics, Rollins School of Public Health, Emory University, 1518 Clifton Road NE, Atlanta, GA 30322, USA, ³Department of Human Genetics, Emory University School of Medicine, 1515 Michael Street, Atlanta, GA 30322, USA, ⁴Health Science Department of Biomedical Informatics, University of California San Diego, 9500 Gilman Drive, La Jolla, CA 92093, USA, ⁵Department of Biomedical Informatics, Stony Brook University, HSC L3-043, Stony Brook, NY 11794, USA, ⁶Department of Computer Science, Stony Brook University, Computer Science Building, Stony Brook, NY 11794, USA and ⁷Department of Biomedical Informatics, Emory University School of Medicine, 36 Eagle Row, Atlanta, GA 30322, USA

Received February 20, 2017; Revised March 27, 2017; Editorial Decision March 31, 2017; Accepted April 04, 2017

Ways to handle Big Data

Distributed systems to handle Big Data



TECHNICAL NOTE

Optimized distributed systems achieve significant performance improvement on sorted merging of massive VCF files

Xiaobo Sun ¹, Jingjing Gao², Peng Jin³, Celeste Eng⁴, Esteban G. Burchard⁴, Terri H. Beaty⁵, Ingo Ruczinski⁶, Rasika A. Mathias⁷, Kathleen Barnes⁸, Fusheng Wang^{9,*}, Zhaohui S. Qin ^{10,*,11} and CAAPA consortium¹¹

Summary

- Genomics Big data widely available.
- There are many different ways to utilize these Big Data
- If carefully designed, These Big Data give us opportunity to gain insights and make new discoveries.
- Statistics and ML thinking is required to use these resources effectively.
- Need latest informatics methods to enable the use of Big Data.

References

- Li B, Sun Z, He Q, Sun Z, Zhu Y, Qin ZS (2015) Bayesian inference with historical data-based informative priors improves detection of differentially expressed genes. *Bioinformatics*. **32**. 682-689.
- Chen L, Qin ZS. (2015) traseR: an R package for performing trait-associated SNP enrichment analysis in genomic intervals. *Bioinformatics*. **32**. 1214-1216.
- Chen L, Jin P, Qin ZS. (2016) DIVAN: accurate identification of non-coding disease-specific risk variants using multi-omics profiles. *Genome Biol.* **17**. 252.
- Sun X, Pittard WS, Xu T, Chen L, Zwick ME, Jiang X, Wang F, Qin ZS. (2017) Omicseq: A web-based search engine for exploring omics datasets. *Nucleic Acids Research*. **45**. W445-W452.
- Sun X, Gao J, Jin P, Eng C, Burchard EG, Beaty TH, Ruczinski I, Mathias RA, Barnes KC, Wang F, Qin ZS on behalf of CAAPA consortium (2018) Optimized Distributed Systems Achieve Significant Performance Improvement on Sorted Merging of Massive VCF Files. *Gigascience*. **7**.
- Xu T, Jin P, Qin ZS. (2020) Regulatory annotation of genomic intervals based on tissue-specific expression QTLs. *Bioinformatics*. **36**. 690-697.

Thank you

Questions:

zhaohui.qin@emory.edu