HiTSeq'2017
HIGH THROUGHPUT SEQUENCING
ALGORITHMS AND APPLICATIONS
ISMB/ECCB 2017, July 21-25, Prague, Czech Republic

HiTSeq 2017 Proceedings

Prague, Czech Republic
July 24-25, 2017
http://www.hitseq.org

Organizers:

Can Alkan
Bilkent University, Bilkent, Ankara, Turkey
E-mail: calkan@gmail.com

Valentina Boeva
Institut Cochin/INSERM/CNRS, Paris, France
E-mail: valentina.boeva@gmail.com

Ana Conesa
University of Florida, Gainesville, Florida, USA
E-mail: vickycoce@gmail.com

Francisco M. De La Vega, D.Sc.
Stanford University, and TOMA Biosciences, USA.
E-mail: Francisco.DeLaVega@stanford.edu

Dirk Evers
Molecular Health GmbH, Heidelberg, Germany
E-mail: dirk.evers@gmail.com

Kjong Lehmann
Memorial Sloan-Kettering Cancer Center. New York, NY, USA
E-mail: lehmann@mskcc.org

Gunnar Rätsch
Memorial Sloan-Kettering Cancer Center. New York, NY, USA
E-mail: gunnar.ratsch@ratschlab.org

Chromatin Accessibility Prediction via Convolutional Long Short-Term Memory Networks with k-mer Embedding

Keywords: chromatin accessibility, k-mer embedding, long short-term memory, deep learning

Abstract: Motivation: Experimental techniques for measuring chromatin accessibility are expensive and time consuming, appealing for the development of computational methods to precisely predict open chromatin regions from DNA sequences. Along this direction, existing computational methods fall into two classes: one based on handcrafted k-mer features and the other based on convolutional neural networks. Although both categories have shown good performance in specific applications thus far, there still lacks a comprehensive framework to integrate useful k-mer co-occurrence information with recent advances in deep learning.

Method and results: We fill this gap by addressing the problem of chromatin accessibility prediction with a convolutional Long Short-Term Memory (LSTM) network with k-mer embedding. We first split DNA sequences into k-mers and pre-train k-mer embedding vectors based on the co-occurrence matrix of k-mers by using an unsupervised representation learning approach. We then construct a supervised deep learning architecture comprised of an embedding layer, three convolutional layers and a Bidirectional LSTM (BLSTM) layer for feature learning and classification. We demonstrate that our method gains high-quality fixed-length features from variable-length sequences and consistently outperforms baseline methods. We show that k-mer embedding can effectively enhance model performance by exploring different embedding strategies. We also prove the efficacy of both the convolution and the BLSTM layers by comparing two variations of the network architecture. We confirm the robustness of our model to hyper-parameters by performing sensitivity analysis. We hope our method can eventually reinforce our understanding of employing deep learning in genomic studies and shed light on research regarding mechanisms of chromatin accessibility.

Availability and implementation: The source code can be downloaded from https://github.com/minxueric/ismb2017_lstm.

Authors:

first name	last name	email	country	organization	corresponding?
Xu	Min	minxueric@gmail.com	China	Tsinghua University	
Wanwen	Zeng	zengww14@mails.tsinghua.edu.cn	China	Tsinghua University	
Ning	Chen	ningchen@tsinghua.edu.cn	China	Tsinghua University	
Ting	Chen	tingchen@tsinghua.edu.cn	China	Tsinghua University	✓
Rui	Jiang	ruijiang@tsinghua.edu.cn	China	Tsinghua University	✓

Discovery and genotyping of novel sequence insertions in many sequenced individuals

Keywords: high throughput sequencing, novel sequence insertions, structural variation

Abstract: Despite recent advances in algorithms design to characterize structural variation using high-throughput short read sequencing (HTS) data, characterization of novel sequence insertions longer than the average read length remains a challenging task. This is mainly due to both computational difficulties and the complexities imposed by genomic repeats in generating reliable assemblies to accurately detect both the sequence content and the exact location of such insertions. Additionally, de novo genome assembly algorithms typically require a very high depth of coverage, which may be a limiting factor for most genome studies. Therefore, characterization of novel sequence insertions is not a routine part of most sequencing projects. There are only a handful of algorithms that are specifically developed for novel sequence insertion discovery that can bypass the need for the whole genome de novo assembly. Still, most such algorithms rely on high depth of coverage, and to our knowledge there is only one method (PopIns) that can use multi-sampledata to ”collectively” obtain very high coverage dataset to accurately find insertions common in a given population. Here we present Pamir, a new algorithm to efficiently and accurately discover and genotype novel sequence insertions using either single or multiple genome sequencing datasets. Pamir is able to detect breakpoint locations of the insertions and calculate their zygosity (i.e. heterozygous vs. homozygous) by analyzing multiple sequence signatures, matching one-end-anchored sequences to small-scale de novo assemblies of unmapped reads, and conducting strand-aware local assembly. We test the efficacy of Pamir on both simulated and real data, and demonstrate its potential use in accurate and routine identification of novel sequence insertions in genome projects.

Availability. Pamir is available at https://github.com/vpc-ccg/pamir

*Contact. fhach@sfu.ca, calkan@cs.bilkent.edu.tr

Authors:

first name	last name	email	country	organization	corresponding?
Pınar	Kavak	pinarkavak@gmail.com	Turkey	Department of Computer Engineering, Boğaziçi University, İstanbul, Turkey	
Yen-Yi	Lin	yenyil@sfu.ca	Canada	School of Computing Science, Simon Fraser University, Burnaby, BC	
Ibrahim	Numanagić	inumanag@sfu.ca	Canada	School of Computing Science, Simon Fraser University, Burnaby, BC	
Hossein	Asghari	hasghari@sfu.ca	Canada	School of Computing Science, Simon Fraser University, Burnaby, BC	
Tunga	Güngör	gungort@boun.edu.tr	Turkey	Department of Computer Engineering, Boğaziçi University, İstanbul, Turkey	
Can	Alkan	calkan@cs.bilkent.edu.tr	Turkey	Department of Computer Engineering, Bilkent University, Ankara, Turkey	✓
Faraz	Hach	fhach@sfu.ca	Canada	School of Computing Science, Simon Fraser University, Burnaby, BC	✓

HopLand: Single-cell pseudotime recovery using continuous Hopfield network based modeling of Waddington's epigenetic landscape

Keywords: single-cell, RNA-seq, Waddington's epigenetic landscape, pseudotime, Hopfield network, embryonic development

Abstract: Motivation: The interpretation of transcriptome dynamics in single-cell data, especially pseudotime estimation, could help understand the transition of gene expression profiles. The recovery of pseudotime increases the temporal resolution of single-cell transcriptional data, but is challenging due to the high variability in gene expression between individual cells. Here, we introduce HopLand, a pseudotime recovery method using continuous Hopfield network to map cells in a Waddington's epigenetic landscape. It reveals from the single-cell data the combinatorial regulatory interactions of genes that control the dynamic progression through successive cellular states.

Results: We applied HopLand to different types of single-cell transcriptome data. It achieved high accuracies of pseudotime prediction compared to existing methods. Moreover, a kinetic model can be extracted from each dataset. Through the analysis of such a model, we identified key genes and regulatory interactions driving the transition of cell states. Therefore, our method has the potential to generate fundamental insights into cell fate regulation.

Availability and implementation: The Matlab implementation of HopLand is available at https://github.com/NetLand-NTU/HopLand.

Availability: http://www.ntu.edu.sg/home/zhengjie/

Authors:

first name	last name	email	country	organization	corresponding?
Jing	Guo	GUOJ0020@e.ntu.edu.sg	Singapore	Nanyang Technological University	
Jie	Zheng	ZhengJie@ntu.edu.sg	Singapore	Nanyang Technological University	✓

Improving the performance of minimizers and winnowing schemes

Keywords: minimizers, winnowing, k-mer

Abstract: Motivation: The minimizers scheme is a method for selecting k-mers from sequences. It is used in many bioinformatics software tools to bin comparable sequences or to sample a sequence in a deterministic fashion at approximately regular intervals, in order to reduce memory consumption and processing time. Although very useful, the minimizers selection procedure has undesirable behaviors (e.g., too many k-mers are selected when processing certain sequences). Some of these problems were already known to the authors of the minimizers technique, and the natural lexicographic ordering of k-mers used by minimizers was recognized as their origin. Many software tools using minimizers employ ad hoc variations of the lexicographic order to alleviate those issues.

Results: We provide an in-depth analysis of the effect of k-mer ordering on the performance of the minimizers technique. By using small universal hitting sets (a recently defined concept), we show how to significantly improve the performance of minimizers and avoid some of its worse behaviors. Based on these results, we encourage bioinformatics software developers to use an ordering based on a universal hitting set or, if not possible, a randomized ordering, rather than the lexicographic order. This analysis also settles negatively a conjecture (by Schleimer et al.) on the expected density of minimizers in a random sequence.

Availability: The software used for this analysis is available on

GitHub: https://github.com/gmarcais/minimizers.git.

Contact: gmarcais@cs.cmu.edu

Authors:

first name	last name	email	country	organization	corresponding?
Guillaume	Marçais	gmarcais@cs.cmu.edu	USA	Carnegie Mellon University	✓
David	Pellow	d.pellowdavid@gmail.com	Israel	Tel Aviv University	
Daniel	Bork	DKB34@pitt.edu	USA	University of Pittsburgh	
Yaron	Orenstein	yaronore@mit.edu	USA	MIT	
Ron	Shamir	rshamir@post.tau.ac.il	Israel	School of Computer Science, Tel Aviv University	
Carl	Kingsford	carlk@cs.cmu.edu	USA	Carnegie Mellon University	✓

Modelling haplotypes with respect to reference cohort variation graphs

Keywords: Variation graph, Sequence graph, Haplotype, Recombination

Abstract: Motivation: Current statistical models of haplotypes are limited to panels of haplotypes whose genetic variation can be represented by arrays of values at linearly ordered bi- or multiallelic loci. These methods cannot model structural variants or variants that nest or overlap.

Results: A variation graph is a mathematical structure that can encode arbitrarily complex genetic variation. We present the first haplotype model that operates on a variation graph-embedded population reference cohort. We describe an algorithm to calculate the likelihood that a haplotype arose from this cohort through recombinations and demonstrate time complexity linear in haplotype length and sublinear in population size. We furthermore demonstrate a method of rapidly calculating likelihoods for related haplotypes. We describe mathematical extensions to allow modelling of mutations. This work is an important incremental step for clinical genomics and genetic epidemiology since it is the first haplotype model which can represent all sorts of variation in the population.

Authors:

first name	**last name**	**email**	**country**	**organization**	**corresponding?**
Yohei	Rosen	yohei@ucsc.edu	USA	University of California, Santa Cruz	✓
Jordan	Eizenga	jeizenga@ucsc.edu	USA	University of California Santa Cruz	✓
Benedict	Paten	benedictpaten@gmail.com	USA	UCSC	✓

Abundance estimation and differential testing on strain level in metagenomics data

Keywords: - metagenomics, - strain level resolution, - quantification, - differential taxa abundance

Abstract: Motivation: Current metagenomics approaches allow analyzing the composition of microbial communities at high resolution. Important changes to the composition are known to even occur on strain level and to go hand in hand with changes in disease or ecological state. However, specific challenges arise for strain level analysis due to highly similar genome sequences present. Only a limited number of tools approach taxa abundance estimation beyond species level and there is a strong need for dedicated tools for strain resolution and differential abundance testing.

Methods: We present DiTASiC (Differential Taxa Abundance in-cluding Similarity Correction) as a novel approach for quantification and differential assessment of individual taxa in metagenomics samples. We introduce a generalized linear model for the resolution of shared read counts which cause a significant bias on strain level. Further, we capture abundance estimation uncertainties, which play a crucial role in differential abundance analysis. A novel statistical framework is built, which integrates the abundance variance and infers abundance distributions for differential testing sensitive to strain level.

Results: As a result, we obtain highly accurate abundance estimates down to sub-strain level and enable fine-grained resolution of strain clusters. We demonstrate the relevance of read ambiguity resolution and integration of abundance uncertainties for differential analysis. Accurate detections of even small changes are achieved and false-positives are significantly reduced. Superior performance is shown on latest benchmark sets of various complexities and in comparison to existing methods. DiTASiC code is freely available from https://rki_bioinformatics.gitlab.io/ditasic.

Authors:

first name	**last name**	**email**	**country**	**organization**	**corresponding?**
Martina	Fischer	fischerm@rki.de	Germany	Robert Koch Institute	✓
Benjamin	Strauch	strauchb@rki.de	Germany	Robert Koch Institute	
Bernhard Y.	Renard	renardb@rki.de	Germany	Robert Koch Institute	✓

deBGR: An Efficient and Near-Exact Representation of the Weighted de Bruijn Graph

Keywords: de Bruijn graph, transcriptome, Bloom filter, High Throughput Sequencing (HTS), kmer, de novo

Abstract: Motivation: Almost all de novo short-read genome and transcriptome assemblers start by building a representation of the de Bruijn Graph of the reads they are given as input (Compeau et al., 2011; Pevzner et al., 2001; Simpson et al., 2009; Schulz et al., 2012; Zerbino and Birney, 2008; Grabherr et al., 2011; Chang et al., 2015; Liu et al., 2016; Kannan et al., 2016). Even when other approaches are used for subsequent assembly (e.g., when one is using "long read" technologies like those offered by PacBio or Oxford Nanopore), efficient k-mer processing is still crucial for accurate assembly (Carvalho et al., 2016; Koren et al., 2017), and state-of-the-art long-read error-correction methods use de Bruijn Graphs Salmela et al. (2016). Because of the centrality of de Bruijn Graphs, researchers have proposed numerous methods for representing de Bruijn Graphs compactly (Pell et al., 2012; Pellow et al., 2016; Chikhi and Rizk, 2013; Salikhov et al., 2013). Some of these proposals sacrifice accuracy to save space. Further, none of these methods store abundance information, i.e., the number of times that each k-mer occurs, which is key in transcriptome assemblers.

Results: We present a method for compactly representing the weighted de Bruijn Graph (i.e., with abundance information) with essentially no errors. Our representation yields zero errors while increasing the space requirements by less than 18%–28%. Our technique is based on a simple invariant that all weighted de Bruijn Graphs must satisfy, and hence is likely to be of general interest and applicable in most weighted de Bruijn Graph-based systems.

Availability: https://github.com/splatlab/debgr

Availability: http://www3.cs.stonybrook.edu/~ppandey/

Authors:

first name	last name	email	country	organization	corresponding?
Prashant	Pandey	ppandey@cs.stonybrook.edu	USA	Stony Brook University	✓
Michael A.	Bender	bender@cs.stonybrook.edu	USA	Stony Brook University	✓
Rob	Johnson	rob@cs.stonybrook.edu	USA	Stony Brook University	✓
Rob	Patro	rob.patro@cs.stonybrook.edu	USA	Stony Brook University	✓

Improved Data-Driven Likelihood Factorizations for Transcript Abundance Estimation

Keywords: RNA-seq, data-driven factorization, Quantification

Abstract:

Motivation: Many methods for transcript-level abundance estimation reduce the computational burden associated with the iterative algorithms they use by adopting an approximate factorization of the likelihood function they optimize. This leads to considerably faster convergence of the optimization procedure, since each round of e.g., the EM algorithm, can execute much more quickly. However, these approximate factorizations of the likelihood function simplify calculations at the expense of discarding certain information that can be useful for accurate transcript abundance estimation. Results: We demonstrate that model simplifications (i.e., factorizations of the likelihood function) adopted by certain abundance estimation methods can lead to a diminished ability to accurately estimate the abundances of highly-related transcripts. In particular, considering factorizations based on transcript- fragment compatibility alone can result in a loss of accuracy compared to the per-fragment, unsimplified model. However, we show that such shortcomings are not an inherent limitation of approximately factorizing the underlying likelihood function. By considering the appropriate conditional fragment probabilities, and adopting improved, data-driven factorizations of this likelihood, we demonstrate that such approaches can achieve performance nearly indistinguishable from methods that consider the complete (i.e., per-fragment) likelihood.

Availability: Our data-driven factorizations are incorporated into a branch of the Salmon transcript quantification tool: https://github.com/COMBINE-lab/salmon/tree/factorizations

Authors:

first name	last name	email	country	organization	corresponding?
Mohsen	Zakeri	mzakeri@cs.stonybrook.edu	USA	Stony Brook University	✓
Avi	Srivastava	asrivastava@cs.stonybrook.edu	USA	Stony Brook University	✓
Fatemehalsadat	Almodaresi T S	falmodaresit@cs.stonybrook.edu	USA	Stony Brook University	✓
Rob	Patro	rob.patro@cs.stonybrook.edu	USA	Stony Brook University	✓

Tumor Phylogeny Inference Using Tree-Constrained Importance Sampling

Keywords: phylogeny, cancer, algorithm

Abstract: Motivation:

A tumor arises from an evolutionary process that can be modeled as a phylogenetic tree. However, reconstructing this tree is challenging as most cancer sequencing uses bulk tumor tissue containing heterogeneous mixtures of cells.

Results:

We introduce PASTRI (Probabilistic Algorithm for Somatic Tree Inference), a new algorithm for bulk-tumor sequencing data that clusters somatic mutations into clones and infers a phylogenetic tree that describes the evolutionary history of the tumor. PASTRI uses an importance sampling algorithm that combines a probabilistic model of DNA sequencing data with a enumeration algorithm based on the combinatorial constraints defined by the underlying phylogenetic tree. As a result, tree inference is fast, accurate and robust to noise. We demonstrate on simulated data that PASTRI outperforms other cancer phylogeny algorithms in terms of runtime and accuracy. On real data from a chronic lymphocytic leukemia (CLL) patient, we show that a simple linear phylogeny better explains the data the complex branching phylogeny that was previously reported. PASTRI provides a robust approach for phylogenetic tree inference from mixed samples.

Authors:

first name	last name	email	country	organization	corresponding?
Gryte	Satas	gryte_satas@brown.edu	USA	Brown University	✓
Benjamin	Raphael	braphael@cs.princeton.edu	USA	Princeton University	✓

HiTSeq 2017 Oral Presentations

Deconvolution of heterogeneous bulk tumor genomic data via structured mixed membership models

Keywords: heterogeneity, unmixing, cancer, manifold learning, deconvolution

Abstract: Single-cell analysis of tumor genomics has shown extraordinary cell-to-cell heterogeneity to be a pervasive feature of solid tumors, yet cancer genomics for the moment is largely limited to more cost-effective bulk genomic analyses that largely obscure that heterogeneity. This limitation has led to widespread interest in genomic deconvolution methods, which use computational inferences to separate signals from distinct clonal subpopulations in bulk genomic data. Here, we present advances on the use of mixture substructure, i.e., heterogeneity in clonal composition across genomic samples in the subsets of a broader pool of clones, in improving deconvolution accuracy and resolution. Such substructure might be expected to arise in single tumors from shared ancestry or regional bias in clonal composition or in cross-sectional analysis of multiple tumors due to common progression pathways across tumors of similar subtype. Our methods draw on prior work using a geometric interpretation of deconvolution [1,2] generalized to represent the problem of structured mixture deconvolution as a special form of manifold learning problem [3]. We specifically describe work in progress to improve mathematical models and algorithmic workflow to handle better complex mixture substructure, improve inference automation, and allow simultaneous inference from distinct but complementary forms of genomic data. We validate the methods on breast tumor data from the Cancer Genome Atlas (TCGA) [4] for cross-sectional deconvolution to identify common progression states across tumor samples. We demonstrate that a common likelihood model and algorithmic pipeline can provide biologically meaningful decompositions from DNA, RNA, or joint DNA/RNA data, as assessed by term enrichment analysis and coincidence with alternative deconvolution methods, data types, and other classifications of tumor subtype. We then further evaluate the methods and demonstrate their utility for use in predicting future progression and related clinical outcomes via machine learning classification from inferred mixture compositions. The results demonstrate the value of accounting for substructure in genomic deconvolution, a general strategy that might be incorporated into many forms of deconvolution model and algorithm now in use.

[1] R. Schwartz and S. Shackney. “Applying unmixing to gene expression data for tumor phylogeny inference.” BMC Bioinformatics, 11:42, 2010.

[2] D. Tolliver, C. Tsourakakis, A. Subramanian, S. Shackney, and R. Schwartz. “Robust unmixing of tumor states in array comparative genomic hybridization data.” Bioinformatics, 26(12):i106-i114, 2010.

[3] T. Roman, B. Fasy, A. Nayyeri, and R. Schwartz. ”A simplicial complex-based approach to unmixing tumor progression data.” BMC Bioinformatics, 16:254, 2015.

[4] The Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumors. Nature. 2012;490(7416):61-70.

Availability: http://www.cmu.edu/bio/contacts/faculty/schwartz.shtml

Authors:

first name	last name	email	country	organization	corresponding?
Theodore	Roman	troman@andrew.cmu.edu	USA	Carnegie Mellon University	
Brenda	Xiao	brendax@andrew.cmu.edu	USA	Carnegie Mellon University	
Russell	Schwartz	russells@andrew.cmu.edu	USA	Carnegie Mellon University	✓

De novo assembly of viral quasispecies using overlap graphs

Keywords: de novo assembly, viral quasispecies, polyploid genome, haplotypes, deep sequencing, low-frequency mutations, overlap graph, error correction

Abstract: A viral quasispecies, the ensemble of viral strains populating an infected person, can be highly diverse. For optimal assessment of virulence, pathogenesis and therapy selection, determining the hap- lotypes of the individual strains can play a key role. As many viruses are subject to high mutation and recombination rates, high-quality reference genomes are often not available at the time of a new disease outbreak. We present SAVAGE, a computational tool for reconstructing individual haplotypes of intra- host virus strains without the need for a high-quality reference genome. SAVAGE makes use of either FM-index based data structures or ad-hoc consensus reference sequence for constructing overlap graphs from patient sample data. In this overlap graph, nodes represent reads and/or contigs, while edges reflect that two reads/contigs, based on sound statistical considerations, represent identical haplotypic sequence. Following an iterative scheme, a new overlap assembly algorithm that is based on the enumeration of sta- tistically well-calibrated groups of reads/contigs then efficiently reconstructs the individual haplotypes from this overlap graph. In benchmark experiments on simulated and on real deep coverage data, SAV- AGE drastically outperforms generic de novo assemblers as well as the only specialized de novo viral quasispecies assembler available so far. When run on ad-hoc consensus reference sequence, SAVAGE performs very favorably in comparison with state-of-the-art reference genome guided tools. We also apply SAVAGE on two deep coverage samples of patients infected by the Zika and the hepatitis C virus, respectively, which sheds light on the genetic structures of the respective viral quasispecies.

Availability: http://homepages.cwi.nl/~as

Authors:

first name	last name	email	country	organization	corresponding?
Jasmijn	Baaijens	baaijens@cwi.nl	Netherlands	Centrum Wiskunde & Informatica	
Amal Zine	El Aabidine	amal_zine@yahoo.fr	France	University of Montpellier	
Eric	Rivals	rivals@lirmm.fr	France	University of Montpellier	✓
Alexander	Schoenhuth	as@cwi.nl	Netherlands	Centrum Wiskunde &amp; Informatica	✓

Diversity in non-repetitive human sequences not found in the reference genome

Keywords: Variant calling, Structural variation, Novel-sequence insertions, Genome-wide association studies, Myocardial infarction

Abstract: We used whole-genome sequencing data of 15,219 Icelanders for discovering these variants that have been largely hidden until now. Previous large-scale variant detection efforts have mostly neglected this type of variation or reported non-reference sequences without placing them in the reference genome. Hence, a key step in our analysis was to scale up our multi-sample caller PopIns to this large number of samples.

We characterized the variants in terms of allele frequencies, novelty, and origin. Interestingly, the vast majority of variants that involve non-reference sequence of 200 bp or longer are present in the chimpanzee genome, implying that the non-reference allele is ancestral and the sequences were deleted in the individuals used to construct the reference genome.

Furthermore, we identified a larger number of loci where a variant is in linkage disequilibrium with a marker listed in the GWAS catalog than other SV detection efforts, an unexpected result in light of the fact that the other efforts included more types of variants. As an additional example for a potential disease impact, we report an association of an intronic 766 bp sequence with myocardial infarction.

We believe that our work is of great interest for the HiTSeq community as it reports the successful application of a variant caller to one of the largest sequencing data sets to date and, in addition, demonstrates how important it is to include variation of all complexity levels when studying the genetics of human disease.

Availability: https://www.bihealth.org/en/research/recruitment/junior-groups/birte-kehr/

Authors:

first name	last name	email	country	organization	corresponding?
Birte	Kehr	birte.kehr@bihealth.de	Germany	Berlin Institute of Health	✓
Anna	Helgadóttir	anna.helgadottir@decode.is	Iceland	deCODE Genetics/Amgen, Inc.	
Pall	Melsted	pmelsted@gmail.com	Iceland	University of Iceland	
Hákon	Jónsson	hakon.jonsson@decode.is	Iceland	deCODE Genetics/Amgen, Inc.	
Hannes	Helgason	hannes.helgason@decode.is	Iceland	deCODE Genetics/Amgen, Inc.	
Adalbjörg	Jonasdottir	adalbjorg.jonasdottir@decode.is	Iceland	deCODE Genetics/Amgen, Inc.	
Aslaug	Jonasdottir	aslaug.jonasdottir@decode.is	Iceland	deCODE Genetics/Amgen, Inc.	
Asgeir	Sigurdsson	asgeir.sigurdsson@decode.is	Iceland	deCODE Genetics/Amgen, Inc.	
Arnaldur	Gylfason	arnaldur.gylfason@decode.is	Iceland	deCODE Genetics/Amgen, Inc.	
Gisli H.	Halldorsson	gisli.halldorsson@decode.is	Iceland	deCODE Genetics/Amgen, Inc.	
Snaedis	Kristmundsdottir	snaedis.kristmundsdottir@decode.is	Iceland	deCODE Genetics/Amgen, Inc.	
Gudmundur	Thorgeirsson	gudmth@landspitali.is	Iceland	Faculty of Medicine, School of Health Sciences, University of Iceland	
Isleifur	Olafsson	isleifur@landspitali.is	Iceland	Department of Clinical Biochemistry, Landspitali–National University Hospital	
Hilma	Holm	hilma.holm@decode.is	Iceland	deCODE Genetics/Amgen, Inc.	
Unnur	Thorsteinsdottir	unnur.thorsteinsdottir@decode.is	Iceland	deCODE Genetics/Amgen, Inc.	
Patrick	Sulem	patrick.sulem@decode.is	Iceland	deCODE Genetics/Amgen, Inc.	
Agnar	Helgason	agnar.helgason@decode.is	Iceland	deCODE Genetics/Amgen, Inc.	
Daniel F.	Gudbjartsson	daniel.gudbjartsson@decode.is	Iceland	deCODE Genetics/Amgen, Inc.	
Bjarni	Halldorsson	bjarnivh@ru.is	Iceland	deCODE genetics and Reykjavik University	
Kári	Stefánsson	kari.stefansson@decode.is	Iceland	deCODE Genetics/Amgen, Inc.	

Clarice - Fast, Accurate and Secure Metagenomic Profiler

Keywords: Microbiome, metagenomics, microbial composition, privacy-preserving, secure profiler, sequence analysis, k-mers, algorithms, genome-centeric metagenomics, classification and taxonomy

Abstract: The number and types of metagenomic studies in microbiology and clinical genomics are increasing at an unprecedented rate, leading to computational challenges in the analysis pipeline. For example, the analysis of samples from healthcare environments requires methods capable of (i) accurately detecting pathogenic organisms, (ii) running with high speed to allow short response-time and diagnosis, and (iii) scaling to ever growing databases of reference genomes. While cloud-computing has the potential to offer low-cost solutions to these needs, serious concerns regarding the protection of genomic data exist due to the lack of control and security in remote genomic databases.

We present Clarice, the first metagenomic analysis tool capable of performing privacy-preserving queries on cloud-based databases. Clarice identifies and estimates the abundance of organisms present in environmental or clinical samples with high accuracy, high speed and a low memory footprint on the client side. Using an extensive set of synthetic and real datasets, we show that Clarice outperforms in accuracy and speed state-of-the-art programs MetaPhlAn and MetaFlow.

Software and datasets used in this paper are available at https://goo.gl/wsEqcC

Availability: http://www.cs.ucr.edu/~rouni001/

Authors:

first name	last name	email	country	organization	corresponding?
Rachid	Ounit	ROUNI001@UCR.EDU	USA	University of California, Riverside, CA, USA	
Niamh B.	O'Hara	niamh.ohara@cornell.edu	USA	Jacobs Technion-Cornell Institute, Cornell Tech, New York, NY, USA	
Stefano	Lonardi	stelo@cs.ucr.edu	USA	University of California, Riverside, CA, USA	✓
Christopher E.	Mason	chm2042@med.cornell.edu	USA	Department of Physiology and Biophysics, Weill Cornell Medicine, New York, NY, USA	✓

Outlier detection for improved differential splicing quantification from RNA-Seq experiments with replicates

Keywords: RNA-Seq, Alternative Splicing, Transcriptome

Abstract: Motivation: A key component in many RNA-Seq based studies is the production of multiple replicates for varying experimental conditions. Such replicates allow to capture underlying biological variability and control for experimental ones. However, during data production researchers often lack clear definitions to what constitutes a "bad" replicate which should be discarded and if data from failed replicates is published downstream analysis by groups using this data can be hampered.

Results: Here we develop a probability model to weigh a given RNA-Seq experiment as a representative of an experimental condition when performing alternative splicing analysis. Using both synthetic and real-life data we demonstrate that this model detects outlier samples which are consistently and significantly different compared to samples from the same condition. Using both synthetic and real-life data we perform extensive evaluation of the algorithm in different scenarios involving perturbed samples, mislabeled samples, no-signal groups, and different levels of coverage, and show it compares favorably with current state of the art tools.

Availability: Program and code will be available at majiq.biociphers.org

Authors:

first name	**last name**	**email**	**country**	**organization**	**corresponding?**
Scott	Norton	scnorton@mail.med.upenn.edu	USA	University of Pennsylvania	
Jorge	Vaquero-Garcia	jvaq@mail.med.upenn.edu	USA	University of Pennsylvania	
Yoseph	Barash	yosephb@mail.med.upenn.edu	USA	University of Pennsylvania	✓

Quantitative assessment of genome integrity from whole genome sequencing data

Keywords: genome integrity, structural variation, copy number variation, wavelet transform, clinical standards and guidelines, R package

Abstract:

Background

Genomic instability and structural alterations are characteristics of most cancers, driving tumor evolution and facilitating the identification of other cancer hallmarks. Whilst tumor genomic profiling using whole-genome sequencing has the potential to identify nearly all forms of genetic variation in one experiment, the numbers of therapeutically actionable genomic rearrangement events identified to date have been relatively small. Recent integrative approaches for structural variant discovery, combining read depth, orientation, read-pair and split-read analyses have improved greatly detection accuracy. However, the high rate of false positives and low concordance between available methods limit their potential to inform on clinical decision-making. Sequence coverage and signal uniformity are prominent features that influence the accuracy and performance of algorithms for structural variant discovery. Therefore, rigorous assessment of these features prior to embarking on detailed analyses is essential to identify potentially problematic datasets and genomic loci for which downstream results will be unreliable.

Results

Here we introduce a computational framework for the assessment of read coverage signal uniformity and guidelines for evaluation of whole genome sequencing data based on the well-established theory of wavelets. The algorithm estimates and quantifies aberrations in sequencing coverage either on the whole genome scale or within predefined reference regions. The analytical framework produces a local aberration score (LAS), a local measure, designed to quantify abnormalities in read coverage signal, and an overall genome integrity metric (GIM) that allows for comparison of whole genome sequence datasets. By rigorous, exhaustive analysis of genomic data and annotations generated by the Genome in a Bottle (NA12878) consortium, as well as genomes of patients enrolled in the Genomics England BRC Cancer Pilot and WGS500 studies, we have created a set of guidelines based on scoring coverage signal roughness both in genomic regions and on the whole-genome level.

The metric and effectiveness of our guidelines were tested using WGS data from 52 patients enrolled in Genomics England BRC Cancer Pilot study. DNA samples (Fresh Frozen, Formalin-Fixed-Paraffin-Embedded and germline) were collected in accordance with routine diagnostics procedures. Our results demonstrate that GIM provides a reliable estimate of coverage uniformity in whole genome sequencing data, which can be associated with variability in sample preparation, storage, quality and in certain cases, disease state. We further show that GIM and LAS values outside of the recommended ranges indicate that accuracy of discovery methods and therefore the detection of structural variants can be severely compromised. Furthermore, our method facilitates the identification of genomic elements responsible for the irregularities in sequence coverage profiles and opens up further avenues of study designed to overcome these.

Conclusions

We have developed a flexible analytical framework that provides a quantitative estimate of genomic deterioration induced by biological or technical factors. The GIM and LAS measures together with corresponding guidelines, can be used as a proxy for the reliability of genomic rearrangement detection in whole genome sequencing data and inform upon the suitability of a datasets for analy-

ses. Importantly, our results highlight also the necessity of introducing industry-grade standards in sequencing applications for molecular diagnostics, prognostics and therapeutics in clinical oncology. The framework is implemented as R package and available from developers upon request.

Acknowledgements

We are grateful to the WGS500 Project Consortium for access to the data. The WGS500 project was funded by the Wellcome Trust, the Oxford NIHR Biomedical Research Centre, and Illumina Inc. SK is supported by the NIHR Biomedical Research Centre, Oxford The views expressed are those of the authors and not necessarily those of the Department of Health or Wellcome Trust. The work was supported also by funding from the Wellcome Trust Core Award Grant Number 090532/Z/09/Z.

Authors:

first name	last name	email	country	organization	corresponding?
Anas	Rana	anas.rana@oncology.ox.ac.uk	United Kingdom	Department of Oncology, University of Oxford	
Alexander	Kanapin	alexander.kanapin@oncology.ox.ac.uk	United Kingdom	Department of Oncology, University of Oxford	
Dimitris	Vavoulis	dimitris.vavoulis@oncology.ox.ac.uk	United Kingdom	Department of Oncology, University of Oxford	
Samantha Jl	Knight	sknight@well.ox.ac.uk	United Kingdom	NIHR Biomedical Research Centre, Wellcome Trust Centre for Human Genetics, University of Oxford	
Anna	Schuh	anna.schuh@oncology.ox.ac.uk	United Kingdom	Department of Oncology, University of Oxford	
Anastasia	Samsonova	a.a.samsonova@gmail.com	United Kingdom	Department of Oncology, University of Oxford	✓

DNA Methylation Dynamics of Human Hematopoietic Stem Cell Differentiation

Keywords: DNA methylation, epigenome maps, single-cell methylome sequencing, hematopoietic stem cell differentiation, human blood, cell type prediction, machine learning, computational epigenetics

Abstract: Although virtually all cells in an organism share the same genome sequence, regulatory mechanisms give rise to hundreds of different, highly specialized cell types. These mechanisms are governed by epigenetic patterns, such as DNA methylation, which determine DNA packaging, spatial organization, interactions with regulatory enzymes as well as RNA expression and which ultimately reflect the state of each individual cell.

Using low-input and single-cell whole genome bisulfite sequencing, we generated genome-wide DNA methylation maps of blood stem and progenitor cells (Farlik et al., 2016). These maps enabled us to characterize cell-type heterogeneity, and aggregating methylation levels of small pools of cells and single cells across putative regulatory regions, we dissected the DNA methylation dynamics of human hematopoiesis. We observed lineage-specific DNA methylation patterns between myeloid and lymphoid progenitors and associate these patterns to regulatory elements, gene expression and chromatin accessibility. Using statistical learning, we could accurately infer cell types from DNA methylation signatures and the resulting models could be used for a data-driven reconstruction of the human hematopoietic system.

Our observations illustrate the power of DNA methylation analysis for the in vivo dissection of differentiation landscapes as a complementary approach to lineage tracing and in vitro differentiation assays. The generated methylome maps and analysis methods provide a comprehensive framework for studying epigenetic regulation of cell differentiation and blood-linked diseases.

References:

Farlik, M., Halbritter, F., Müller, F., Choudry, F. A., Ebert, P., Klughammer, J., et al. (2016). DNA Methylation Dynamics of Human Hematopoietic Stem Cell Differentiation. Cell Stem Cell, 19(6), 808–822. http://doi.org/10.1016/j.stem.2016.10.019

Authors:

first name	last name	email	country	organization	corresponding?
Fabian	Müller	fmueller@mpi-inf.mpg.de	Germany	Max Planck Institute for Informatics	✓
Matthias	Farlik	mfarlik@cemm.oeaw.ac.at	Austria	CeMM Research Center for Molecular Medicine	
Florian	Halbritter	fhalbritter@cemm.oeaw.ac.at	Austria	CeMM Research Center for Molecular Medicine	
Mattia	Frontini	mf471@cam.ac.uk	United Kingdom	University of Cambridge	
Thomas	Lengauer	lengauer@mpi-inf.mpg.de	Germany	Max Planck Institute for Informatics	
Christoph	Bock	cbock@cemm.oeaw.ac.at	Austria	CeMM Research Center for Molecular Medicine	

SeqOthello: A Novel Indexing Structure to Support Accurate and Scalable Query over Large Scale Sequencing Reads

Keywords: SeqOthello, Othello, sequence query, large scale sequencing data, hierachical structure

Abstract: Huge amount of sequencing data shared through public databases, such as Sequencing Read Archive (SRA), provides invaluable resources for researchers to test hypotheses by reusing existing datasets. However, computational solutions to index and query large-scale sequencing data remained to be an unmet need until the recent development of tools pioneered by Sequence Bloom Tree.

In this paper, we introduced SeqOthello, a novel computational method to support accurate and scalable query over large-scale RNA-seq data. Deployable on both cloud-based and standalone machines, SeqOthello employs a two-layer hierarchical structure, with the first layer mapping k-mers to their frequency buckets and the second layer mapping k-mer to their occurrence map across all samples. The mapping at each node can be reduced into a many-to-one mapping problem between (hundreds of) millions of k-mers and up to millions of disjoint categories. These mappings are efficiently implemented using Othello, a core data structure described in the paper.

Comparison of SeqOthello with SBT demonstrated its superior performance in both query speed, disk, and memory usage. All the SeqOthello implementation variants were around two orders of magnitude faster than SBT on a large query (36,076 sequences) without loss of accuracy and compromises in memory usage. In the meantime, SeqOthello achieves a compression ratio of 165:1 (913:1 with compressed kmer), compared to that of 25:1 for SBT. SeqOthello is a stand-alone program implemented in C++, and can be downloaded via URL: https://github.com/sdyy1990/SeqOthello.

Availability: http://www.cs.uky.edu/~qian/

Authors:

first name	last name	email	country	organization	corresponding?
Xinan	Liu	xinan.liu@uky.edu	USA	University of Kentucky	
Ye	Yu	ye.yu@uky.edu	USA	University of Kentucky	
James N.	MacLeod	jnmacleod@uky.edu	USA	University of Kentucky	
Chen	Qian	cqian12@ucsc.edu	USA	University of California Santa Cruz	
Jinze	Liu	liuj@cs.uky.edu	USA	UKY	✓

Computing Optimal Flow Decompositions for Assembly

Keywords: flow decomposition, assembly, RNA-seq, transcript, parameterized algorithm

Abstract: Transcript and metagenomic assembly problems require disentangling overlapping reads into linear sequences in order to recover the constituent transcripts or genomes from a set of shotgun sequencing reads. This can be done mathematically by finding a flow decomposition of a splice graph using few paths. We introduce an algorithm that provably finds a decomposition with the minimum number of paths and demonstrate that it outperforms the previous best heuristics.

Availability: http://www.csc.ncsu.edu/faculty/bdsullivan

Authors:

first name	last name	email	country	organization	corresponding?
Kyle	Kloster	kakloste@ncsu.edu	USA	North Carolina State University	✓
Philipp	Kuinke	kuinke@cs.rwth-aachen.de	Germany	RWTH Aachen	✓
Michael P.	O'Brien	mpobrie3@ncsu.edu	USA	North Carolina State University	✓
Felix	Reidl	felix.reidl@gmail.com	Germany	RWTH Aachen	✓
Andrew	van der Poel	ajvande4@ncsu.edu	USA	North Carolina State University	✓
Fernando	Sánchez Villaamil	fernando.sanchez@cs.rwth-aachen.de	Germany	RWTH Aachen	✓
Blair D.	Sullivan	blair_sullivan@ncsu.edu	USA	NC State University	✓

A Pan-cancer Analysis of Alternative Transcription Start Sites

Keywords: Gene Regulation, RNA-Seq, Gene Promoters, Regulatory elements, Transcriptomics

Abstract:

BACKGROUND

Cancer is a disease of the genome where alterations in the DNA lead to uncontrollable cell proliferation and division. These modifications in a cell's behavior are reflected at the transcriptome level. Transcriptional regulation, whose central element is the promoter, is responsible for controlling these changes in the expression. International consortia efforts such as The Cancer Genome Atlas (TCGA)[1] and the International Cancer Genomics Consortium (ICGC)[2] produced vast amounts of publicly available RNA-Seq data to map changes in the cancer transcriptome. However due to lack of ChIP-Seq[3] and CAGE[4] data the role of the promoters in controlling transcriptional changes in cancer is still mostly unexplored.

RESULTS

Here, we developed a framework for estimating promoter activity from RNA-Seq data and used this framework to study the transcriptional regulatory changes that are associated with cancer. We have analyzed 1359 samples from the Pan-Cancer Analysis of Whole Genomes (PCAWG) and 6677 samples from the Genotype-Tissue Expression (GTEX)[5] projects encompassing 27 cancer types. We demonstrated that our approach accurately identifies active promoters by comparing our promoter activity estimations with H3K4me3 data from the Encyclopedia of DNA Elements (ENCODE)[6] project. We found hundreds of tissue specific alternative promoters that are not observable at the gene expression level. Furthermore, we identified promoters with significant activity changes in cancer compared to normal samples for individual cancer types. We examined the associations between noncoding promoter mutations and promoter activity levels per cancer type and pan-cancer. We found that filtering mutated promoters by promoter activity leads to further enrichment of known cancer genes.

CONCLUSIONS

In summary, we showed that promoter activity can be estimated using RNA-Seq data and used this approach to identify cancer associated alternative promoters for 27 cancer types. We anticipate that the promoter activity estimation using RNA-seq data will broaden our understanding of the promoters' role in cancer by enabling the use of widely available RNA-Seq data. Furthermore, the catalogue of cancer associated promoters identified here will be a useful resource to uncover the regulatory and transcriptional changes in cancer.

REFERENCES

1. Cancer Genome Atlas Research, N., et al., The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet, 2013. 45(10): p. 1113-20.

2. International Cancer Genome, C., et al., International network of cancer genome projects. Nature, 2010. 464(7291): p. 993-8.

3. Johnson, D.S., et al., Genome-wide mapping of in vivo protein-DNA interactions. Science, 2007. 316(5830): p. 1497-502.

4. Kodzius, R., et al., CAGE: cap analysis of gene expression. Nat Methods, 2006. 3(3): p. 211-22.

5. Consortium, G.T., The Genotype-Tissue Expression (GTEx) project. Nat Genet, 2013. 45(6): p. 580-5.

6. Consortium, E.P., The ENCODE (ENCyclopedia Of DNA Elements) Project. Science, 2004. 306(5696): p. 636-40.

Availability: http://cracs.fc.up.pt/~nf/

Authors:

first name	last name	email	country	organization	corresponding?
Deniz	Demircioğlu	ddemircioglu@gis.a-star.edu.sg	Singapore	Genome Institute of Singapore	
Tannistha	Nandi	nandit@gis.a-star.edu.sg	Singapore	Genome Institute of Singapore	
Claudia	Calabrese	ccalabre@ebi.ac.uk	United Kingdom	EMBL-EBI	
Engin	Cukuroglu	cukuroglue@gis.a-star.edu.sg	Singapore	Genome Institute of Singapore	
Nuno A.	Fonseca	nunofonseca@acm.org	United Kingdom	EMBL-EBI, European Bioinformatics Institute	
Andre	Kahles	andre.kahles@ratschlab.org	Switzerland	ETH Zurich	
Kjong	Lehmann	kjong.lehmann@inf.ethz.ch	Switzerland	ETH Zurich	
Steve	Rozen	steve.rozen@duke-nus.edu.sg	Singapore	Duke-NUS Graduate Medical School Singapore	
Bin Tean	Teh	teh.bin.tean@singhealth.com.sg	Singapore	National Cancer Centre Singapore	
Oliver	Stegle	oliver.stegle@ebi.ac.uk	United Kingdom	EMBL-European Bioinformatics Institute	
Alvis	Brazma	brazma@ebi.ac.uk	United Kingdom	European Bioinformatics Institute	
Angela	Brooks	anbrooks@ucsc.edu	USA	University of California	
Gunnar	Rätsch	raetsch@inf.ethz.ch	Switzerland	ETH Zurich	
Patrick	Tan	tanbop@gis.a-star.edu.sg	Singapore	Genome Institute of Singapore, Duke-NUS Graduate Medical School	✓
Jonathan	Göke	gokej@gis.a-star.edu.sg	Singapore	Genome Institute of Singapore	✓

Metagenome representation with Scalable Reference Graphs

Keywords: genomics, metagenomics, genome graph, data structure

Abstract: The accurate and comprehensive annotation but yet sparse representation of a sample of mixed high-throughput sequencing reads remains an unsolved problem for metagenome and metatranscriptome sequencing projects. This becomes even more important, as the characterization of microbial communities from sequencing data becomes an increasingly relevant task in clinical research. Currently established methods for the analysis of microbiota assign variants of 16S ribosomal RNA (rRNA) or reads from whole genome shotgun sequencing (WGS) to single entities in a given taxonomy tree or to elements in functional databases. These approaches are limited through incomplete taxonomies and annotation biases and, importantly, waste a large fraction of the raw sequence data that cannot be assigned to any existing reference. To address these shortcomings, we have implemented a new, highly sensitive approach to combine, represent and identify the microbial and/or functional composition of a large set of metagenome samples with a major focus on taking previous knowledge into account. Building on techniques from genome assembly and text compression, we use succinct data structures to efficiently represent all sequence information in a k-mer based assembly graph, which not only represents single species and their individual relationships but also captures intra-species variability. A set of more than 50,000 different viral genome sequences is compressed by over 80% when stored in the graph instead as raw sequences and by over 50% when including strain annotations. Importantly, the graph is structured as a self-index that can be used for alignment and annotation of reads arising from metagenome sequencing experiments. Hence, our representation is not only sparse but also efficiently searchable. In addition, the index is dynamic and allows efficient extension to new genome sequences without recomputing the whole index. To keep the graph accessible for fast alignment, we have developed a concept to distribute the index over a set of compute nodes minimizing inter-node communication. The nodes of the graph are colored using compressed binary annotation vectors, encoding information such as species, functional elements or other metadata associated to the underlying sequence. The graph's full utility is tailored for its use with WGS data, where not only unknown species can be represented in the correct relationship to known species but also single functional entities, e.g., single genes can be identified. The reference graph leverages information from known genomes as well as from the many previous studies, giving access to rare observations not yet present in reference databases. It is designed to integrate further knowledge over time, e.g., to accumulate information over many patients and studies. Thus, it will have a greater sensitivity to detect unseen or rarely seen species and inherently represents nearest neighbors with less bias towards species overrepresented in existing databases.

Authors:

first name	last name	email	country	organization	corresponding?
Andre	Kahles	andre.kahles@ratschlab.org	Switzerland	ETH Zurich	✓
Gunnar	Rätsch	raetsch@inf.ethz.ch	Switzerland	ETH Zurich	✓

Rapid phylogenetic placement via ancestral reconstruction

Keywords: Phylogenetic placement, taxonomic identification, metagenomics, ancestral sequence reconstruction, high throughput sequencing

Abstract: Metagenomic projects are scaling up and have spread to many fields such as ecology, environmental monitoring and agronomy. They have also spread to medical fields, in particular through medical diagnostics based on microbiomes and viromes. Common to all these area remains the challenge of taxonomical classification of the metagenomic reads (also called sequence "binning"). The limited contents of reference databases often limit the taxonomical identification of reads to specific marker genes, leaving a large proportion of the metagenome to unprecise identification or to the state of "dark matter" (DNA fragments which cannot be associated to specific clades). Nevertheless, the democratization of high-throughput sequencing approaches shifted the dark matter paradigm from a "side product" to a new source of "novel genomes and biodiversity", as shown by the recent discovery of novel prokaryotic clades [1] and first glimpses into the unexpected richness of the panvirome [2]. Strikingly, approaches dedicated to the exploration of poorly identified metagenomic fractions remain limited. While composition-based classification criteria can produce interesting results at high taxonomic levels [3], large-scale comparative genomics made through local alignment to very large references databases often remains a widely used solution when lower classification levels are required [4].

A recent alternative appeared with "phylogenetic placements" algorithms, which benefits from the power of phylogenetic theory, while attempting to scale the process to NGS datasets. Phylogenetic placement can identify sequences which are relatively distant to the clades composing a reference phylogeny, accelerating the understanding of a large collection of metagenomic sequences [5-6]. Still, this remains an alignment-dependant analysis: relevant metagenomic reads need first to be selected and aligned to the reference phylogenies (which is sometimes limited to pairwise HMM-based alignments, for speed purpose).

We are currently developing a new approach of metagenomic read placement, designed to combine alignment and placement steps and accelerate the process of taxonomic identifications in metagenomes. To do so, we use a probabilistic framework: our algorithm is based on the posterior probabilities generated though (marginal) ancestral sequence reconstruction on all internal nodes of a reference phylogeny. From these, a database aggregating the most probable ancestral short words (k-mers) is generated and stored in a hash table. Metagenomic reads can then be cut to short k-mers and compared to this database. The mapping result is then evaluated through the corresponding posterior probabilities, producing both a genomic localization on the reference alignment and the phylogenetic placement itself on the reference tree (alignment + placement). I will describe the algorithm of this approach, its performance compared to other phylogenetic placement pipelines and its preliminary application on benchmark datasets of typical environmental and clinical data (bacterial and viral metagenomes). Finally, I will briefly discuss the potential of the approach for the immediate detection of key reads holding rearrangements that may differentiate the query sequences from the user references. In the case of virus evolution, this may be extended to detect the common phenomena of recombination and reassortment, through the detection of reads which can be placed partially in different clades.

[1] Solden L, Lloyd K, Wrighton K. The bright side of microbial dark matter: lessons learned from the uncultivated majority. Curr Opin Microbiol. 2016

[2] Krishnamurthy SR, Wang D. Origins and challenges of viral dark matter. Virus Res. 2017

[3] Sedlar K, Kupkova K, Provaznik I . Bioinformatics strategies for taxonomy independent binning and visualization of sequences in shotgun metagenomics . Comput Struct Biotechnol J. 2017

[4] Huson DH, Beier S, Flade I, Górska A, El-Hadidi M, Mitra S, Ruscheweyh HJ, Tappu R. MEGAN Community Edition - Interactive Exploration and Analysis of Large-Scale Microbiome Sequencing Data. PLoS Comput Biol. 2016

[5] Matsen FA, Kodner RB, Armbrust EV. pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. BMC Bioinformatics. 2010

[6] Stark M, Berger SA, Stamatakis A, von Mering C. MLTreeMap–accurate Maximum Likelihood placement of environmental DNA sequences into taxonomic and functional reference phylogenies. BMC Genomics. 2010

Availability: http://www.lirmm.fr

Authors:

first name	last name	email	country	organization	corresponding?
Benjamin	Linard	linard@lirmm.fr	France	LIRMM - CNRS	✓
Krister	Swenson	swenson@lirmm.fr	France	LIRMM - CNRS	
Fabio	Pardi	pardi@lirmm.fr	France	LIRMM - CNRS	

SQANTI: extensive characterization of long read transcript sequences to remove artifacts in transcriptome identification and quantification

Keywords: Full-length isoform sequencing, Quality Control, Isoform Quantification, Proteogenomics

Abstract: The possibility of massively sequencing full-length transcripts has paved the way for the discovery of thousands of novel transcripts, even in very well annotated organisms as mice and humans. With the increasing utilization of long read technologies such as Pacbio, the necessity for a tool that provides a comprehensive classification of these novel transcripts as well as their exhaustive characterization is ever more pressing.

Here we present SQANTI, an automated pipeline for the classification of long-read transcripts that allows for identification and classification of transcripts and computes over 30 descriptors that can be used to assess the quality of the data and of the preprocessing pipelines.

We applied SQANTI to a neuronal mouse transcriptome using PacBio long reads and illustrate how the tool is effective in readily describing the composition of and characterizing the full-length transcriptome. We perform extensive evaluation of ToFU PacBio transcripts by PCR to reveal that an important number of the novel transcripts are technical artifacts of the sequencing approach, and that SQANTI quality descriptors can be used to engineer a filtering strategy to remove them. Bedsides, by comparing our iso-transcriptome with public proteomics databases we find that alternative isoforms are elusive to proteogenomics detection and are abundant in major protein changes with respect to the principal isoform of their genes. Finally, a comparison of Iso-Seq over the classical RNA-seq approaches solely based on short-reads demonstrates that the PacBio transcriptome not only succeeds in capturing the most robustly expressed fraction of transcripts, but also avoids quantification errors caused by unaccounted 3' end variability in the reference.

In conclusion, SQANTI allows the user to maximize the analytical outcome of long read technologies by providing the tools to deliver quality-evaluated and curated full-length transcriptomes.

Availability: http://bioinfo.cipf.es/aconesawp/

Authors:

first name	last name	email	country	organization	corresponding?
Lorena	de La Fuente Lorente	lfuente@cipf.es	Spain	Centro de Investigación Príncipe Felipe (CIPF)	✓
Manuel	Tardaguila	manueltar@ufl.edu	USA	UNIVERSITY OF FLORIDA	✓
Cristina	Marti	mcmarti@cipf.es	Spain	CIPF	
Hector	Del Risco	hdelrisco@ufl.edu	USA	University of Florida - Dr. Conesa's Lab	
Cecile	Pereira	cecilepereira@ufl.edu	USA	University of Florida - Dr. Conesa's Lab	
Marissa	Macchietto	mmacchie@uci.edu	USA	University of California Irvine	
Maravillas	Mellado	mmellado@cipf.es	Spain	Centro de Investigación Príncipe Felipe	
Ali	Mortazavi	ali.mortazavi@uci.edu	USA	University of California, Irvine	
Susana	Rodriguez	srodriguez@cipf.es	Spain	Centro de Investigación Príncipe Felipe	
Victoria	Moreno	vmorenom@cipf.es	Spain	CENTRO DE INVESTIGACION PRINCIPE FELIPE	
Ana	Conesa	aconesa@cipf.es	Spain	Genomics of Gene Expression Lab	✓

Integrating Diverse Transcriptomic Alterations to Identify Cancer-Relevant Genes and Signatures

Keywords: RNA-Seq, Recurrence analysis, Pan-cancer, Transcriptomics, signatures

Abstract: We present a novel analysis that 1) identifies cancer driver genes through a recurrence analysis over diverse types of transcriptomic alterations 2) identifies frequent and heterogeneous transcriptomic alteration signatures in 1190 samples across 25 cancer types as part of the PanCancer Analysis of Whole Genomes (PCAWG) of the International Cancer Genome Consortium (ICGC). We integrated the following alteration types: expression outliers, alternative splicing outliers, gene fusions, alternative promoters, somatic variants, RNA-editing, and allele-specific expression.

Previous multi-cancer genomic studies have focused on the analysis of somatic mutations as the driver of phenotypic changes. Here, we propose a method to integrate a wide variety of transcriptomic aberrations in combination with DNA-level changes to redefine the concept of driver events and account for the role of the transcriptome in tumorigenesis.

Availability: http://www.raetschlab.org/members/akahles/

Authors:

first name	last name	email	country	organization	corresponding?
Natalie	Davidson	nrd44@cornell.edu	USA	MSKCC	✓
Kjong-Van	Lehmann	kjong.lehmann@inf.ethz.ch	Switzerland	ETH Zurich	
Andre	Kahles	andre.kahles@ratschlab.org	Switzerland	ETH Zurich	
Alvis	Brazma	brazma@ebi.ac.uk	United Kingdom	European Bioinformatics Institute	
Angela	Brooks	anbrooks@ucsc.edu	USA	University of California, Santa Cruz	
Claudia	Calabrese	ccalabre@ebi.ac.uk	United Kingdom	EMBL-EBI	
Nuno A.	Fonseca	nunofonseca@acm.org	United Kingdom	EMBL-EBI, European Bioinformatics Institute	
Jonathan	Göke	gokej@gis.a-star.edu.sg	Singapore	Genome Institute of Singapore	
Roland	Schwarz	roland.schwarz@mdc-berlin.de	Germany	Max Delbrueck Center for Molecular Medicine	
Gunnar	Rätsch	gunnar.ratsch@ratschlab.org	USA	Memorial Sloan Kettering Center	
Zemin	Zhang	zemin@pku.edu.cn	China	Peking University	

Profiling immunoglobulin repertoires across multiple human tissues by RNA Sequencing

Keywords: Immunoglobulin (Ig) receptor repertoires, RNA-Seq, unmapped reads, immune response

Abstract: Assay-based approaches provide a detailed view of the adaptive immune system by profiling immunoglobulin (Ig) receptor repertoires. However, these methods carry a high cost and lack the scale of standard RNA sequencing (RNA-Seq). Here we report the development of ImReP, a novel computational method for rapid and accurate profiling of the immunoglobulin repertoire from regular RNA-Seq data. ImReP can also accurately assemble the complementary determining regions 3 (CDR3s), the most variable regions of Ig receptors. We applied our novel method to 8,555 samples across 53 tissues from 544 individuals in the Genotype-Tissue Expression (GTEx v6) project. ImReP is able to efficiently extract Ig-derived reads from RNA-Seq data. Using ImReP, we have created a systematic atlas of Ig sequences across a broad range of tissue types, most of which have not been studied for Ig receptor repertoires. We also compared the GTEx tissues to track the flow of Ig clonotypes across immune-related tissues, including secondary lymphoid organs and organs encompassing mucosal, exocrine, and endocrine sites, and we examined the compositional similarities of clonal populations between these tissues. The Atlas of Immune Immunoglobulin repertoires (The AIR), is freely available at https://smangul1.github.io/TheAIR/ , is one of the largest collection of CDR3 sequences and tissue types. We anticipate this recourse will enhance future immunology studies and advance development of therapies for human diseases. ImReP is freely available at https://sergheimangul.wordpress.com/imrep/ . The full paper is available on biorxiv : http://biorxiv.org/content/early/2017/03/25/089235

Authors:

first name	last name	email	country	organization	corresponding?
Serghei	Mangul	smangul@ucla.edu	USA	UCLA	✓

Tigmint: Correct Misassemblies Using Linked Reads From Large Molecules

Keywords: genome, assembly, misassembly correction, scaffolding, 10x, Chromium, linked reads, ABySS, gymnosperm, mitochondrion

Abstract: Long-read sequencing technologies have greatly improved assembly contiguity, but at a cost roughly ten times that of short-read sequencing technology. For population studies and when sequencing large genomes, such as conifer genomes and other economically important crop species, this cost may be prohibitive. The 10x Genomics Chromium technology generates linked reads from large DNA molecules at a cost comparable to standard short-read sequencing technologies. Whereas paired-end sequencing gives two reads from a small DNA fragment, linked reads yield roughly a hundred reads from molecules with a typical size of 10 to 100 kilobases. Linked reads indicate which reads were derived from the same DNA molecule, and so should be in close proximity in the underlying genome. Linked reads have been used previously to phase diploid genomes using a reference, de novo assemble complex genomes in the gigabase scale, and further scaffold draft assemblies.

In de novo sequencing projects, it is challenging yet important to measure the correctness of the resulting assemblies. Linked reads from technologies such as Chromium offer an opportunity to algorithmically address this problem. Here we introduce a software tool, Tigmint, to identify misassemblies using linked reads. The reads are first aligned to the assembly, and the extents of the large DNA molecules are inferred from the alignments of the reads. The physical coverage of the large molecules is more consistent and less prone to coverage dropouts than that of the short read sequencing data. Atypical drops in physical molecule coverage, less than the median minus 1.5 times the inter-quartile range, reveal possible misassemblies. Clipped alignments of the first and last reads of a molecule are used to refine the coordinates of the misassembly with base-pair accuracy.

No software tool currently exists for the specific purpose of identifying misassemblies using linked reads. The tool Long Ranger by 10x Genomics detects structural variants, which is a similar task. It requires however a reference genome assembled in fewer than 500 contigs, whereas a de novo assembly is often more fragmented. Tigmint addresses specifically the unaddressed problem of identifying misassemblies using linked reads.

Assemblies of short read sequencing data are easily confounded by repetitive sequence larger than the fragment size of the sequencing library. When the size of a repeat exceeds the library fragment size, the contig comes to an end in the best case, or results in misassembled sequence in the worst case. Tigmint is particularly useful in correcting these misassemblies when the initial assembly of the Illumina paired-end reads did not employ the barcodes of the linked reads, and this rich source of evidence is yet untapped.

Misassemblies not only complicate downstream analyses, but also limit the contiguity of the assembly, when incorrectly assembled sequences prevent joining their adjacent and correctly assembled sequences. To demonstrate the utility of Tigmint, we assemble the six megabase mitochondrial genome of Sitka spruce (Picea sitchensis) from 10x Genomics Chromium data using ABySS 2.0, identify and correct misassemblies using Tigmint, and scaffold using ARCS. Tigmint identifies 16 structural misassemblies in this case. After scaffolding with ARCS, the mitochondrial genome is assembled in 12 scaffolds larger than 100 kbp, with an N50 of 493 kbp. We plan to apply this method to assemble the twenty gigabase nuclear genome of Sitka spruce. Chromium reads permit cost-effective assembly of large genomes with high-throughput, short-read sequencing technology, while also providing large-molecule scaffolding data.

Availability: http://sjackman.ca

Authors:

first name	**last name**	**email**	**country**	**organization**	**corresponding?**
Shaun D	Jackman	sjackman@bcgsc.ca	Canada	BC Cancer Agency Genome Sciences Centre	✓
Benjamin P	Vandervalk	benv@bcgsc.ca	Canada	BC Cancer Agency Genome Sciences Centre	
Rene L	Warren	rwarren@bcgsc.ca	Canada	BC Cancer Agency Genome Sciences Centre	
Hamid	Mohamadi	hmohamadi@bcgsc.ca	Canada	BC Cancer Agency Genome Sciences Centre	
Justin	Chu	cjustin@bcgsc.ca	Canada	BC Cancer Agency Genome Sciences Centre	
Sarah	Yeo	syeo@bcgsc.ca	Canada	BC Cancer Agency Genome Sciences Centre	
Lauren	Coombe	lcoombe@bcgsc.ca	Canada	BC Cancer Agency Genome Sciences Centre	
Joerg	Bohlmann	bohlmann@mail.ubc.ca	Canada	Department of Forest and Conservation Sciences, University of British Columbia	
Steven Jm	Jones	sjones@bcgsc.ca	Canada	BC Cancer Agency Genome Sciences Centre	
Inanc	Birol	ibirol@bcgsc.ca	Canada	BC Cancer Agency Genome Sciences Centre	✓

HiTSeq 2017 Poster Abstracts

Gene-gene interaction analysis for rare variants

Keywords: epistatis, gene-gene interaction, rare variants

Abstract: With the rapid advancement of array-based genotyping techniques, genome-wide association studies (GWAS) have successfully identified common genetic variants associated with common complex diseases. However, it has been shown that only a small proportion of the genetic etiology of complex diseases could be explained by the genetic factors identified from GWAS. This missing heritability could possibly be explained by gene-gene interaction and rare variants. There has been an exponential growth of gene-gene interaction analysis for common variants in terms of methodological developments and practical applications. Also, the recent advancement of high-throughput sequencing technologies makes it possible to conduct rare variant analysis. However, little progress has been made in gene-gene interaction analysis for rare variants.

Here, we propose a new gene-gene interaction method for the rare variants in the frame-work of the multifactor dimensionality reduction (MDR) analysis. The proposed method consists of three steps. The first step is to collapse the rare variants according to their biological characteris-tics such as allele frequency or functional regions; this step utilizes known biological information to redefine the given genotypes to a more biologically meaningful categorical variable. An example would be a gene having no exonic rare variants given a value close to 0, and 1 otherwise, since non-exonic variants have weak or no effect on the function of a gene. The second step is to perform MDR analysis for the collapsed rare variants. The last is to use several evaluation measures to detect top candidate interaction pairs. This proposed method can be used for the detection of not only gene-gene interactions, but also interactions within a single gene. The proposed method is illustrated with 1080 whole exome sequencing data of the Korean population in order to identify causal gene-gene interaction for rare variants for type 2 diabetes.

Availability: http://www.bibs.snu.ac.kr

Authors:

first name	last name	email	country	organization	corresponding?
Min-Seok	Kwon	intellims@gmail.com	South Korea	Seoul National University	
Sangseob	Leem	leemss@snu.ac.kr	South Korea	Seoul National University	
Joon	Yoon	joonyoon.jay@gmail.com	South Korea	Seoul National University	
Taesung	Park	tspark@stats.snu.ac.kr	South Korea	Seoul National University	✓

Abundance estimation and differential testing on strain level in metagenomics data

Keywords: metagenomics, quantification, differential taxa abundance, strain level

Abstract: Motivation: Current metagenomics approaches allow analyzing the composition of microbial communities at high resolution. Important changes to the composition are known to even occur on strain level and to go hand in hand with changes in disease or ecological state. However, specific challenges arise for strain level analysis due to highly similar genome sequences present. Only a limited number of tools approach taxa abundance estimation beyond species level and there is a strong need for dedicated tools for differential abundance testing.

Methods: We present DiTASiC (Differential Taxa Abundance including Similarity Correction) as a novel approach for quantification and differential assessment of individual taxa in metagenomics samples. We introduce a generalized linear model for the resolution of shared read counts which cause a significant bias on strain level. Further, we capture abundance estimation uncertainties, which play a crucial role in differential abundance analysis. A novel statistical framework is built, which integrates the abundance variance and infers abundance distributions for differential testing sensitive to strain level.

Results: As a result, we obtain highly accurate abundance estimates of taxa. We demonstrate the relevance of read ambiguity resolution and integration of abundance uncertainties for differential analysis on strain level. Accurate detections of even small changes are achieved and false-positives are significantly reduced. Performance is evaluated on different data sets of various complexities and in comparison to existing methods. DiTASiC code is freely available from https://rki_bioinformatics.gitlab.io/ditasic.

Authors:

first name	last name	email	country	organization	corresponding?
Martina	Fischer	fischerm@rki.de	Germany	Robert Koch Institute	✓
Benjamin	Strauch	strauchb@rki.de	Germany	Robert Koch Institute	
Bernhard Y.	Renard	renardb@rki.de	Germany	Robert Koch Institute	✓

seq-seq-pan: Rapid construction of a pan-genome data structure with iterative updates

Keywords: pan-genome, data structure, sequence analysis, comparative genomics

Abstract: Motivation: The increasing application of Next Generation Sequencing technologies has led to the availability of thousands of reference genomes, often providing multiple genomes for the same or closely related species. The current approach to represent a species or a population with a single reference sequence and a set of variations cannot represent their full diversity and introduces bias towards the chosen reference. There is a need for the representation of multiple sequences in a composite way that is compatible with existing data sources for annotation and suitable for established sequence analysis methods. At the same time, this representation needs to be easily accessible and extendable to account for the constant change of available genomes.

Results: We introduce seq-seq-pan, a sequential genome aligning workflow for the rapid construction of a pan-genome data structure from multiple genomes. The flexible data structure provides methods for adding a new genome to or removing one from the set of genomes. It further allows the extraction of sequences and the fast generation of a consensus sequence and provides a global coordinate system. All these features form the basis for the usage of pan-genomes in downstream analyses.

Availability: https://gitlab.com/groups/rki_bioinformatics

Authors:

first name	last name	email	country	organization	corresponding?
Christine	Jandrasits	JandrasitsC@rki.de	Germany	Robert Koch Institute	✓
Piotr W.	Dabrowski	DabrowskiW@rki.de	Germany	Robert Koch Institute	
Stephan	Fuchs	FuchsS@rki.de	Germany	Robert Koch Institute	
Bernhard Y.	Renard	RenardB@rki.de	Germany	Robert Koch Institute	✓

Anaconda: AN Automated pipeline for somatic COpy Number variation Detection and Annotation from tumor exome sequencing data

Keywords: Copy number variation, Whole exome sequencing, Sequence analysis application

Abstract: Background: Copy number variations(CNVs) are one of the main genetic structural variations of cancer genomes. De-tecting CNAs in genetic exome region is efficient and cost-effective in identifying cancer associated genes. Many tools have been developed accordingly and yet these tools lack of reliability as a consequence of high false negative rate, which is intrinsically caused by genome exonic bias.

Results: To provide an alternative option, here, we report Anaconda, a comprehensive pipeline that allows flexible in-tegration of multiple CNV-calling methods and systematic annotation of CNVs in analyzing WES data. Just by one command, Anaconda can generate CNV detection result by up to four CNV detecting tools. Associated with compre-hensive annotation analysis of genes involved in shared CNV regions, Anaconda is able to deliver a more reliable and useful report in assistance with CNV-associate cancer researches.

Availability: Anaconda package and manual can be freely accessed at http://mcg.ustc.edu.cn/bsc/ANACONDA/.

Contact: zyuanwei@ustc.edu.cn

Authors:

first name	last name	email	country	organization	corresponding?
Jianing	Gao	gjn1106@mail.ustc.edu.cn	China	University of Science and Technology of China	
Changlin	Wan	omgwcl@mail.ustc.edu.cn	China	University of Science and Technology of China	
Qiguang	Zang	qiguang@mail.ustc.edu.cn	China	University of Science and Technology of China	
Rongjun	Ban	banrongjun1101@gmail.com	China	University of Science and Technology of China	
Huan	Zhang	likemoonriver@126.com	China	University of Science and Technology of China	
Yuanwei	Zhang	zyuanwei@ustc.edu.cn	China	University of Science and Technology of China	✓

Quality assessment of high-throughput DNA sequencing data via range analysis

Keywords: sequencing data quality assessment, range selection, quality scores, high-throughput sequencing, DNA sequencing platforms

Abstract: Motivation:

In the recent literature there appeared a number of studies for the quality assessment of sequencing data. These efforts, to a great extent, focused on reporting the statistical parameters regarding to the distribution of the quality scores and/or the base-calls in a FASTQ file. We investigate another dimension for the quality assessment motivated with the fact that reads including long intervals having fewer errors improve the performances of the post-processing tools in the down-stream analysis. Thus, the quality assessment procedures proposed in this study aim to analyze the segments on the reads that are above a certain quality. We define an interval of a read to be of desired quality when there are at most k quality scores less than or equal to a threshold value v, for some v and k provided by the user.

Results:

We present the algorithm to detect those ranges and introduce new metrics computed from their lengths. These metrics include the mean values for the longest, shortest, average, cubic average, and average variation coefficient of the fragment lengths that are appropriate according to the v and k input parameters. We provide a new software tool QASDRA for quality assessment of sequencing data via range analysis. QASDRA, implemented in Python, and publicly available at https://github.com/ali-cp/QASDRA.git, creates the quality assessment report of an input FASTQ file according to the user specified k and v parameters. It also has the capabilities to filter out the reads according to the metrics introduced.

Contact: kulekci@itu.edu.tr

Availability: http://www.busillis.com/o_kulekci

Authors:

first name	last name	email	country	organization	corresponding?
M. Oguzhan	Kulekci	oguzhankulekci@gmail.com	Turkey	Istanbul Technical University - Informatics Institute	✓
Ali	Fotouhi	fotouhi@itu.edu.tr	Turkey	Istanbul Technical University, Biomedical Engineering Program	
Mina	Majidi	minamajidi22@gmail.com	Iran	Dept. of Mathematics, University of Tehran	

Representing Genetic Determinants in Bacterial GWAS with Compacted De Bruijn Graphs

Keywords: Bacterial GWAS, Antibiotics microbial resistance, De Bruijn graph, Accessory genome, kmer

Abstract: Motivation: Antimicrobial resistance has become a major worldwide public health concern, calling for a better characterization of existing and novel resistance mechanisms. GWAS methods applied to bacterial genomes have shown encouraging results for new genetic marker discovery. Most existing approaches either look at SNPs obtained by sequence alignment or consider sets of kmers, whose presence in the genome is associated with the phenotype of interest. While the former approach can only be performed when genomes are similar enough for an alignment to make sense, the latter can lead to redundant descriptions and to results which are hard to interpret. Results: We propose an alignment-free GWAS method detecting haplotypes of variable length associated to resistance, using compacted De Bruijn graphs. Our representation is flexible enough to deal with very plastic genomes subject to gene transfers while drastically reducing the number of features to explore compared to kmers, without loss of information. It accomodates polymorphisms in core genes, accessory genes and noncoding regions. Using our representation in a GWAS leads to the selection of a small number of entities which are easier to visualize and interpret than fixed-length kmers. We illustrate the benefit of our approach by describing known as well as potential novel determinants of antimicrobial resistance in P. aeruginosa, a pathogenic bacteria with a highly plastic genome. Availability and implementation: The code and data used in the experiments are available at ftp://pbil.univ-lyon1.fr/pub/datasets/jacob/supp-bioinfo-dbgNodes.tgz under the GPL licence Contact: magali.dancette@biomerieux.com Supplementary information: Supplementary data are available at Bioinformatics online.

Authors:

first name	last name	email	country	organization	corresponding?
Magali	Jaillard	magali.dancette@biomerieux.com	France	Biomérieux	✓
Maud	Tournoud	maud.tournoud@biomerieux.com	France	Biomérieux	
Leandro	Lima	leandro.ishi-soares-de-lima@etu.univ-lyon1.fr	France	Université Lyon 1	
Vincent	Lacroix	vincent.lacroix@univ-lyon1.fr	France	Université Lyon 1	
Jean-Baptiste	Veyrieras	jb.veyrieras@gmail.com	France	Biomérieux	
Laurent	Jacob	laurent.jacob@univ-lyon1.fr	France	Université Lyon 1/CNRS	

Identification of epigenetic control markers for predicting stem cell pluripotency

Keywords: pluripotency, embryonic stem cell, regression modeling

Abstract: Abstract

Motivation: To understand the molecular mechanisms associated with cellular differentiation, re-search on pluripotent stem cells has become of great importance, and many studies of stem cells have been conducted in an effort to develop regenerative medicines. Studying genome-wide epi-genetic landscapes through DNA methylation offers significant insight into understanding the het-erogeneity of the pluripotent embryonic stem cells (ESCs).

Results: We constructed a predictive model of the pluripotency of single mouse ESCs using tran-scriptome and methylome sequencing data. The constructed model enables us to understand the extent to which each single ESC has more pluripotency or exists in a more differentiated status. In this study, 22 genomic loci were selected, and those methylation levels were used in our predictive model. The Pearson's correlation coefficient between the cell pseudo-time and the predicted pseudo-time is 0.76 according to an internal validation and RMSE is 18.75. In addition, compared to independent dataset of mESCs grown in serum or in a 2i medium using predicted model, the AUC value is 0.978. Our results show that exploring the epigenetic differences of each ESC can demonstrate different biological states.

Authors:

first name	**last name**	**email**	**country**	**organization**	**corresponding?**
Soobok	Joe	soobok@gist.ac.kr	South Korea	School of Electrical Engineering and Computer Science, Gwangju Institute of Science and Technology	
Hojung	Nam	hjnam@gist.ac.kr	South Korea	School of Electrical Engineering and Computer Science, Gwangju Institute of Science and Technology	✓

Iteratively Adjusted Surrogate Variable Analysis (IA- SVA) uncovers hidden heterogeneity in bulk and single cell transcriptomic data

Keywords: hidden factors, batch correction, gene expression, surrogate variable analysis, single cell RNA-Seq, hidden heterogeneity, clustering

Abstract: Motivation: High-throughput sequencing data typically harbor unwanted variation from diverse sources. Existing statistical methods for parsing the sources of unwanted variation assume that these multiple sources are uncorrelated with each other, an assumption that is frequently not met in sequencing data due to poor experimental design or technical limitations. We present a statistical framework to uncover hidden sources of variation even when these sources are cor- related, namely Iteratively Adjusted Surrogate Variable Analysis (IA-SVA). IA-SVA provides a flexible methodology to i) identify a hidden factor for unwanted heterogeneity while adjusting for all known factors; ii) test the significance of the putative hidden factor for explaining the variation in the data; iii) adjust the data for the detected factor if the factor is significant; and iv) iterate the procedure to uncover further potentially correlated hidden factors.

Results: Using simulated and real-world bulk and single-cell RNA-Seq data, we studied the efficacy of IA-SVA for un- covering sources of unwanted variation in transcriptomic data and compared against existing supervised (i.e., methods based on a control set of genes) and unsupervised methods. IA-SVA outperformed alternative methods in terms of statistical power, Type I error rate, and accuracy in detecting/estimating the hidden factors and proved to be an effective method in the absence of a negative control set. As a case study, we applied IA-SVA to uncover variation in single cell RNA-Seq data from human islets and showed that our method can capture cell types within a cell composition with high accuracy and detect variation that only affects a subset of alpha cells due to the high expression of a small number of genes.

Availability: An R package for IA-SVA with example case scenarios is freely available from https://github.com/UcarLab Contact: donghyung.lee@jax.org and duygu.ucar@jax.org

Supplementary information: Supplementary Data are available at Bioinformatics online.

Authors:

first name	last name	email	country	organization	corresponding?
Donghyung	Lee	donghyung.lee@jax.org	USA	The Jackson Laboratory for Genomic Medicine	✓
Anthony	Cheng	anthony.cheng@jax.org	USA	The Jackson Laboratory for Genomic Medicine	
Duygu	Ucar	duygu.ucar@jax.org	USA	The Jackson Laboratory for Genomic Medicine	✓

Improving the performance of minimizers and winnowing schemes

Keywords: minimizers, winnowing, de Bruijn graph, universal hitting set

Abstract: Motivation: The minimizers scheme is a method for selecting k-mers from sequences. It is used in many bioinformatics software tools to bin comparable sequences or to sample a sequence in a deterministic fashion at approximately regular intervals, in order to reduce memory consumption and processing time. Although very useful, the minimizers selection procedure has undesirable behaviors (e.g., too many k-mers are selected when processing certain sequences). Some of these problems were already known to the authors of the minimizers technique, and the natural lexicographic ordering of k-mers used by minimizers was recognized as their origin. Many software tools using minimizers employ ad hoc variations of the lexicographic order to alleviate those issues. Results: We provide a more in-depth analysis of the effect of the k-mer ordering on the performance of the minimizers technique. By using small universal hitting sets (a recently defined concept), we show how to significantly improve the performance of minimizers and avoid some of its worse behaviors. Based on these results, we encourage bioinformatics software developers to use an ordering based on a universal hitting set or, if not possible, a randomized ordering, rather than the lexicographic order. This analysis also settles negatively a conjecture (by Schleimer et al.) on the expected density of minimizers in a random sequence.

Availability: http://people.csail.mit.edu/yaronore/

Authors:

first name	last name	email	country	organization	corresponding?
Guillaume	Marcais	gmarcais@cs.cmu.edu	USA	Carnegie Mellon University	✓
David	Pellow	d.pellowdavid@gmail.com	Israel	Tel Aviv University	
Daniel	Bork	DKB34@pitt.edu	USA	University of Pittsburgh	
Yaron	Orenstein	yaronore@mit.edu	USA	MIT	
Ron	Shamir	rshamir@post.tau.ac.il	Israel	School of Computer Science, Tel Aviv University	
Carl	Kingsford	carlk@cs.cmu.edu	USA	Carnegie Mellon University	✓

Poly-Harsh: A Gibbs-sampling based Algorithm for Polyploid Haplotype Phasing

Keywords: Polyploid Haplotype Phasing, Gibbs Sampling, Minimum Error Correction, Alternative Sampling

Abstract: Inference of haplotypes, or the sequence of alleles along the same chromosomes, is a fundamental problem in genetics and is a key component for many analyses including admixture mapping, identifying regions of identity by descent and imputation. Haplotype phasing based on sequencing reads has attracted lots of attentions. Diploid haplotype phasing where the two haplotypes are complimentary have been studied extensively. In this work, we focused on Polyploid haplotype phasing where we aim to phase more than two haplotypes at the same time from sequencing data. The problem is much more complicated as the search space becomes much larger and the haplotypes do not need to be complimentary any more. We proposed a Gibbs Sampling based algorithm Poly-Harsh which alternatively samples haplotypes and the read assignments to minimize the mismatches between the reads and the phased haplotypes. Our experiments on both simulated data and real data showed that our method achieves a better performance than the state-of-the-art methods.

Availability: http://www.cs.ucla.edu/~dhe

Authors:

first name	last name	email	country	organization	corresponding?
Dan	He	dhe@us.ibm.com	USA	IBM T.J. Watson	✓

DiscoSnp++: de novo detection of small variants from raw unassembled read set(s)

Keywords: variant calling, reference free, VCF, de Bruijn Graph, SNP, indel

Abstract: Motivation: Next Generation Sequencing (NGS) data provide an unprecedented access to life mechanisms. In particular, these data enable to detect polymorphisms such as SNPs and indels. As these polymorphisms represent a fundamental source of information in agronomy, environment or medicine, their detection in NGS data is now a routine task. The main methods for their prediction usually need a reference genome. However, non-model organisms and highly divergent genomes such as in cancer studies are extensively investigated.

Results: We propose DiscoSnp++, in which we revisit the DiscoSnp algorithm. DiscoSnp++ is designed for detecting and ranking all kinds of SNPs and small indels from raw read set(s). It outputs files in fasta and VCF formats. In particular, predicted variants can be automatically localized afterwards on a reference genome if available. Its usage is extremely simple and its low resource requirements make it usable on common desktop computers. Results show that DiscoSnp++ performs better than state-of-the-art methods in terms of computational resources and in terms of results quality. An important novelty is the de novo detection of indels, for which we obtained 99% precision when calling indels on simulated human datasets and 90% recall on high confident indels from the Platinum dataset.

License: GNU Affero general public license

Availability: http://www.irisa.fr/symbiose/pierre_peterlongo

Authors:

first name	last name	email	country	organization	corresponding?
Pierre	Peterlongo	pierre.peterlongo@inria.fr	France	EPI Symbiose, Centre INRIA Rennes-Bretagne-Atlantique, France	✓
Chloé	Riou	chloe.riou@inria.fr	France	INRIA	
Erwan	Drezen	erwandrezen@yahoo.fr	France	CHU Pontchaillou	
Claire	Lemaitre	claire.lemaitre@inria.fr	France	INRIA	

ReMILO: reference assisted misassembly detection algorithm using short and long reads

Keywords: Misassembly errors, Reference genome, PacBio long reads

Abstract:

Motivation: Contigs assembled from the second generation sequencing short reads may contain misassembly errors, and thus complicate downstream analysis or even lead to incorrect analysis results. Fortunately, with more and more sequenced species available, it becomes possible to use the reference genome of a closely related species to detect misassembly errors. In addition, long reads of the third generation sequencing technology have been more and more widely used, and can also help detect misassembly errors.

Results: Here, we introduce ReMILO, a reference assisted misassembly detection algorithm that uses both short reads and PacBio SMRT long reads. ReMILO aligns the initial short reads to both the contigs and reference genome, and then constructs a novel data structure called red black multipositional de Bruijn graph to detect misassembly errors. In addition, ReMILO also aligns the contigs to long reads and find their differences from the long reads to detect more misassembly errors. In our performance tests on contigs assembled from short reads of human chromosome 14 and japonica rice data, ReMILO can detect 16.1-84.1% extensive misassembly errors and 14.0-100.0% local misassembly errors. On hybrid A. thaliana contigs assembled from both short and long reads, ReMILO can also detect 10.7-25.7% extensive misassembly errors and 8.7-23.8% local misassembly errors.

Availability: The ReMILO software can be downloaded for free from this site:
https://github.com/songc001/remilo

Authors:

first name	last name	email	country	organization	corresponding?
Ergude	Bao	baoe@bjtu.edu.cn	China	School of Software Engineering, Beijing Jiaotong University	✓
Changjin	Song		China	School of Software Engineering, Beijing Jiaotong University	
Lingxiao	Lan		China	School of Software Engineering, Beijing Jiaotong University	

Modelling haplotypes with respect to reference cohort variation graphs

Keywords: Variation graph, Sequence graph, Haplotype, Recombination

Abstract: Current statistical models of haplotypes are limited to panels of haplotypes whose genetic variation can be represented by arrays of values at linearly ordered bi- or multiallelic loci. These methods cannot model structural variants or variants that nest or overlap. A variation graph is a mathematical structure that can encode arbitrarily complex genetic variation. We present the first haplotype model that operates on a variation graph-embedded population reference cohort. We describe an algorithm to calculate the likelihood that a haplotype arose from this cohort through recombinations and demonstrate time complexity linear in haplotype length and sublinear in population size. We furthermore demonstrate a method of rapidly calculating likelihoods for related haplotypes. We describe mathematical extensions to allow modelling of mutations. This work is an essential step forward for clinical genomics and genetic epidemiology since it is the first haplotype model which can represent all sorts of variation in the population.

Authors:

first name	last name	email	country	organization	corresponding?
Yohei	Rosen	yohei@ucsc.edu	USA	University of California, Santa Cruz	✓
Jordan	Eizenga	jeizenga@ucsc.edu	USA	University of California Santa Cruz	
Benedict	Paten	benedictpaten@gmail.com	USA	UCSC	✓

2-Way Gaussian Mixtures as a Model for Microbiome Samples

Keywords: Microbiome, Expectation-Maximization, Clustering, Gaussian Mixture, Admixture

Abstract: Microbiome sequencing allows defining clusters of samples with shared composition.

However, this paradigm poorly accounts for samples whose composition is a mixture of cluster-characterizing ones, and therefore lie in between them in cluster space.

This paper addresses unsupervised learning of 2-way Gaussian mixtures.

It defines a Gaussian mixture model that allows 2-way cluster assignment and describes a variant of expectation-maximization for learning such a model. We demonstrate applicability to microbial 16S rDNA sequencing data from the Human Vaginal Microbiome Project.

Availability: http://www.cs.columbia.edu/~itsik

Authors:

first name	last name	email	country	organization	corresponding?
Weston J	Jackson	wjj2106@columbia.edu	USA	Columbia University	
Ipsita	Agarwl	ia2337@columbia.edu	USA	Columbia University	
Itsik	Pe'Er	itsik@cs.columbia.edu	USA	Columbia University	✓

High-speed and high-ratio lossless compression for NGS genomes

Keywords: genome compression, referential compression, DNA sequence compression, algorithm

Abstract: Motivation: Next-generation sequencing (NGS) technologies have produced huge amounts of genomic data which are revolutionizing biomedical research. But they are accompanied by problems in data storage, compression and communication. Traditional algorithms are unable to meet the compression challenge due to some intrinsic features of DNA sequences such as small alphabet size, frequent repeats and palindromes. Reference-based lossless compression, by which only the differences between two similar genomes are stored, is a promising approach with potentially high compression ratio.

Results: We have developed here a high-performance referential genome compression algorithm, HiRGC. It is based on a 2-bit encoding scheme and an advanced greedy-matching strategy operating on a hash table. The performance of HiRGC is compared on six benchmark human genome data sets with four state-of-the-art compression algorithms. HiRGC uses less than 15 minutes to compress about 15 gigabytes of each set of five target genomes into less than 100 megabytes, achieving a compression ratio of 150 times. This speed is significant faster (at least 4 times) than the state-of-the-art algorithms, and the compression ratio is improved by one to two orders of magnitude. HiRGC also exhibits a very stable and robust performance when tested on different reference genomes, greatly narrowing down the wide performance variation of other methods reported in the literature.

Authors:

first name	last name	email	country	organization	corresponding?
Yuansheng	Liu	yyuanshengliu@gmail.com	Australia	University of Technology Sydney	✓
Hui	Peng	Hui.Peng-2@student.uts.edu.au	Australia	University of Technology Sydney	
Limsoon	Wong	wongls@comp.nus.edu.sg	Singapore	National University of Singapore	
Jinyan	Li	Jinyan.Li@uts.edu.au	Australia	University of Technology Sydney	

DecontaMiner: a tool for the investigation of contaminating sequences in human NGS data

Keywords: NGS, contaminating sequences, unmapped reads

Abstract:

Motivation: Next Generation Sequencing (NGS) experiments produce millions of short sequences that, mapped to a reference genome, provide biological insights at genomic, transcriptomic and epigenomic level. Nonetheless, a variable number of reads fails to correctly align to the reference. In most of the cases this failure is due to the low quality of the bases called during the sequencing, but very often this 'misalignment' is due to sequence differences between the reads and the corresponding genome. Investigating the provenance of these reads is definitely important to better assess the quality of the whole experiment and to look for possible downstream or upstream 'contamination' from exogenous nucleic acids. DecontaMiner is a tool designed and developed for the detection and analysis of these contaminating sequences. Our aim is to help researcher in obtaining more information from the data, and, in particular, to check for microorganisms presence that can not only affect the reliability of the whole experiment, but also foster the evaluation of the samples and the conditions under an additional perspective.

Results: DecontaMiner is a novel tool, that, through a subtraction approach, detects contaminating sequences among the reads discarded from the alignment to the reference genome. It provides a pipeline that takes in input the files containing the unmapped reads, allowing a post-alignment investigation that can be integrated to the standard procedures used in NGS data analysis. The functionality and utility of this tool have been tested on RNA-seq data.

Availability: DecontaMiner is available as command-line tool at the URL:
http://www-labgtp.na.icar.cnr.it/DecontaMiner

Authors:

first name	last name	email	country	organization	corresponding?
Ilaria	Granata	ilaria.granata@na.icar.cnr.it	Italy	National Research Council	✓
Mara	Sangiovanni	mara.sangiovanni@icar.cnr.it	Italy	ICAR-CNR	✓
Amarinder	Thind	thind.amarinder@gmail.com	Italy	ICAR-CNR	
Mario Rosario	Guarracino	mario.guarracino@cnr.it	Italy	ICAR-CNR	

Eliminating redundancy among protein sequences using submodular optimization

Keywords: optimization, submodular optimization, representative set selection, sequence clustering

Abstract: Motivation: Removing redundancy in protein sequence data sets by selecting a representative subset of sequences is a common step in many bioinformatics workflows, such as the creation of non-redundant training sets for sequence and structural models or selection of "operational taxonomic units" from metagenomics data. Previous methods for this task apply a heuristic threshold-based algorithm. This threshold algorithm attempts to optimize an objective function that may not accurately represent the quality of a representative set, and the algorithm is not guaranteed to produce a good solution relative to this objective function. We propose an new approach based on submodular optimization. Submodular optimization, a discrete analogue to continuous convex optimization, has been used with great success for other representative set selection problems.

Results: We demonstrate that the submodular optimization approach results in representative protein sequence subsets with greater structural diversity than sets chosen by existing methods, using as a gold standard the SCOPe library of protein domain structures. In this setting, submodular optimization consistently yields protein sequence subsets that include more SCOPe domain families than sets of the same size selected by competing approaches. We also show how the optimization framework allows us to design a mixture objective function that performs well for both large and small representative sets. The framework we describe is theoretically optimal under some assumptions, and it is flexible and intuitive because it applies generic methods to optimize one of a variety of objective functions.

Availability: http://noble.gs.washington.edu/~maxwl/

Authors:

first name	last name	email	country	organization	corresponding?
Maxwell	Libbrecht	maxwl@cs.washington.edu	USA	University of Washington	
Jeffrey	Bilmes	bilmes@ee.washington.edu	USA	University of Washington	
William	Noble	william-noble@uw.edu	USA	University of Washington	✓

Scalable Genomic Assembly through Parallel de Bruijn Graph Construction for Different K-mers

Keywords: de Bruijn graph, iterative graph construction, de novo genome assembly, parallel assembler, multi-level parallelism

Abstract: Extraordinary progress in genome sequencing technologies has brought down the cost of sequencing a human genome by a factor of 450,000 since the first finished human genome was generated by the Human Genome Project in 2003. It has also led to a tremendous increase in the number of sequenced genomes. However, the lack of efficient de novo assembly algorithms to put together the sequences coming from the sequencing instruments is hindering research efforts to determine genetic variation and its mapping to human diseases. Among several approaches to assembly, the iterative de Bruijn graph assemblers, such as IDBA, generates high-quality assemblies by sequentially iterating from small to large k values used in graph construction. However, this approach is time intensive because the creation of the graphs for increasing k-values proceeds sequentially. For example, with just sixteen k-values, the graph constructions take 95% of the total time to assemble an SAR 324 genome with 55M paired-end reads. In this paper, we propose BootStrapDBG, which transforms the sequential process of de Bruijn graph construction for a range of k values, to one where each graph is built independently and in parallel. We develop a novel mechanism whereby the graph for the higher k value can be "patched" with contigs generated from the graph with the lower k value. We show that our technique achieves higher parallelism and similar accuracy as IDBA for the assembled genome, for a variety of datasets. Morever, BootStrapDBG's multi-level parallelism allows it to simultaneously leverage the power of mighty server machines by using all its cores and of compute clusters by scaling out.

Authors:

first name	last name	email	country	organization	corresponding?
Kanak	Mahadik	kmahadik@purdue.edu	USA	Purdue University	✓
Christopher	Wright	christopherwright@purdue.edu	USA	Purdue University	✓
Saurabh	Bagchi	sbagchi@purdue.edu	USA	Purdue University	✓
Milind	Kulkarni	milind@purdue.edu	USA	Purdue University	✓
Somali	Chaterji	schaterji@purdue.edu	USA	Purdue University	✓

Large-scale Population Genotyping from Low-coverage Sequencing Data using a Reference Panel

Keywords: genotype calling, population genomics, human genetic variation, whole-genome sequencing, low-coverage sequencing data, large-scale population genotyping

Abstract: In recent years, several large-scale whole-genome sequencing projects were completed, including the 1000 Genomes Project, and the UK10K Cohorts Project. These projects aim to provide high-quality genotypes for a large number of whole genomes in a cost-efficient manner, by sequencing each genome at low coverage and subsequently identifying alleles jointly in the entire cohort. The resulting variant data are critical for the broad characterization of human genome variation within and across populations in the original projects, supporting many downstream applications, such as genome-wide association studies. The same datasets carry the potential to increase the quality of genotype calling in other low-coverage sequencing data in future sequencing projects, because the existing genotype calls capture the linkage disequilibrium structures of the cohorts they represent. In this paper we present reference-based Reveel (or Ref-Reveel in short), a novel method for large-scale population genotyping. Ref-Reveel is based on our earlier method, Reveel, which has been demonstrated to be an effective tool for variant calling from low-coverage sequencing data. Our new method introduces several novel algorithmic and design improvements, leveraging genotype calls from a sequenced cohort to enhance the genotyping accuracy of new datasets, while main-taining high computational efficiency. In particular, Ref-Reveel learns the most effective yet low computational cost linkage disequilibrium metrics and genotype inference model for every marker directly from the input data. We show that using a reference panel improves the quality of genotype calling via extensive experiments on simulated as well as real whole-genome data.

Authors:

first name	last name	email	country	organization	corresponding?
Lin	Huang	linhuang@cs.stanford.edu	USA	Stanford University	✓
Petr	Danecek	pd3@sanger.ac.uk	United Kingdom	Wellcome Trust Sanger Institute	
Sivan	Bercovici	sberco@gmail.com	USA	Stanford University	
Serafim	Batzoglou	serafim@cs.stanford.edu	USA	Stanford University	✓

GapReduce: a gap filling algorithm based on partitioned read sets

Keywords: gap filling, de bruijn graph, paired reads

Abstract: With the advances in technologies of sequencing and assembly, draft sequences of more and more genomes are available. However, there commonly exist gaps in these draft sequences which influence various downstream analysis of biological studies. Gap filling methods can shorten the length of gaps and improve the completion of these draft sequences of genomes. Although some gap filling tools have been presented, their effectiveness and accuracy need to be improved. In this study, we develop a novel tool, called GapReduce, which can fill the gaps using the paired reads. For a gap, GapReduce selects the reads whose mate reads are aligned on the left or the right flanking region, and partitions the reads to two sets. Then GapReduce adopts different *k* values and k-mer frequency thresholds to iteratively construct de bruijn graphs, which are used for finding the correct path to fill the gap. Aiming to overcome the branching problems caused by repetitive regions and sequencing errors in the procedure of path selection, GapReduce designs a novel approach simultaneously considering *k-mer* frequency and distribution of paired reads based on the partitioned read sets. We compare the performance of GapReduce with current popular gap filling tools. The experimental results demonstrate that GapReduce can produce more satisfactory gap filling results.

Availability: http://netlab.csu.edu.cn/WJX/index.html

Authors:

first name	last name	email	country	organization	corresponding?
Junwei	Luo	luojunwei@csu.edu.cn	China	Central South University	
Jianxin	Wang	jxwang@mail.csu.edu.cn	China	Central South University	✓
Juan	Shang	908710583@qq.com	China	Central South University	
Huimin	Luo	luohuimin@csu.edu.cn	China	Central South University	
Min	Li	limin@csu.edu.cn	China	Central South University	
Fangxiang	Wu	faw341@mail.usask.ca	Canada	University of Saskatchewan	
Yi	Pan	yipan@gsu.edu	USA	Georgia State University	

The conservation of DNA replication origins between S. cerevisiae and S. uvarum

Keywords: ARS, DNA replication origins, sensu strict, EM, origin conservation, ACS

Abstract: Motivation: Eukaryotic chromosomes initiate DNA synthesis from multiple replication origins. The sites where the replication initiation proteins bind have diverged significantly and little is known about their evolution.

Results: We construct a comprehensive map of \bay\ origins and demonstrate that in spite of the redundant nature of the entire set of origins, the set as a whole is under selective pressure. Moreover, we demonstrate a mechanism of origin conservation even as the specific binding site of the replication initiation proteins is not conserved.

Availability: Raw sequencing and microarray data will be made available through SRA, GEO, and PUMA Databases upon publication.

Contact: uri@maths.usyd.edu.au and maitreya@uw.edu

Supplementary information: Supplementary data are available at Bioinformatics online.

Authors:

first name	last name	email	country	organization	corresponding?
Ivan	Liachko	liachko.ivan@gmail.com	USA	University of Washington	
David	Manescu	dman2626@uni.sydney.edu.au	Australia	University of Sydney	
Gina	Alvino	alvino@uw.edu	USA	University of Washington	
Brona	Brejova	brejova@dcs.fmph.uniba.sk	Slovakia	Comenius University	
Tomas	Vinar	tomas.vinar@fmph.uniba.sk	Slovakia	Comenius University	
Thomas	Pohl	tpohl@princeton.edu	USA	University of Washington	
M. K.	Raghuraman	raghu@u.washington.edu	USA	University of Washington	
Bonita	Brewer	bbrewer@uw.edu	USA	University of Washington	
Maitreya	Dunham	maitreya@uw.edu	USA	University of Washington	✓
Uri	Keich	uri@maths.usyd.edu.au	Australia	University of Sydney	✓

HGT-ID: An efficient and sensitive workflow to detect human-viral insertion sites using next-generation sequencing data

Keywords: Viral integration, Cancer, Horizontal gene transfer, Next generation sequencing

Abstract:

Motivation: Transfer of genetic material from microbes or viruses into the host genome is known as horizontal gene transfer (HGT). The integration of viruses into the human genome is associated with multiple cancers.

Methods: We designed a novel computational workflow (HGT-ID) to identify the integration of viruses into the human genome using the sequencing data. The HGT-ID workflow primarily follows a four-step procedure: i) pre-processing of unaligned reads, ii) virus detection using subtraction approach, iii) identification of virus integration site using discordant and soft-clipped reads and iv) HGT candidates prioritization through a scoring function. Annotation and visualization of the events, as well as primer design for experimental validation are also provided in the final report.

Results: We evaluated the tool performance with the well-understood cervical cancer samples. The workflow accurately detected known human papilloma virus (HPV) integration sites with high sensitivity and specificity compared to previous HGT methods. We applied HGT-ID to The Cancer Genome Atlas (TCGA) whole-genome sequencing data (WGS) from liver tumor-normal pairs. Multiple hepatitis B virus (HBV) integration sites were identified in TCGA liver samples and confirmed by HGT-ID using the RNA-Seq data from the matched liver pairs. This shows the applicability of the method in both the data types and cross-validation of the HGT events in liver samples. We also processed 220 breast tumor WGS data through the workflow; however, there were no HGT events detected in those samples.

Availability: The HGT-ID workflow is available at
http://bioinformaticstools.mayo.edu/research/hgt-id.

Contact: Baheti.Saurabh@mayo.edu

Authors:

first name	last name	email	country	organization	corresponding?
Baheti	Saurabh	Baheti.Saurabh@mayo.edu	USA	Mayo Clinic	
Xiaojia	Tang	Tang.Xiaojia@mayo.edu	USA	Mayo Clinic	
Daniel	O'Brien	OBrien.Daniel@mayo.edu	USA	Mayo Clinic	
Nicholas	Chia	Chia.Nicholas@mayo.edu	USA	Mayo Clinic	
Lewis	Roberts	Roberts.Lewis@mayo.edu	USA	Mayo Clinic	
Heidi	Nelson	nelsonh@mayo.edu	USA	Mayo Clinic	
Judy	Boughey	boughey.judy@mayo.edu	USA	Mayo Clinic	
Liewei	Wang	Wang.Liewei@mayo.edu	USA	Mayo Clinic	
Matthew	Goetz	Goetz.Matthew@mayo.edu	USA	Mayo Clinic	
Jean-Pierre	Kocher	kocher.jeanpierre@mayo.edu	USA	Mayo Clinic	
Krishna	Kalari	kalari.krishna@mayo.edu	USA	Mayo Clinic	✓

deBGR: An Efficient and Near-Exact Representation of the Weighted de Bruijn Graph

Keywords: de Bruijn graph, transcriptome, Bloom filter, High Throughput Sequencing (HTS), kmer, de novo

Abstract: Motivation: Almost all short-read genome and transcriptome assemblers start by building a representation of the de Bruijn Graph of the reads they are given as input. Even when other approaches are used for subsequent assembly (e.g., when one is using "long read" technologies like those offered by PacBio or Oxford Nanopore), efficient k-mer processing is still crucial for accurate assembly, and state-of-the-art long read error correction methods use de Bruijn Graphs.

Because of the centrality of de Bruijn Graphs, researchers have proposed numerous methods for representing de Bruijn Graphs compactly. Many of these proposals sacrifice accuracy to save space. None of these methods store abundance information, i.e. the number of times that each k-mer occurs, which is key in transcriptome assemblers.

Results: We present a method for compactly representing the weighted de Bruijn Graph (i.e. with abundance information) with almost no errors. Our representation yields five-orders-of-magnitude reduction in the number of errors made by a state-of-the-art approximate weighted de Bruijn Graph representation, while increasing the space requirements by less than 10%. Our technique is based on a simple invariant that all weighted de Bruijn Graphs must satisfy, and hence is likely to be of general interest and applicable in most weighted de Bruijn Graph-based systems.

Availability: http://www3.cs.stonybrook.edu/~ppandey/

Authors:

first name	last name	email	country	organization	corresponding?
Prashant	Pandey	ppandey@cs.stonybrook.edu	USA	Stony Brook University	✓
Michael A.	Bender	bender@cs.stonybrook.edu	USA	Stony Brook University	✓
Rob	Johnson	rob@cs.stonybrook.edu	USA	Stony Brook University	✓
Rob	Patro	rob.patro@cs.stonybrook.edu	USA	Stony Brook University	✓

GeneMarkS-2: Raising Standards of Accuracy in Gene Recognition

Keywords: HTP data analysis, gene prediction and annotation, machine learning for bioinformatics algorithms

Abstract: Motivation: Ab initio gene prediction in prokaryotic genomes is supposed to be so accurate that RNA-Seq data are rarely produced to bring in an additional layer of evidence. In 2016 more than 60,000 prokaryotic genomes were re-annotated by the NCBI pipeline. Given the sheer volume of prokaryotic DNA data flowing from next generation sequencing facilities into genome databases, the annotation accuracy should be at the highest level possible. Still, the prevalence of horizontal gene transfer as well as ubiquitous leaderless transcription observed in prokaryotic species call for introducing more complex models of genes and regulatory regions than it was thought to be sufficient earlier.

Results: We describe a new algorithm and software tool GeneMarkS-2. The new multi-model tool has an option to select parameters best matching local genomic GC content that may vary widely due to horizontal gene transfer. Genomes are automatically classified by the inferred types of organization of gene starts neighborhoods which evolution is directed by species specific transcription and translation mechanisms. A new motif search algorithm, LFinder, introduced to reach higher accuracy in detecting conserved motifs in regulatory regions upstream to predicted gene starts uses objective function depending on motif localization. In performance assessments made on test sets validated by proteomics experiments and other sources of evidence we have demonstrated superior accuracy of GeneMarkS-2 in comparison with other state-of-the-art gene prediction tools including GeneMarkS which "plus" version is currently used by the NCBI prokaryotic genome annotation pipeline.

Availability: http://topaz.gatech.edu/GeneMark/genemarks2.cgi

Contact: borodovsky@gatech.edu

Authors:

first name	last name	email	country	organization	corresponding?
Alexandre	Lomsadze	alexandre.lomsadze@bme.gatech.edu	USA	Joint Georgia Tech and Emory Wallace H. Coulter Department of Biomedical Engineering, Georgia Institution of Technology	✓
Shiyuyun	Tang	tangshi06@gatech.edu	USA	Georgia Institute of Technology	✓
Karl	Gemayel	karl@gatech.edu	USA	Georgia Tech	✓
Mark	Borodovsky	borodovsky@gatech.edu	USA	Georgia Tech	✓

Locating CNV candidates in WGS data using wavelet-compressed Bayesian HMM

Keywords: HMM, WGS, CNV, wavelet, Bayesian inference

Abstract: Motivation: Detection of copy-number variants from whole-genome (WGS) read-depth data suffers from confounding factors such as amplification and mapping bias, which are difficult to separate from the signal, and the sheer data size, which mandates simplifying assumptions. Methodological problems concern correspondence between segment means and copy numbers, implicit biases imposed by the modeling, and uncertainty in the number of copy numbers called. Additionally, methods often lack a calibrated measure of uncertainty in their CNV calls owing to computational effort. It is notable that fully Bayesian methods have not been used for WGS data.

Results: We revisit multiplexed sequencing of multiple individuals from two populations as a method to identify recurrent copy number differences. Our novel implementation of Forward-Backward Gibbs (FBG) sampling for Bayesian Hidden Markov Models (HMM) is based on wavelet compression, and can analyze sequences of hundreds of millions of observations in a few minutes. Using two rat populations divergently selected for tame and aggressive behavior as an example, we demonstrate that multiplexed sequencing addresses the bias and normalization problems. Algorithmic improvements include a novel data structure called a breakpoint array. It allows for efficient dynamic compression of the data into blocks for which the underlying mean-signal is constant and without discontinuities. We show that a breakpoint array can be obtained from Haar wavelet regression at arbitrary noise levels in-place and in linear time. We discuss the discovery of several CNV of varying sizes consistent with earlier results concerning the domestication syndrome as well as experimental observations.

Availability: http://schlieplab.org

Authors:

first name	last name	email	country	organization	corresponding?
John	Wiedenhoeft	john@wiedenhoeft.biz	Sweden	Chalmers University of Technology	✓
Alex	Cagan		United Kingdom	Wellcome Trust Sanger Institute	
Rimma	Kozhemjakina		Russia	Institute of Cytology and Genetics of the Russian Academy of Sciences	
Rimma	Gulevich		Russia	Institute of Cytology and Genetics of the Russian Academy of Sciences	
Alexander	Schliep	alexander@schlieplab.org	Sweden	Gothenburg University	

Kullback-Leibler divergence criterion in detecting genome-wide Copy Number Variations

Keywords: Copy Number Variation, Next Generation Sequencing, Kullback-Leibler divergence measure, Jeffrey's divergence measure, Mate pair reads, Bootstrap sampling, k-means clustering algorithm

Abstract: Motivation: Copy Number Variation (CNV) is a type of large structural variation in the human genome. CNVs are envisaged to be associated with cancer, schizophrenia, autism and developmental disabilities. Although some methods are previously developed for the genome-wide CNV detection using next generation sequencing technology, there is a necessity to develop more powerful computational tools for this purpose.

Results: In this study, we introduce a non-parametric statistical tool (KL-CNV) based on information theory for detecting genome-wide CNVs. In the KL-CNV, after mapping mate pairs to the reference genome, insertion size distributions and also read count distributions for the sample and reference genome reads are compared to each other in every genomic segment. For this aim, a Kullback-Leibler divergence measure is applied for modeling discrepancies in the insertion size distributions and also discrepancies in the read count distributions, for the sample and reference genome reads. A high Kullback-Leibler divergence, either in insertion size or in read count distributions, is an indication for the CNV region.

The performance for this new algorithm is compared to the state of the art CNV detection tools in synthetized data. Also, the KL-CNV is applied for detecting genomic regions with copy gain or loss in the real data of HapMap individual NA12878. We also report a set of 6,530 CNV regions that are detected from the NA12878.

Availability: MATLAB programs of KL-CNV are available at https://github.com/CNVdetection/KL-CNV

Authors:

first name	last name	email	country	organization	corresponding?
Seyed Amir	Malekpour	a.malekpour@ut.ac.ir	Iran	University of Tehran	✓
Hamid	Pezeshk	pezeshk@ut.ac.ir	Iran	University of Tehran	✓
Mehdi	Sadeghi	sadeghi@nigeb.ac.ir	Iran	National Institute of Genetic Engineering and Biotechnology	

Tumor Phylogeny Inference Using Tree-Constrained Importance Sampling

Keywords: cancer, DNA sequencing, phylogeny

Abstract: Motivation: A tumor arises from an evolutionary processes that can be modeled as a phylogenetic tree. However, reconstructing this phylogeny is challenging as most cancer sequencing is done using bulk-sequencing data. Thus the observed samples are mixtures of thousands of heterogeneous cells.

Results: We introduce PASTRI (Probabilistic Algorithm for Somatic Tree Inference), a new algorithm for bulk-tumor sequencing data that clusters somatic mutations into clones and infers a phylogenetic tree that describes the evolutionary history of the tumor. PASTRI combines the combinatorial structure described in past work with a probabilistic model of allele counts from bulk-sequencing data. As a result, tree inference is fast, accurate and robust to noise. We demonstrate on simulated data that PASTRI outperforms other phylogenetic reconstruction algorithms in terms of runtime and accuracy. On a chronic lymphocytic leukemia (CLL) patient that had previously been categorized as having a complex branching phylogeny, we find a linear phylogeny that explains the data. PASTRI provides a robust approach for phylogenetic tree inference from mixed samples

Authors:

first name	last name	email	country	organization	corresponding?
Gryte	Satas	gryte_satas@brown.edu	USA	Brown University	✓
Benjamin	Raphael	braphael@cs.princeton.edu	USA	Princeton University	✓

Distributed Variant Calling with Avocado

Keywords: Variant calling, Genotyping, Parallel/Distributed Computing, Apache Spark

Abstract: Motivation: With the increasing size of genomic datasets, computational analysis is a rate limiting step. Variant calling is a particularly expensive step, especially when working with whole genome data. To improve performance, we implement variant calling on top of the horizontally scalable Apache Spark distributed computing framework.

Results: Avocado achieves linear scalability with the size of the computing cluster, while achieving state- of-the-art accuracy in SNP calling, and competitive accuracy at INDEL calling.

Availability: Avocado is open source software, released under the Apache 2 license. The Avocado source code is hosted at https://github.com/bigdatagenomics/avocado.

Availability: http://www.fnothaft.net

Authors:

first name	last name	email	country	organization	corresponding?
Frank	Nothaft	fnothaft@berkeley.edu	USA	UC Berkeley	✓
David	Patterson	pattrsn@berkeley.edu	USA	UC Berkeley	
Anthony	Joseph	adj@berkeley.edu	USA	UC Berkeley	

Identifying structural variants using 10X Genomics' linked-read sequencing data

Keywords: Whole genome sequencing, Variant calling, Structural variation, Probabilistic model

Abstract: Structural variants (SVs) are large-scale differences between an individual genome and a reference genome, and include deletions, duplications, inversions, and translocations. A number of methods have been developed to identify structural variants from short-read whole genome sequencing data. However, SV identification still remains challenging because many SVs have breakpoints in repetitive regions that may not be identified by short reads. A new sequencing technology called linked-read sequencing was recently commercialized by 10X Genomics. This technology encodes long-range linking information into paired-end short-reads by tagging reads originating from the same long molecule 50Kb with a molecular a barcode. We present a new algorithm: Linked-read Structural Variants (LRSV) that utilizes signals unique to linked-read sequencing to identify large structural variants. LRSV combines signals from paired-end reads and linked-reads using a probabilistic model to identify and rank candidate SVs from linked-read sequencing data. We compare LRSV to several other methods, including two recent methods that also analyze linked-reads, on both simulated data and human whole-genome sequencing data. On both simulated and real data, LRSV identifies known SVs with higher recall and precision than other methods.

Authors:

first name	last name	email	country	organization	corresponding?
Rebecca	Elyanow	rebecca_elyanow@brown.edu	USA	Brown University	
Hsin-Ta	Wu	hsin-ta_wu@brown.edu	USA	Brown University	
Benjamin	Raphael	braphael@princeton.edu	USA	Princeton University	✓

Fast accurate sequence alignment using Maximum Exact Matches

Keywords: Sequence Alignment, Algorithm, Dynamic Programming

Abstract: Motivation: Sequence alignment is a central technique in biological sequence analysis, and dynamic programming is widely used to perform an optimal alignment of two sequences. While efficient, dynamic programming is still costly in terms of time and memory when aligning very large sequences.

Results: We describe MEM-Align, an optimal alignment algorithm that focuses on Maximal Exact Matches (MEMs) between two sequences, instead of processing every symbol individually. In its original definition, MEM-Align is guaranteed to find the optimal alignment but its execution time is not manageable unless optimisations are applied that decrease its accuracy. However it is possible to configure these optimisations to balance speed and accuracy. The resulting algorithm outperforms existing solutions such as GeneMyer and Ukkonen. MEM-Align can replace edit distance-based aligners or provide a faster alternative to Smith-Waterman alignment for most of their applications including in the final stage of short read mapping.

Authors:

first name	last name	email	country	organization	corresponding?
Arash	Bayat	a.bayat@unsw.edu.au	Australia	University of New South Wales	✓
Aleksandar	Ignjatovic	ignjat@cse.unsw.edu.au	Australia	University of New South Wales	
Bruno	Gaeta	bgaeta@unsw.edu.au	Australia	University of New South Wales	
Sri	Parameswaran	sri.parameswaran@unsw.edu.au	Australia	University of New South Wales	✓

HopLand: Single-cell pseudotime recovery using continuous Hopfield network based modeling of Waddington's epigenetic landscape

Keywords: Pseudotime recovery, Single-cell data, Continuous Hopfield network, Waddington's epigenetic landscape, Mouse embryonic development

Abstract: Motivation: The interpretation of transcriptome dynamics in single-cell data, especially pseudotime estimation, could help understand the transition of genetic profiles. The recovery of pseudotime increases the temporal resolution of single-cell transcriptional data, but is challenging due to the high variability in gene expression between individual cells. Here, we introduce HopLand, a pseudotime recovery method using continuous Hopfield network to map cells in a Waddington's epigenetic landscape. It reveals the combinatorial regulatory interactions of genes from the data that control the dynamic progression through successive cellular states.

Results: We applied HopLand to different types of single-cell transcriptome data. It achieved a high accuracy of pseudotime prediction compared to existing methods. A kinetic model was extracted from the data. Through the analysis of the model, we identified key genes and regulatory interactions driving the transition of cell states. Therefore, our method has the potential to generate fundamental insights into cell fate regulation.

Availability: http://www.ntu.edu.sg/home/zhengjie/

Authors:

first name	last name	email	country	organization	corresponding?
Jing	Guo	GUOJ0020@e.ntu.edu.sg	Singapore	Nanyang Technological University	
Jie	Zheng	ZhengJie@ntu.edu.sg	Singapore	Nanyang Technological University	✓

Universal Denoising of Aligned Genomic Sequences

Keywords: Bioinformatics, Denoising, High-Throughput Sequencing

Abstract: Advances in sequencing technology have spurred the production of sequencing data at ever increasing speed and volume. Despite the proliferation of new technologies and improvements in sequencing techniques, noise in sequencing data remains a confounding factor in data analysis and interpretation. We propose a denoising scheme that denoises nucleotide bases in the reads while also updating quality score values when necessary. We show that this scheme results in more accurate variant calling in high-coverage datasets.

Authors:

first name	last name	email	country	organization	corresponding?
Irena	Fischer-Hwang	ihwang@stanford.edu	USA	Stanford University	✓
Mikel	Hernaez	mhernaez@stanford.edu	USA	Stanford University	✓
Idoia	Ochoa	idoia@illinois.edu	USA	UIUC	✓
Tsachy	Weissman	tsachy@stanford.edu	USA	Stanford University	

GenomeHash - Database-less Microbial Genotyping

Keywords: Genotyping, Multi-locus sequencing typing, MLST, Bacterial typing, Bacteria, Sequencing, Microbiology

Abstract: Background: Exchanging information about the genotypes of microbial isolates is a cornerstone of many clinical and public-health-related genomics applications. Genotyping methods such as Multilocus Sequence Typing (MLST) and its variants have proven to be useful tools for this purpose, labelling genotypes by the alleles observed at a chosen set of conserved loci. However, as more isolates are sequenced across a wider range of species, curating the database of observed alleles and allele combinations is becoming an increasingly significant bottleneck in the timely communication of novel genotypes Results: We propose a simple method for eliminating the majority of the curation burden through the use of cryptographic hashing and a minimal allele database, allowing a hash of the allele to be used as a self-identifying signature. Combining these hash-based signatures yields a genotypic identifier, allowing unambiguous discourse about the genotypes of microbial isolates with minimal recourse to curated databases. We demonstrate the effectiveness of this approach by showing that the use of standard hashing algorithms result in unambiguous labelling of alleles in practice, and provide a simple reference implementation that replicates the information provided by MLST and cgMLST across four sets of bacterial isolates. Finally, we discuss several implications of enabling allelic genotyping without the creation and maintenance of a large database of observed alleles including improved scalability, easier augmentation of existing schemes and simplifying development of novel schemes. Availability: GenomeHash is open-source, requires R and BLAST, and is available from https://github.com/bgoudey/genomehash. Contact: Address correspondence to Dr. Benjamin Goudey (bgoudey@au1.ibm.com)

Authors:

first name	last name	email	country	organization	corresponding?
Benjamin	Goudey	bwgoudey@gmail.com	Australia	IBM Research - Australia	✓
Hannah	Huckstep		Canada	IBM Research - Australia	
Kelly	Wyres		Australia	IBM Research - Australia	
Thomas	Conway		Australia	IBM Research - Australia	

Data Driven Likelihood Factorizations Improve Lightweight Transcript Abundance Estimation

Keywords: RNA-seq, data-driven factorization, Quantification

Abstract: Motivation: Recently, lightweight methods for transcript abundance estimation have gained considerable favor in the community as an alternative to more traditional (alignment-based) approaches for transcript quantification. These methods rely on techniques such as k-mer counting, pseudo-alignment, and quasi-mapping to produce abundance estimates much more quickly than is typically possible with alignment-based techniques. Crucially, these methods also adopt approximate factorizations of the likelihood function, which simplify calculations at the expense of discarding certain information.

Results: We demonstrate that model simplifications adopted by certain lightweight methods can lead to the inability to accurately estimate the abundances of related transcripts. In particular, considering transcript-fragment compatibility alone can lead to highly sub-optimal estimates. However, we show that such shortcomings are not an inherent limitation of lightweight methods. By considering the appropriate conditional fragment probabilities, and adopting improved data-driven factorizations of the likelihood function, we demonstrate that certain lightweight methods can achieve performance nearly indistinguishable from methods that consider the complete (i.e., per-fragment) likelihood function.

Availability: Our data-driven factorizations are incorporated into the Salmon transcript quantification tool.

Authors:

first name	**last name**	**email**	**country**	**organization**	**corresponding?**
Mohsen	Zakeri	mzakeri@cs.stonybrook.edu	USA	Stony Brook University	✓
Avi	Srivastava	asrivastava@cs.stonybrook.edu	USA	Stony Brook university	✓
Fatemehalsadat	Almodarresi Ts	fatemehalsadat.almodaresits@stonybrook.edu	USA	Stony Brook University	✓
Rob	Patro	rob.patro@cs.stonybrook.edu	USA	Stony Brook University	✓

Graph regularized, semi-supervised learning improves annotation of de novo transcriptomes

Keywords: Semi-supervised learning, Annotation, denovo transcriptome

Abstract: Motivation: Accurate annotation of denovo assembled contigs is important for the interpretation of quantification results and differential expression analyses. Predominant techniques for performing such annotation include running BLAST against extensive databases or using tools which customize the database and search parameters. However, this process often leaves a substantial fraction of contigs unannotated and does not differentiate between contigs that simply have no good annotation, and those that are likely spurious assembly results that should be removed from subsequent analysis.

Results: We introduce a novel methodology, GRASS (Graph Regularized Annotation via Semi-Supervised learning), for the annotation of contigs from a denovo transcriptome assembly, using information from the annotated genome of a closely related species. The graph-based approach makes the shared-sequence relationships between assembled contigs explicit in the form of a graph, and applies an algorithm that performs label propagation to transfer annotations between related contigs and modifies the graph topology iteratively. GRASS also incorporates quantification information to remove spurious contigs from the assembly. Our tests on data from three different species show that GRASS improves the contig annotation and clustering and correctly annotates a larger number of contigs. We find that our proposed pipeline for quantification, filtering, contig clustering, initial annotation and annotation refinement can be completed in a matter of tens of minutes on datasets where other approaches require hours. Finally, we believe that GRASS represents a new conceptual direction for further improving denovo transcriptome annotation and analysis.

Availability: GRASS is written in Python, and is freely-available under an open-source (BSD) license at https://github.com/COMBINE-lab/GRASS

Authors:

first name	last name	email	country	organization	corresponding?
Laraib Iqbal	Malik	laraib.malik@stonybrook.edu	USA	Stony Brook University	✓
Shravya	Thatipally	shravyarani.thatipally@stonybrook.edu	USA	Stony Brook University	
Nikhil	Junneti	nikhil.junneti@stonybrook.edu	USA	Stony Brook University	
Rob	Patro	rob.patro@cs.stonybrook.edu	USA	Stony Brook University	✓

Haplotype resolved assembly and structural variant detection with long-reads

Keywords: Genomics, Structural Variation, Genome Assembly, Long-read sequencing

Abstract: Motivation: The last decade has seen a dramatic expansion in the number of sequenced human genomes. While SNV and indel calling methods have proven highly robust with shortread data, providing high-quality SV calls and high-quality phased genome assemblies has been largely out of reach due to inherent limitations of sequencing technologies. Recent advances in long-read sequencing and barcoding approaches from Pacific Biosciences (PacBio) and 10X Genomics (10X) have the potential to provide fully resolved SV sets for any individual, but require novel tools to fully take advantage of the data.

Results: Here, we present a novel haplotype resolved SV caller, Slap, for combining raw PacBio reads and 10X phased SNV information to provide high-quality assembled haplotigs and SV calls. The tool provides a framework for partitioning reads, assembling haplotypes (via existing software), and converting these assemblies into SV calls with zygosity calls by jointly calling the assembled haplotigs and validating via a novel community detection based approach. The approach yields increased SV coverage relative to other methods while also providing more accurate zygosity information than running haplotigs independently against the reference. This approach represents a framework for haplotype-resolved SV calls and moves one step closer to the goal of completely resolved diploid genomes.

Availability: Slap is available at https://bitbucket.org/oscarlr/slap

Authors:

first name	last name	email	country	organization	corresponding?
Oscar	Rodriguez	oscar.rodriguez@icahn.mssm.edu	USA	MSSM	
Matthew	Pendleton	matthew.pendleton@mssm.du	USA	MSSM	
Alex	Ledger	a.led1027@gmail.com	USA	Reed College	
Anna	Ritz	aritz@reed.edu	USA	Reed College	✓
Ali	Bashir	ali.bashir@gmail.com	USA	MSSM	✓

Predicting genome accessibility with a hybrid deep convolutional neural network

Keywords: convolutional neural network, DNA accessibility, classification and regression, genetic variants

Abstract: Motivation: The majority of known genetic variants associated with human inherited diseases lie in non-coding regions that lack adequate interpretation, making it indispensable to systematically discover functional elements at the whole genome level and precisely decipher their implications in a comprehensive manner. Although computational approaches have been complementing high-throughput biological experiments towards the annotation of the genome, it still remains a big challenge to characterize the risk of a genetic variant in the context of a specific cell type via automatic learning of the DNA sequence code from large-scale sequencing data. Indeed, the development of an accurate and interpretable model to learn the DNA sequence signature and further enable the identification of causative genetic variants has become essential in both genomic and genetic studies.

Results: We propose DeepNet, a machine learning framework based on the deep convolutional neural network, to automatically learn the regulatory code of a DNA sequence and predict its chromatin accessibility. In a series of comparison with existing methods, we show the superior performance of our model in not only the classification of accessible regions against random sequences, but also the regression of DNase-seq signals. Besides, we find that the integration of features discovered by the convolutional layers and extracted from the k-mer spectrum could further improve the prediction performance. We further visualize the convolutional kernels and show the match of identified sequence signatures and known motifs. We finally demonstrate the sensitivity of our model in finding causative noncoding variants in the analysis of a breast cancer dataset. We expect to see wide applications of DeepNet with either public or in-house chromatin accessibility data in the annotation of the genome and the identification of non-coding variants associated with diseases.

Authors:

first name	last name	email	country	organization	corresponding?
Qiao	Liu	liu-q16@mails.tsinghua.edu.cn	China	Tsinghua University	
Fei	Xia	feixia@stanford.edu	USA	Stanford University	
Qijin	Yin	yinqijin@buaa.edu.cn	China	Beihang University	
Rui	Jiang	ruijiang@tsinghua.edu.cn	China	Tsinghua University	✓

Detecting virus integration sites based on mul-tiple related sequencing data by VirTect

Keywords: High-throughput sequencing, multi-region sequencing, virus integration sites

Abstract: Abstract

Motivation: Since tumor often has a high level of intra-tumor heterogeneity, multiple tumor samples from the same patient at different locations or different time points are often sequenced to study tumor intra-heterogeneity or tumor evolution. In virus-related tumors such as human papillomavirus- and Hepatitis B Virus-related tumors, since virus genome integrations can be critical driving events, an important aspect is to investigate the integration sites of the virus genomes. Currently, a few algorithms for detecting virus integration sites based on high-throughput sequencing have been developed, but their insufficient performance in their sensitivity, specificity and computational complexity greatly hinders their applications in multiple related tumor sequencing.

Results: We develop VirTect for detecting virus integration sites simultaneously from multiple related-sample data. This algorithm is mainly based joint analysis of short reads spanning breakpoints of integration sites from multiple samples. To achieve high specificity and breakpoint accuracy, a local precise sandwich alignment algorithm is used. Simulation and real data analyses show that, compared with other algorithms, VirTect is significantly more sensitive and has similar or lower false discovery rate. In addition, VirTect can provide more accurate breakpoint position and is computationally much more efficient in terms both memory requirement and computational time.

Availability: VirTect is implemented in python and is available at https://github.com/xyc0813/VirTect/.

Contact: ruibinxi@math.pku.edu.cn

Supplementary information: Supplementary data are available at Bioinformatics online.

Authors:

first name	last name	email	country	organization	corresponding?
Yuchao	Xia	xiayuchao@pku.edu.cn	China	School of Mathematics Science, Peking University	
Hongjun	Wang	steves880106@gmail.com	Taiwan	Department of Electrical Engineering, National Taiwan University	
Yun	Liu	liuyun_math@pku.edu.cn	China	School of Mathematics Science, Peking University	
Minghua	Deng	dengmh@math.pku.edu.cn	China	School of Mathematics Science, Peking University	
Ruibin	Xi	ruibinxi@math.pku.edu.cn	China	School of Mathematics Science, Peking University	✓

Meta-TMP: optimized network mining and efficient analysis towards a refined platform for microbiome researches

Keywords: metagenomics, co-ocurrence network, parallel computing

Abstract: Motivation: With the increasing concerns about human health and their association with human microbiome, metagenomics become one of the mainstreams for omics researches. However, faced with the paramount of metagenomic data, current metagenomic data processing methods could still not keep pace, especially in efficiency and ecology network analysis.

Results: In this work, we have proposed the Meta-TMP (the Trans-Meta Platform), which is the completely updated version of the Parallel-Meta pipeline, towards a more effective platform, for more efficient and comprehensive taxonomical and functional analyses of microbial communities. Firstly, species-species co-occurrence network establishment, clustering and visualization modules have been developed and optimized. Secondly, the multi-thread CPU and GPU computation have been optimized for magnitudes of speed-ups. Thirdly, the analytical units in the package were better modularized and with clearer connections by using configuration file, as well as APIs for connecting external databases and tools, towards more flexible operation on metagenome data. Thus, Meta-TMP could facilitate better understanding of ecological patterns presented in metagenomic data, in an efficient manner.

Authors:

first name	last name	email	country	organization	corresponding?
Shaojun	Yu	yushaojun@hust.edu.cn	China	Huazhong University of Science and Technology	
Pengshuo	Yang	u201312508@hust.edu.cn	China	Huazhong University of Science and Technology	
Maozhen	Han	bioinformatics@hust.edu.cn	China	Huazhong University of Science and Technology	
Kang	Ning	ningkang@hust.edu.cn	China	Huazhong University of Science and Technology	✓

Chromatin Accessibility Prediction via Convolutional Long Short-Term Memory Networks with k-mer Embedding

Keywords: chromatin accessibility, k-mer embedding, long short-term memory, deep learning, variable-length input

Abstract: Motivation: Biological experiment techniques for measuring chromatin accessibility are expensive and time consuming, prompting development of computational methods for DNA sequence analysis. Meantime, current methods fall into two classes: one is based on handcrafted k-mer features and the other is based on convolutional neural networks, each having their own limitations. Hence there still lacks a comprehensive deep learning framework that utilizes the useful k-mer co-occurrence information and is capable of handling variable-length input sequences.

Method and results: We fill this gap by proposing a convolutional Long Short-Term Memory (LSTM) network with k-mer embedding, to address the problem of chromatin accessibility prediction. We first split DNA sequences into k-mers and pre-train k-mer embedding vectors based on co-occurrence matrix by unsupervised representation learning. We then construct a supervised deep learning architecture mainly comprised of an embedding layer, three convolutional layers and a Bidirectional LSTM (BLSTM) layer for feature learning and classification tasks. We demonstrate that our method gains high-quality fixed-length features from variable-length sequences and consistently outperforms baseline methods. We show that k-mer embedding can effectively enhance model performance by exploring different embedding strategies. We also prove the efficacy of both the convolution and BLSTM by comparing two variant deep architectures. We confirm the robustness of our model to hyper-parameters by performing sensitivity analysis. Eventually, we hope our method can reinforce our understanding of employing deep learning in genomics study and help shed light on the chromatin accessibility mechanism.

Authors:

first name	last name	email	country	organization	corresponding?
Xu	Min	minxueric@gmail.com	China	Tsinghua University	
Wanwen	Zeng	zengww14@mails.tsinghua.edu.cn	China	Tsinghua University	
Ning	Chen	ningchen@tsinghua.edu.cn	China	Tsinghua University	
Ting	Chen	tingchen@tsinghua.edu.cn	China	Tsinghua University	✓
Rui	Jiang	ruijiang@tsinghua.edu.cn	China	Tsinghua University	✓

Secure Cloud Computing for Pairwise Sequence Alignment

Keywords: privacy and security, cloud computing, sequence alignment, big data

Abstract: The recent exponential growth of biological sequence data sets has the potential to rapidly advance our understanding of life's processes. However, since analyzing biological sequences is a very expensive computing task, users face a formidable challenge in trying to analyze these data on their own (e.g., hundred of thousands of sequences alignments are needed to perform phylogenetic analysis). Cloud computing offers access to a large amount of computing resources in an on-demand and pay-per-use fashion, which is a practical way for people to analyze these huge data sets. Although the advantages of cloud computing are widely recognized, many people are still reluctant to outsource biological sequences to the cloud because they contain sensitive information that should be kept secret for ethical, security, and legal reasons. In this paper, we develop a secure outsourcing algorithm that optimally solves a PSA, which is one of the most fundamental and frequently used computational tools in biological sequence analysis. Specifically, we design a low-complexity sequence transformation that conceals the characters of the sequences. We then develop a secure outsourcing algorithm to solve the PSA under the transformed sequences at the cloud, and efficiently find the solution to the original PSA at the user. We implement the proposed algorithm on the Amazon Elastic Compute Cloud (EC2) and a laptop. We find that our proposed algorithm offers significant time savings compared with previous works.

Authors:

first name	last name	email	country	organization	corresponding?
Sergio	Salinas	salinas@cs.wichita.edu	USA	Wichita State University	✓
Pan	Li	lipan@case.edu	USA	Case Western Reserve University	✓

EAGLE: Explicit Alternative Genome Likelihood Evaluator

Keywords: Genome Variant Calling, Probabilistic Model, DNA Sequencing Data Analysis

Abstract: Motivation: Reliable detection of genome variations, especially insertions and deletions (indels), from single sample DNA sequencing data remains challenging, partially due to the inherent uncertainty involved in aligning sequencing reads to the reference genome.

Results: We present EAGLE, a method for evaluating the degree to which sequencing data supports a given candidate genome variant. EAGLE incorporates candidate variants into explicit hypotheses about the individual's genome, and then computes the probability of the observed data (the sequencing reads) under each hypothesis. In comparison with methods which rely heavily on a particular alignment of the reads to the reference genome, EAGLE readily accounts for uncertainties that may arise from multi-mapping or local misalignment and uses the entire length of each read.

We compared the scores assigned by several well-known variant callers to EAGLE for the task of ranking true putative variants over false ones on both simulated data and real genome sequencing based benchmarks. Overall EAGLE tended to rank true variants higher than the scores reported by the callers. For indels (mostly ¡50bp) EAGLE obtained marked improvement on simulated data and a whole genome sequencing benchmark, and modest but statistically significant improvement on an exome sequencing benchmark.

Authors:

first name	last name	email	country	organization	corresponding?
Tony	Kuo	kuo.tony@aist.go.jp	Japan	AIST, Computational Biology Research Center	
Martin	Frith	mcfrith@edu.k.u-tokyo.ac.jp	Japan	AIST, Computational Biology Research Center	
Jun	Sese	sese.jun@aist.go.jp	Japan	AIST, Computational Biology Research Center	
Paul	Horton	paulh@iscb.org	Japan	AIST, Computational Biology Research Center	✓

Compression of genomic sequencing reads with and without preserving the order: Analysis and improvements

Keywords: FASTQ compression, Read compression, Reordering reads, Entropy of reads, High-throughput sequencing

Abstract: Motivation: New Generation Sequencing (NGS) technologies for genome sequencing produce large amounts of short genomic reads per experiment, which is highly redundant and compressible. However, general-purpose compressors are unable to exploit this redundancy, due to the special structure present in the data. Although numerous specialized FASTQ compressors have been presented in the past few years, there has been no systematic analysis of the read compression problem and its limits in the literature.

Results: In this work, we first analyze the limits of read compression by computing bounds on the entropy of the reads. We then analyze the performance of some of the state-of-the-art read compressors and understand their shortcomings with respect to read compression. Finally, guided by the analysis, we present a simple algorithm for read compression geared towards achieving compression ratios close to the fundamental limit. The algorithm achieves compression ratios which are 1.3-2x better than the state-of-the-art compressors on the analyzed datasets. The algorithm compresses only the read sequence, works with unaligned FASTQ files, and does not need a reference.

Authors:

first name	last name	email	country	organization	corresponding?
Shubham	Chandak	schandak@stanford.edu	USA	Stanford University	✓
Kedar	Tatwawadi	kedart@stanford.edu	USA	Stanford University	✓
Tsachy	Weissman	tsachy@stanford.edu	USA	Stanford University	

MIPUP: Minimum perfect unmixed phylogenies for multi-sampled tumors via branchings and ILP

Keywords: tumor evolution, tumor heterogeneity, tumor subclones, high-throughput sequencing, perfect phylogeny, graph problem, integer linear programming

Abstract: Motivation: Discovering the subpopulations of a tumor and their evolution may help identify driver mutations and provide a more comprehensive view on the history of the tumor. Recent studies have tackled this problem using multiple samples sequenced from a tumor, and due to clinical implications, this has attracted great interest. However, such samples usually mix several distinct tumor subpopulations, which confounds the analysis.

Results: We address two natural problems requiring to decompose the tumors samples into several subclones with the objective of forming a minimum perfect phylogeny. Our method, MIPUP (minimum perfect unmixed phylogenies), exploits a recent theoretical development connecting the two perfect phylogeny problems to two problems on finding branchings in a directed acyclic graph. The latter ones can be solved efficiently using Integer Linear Programming. We tested MIPUP on four real datasets with accurate results, and typically more reliable than those of an existing method.

Availability: MIPUP is available at https://github.com/zhero9/MIPUP as open source.

Authors:

first name	**last name**	**email**	**country**	**organization**	**corresponding?**
Ademir	Hujdurović	ademir.hujdurovic@upr.si	Slovenia	University of Primorska, UP IAM, Koper, Slovenia	✓
Edin	Husić	edinehusic@gmail.com	France	ENS Lyon (also UP FAMNIT, University of Primorska, Koper, Slovenia)	✓
Miika	Mehine	miika.mehine@helsinki.fi	Finland	University of Helsinki	✓
Martin	Milanič	martin.milanic@upr.si	Slovenia	UP IAM and UP FAMNIT, University of Primorska	✓
Romeo	Rizzi	romeo.rizzi@univr.it	Italy	University of Verona, Italy	✓
Alexandru I.	Tomescu	tomescu@cs.helsinki.fi	Finland	Helsinki Institute for Information Technology HIIT, Department of Computer Science, University of Helsinki	✓

Amino Acid Confidence Evaluation to Improve the Quality Control of De Novo Peptide Sequencing and Modification Site Localization

Keywords: tandem mass spectrometry, de novo peptide sequencing, modification site localization, phosphorylation

Abstract: De novo peptide sequencing has been improved remarkably with the development of mass spectrometry and computational approaches but still lacks quality control methods. Here we proposed a novel method pPredictAA to evaluate the confidence of each amino acid rather than the full-length peptides obtained by de novo sequencing. A semi-supervised learning approach is used for discriminating correct amino acids from the random ones. Furthermore, an expectation-maximization algorithm is used to precisely and robustly control the false discovery rate (FDR). On three test data sets, pPredictAA recalled 124.8% more amino acids on average compared with the second-best method at the FDR of 5%, and the estimated FDRs of pPredictAA were very close to the real ones. On the other hand, pPredictAA also performed superiorly on the modification site localization problem, which is essentially a special case of amino acid confidence evaluation. On three phosphorylation enrichment data sets, the recall rate of pPredictAA was on average 9.6% more than that of phosphoRS at the 5% false localization rate. In addition, pPredictAA can cover 94% of the results of phosphoRS and contributed 14% more phosphorylation sites. Further analysis showed that the use of the distinct fragmentation features in the high resolution MS/MS spectra, such as the neutral loss ions, played an important role for the improvement of pPredictAA. In summary, the effective universal model together with the extensive use of spectral information makes pPredictAA an excellent tool for the quality assessment of amino acids in de novo sequencing and modification site localization.

Authors:

first name	**last name**	**email**	**country**	**organization**	**corresponding?**
Hao	Yang	yanghao01@ict.ac.cn	China	Institute of Computing Technololgy, Chinese Academy of Sciences, Beijing	
Hao	Chi	chihao@ict.ac.cn	China	Institute of Computing Technololgy, Chinese Academy of Sciences, Beijing	✓
Chao	Liu	liuchao1016@ict.ac.cn	China	Institute of Computing Technololgy, Chinese Academy of Sciences, Beijing	
Wen-Feng	Zeng	zengwenfeng@ict.ac.cn	China	Institute of Computing Technololgy, Chinese Academy of Sciences, Beijing	
Wen-Jing	Zhou	zhouwenjing@ict.ac.cn	China	Institute of Computing Technololgy, Chinese Academy of Sciences, Beijing	
Zhao-Wei	Wang	wangzhaowei@ict.ac.cn	China	Institute of Computing Technololgy, Chinese Academy of Sciences, Beijing	
Rui-Min	Wang	wangruimin@ict.ac.cn	China	Institute of Computing Technololgy, Chinese Academy of Sciences, Beijing	
Xiu-Nan	Niu	niuxiunan@ict.ac.cn	China	Institute of Computing Technololgy, Chinese Academy of Sciences, Beijing	
Zhen-Lin	Chen	chenzhenlin@ict.ac.cn	China	Institute of Computing Technololgy, Chinese Academy of Sciences, Beijing	
Si-Min	He	smhe@ict.ac.cn	China	Institute of Computing Technololgy, Chinese Academy of Sciences, Beijing	✓

An omnibus test for differential distribution analysis of microbiome sequencing data

Keywords: mcirobiome, differential distribution analysis, statistical test

Abstract: Motivation: One objective of human microbiome studies is to identify differentially abundant microbes across biological conditions. Previous statistical methods focus on detecting the shift in the abundance and/or prevalence of the microbes and treat the dispersion (spread of the data) as a nuisance. These methods also assume that the dispersion is the same across conditions, an assumption which may not hold in presence of sample heterogeneity. Moreover, the widespread outliers in the microbiome sequencing data make existing parametric models not overly robust. Therefore, a robust and powerful method that allows covariate-dependent dispersion and addresses outliers is still needed for differential abundance analysis.

Results: We introduce a novel test for differential distribution analysis of microbiome sequencing data by jointly testing the abundance, prevalence, and dispersion. The test is built on zero-inflated negative binomial regression model and winsorized count data to account for zero-inflation and outliers. Using simulated data and real microbiome sequencing datasets, we show that our test is robust across various biological conditions and overall more powerful than previous methods.

Availability: R package is available at https://github.com/jchen1981/MicrobiomeDDA

Authors:

first name	last name	email	country	organization	corresponding?
Jun	Chen	chen.jun2@mayo.edu	USA	Mayo Clinic	✓
Emily	King		USA	Iowa State University	
Rebecca	Deek		USA	Columbia University	
Zhi	Wei	zhiwei@njit.edu	USA	New Jersey Institute of Technology	✓
Yue	Yu		USA	Mayo Clinic	
Diane	Grill		USA	Mayo Clinic	
Karla	Ballman		USA	Weill Cornell Medical College	

Insilico Identification of Protein-Coding and Non-Coding Regions in Next-generation Technology Transcriptome Sequence Data: A Machine Learning Approach

Keywords: Protein-Coding, Non-Coding, Next-generation Technology, Transcriptome Sequence Data, Machine Learning

Abstract: With the rapid increase in the volume of sequence data and multi-species transcripts generated using next-generation sequencing technologies, designing algorithms to process these data in an efficient manner and gaining biological insight is becoming a significantly growing challenge.

But there is still no known effective method to discriminate between non-coding and protein-coding regions in transcriptomes because RiboNucleic Acid (RNA) types show similar features to each other.

We describe a few techniques that discriminate between non-coding and protein-coding regions in transcripts, but all of these involve slow computational speed for small datasets or multi-threading in large datasets just to achieve small difference in computational speed, and thus risking a high execution time of the tool.

To solve this problem, we propose a fast, accurate and robust alignment-free predictor based on multiple feature groups using Logistic Regression, for the discrimination of protein-coding regions in multispecies transcriptome sequence data, where the predictive performance is influenced by Open Reading Frame(ORF)-Related and ORF-unrelated features used in the model rather than the training datasets, thereby achieving a relatively high performance and computational speed in processing small and large datasets of full-length and partial-length protein-coding and non-coding transcripts derived from transcriptome sequencing.

We describe a series of experiments on each of the datasets for human, mouse and fly species, with a goal that, our predictor generally performs better than competing techniques.

We expect this new approach to dramatically reduce the computational cost of identifying non-coding and protein-coding regions in transcripts, and hence make transcriptome analysis easier.

Authors:

first name	last name	email	country	organization	corresponding?
Olaitan	Awe	laitanawe@gmail.com	Nigeria	University of Ibadan	✓
Angela	Makolo		Nigeria	University of Ibadan	
Segun	Fatumo		United Kingdom	Wellcome Trust Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge	

Kart-RNA – a fast and accurate RNA-seq mapper

Keywords: RNA-seq, spliced alignment, sequence partitioning

Abstract: Motivation: In recent years, the massively-parallel cDNA sequencing (RNA-Seq) technologies have become a powerful tool to provide high resolution measurement of expression and high sensitivity in detecting low abundance transcripts. However, RNA-seq data requires a huge amount of computational analysis. The very first and critical step in the analysis is to align each sequence fragment against the reference genome. Various de novo spliced RNA aligners have been developed in re-cent years. Though these aligners can handle spliced alignment and detect splice junctions, some issues still remain to be resolved. With the advances in sequencing technologies and the ongoing collection of sequencing data in the ENCODE project, more efficient alignment algorithms are highly demanded to handle longer reads and higher error rates.

Results: We proposed a novel RNA-seq de novo mapping algorithm, call Kart-RNA to handle spliced alignment with a sequence partitioning strategy. Kart-RNA can handle both short and long RNA-seq reads. The experiment results on synthetic datasets and real NGS datasets showed that Kart-RNA is a highly efficient aligner that yields the highest or comparable sensitivity and accuracy, and it spends the least amount of time among the selected aligners.

Availability: https://github.com/hsinnan75/Kart-RNA/

Authors:

first name	last name	email	country	organization	corresponding?
Hsin-Nan	Lin	arith@iis.sinica.edu.tw	Taiwan	Academia Sinica	
Wen-Lian	Hsu	hsu@iis.sinica.edu.tw	Taiwan	Academia Sinica	✓

Predictive Genome Analysis Using Partial DNA Sequencing Data

Keywords: Variant Calling Pipeline, GATK, Read Prediction

Abstract: Motivation: Much research has been dedicated to reducing the computational time associated with the analysis of genome data, which resulted in shifting the bottleneck from the time needed for the computational analysis part to the actual time needed for sequencing of DNA information. DNA sequencing is a time consuming process, and all existing DNA analysis methods have to wait for the DNA sequencing to completely finish before starting the analysis.

Results: In this paper, we propose a new DNA analysis approach where we start the genome analysis before the DNA sequencing is completely finished. The genome analysis is started when the DNA reads are still in the process of being sequenced. We use algorithms to predict the unknown bases and their corresponding base quality scores of the incomplete read. Results show that our method of predicting the unknown bases and quality scores achieves more than 90% similarity with the full dataset for 50 unknown bases (slashing more than a day of sequencing time). We also show that our base quality value prediction scheme is highly accurate, only reducing the similarity of the detected variants by 0.45%. However, there is still room to introduce more accurate prediction schemes for the unknown bases to increase the effectiveness of the analysis by up to 5.8%.

Availability: http://www.ce.ewi.tudelft.nl/ahmed/

Authors:

first name	last name	email	country	organization	corresponding?
Nauman	Ahmed	n.ahmed@tudelft.nl	Netherlands	Delft University of Technology	✓
Koen	Bertels	k.l.m.bertels@tudelft.nl	Netherlands	Delft University of Technology	
Zaid	Al-Ars	z.al-ars@tudelft.nl	Netherlands	Delft University of Technology	✓

Self-refining Sorting of Single-cell RNA-seq data us-ing Unsupervised Gene Selection (uSORT)

Keywords: Gene selection, Pseudo-time inference, Single-cell RNA-sequencing

Abstract: Motivation: Recent technologies allow for transcriptomic characterization of biological samples at the single cell level. This offers a rich amount of information for various important studies, especially the re-construction of cell progression paths. Accurate path re-construction permits interrogation of the mechanisms governing important cellular events including cell differentiation and proliferation. Due to the high dimensionality of transcriptomic data, appropriate gene selection for the process of constructing cellular progression path is crucial. Single-cell transcriptomic data is intrinsically noisy and heterogeneous, thus the task of gene selection remains challenging. Exiting methods are largely restricted to differential expression analysis, or solely depend on variable genes e.g. principal component analysis.

Results: We develop a self-refining approach, named as uSORT, which progressively improves the performance of gene selection and also the accuracy of pseudo-temporal ordering of individual cells along the progression path (a.k.a. cell ordering) in an unsupervised manner. The current form of uSORT works on non-branching cell progression. Cell progression path constructed using uSORT detected genes is highly comparable to that built on supervised genes i.e. differentially expressed genes. The R package built for uSORT provides a number of state-of-the-art cell ordering algorithms together with two in-house enhanced algorithms to fit different types of data. Experimental data demonstrates that uSORT correctly recapitulates an accurate dendritic cell precursor developmental path, without relying on the differentially expressed genes. We also show that uSORT manages to deal with cyclic cell progression based on a human cell-cycle data. In both of these public datasets, the cell progression path re-constructed by uSORT demonstrates signature gene profiles that are consistent with those showed in previous studies. Furthermore, we show that genes selected by uSORT are biological relevant.

Authors:

first name	last name	email	country	organization	corresponding?
Mai Chan	Lau	lau_mai_chan@immunol.a-star.edu.sg	Singapore	Singapore Immunology Network	
Hao	Chen	chen_hao@immunol.a-star.edu.sg	Singapore	Singapore Immunology Network	
Michael	Poidinger	Michael_Poidinger@immunol.a-star.edu.sg	Singapore	Singapore Immunology Network	
Jinmiao	Chen	Chen_Jinmiao@immunol.a-star.edu.sg	Singapore	Singapore Immunology Network	✓

Descriptors for the compression of aligned high-throughput sequencing data

Keywords: High-Throughput Sequencing, Compression, Alignment

Abstract: MOTIVATION: Recent advancements in high-throughput sequencing technology have led to a rapid growth of genomic data. Especially high-coverage data generated by these technologies contains highly redundant information. Several compression schemes have been proposed for the coding of such data present in the form of raw FASTQ files and aligned SAM/BAM files. The alignment of such data allows for the efficient compression by exploitation of this redundancy. We describe the compression tool TSC 2 for aligned sequence reads. TSC 2 uses the alignment information to implicitly assemble local parts of the donor genome to compress the sequence alignments. In contrast to other algorithms, TSC 2 does not rely on the reference sequences used for alignment. Compression is performed using solely a permanently updated short-time memory as context for the prediction of sequence alignments.

RESULTS: We represent sequence alignments with a novel set of descriptors. Detailed investigation of the compressibility of these descriptors reveals characteristic properties of the compressed datasets and the used sequencing technologies. Moreover, by introducing the concept of a detailed analysis of sequence alignment descriptors, we outline the path for a well-structured optimization of sequence alignment encoders. Finally, the TSC 2 algorithm only requires negligible amount of memory and exhibits compression performances on par with state-of-the-art tools and compresses sequence alignment information down to 1.12% of the uncompressed size.

Availability: https://www.tnt.uni-hannover.de/~voges/

Authors:

first name	last name	email	country	organization	corresponding?
Jan	Voges	voges@tnt.uni-hannover.de	Germany	Leibniz Universitaet Hannover	✓
Marco	Munderloh	munderl@tnt.uni-hannover.de	Germany	Leibniz Universitaet Hannover	
Joern	Ostermann	office@tnt.uni-hannover.de	Germany	Leibniz Universitaet Hannover	

CALQ: compression of quality values of aligned sequencing data

Keywords: High-Throughput Sequencing, Compression, Alignment, Quantization, Variant Calling, Quality Values, Downstream Analysis

Abstract: MOTIVATION: Recent advancements in high-throughput sequencing technology have led to a rapid growth of genomic data. Several lossless compression schemes have been proposed for the coding of such data present in the form of raw FASTQ files and aligned SAM/BAM files. However, due to their high entropy, losslessly compressed quality values account for about 80% of the size of compressed files. We present a novel lossy compression scheme for the quality values named CALQ. By controlling the coarseness of quality value quantization with a statistical genotyping model, we minimize the impact of the introduced distortion on downstream analyses.

RESULTS: We analyze the performance of several lossy compressors of quality values in terms of trade-off between the achieved compressed size (in bits per quality value) and the Precision and Recall achieved after running a variant calling pipeline over sequencing data of the well known NA12878 individual. We show that CALQ can achieve, on average, better performance than with the original data with a size reduction of more than an order of magnitude with respect to the state-of-the-art lossless compressors.

Availability: https://www.tnt.uni-hannover.de/~voges/

Authors:

first name	last name	email	country	organization	corresponding?
Jan	Voges	voges@tnt.uni-hannover.de	Germany	Leibniz Universitaet Hannover	✓
Mikel	Hernaez	mikelhernaez@gmail.com	USA	Stanford University	✓
Joern	Ostermann	office@tnt.uni-hannover.de	Germany	Leibniz Universitaet Hannover	

Assembly and Scaffolding of Large Genomes With ABySS 2.0 Using Short Reads and Long Range Linkage Information

Keywords: de novo genome assembly, de Bruijn graph, high throughput sequencing, 10X Genomics, Chromium, Homo sapiens

Abstract: Owing to the efficiency and simplicity of its k-mer based algorithm, the de Bruijn graph assembly method remains the dominant approach for de novo genome assembly. Nonetheless, de novo assembly of large genomes such as human is a challenging computing task, due to the large amount of RAM required to store the full k-mer set of the sequencing reads in memory. Previously, ABySS 1.0 addressed the issue of large memory requirements by distributing the de Bruijn graph across cluster nodes using MPI. Here we present our recently published work on ABySS 2.0, a re-design of ABySS that replaces our earlier distributed approach with a compact, Bloom filter based representation of the de Bruijn graph.

Here, we compare benchmarking results of ABySS 2.0 against other leading assemblers using a recent Genome in a Bottle dataset, consisting of 250 bp paired-end reads and 6 kbp mate-pair libraries. The ABySS 2.0 assembly for this data ran in under 24 hours, required less than 35 GB of RAM, and achieved a scaffold NG50 of 3.5 Mbp. We describe further scaffolding steps we performed on this assembly, using a combination of 10X Chromium data and a BioNano optical map, resulting in the reconstruction of complete chromosome arms and a scaffold NG50 of 42 Mbp.

Recently, 10X Genomics introduced the Chromium library protocol, which augments standard Illumina sequencing with valuable long range sequence information by associating reads to molecule barcodes. We present further ABySS developments that make use of this information to navigate the compacted de Bruijn graph.

Authors:

first name	last name	email	country	organization	corresponding?
Benjamin P	Vandervalk	benv@bcgsc.ca	Canada	BC Cancer Agency Genome Sciences Centre	✓
Shaun D	Jackman	sjackman@bcgsc.ca	Canada	BC Cancer Agency Genome Sciences Centre	✓
Hamid	Mohamadi	hmohamadi@bcgsc.ca	Canada	BC Cancer Agency Genome Sciences Centre	✓
Justin	Chu	cjustin@bcgsc.ca	Canada	BC Cancer Agency Genome Sciences Centre	✓
Sarah	Yeo	syeo@bcgsc.ca	Canada	BC Cancer Agency Genome Sciences Centre	✓
S Austin	Hammond	shammond@bcgsc.ca	Canada	BC Cancer Agency Genome Sciences Centre	✓
Hamza	Khan	hkhan@bcgsc.ac	Canada	BC Cancer Agency Genome Sciences Centre	✓
Lauren	Coombe	lcoombe@bcgsc.ca	Canada	BC Cancer Agency Genome Sciences Centre	✓
Rene L	Warren	rwarren@bcgsc.ca	Canada	BC Cancer Agency Genome Sciences Centre	✓
Inanc	Birol	ibirol@bcgsc.ca	Canada	BC Cancer Agency Genome Sciences Centre	✓

Measuring the spatial correlations of protein binding sites

Keywords: Transcription factor binding, Correlations, Simpson's paradox, Next generation sequencing, ChIP-seq

Abstract: Understanding the interactions of different DNA binding proteins is a crucial first step toward deciphering gene regulatory mechanism. With advances of high-throughput sequencing technology such as ChIP-seq, the genome-wide binding sites of many proteins have been profiled under different biological contexts. It is of great interest to quantify the spatial correlations of the binding sites, such as their overlaps, to provide information for the interactions of proteins. Analyses of the overlapping patterns of binding sites have been widely performed, mostly based on Ad hoc methods. Due to the heterogeneity and the tremendous size of the genome, such methods often lead to biased even erroneous results.

In this work, we discover a Simpson's paradox phenomenon in assessing the genome-wide spatial correlation of protein binding sites: two proteins could be completely independent at different segments of genome but appear correlated genome-wide. Leveraging information from publicly available data, we propose a testing procedure for evaluating the significance of overlapping from a pair of proteins, which accounts for background artifacts and genome heterogeneity. Real data analyses demonstrate that the proposed method provide more biologically meaningful results. We implement the method as an R package available at http://www.sta.cuhk.edu.hk/YWei/ChIPCor.html.

The paper has been published in Bioinformatics (2016) 32 (12):
1766-1772, https://doi.org/10.1093/bioinformatics/btw058.

Authors:

first name	last name	email	country	organization	corresponding?
Yingying	Wei	yweicuhk@gmail.com	Hong Kong	The Chinese University of Hong Kong	✓
Hao	Wu		USA	Emory University	

Reconstructing Reprogramming Dynamics Using High-throughput Single-cell Transcriptomes

Keywords: Reprogramming, Drop-seq, Transcriptional dynamics, Computational models

Abstract: The ability to generate all cell types of the body has thrust pluripotent stem cells (PSCs) into the spotlight as panacea for regenerative medicine. Creating PSCs by reprogramming (yielding iPSCs) has ushered in an era of personalized medicine allowing disease in a dish modeling and ultimately cell replacement strategies. However, despite their potential, reprogramming of somatic cells to iPSCs, a process triggered by ectopic expression of few transcription factors, is limited by its low efficiency. Because reprogramming takes place over a period of weeks in densely populated cultures in which only 1% of cells progress to pluripotency, population studies cannot be used to measure gene expression dynamics. Instead, we apply the Drop-seq method to measure mRNA-transcripts from thousands of individual cells simultaneously at different time-points during reprogramming. We uncover the temporal sequence of cell state transitions as somatic cells reprogram to the pluripotent state by identifying the genes that are activated or inactivated during these transitions, modeling alternate reprogramming paths toward pluripotency and defining barriers in the reprogramming process at single-cell resolution. We are able to deconstruct cell population and infer gene regulatory linkages by utilizing ordered expression profiles. Our results provide fundamental insights into epigenetic reprogramming and help reduce the time needed for reprogramming while increasing its efficiency.

Authors:

first name	last name	email	country	organization	corresponding?
Shan	Sabri	ShanASabri@gmail.com	USA	University of California, Los Angeles	✓
Justin	Langerman		USA	University of California, Los Angeles	
Jason	Ernst		USA	University of California, Los Angeles	
Kathrin	Plath		USA	University of California, Los Angeles	

Systems human genome and metagenome analysis on circulating proteins in a population cohort

Keywords: proteomics, metagenomics, disease biomarkers, pQTL mapping

Abstract: Both genetics and microbiome are known to be crucial factors in determining individual's susceptibility risk for complex diseases, including immune diseases, cancers and cardiovascular diseases (CVD). However, the contribution of these factors to inter-individual variation of intermediate molecular phenotypes, for instance of circulating proteins, in the general population is largely unknown. These circulating proteins are often measured as biomarkers, holding a promise for early disease diagnosis and monitoring therapeutics. Understanding the impact of genetics and microbiome on circulating proteins can provide a better understanding of the underlying disease etiology.

We have now measured serum levels of 92 CVD-relevant proteins in 1,294 individuals from a general Dutch population cohort (LifeLines-DEEP) for whom we also have data on the human genome and "the second human genome": the metagenome. For each protein, we performed a genome-wide pQTL mapping analysis with 8 million SNPs and metagenome-wide association analysis with 340 bacterial species and 702 functional pathways determined by metagenomics sequencing. At FDR 0.05, we identified 72 proteins that were genetically controlled and 51 proteins associated with the gut microbiome. Serum levels of 37 proteins were affected by both genetics and microbiome. C-C motif chemokine 15 (CCL15), for example, is a liver-derived chemokine involved in immunoregulatory and inflammatory processes. In addition to its strong genetic regulation (association to rs854626 at P=2.5x10-136), an elevated serum level of CCL15 was also associated to higher bacterial capacity for fatty acid biosynthesis (P=5.3x10-4). We further confirmed the causal effect of fatty acids on CCL15 production by stimulating hepatocytes (HepG2) with free fatty acids, observing a 40% increase in CCL15 expression 24 hours after stimulation.

Fourteen proteins were more affected by the gut microbiome than by genetics. For instance, adipose-derived cytokine PAI-1 is strongly associated with obesity and its elevation is also a risk factor for atherosclerosis. While we did not detect significant associations with genetics, serum levels of PAI-1 were not only associated to a lower richness of bacterial species (P=9.7x10-4) but also to 138 bacterial function pathways, in particular to bacterial energy metabolism.

By using 92 CVD-related circulating proteins we demonstrate that serum proteomics are affected by both genetics and gut microbiome. Our data suggests that both the human genome and metagenome should be taken into account when using circulating proteins as potential biomarkers for disease monitoring or as therapeutic targets for personalized medicine.

Authors:

first name	last name	email	country	organization	corresponding?
Daria	Zhernakova	dashazhernakova@gmail.com	Netherlands	University of Groningen, University Medical Center Groningen, Department of Genetics	✓
Alexander	Kurilshikov	alexa.kur@gmail.com	Netherlands	University of Groningen, University Medical Center Groningen, Department of Genetics	
Biljana	Atanasovska	bibi_a_85@yahoo.com	Netherlands	University of Groningen, University Medical Center Groningen, Department of Genetics	
Trang	Le	t.le.4@student.rug.nl	Netherlands	University of Groningen, University Medical Center Groningen, Department of Genetics	
Marc Jan	Bonder	marcj89@gmail.com	Netherlands	University of Groningen, University Medical Center Groningen, Department of Genetics	
Serena	Sanna	aamichigan@gmail.com	Netherlands	University of Groningen, University Medical Center Groningen, Department of Genetics	
Rudolf	Boer	r.a.de.boer@umcg.nl	Netherlands	University of Groningen, Department of Cardiology, Groningen, Netherlands	
Folkert	Kuipers	f.kuipers@umcg.nl	Netherlands	University of Groningen, Department of Pediatrics, Groningen, Netherlands	
Lude	Franke	ludefranke@gmail.com	Netherlands	University of Groningen, University Medical Center Groningen, Department of Genetics	
Cisca	Wijmenga	cisca.wijmenga@gmail.com	Netherlands	University of Groningen, University Medical Center Groningen, Department of Genetics	
Alexandra	Zhernakova	sashazhernakova@gmail.com	Netherlands	University of Groningen, University Medical Center Groningen, Department of Genetics	
Jingyuan	Fu	fjingyuan@gmail.com	Netherlands	University of Groningen, University Medical Center Groningen, Department of Genetics	

f-divergence cutoff index to simultaneously identify differential expression in the integrated transcriptome and proteome

Keywords: Transcriptomics, Proteomics, Differential Expression, Proteogenomics

Abstract: The ability to integrate 'omics' (i.e. transcriptomics and proteomics) is becoming increasingly important to the understanding of regulatory mechanisms. There are currently no tools available to identify differentially expressed genes (DEGs) across different 'omics' data types or multi-dimensional data including time courses. We present fCI (f-divergence Cut-out Index), a model capable of simultaneously identifying DEGs from continuous and discrete transcriptomic, proteomic and integrated proteogenomic data. We show that fCI can be used across multiple diverse sets of data and can unambiguously find genes that show functional modulation, developmental changes or misregulation. Applying fCI to several proteogenomics datasets, we identified a number of important genes that showed distinctive regulation patterns. The package fCI is available at R Bioconductor and http://software.steenlab.org/fCI/.

Authors:

first name	**last name**	**email**	**country**	**organization**	**corresponding?**
Shaojun	Tang	shaojun.tang@georgetown.edu	USA	Georgetown University	✓

A scalable big data platform for human genomics

Keywords: scalable computing, hadoop, mapreduce, spark, hbase, genomic databases, gwas

Abstract: The accelerated rate at which complete human genome sequences are produced has overwhelmed the capacity of traditional computational tools. A promising approach to scalable bioinformatics computing is the use of big data technologies such as the Hadoop ecosystem. Here we present the development of such a platform for the storage, management, and analysis of human genomic data, and show its application to genome-wide association studies (GWAS). Variation data from 2500 complete human genomes were stored in the distributed database HBase. Queries and analysis with existing PLINK software were implemented as MapReduce and/or Spark jobs, which allowed for scalable, parallel processing. A proof-of-concept, web-based user interface was implemented to demonstrate submission of GWAS analyses and viewing of results. Benchmarking of different steps of the computational pipeline showed that query times depended little on the size of the database. The performance of the system could be controlled by increasing the number of computing nodes as the database size increased with the addition of more human genomes. Wrapping PLINK into a MapReduce job achieved a performance boost without the need of modifying the existing application. Processing steps implemented in Spark outperformed MapReduce implementations. Thus this approach proved its potential as a basis for a comprehensive genomic data science platform for clinical and research applications.

Authors:

first name	last name	email	country	organization	corresponding?
Paul	Hodor	paul@aurynia.com	USA	Aurynia Scientific	✓
Ezekiel	Maier	maier_ezekiel@bah.com	USA	Booz Allen Hamilton	
Christopher	Gardner	christopher.gardner@outlook.com	USA	InterSystems Corporation	
Natasha	Sefcovic	sefcovic_natasha@bah.com	USA	Booz Allen Hamilton	
Lauren	Neal	neal_lauren@bah.com	USA	Booz Allen Hamilton	

InFusion: advancing discovery of fusion genes and chimeric transcripts from RNA-sequencing data

Keywords: High-throughput sequencing, RNA-seq, precision medicine, fusion genes, chimeric transcripts

Abstract: Functional fusion genes and chimeric transcripts have been shown to occur in cancers due to genomic rearrangements as well as in non-cancerous cells due to trans-splicing or transcriptome machinery failure. Careful deactivation of fusions can stop further arrangement and growth of cancer. Therefore correct and detailed detection of fusions is important in scientific research and in precision medicine as well. We have designed and implemented a novel toolkit called InFusion for chimeric transcript discovery from RNA-seq data. Our approach introduces several unique features such as discovery of fusions involving intergenic regions and detection of anti-sense chimeras based on the strand-specificity of the sequencing library. Additionally, the toolkit includes several advanced post-analysis steps such as comparison of results among well-known exiting tools and design of sequences for further experimental validation.

Using simulated and public data we demonstrated that InFusion has superior detection sensitivity compared to other existing methods and is able to discover a wider spectrum of fusion events that can occur in the transcriptome. Importantly, we also performed deep RNA sequencing of two prostate cancer cell lines. From this experimental data analysis we discovered in-silico and verified in-vitro 26 novel fusion events, including alternatively spliced fusion isoforms and chimeric RNAs involving non-exonic regions. Moreover, we confirmed four fusions that involve intergenic regions. To our knowledge, discovery of such events has not been addressed previously, despite their potential to encode functional proteins or regulate gene transcription.

The detailed landscape of the chimeric RNAs, mechanisms underlying their genesis and their functional roles are yet to be studied. InFusion may prove to be a useful tool for detecting the whole scope of possible events. The software toolkit is open-source and available for download at: http://bitbucket.org/kokonech/infusion

Authors:

first name	**last name**	**email**	**country**	**organization**	**corresponding?**
Konstantin	Okonechnikov	k.okonechnikov@dkfz-heidelberg.de	Germany	German Cancer Research Center (DKFZ)	✓
Aki	Imai-Matsushima	imai@mpiib-berlin.mpg.de	Germany	Max Planck Institute for Infection Biology	
Lukas	Paul	lukas.paul@lexogen.com	Austria	Lexogen GmbH	
Alexander	Seitz	alexander.seitz@lexogen.com	Austria	Lexogen GmbH	
Thomas	Meyer	meyer@mpiib-berlin.mpg.de	Germany	Max Planck Institute for Infection Biology	
Fernando	Garcia-Alcalde	fernando.garcia-alcalde@roche.com	Switzerland	Roche Innovation Center	✓

Combining multiple tools outperforms individual methods in gene set enrichment analyses

Keywords: RNAseq, Microarray, Gene Set Enrichment Analysis, Pathway Analysis, Systems Biology

Abstract: Gene set enrichment (GSE) analysis allows researchers to efficiently extract biological insight from long lists of differentially expressed genes by interrogating them at a systems level. In recent years, there has been a proliferation of GSE analysis methods and hence it has become increasingly difficult for researchers to select an optimal GSE tool based on their particular data set. Moreover, the majority of GSE analysis methods do not allow researchers to simultaneously compare gene set level results between multiple experimental conditions. The ensemble of genes set enrichment analyses (EGSEA) is a method developed for RNA-sequencing data that combines results from twelve algorithms and calculates collective gene set scores to improve the biological relevance of the highest ranked gene sets. EGSEA's gene set database contains around 25,000 gene sets from sixteen collections. It has multiple visualization capabilities that allow researchers to view gene sets at various levels of granularity. EGSEA has been tested on simulated data and on a number of human and mouse data sets and, based on biologists' feedback, consistently outperforms the individual tools that have been combined. Our evaluation demonstrates the superiority of the ensemble approach for GSE analysis, and its utility to effectively and efficiently extrapolate biological functions and potential involvement in disease processes from lists of differentially regulated genes.

Availability: http://ww.csl.com.au

Authors:

first name	last name	email	country	organization	corresponding?
Monther	Alhamdoosh	m.hamdoosh@gmail.com	Australia	CSL Limited	✓
Milica	Ng	Milica.Ng@csl.com.au	Australia	CSL Limited	✓
Nicholas J.	Wilson	Nick.Wilson@csl.com.au	Australia	CSL Limited	✓
Julie M.	Sheridan	sheridan@wehi.EDU.AU	Australia	Walter and Eliza Hall Institute of Medical Research	✓
Huy	Huynh	Huy.Huynh@csl.com.au	Australia	CSL Limited	✓
Michael	Wilson	Michael.Wilson@csl.com.au	Australia	CSL Limited	✓
Matthew	Ritchie	mritchie@wehi.edu.au	Australia	Walter and Eliza Hall Institute of Medical Research	✓

CIDR: Ultrafast and accurate clustering through imputation for single-cell RNA-seq data

Keywords: Single-cell scRNA-seq, Dropout, Imputation, Dimensionality reduction, Clustering, Cell type

Abstract: Most existing dimensionality reduction and clustering packages for single-cell RNA-seq (scRNA-seq) data deal with dropouts by heavy modeling and computational machinery. Here, we introduce CIDR (Clustering through Imputation and Dimensionality Reduction), an ultrafast algorithm that uses a novel yet very simple implicit imputation approach to alleviate the impact of dropouts in scRNA-seq data in a principled manner. Using a range of simulated and real data, we show that CIDR improves the standard principal component analysis and outperforms the state-of-the-art methods, namely t-SNE, ZIFA, and RaceID, in terms of clustering accuracy. CIDR typically completes within seconds when processing a data set of hundreds of cells and minutes for a data set of thousands of cells. CIDR can be downloaded at https://github.com/VCCRI/CIDR.

Availability: http://bioinformatics.victorchang.edu.au/

Authors:

first name	last name	email	country	organization	corresponding?
Peijie	Lin	p.lin@victorchang.edu.au	Australia	Victor Chang Cardiac Research Institute	
Michael	Troup	m.troup@victorchang.edu.au	Australia	Victor Chang Cardiac Research Institute	
Joshua	Ho	j.ho@victorchang.edu.au	Australia	Victor Chang Cardiac Research Institute	✓

Reliable analyses of circulating tumor cells via single cell whole exome sequencing

Keywords: single cell, CTC, breast cancer, whole genome amplification, ampli1, whole exome sequencing

Abstract: The presence of circulating tumor cells (CTCs) in breast cancer patients is an important prognostic factor as the cells represent a link between the primary tumor and the putative development of a metastasis. A reliable characterization of the mutational landscape of CTCs may therefore provide important insights into the early metastatic processes as well as drug resistance development. In order to establish a workflow that allows the identification of relevant mutations as well as copy number variations (CNVs), we systematically evaluated the reproducibility and reliability of an Ampli1TM based whole genome amplification coupled to Agilent SureSelect XT whole exome target enrichment. In order to cover a broad range of technical and biological variability, we performed our analyses on single cells from two cell lines (NA12878 and MCF7) and a set of peripheral blood lymphocytes (PBL) from a healthy donor under different technical setups. Based on the obtained results, we defined a simple model to describe the reliability of an Ampli1TM based amplification and used it to improve our variant calling and CNVs identification pipelines in single cells. Applying our workflow to CTCs from three metastatic breast cancer patients who participated in the DETECT III study revealed several recurrent mutations and CNVs of known breast cancer marker genes.

Availability: https://www.item.fraunhofer.de/en/services_expertise/tumor_therapy.html

Authors:

first name	last name	email	country	organization	corresponding?
Urs	Lahrmann	urs.lahrmann@item.fraunhofer.de	Germany	Fraunhofer ITEM-R	✓
Bernhard	Polzer	bernhard.michael.polzer@item.fraunhofer.de	Germany	Fraunhofer ITEM-R	
Zbigniew	Czyz	zbigniew.czyz@item.fraunhofer.de	Germany	Fraunhofer ITEM-R	
Christoph	Klein	christoph.andreas.klein@item.fraunhofer.de	Germany	Fraunhofer ITEM-R	
Stefan	Kirsch	stefan.kirsch@item.fraunhofer.de	Germany	Fraunhofer ITEM-R	

Single cell transcriptomics reveals specific RNA editing signatures in the human brain

Keywords: Single cell, RNAseq, RNA editing

Abstract: background

A-to-I RNA editing in human is carried out by members of ADAR family of enzymes that act on double strand RNAs and can alter codon identity, splicing sites or base-pairing interactions within higher-order RNA structures. Recoding RNA editing is essential for normal brain development and regulates important functional properties of neurotransmitter receptors [1, 2]. Indeed, its deregulation has been linked to several nervous diseases such as epilepsy, schizophrenia, Alzheimer, major depression and amyotrophic lateral sclerosis [3, 4]. Recently we have profiled RNA editing in six different human tissues using whole transcriptome sequencing and detected more than three million events [5]. Interestingly, genes undergoing RNA editing were consistently enriched in genes involved in neurological disorders and cancer, confirming the relevant biological role of RNA editing in human.

Although investigations in bulk tissues are extremely useful, they do not capture the transcriptomic heterogeneity of multiple cell types constituting the ensemble tissue.

results

To characterize the complexity of RNA editing at single cell resolution, we investigated this phenomenon in single cells from adult human cortex obtained from living subjects in which transcriptome diversity was already surveyed by single cell RNA sequencing (scRNA-seq) [6]. Using a comprehensive collection of known RNA editing events, we explored inosinome profiles in 466 cortex cells. Individual scRNAseq data were quality checked by FASTQC and poor regions at 3' ends were trimmed by means of trim_galore tool. Cleaned read were then mapped onto the human reference genome by STAR aligner. RNA editing candidates were detected using our REDItools [7] and analyzed by custom scripts. We found that the identification of A-to-I RNA editing in single cells was strongly correlated with the amount of generated RNA reads. RNA editing profiles were quite heterogeneous also inside the same cell population. However, the observed RNA editing profile as well as the Alu editing index were sufficient to discriminate major cell types as neurons, astrocytes and oligodendrocytes, underlining the cell specific nature of RNA editing. Interestingly, recoding RNA editing were mainly detectable in neurons, remarking the primary role of A-to-I editing in modulating brain functions through key modifications in receptors for neurotransmitters.

conclusions

Taken together, our results demonstrate that RNA editing is detectable in single cells and demonstrates that A-to-I patterns reveal specific editing signatures distinguishing major cell types in the human brain. Profiling RNA editing in single cells yields novel and exiting insights into neuronal plasticity and opens up the possibility of deciphering as yet unknown molecular mechanisms in diverse neurological or neurodegenerative disorders. In addition, A-to-I changes in single cells may contribute to the identification of novel therapeutic targets or prognostic markers for innovative approaches of precision medicine.

references

1. Li JB, Levanon EY, Yoon JK, Aach J, Xie B, Leproust E, Zhang K, Gao Y, Church GM: Genome-wide identification of human RNA editing sites by parallel DNA capturing and sequencing. Science 2009, 324(5931):1210-1213.

2. Mehler MF, Mattick JS: Noncoding RNAs and RNA editing in brain development, functional diversification, and neurological disease. Physiological reviews 2007, 87(3):799-823.

3. Maas S, Kawahara Y, Tamburro KM, Nishikura K: A-to-I RNA editing and human disease. RNA biology 2006, 3(1):1-9.

4. Khermesh K, D'Erchia AM, Barak M, Annese A, Wachtel C, Levanon EY, Picardi E, Eisenberg E: Reduced levels of protein recoding by A-to-I RNA editing in Alzheimer's disease. Rna 2016, 22(2):290-302.

5. Picardi E, Manzari C, Mastropasqua F, Aiello I, D'Erchia AM, Pesole G: Profiling RNA editing in human tissues: towards the inosinome Atlas. Sci Rep 2015, 5:14941.

6. Darmanis S, Sloan SA, Zhang Y, Enge M, Caneda C, Shuer LM, Hayden Gephart MG, Barres BA, Quake SR: A survey of human brain transcriptome diversity at the single cell level. Proceedings of the National Academy of Sciences of the United States of America 2015, 112(23):7285-7290.

7. Picardi E, Pesole G: REDItools: high-throughput RNA editing detection made easy. Bioinformatics 2013, 29(14):1813-1814.

Authors:

first name	**last name**	**email**	**country**	**organization**	**corresponding?**
Ernesto	Picardi	ernesto.picardi@uniba.it	Italy	University of Bari & IBIOM-CNR	✓
Anna Maria	D'Erchia	annamaria.derchia@uniba.it	Italy	University of Bari & IBIOM-CNR	
Graziano	Pesole	graziano.pesole@uniba.it	Italy	University of Bari & IBIOM-CNR	

Base Calling and Indexing Oxford Nanopore Reads

Keywords: MinION, recurrent neural networks, dynamic time warping

Abstract: Recently, we have developed an open source DeepNano (Boža et al. 2016) base caller for Oxford Nanopore reads based on recurrent neural networks. On R7 data, our base caller outperforms alternatives, while on R9 data, accuracy of DeepNano is slightly worse than Albacore and Nanonet base callers released by Oxford Nanopore.

The advantage of DeepNano, however, is in its flexibility. Under the default settings, DeepNano is faster, and by adjusting the size of the underlying network, it is possible to further trade accuracy for speed. Fast base calling is essential in applications such as selective on-device sequencing (ReadUntil, Loose et al. 2016) and in settings where using cloud services, as supported by Oxford Nanopore, is impractical. It is also possible to adaptively retrain the network, which can be used to leverage data that is otherwise impossible to base call through standard means (e.g., due to modifications or damage to the DNA).

Finally, we examine the dynamic-time-warp (DTW, Sankoff and Kruskal 1983) scheme for classification of reads and show that for applications such as ReadUntil, the method suffers from low specificity at high sensitivity. We demonstrate that by adjusting methods for scaling raw data, the sensitivity vs. specificity tradeoff can be much improved.

Availability: http://compbio.fmph.uniba.sk/~bbrejova

Authors:

first name	last name	email	country	organization	corresponding?
Vladimír	Boža	usama@ksp.sk	Slovakia	Comenius University	✓
Broňa	Brejová	brejova@dcs.fmph.uniba.sk	Slovakia	Comenius University in Bratislava	✓
Tomas	Vinar	vinar@fmph.uniba.sk	Slovakia	Comenius University in Bratislava	✓

A computational framework to infer the order of accumulating mutations in individual tumors

Keywords: Single-cell sequencing (SCS), Multi-region sequencing, Intra Tumor Heterogeneity (ITH), Phylogenetic Inference, Mutational Trees

Abstract: The increasing availability and reliability of high-resolution data on single tumors, as provided by single cell sequencing (SCS) [1] and multi-region data [2], allows to investigate intra-tumor heterogeneity (ITH) [3] with extraordinary efficiency and accuracy, hence paving the way for the development of new diagnostic and therapeutic (personalized) strategies. Several algorithmic approaches make use of such data to: (i) deconvolve the signature and cellular prevalence of cancer (sub)clones, and the evolutionary relations among them [4-6], or (ii) estimate the ordering of accumulation of (epi)genomic alterations driving cancer progression, in terms of mutational trees [7,8]. The concerted application of such complementary approaches may allow to draw an all-encompassing picture of ITH.

We here introduce a computational framework for the inference of the accumulation ordering of (epi)genomic alterations in single tumors (category (ii)), conceived to overcome some limitations of standard phylogenetic approaches. In fact, most techniques in both categories (i-ii) ground their roots in phylogenetic theories, hence requiring technical assumptions regarding, e.g., sequence substitution models, alleles fixations, topological structure, and noise learning/modeling. This results in models that: (a) presents limitations in terms of expressivity – structures that are more complex than trees cannot be inferred; (b) have a certain degree of arbitrariness – e.g., heuristics are needed to disambiguate equivalent scoring solutions in maximum parsimony approaches; (c) often cannot deal with data-specific errors, unless they include ad-hoc computationally demanding algorithmic strategies.

Our framework provides a statistically robust estimation of mutational orderings in single tumors, in the form of maximum a posteriori probabilistic models, and presents significant advantages regarding model expressivity and computational complexity. In our models, if an edge connects two mutations: (a) it defines the temporal precedence and (b) such mutations are statistically dependent. Statistical confidence is assessed via various bootstrap and cross-validation techniques. The framework includes several polynomial-time algorithms with different goals and properties, which are not restricted to the inference of tree-structures, and are designed to account for phenomena such as the possible presence of: (a) confounding factors in the generative model; (b) multiple independent progressions [9]; (c) distinct evolutionary trajectories converging to the same mutation (i.e., a generalization of convergent evolution [2]). The algorithms efficiently deal with data including missing entries and/or false positives/negatives. Furthermore, the computational complexity of our techniques is limited compared to standard Bayesian approaches, as we do not compute a full posterior over our estimates and we do not include any noise-learning procedure. This improves the method scalability, in anticipation of the mounting accessibility of SCS data. The framework is available within the TRONCO R suite [10].

On synthetic data, we prove that our methods display state-of-the-art accuracy and improved computational efficiency with respect to competing techniques, also on noisy data and with small sample size, on both SCS and multi-region data. Besides, the application on breast cancer SCS data [11] and colorectal cancer multi-region data [12] shows that the coupled application of our framework and standard phylogenetic methods allows one to characterize ITH with unmatched

resolution, also leading to the formulation of new experimental hypotheses.

References

[1] Navin, N. et al. Tumour evolution inferred by single-cell sequencing. Nature, 472(7341):90-94, 2011.

[2] Gerlinger, M. et al. Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. New England Journal of Medicine, 366(10):883-892, 2012.

[3] Vogelstein, B. et al. Cancer genome landscapes. Science, 339(6127):1546-1558, 2013.

[4] Yuan, K. et al. Bitphylogeny: a probabilistic framework for reconstructing intra-tumor phylogenies. Genome Biology, 16(1):1, 2015.

[5] Ross, E. M. and Markowetz, F. Onconem: inferring tumor evolution from single-cell sequencing data. Genome Biology, 17(1):1, 2016.

[6] Roth, A. et al. Clonal genotype and population structure inference from single-cell tumor sequencing. Nature Methods, 13(7):573-576, 2016.

[7] Kim, K. I. and Simon. R. Using single cell sequencing data to model the evolutionary history of a tumor. BMC Bioinformatics, 15(1):27,2014.

[8] Jahn, K. et al. Tree inference for single-cell data. Genome Biology, 17(1):1, 2016.

[9] Parsons, B. L. Many different tumor types have polyclonal tumor origin: evidence and implications. Mutation Research/Reviews in Mutation Research, 659(3):232-247, 2008.

[10] De Sano, L. et al. TRONCO: an R package for the inference of cancer progression models from heterogeneous genomic data. Bioinformatics, 32(12):1911-1913, 2016.

[11] Wang, Y. et al. Clonal evolution in breast cancer revealed by single nucleus genome sequencing. Nature, 512(7513):155-160, 2014.

[12] Lu, Y-W. et al. Colorectal cancer genetic heterogeneity delineated by multi-region sequencing. PloS One, 11(3):e0152673, 2016.

Authors:

first name	last name	email	country	organization	corresponding?
Daniele	Ramazzotti	daniele.ramazzotti@stanford.edu	USA	University of Stanford	✓
Alex	Graudenzi	alex.graudenzi@unimib.it	Italy	University of Milan-Bicocca	✓
Luca	De Sano	luca.desano@gmail.com	Italy	University of Milan-Bicocca	✓
Marco	Antoniotti	marco.antoniotti@unimib.it	Italy	University of Milan-Bicocca	✓
Giulio	Caravagna	giulio.caravagna@ed.ac.uk	United Kingdom	University of Edinburgh	✓

Bioinformatics approaches for RNA editing detection: a comparative study of state of the art software

Keywords: RNA editing, RNAseq, RNAseq simulation

Abstract: background

RNA editing is a widespread post-transcriptional mechanism determining modifications in the nucleotide sequence of RNA molecules, frequently affecting molecular processes in cells. Main RNA editing events include C-to-U deamination and A-to-I deamination, the latter catalyzed by ADAR enzymes and recognized as A-to-G substitutions by cellular machineries [1]. In humans, the A-to-I RNA editing is prominent in the brain and its deregulation has been linked to several neurological and neurodegenerative diseases [2]. The detection of editing sites has been recently improved through RNA-seq, thus requiring accurate bioinformatics pipelines for data analysis.

results

Several bioinformatics approaches for RNA editing detection were tested on simulated data by combining different mappers (BWA, GSNAP, HiSat2 and STAR [3-6]) and variant callers (REDItools, RES-Scanner, RNAEditor, GIREMI and JACUSA [7-11]). We simulated different libraries (unstranded, forward- and reverse-stranded), and matching DNA-seq data. Poly(A) tail were removed from primary transcripts. Generally, BWA and STAR mappers generated the best alignments, with the lowest rate of missing true editing sites and wrong sites. REDItools and JACUSA were shown to be the most sensitive and precise tools, however the latter also showed the highest rates of false positives, proving low efficacy in discriminating real variants from artifacts. In addition, F-scores and correlation coefficients confirmed REDItools performance as the most accurate for all kinds of samples, followed by JACUSA (unstranded) and RES-Scanner (FR-stranded). Editing sites were detected more efficiently by REDItools and JACUSA within Alu regions, by JACUSA and RES-Scanner within non-Alu regions.

conclusions

Here we provide a performance assessment of state of the art software devoted to the bioinformatics detection of RNA editing events. Although several computational approaches have been released, our results demonstrate that the accurate identification of RNA editing changes in massive transcriptome sequencing dataset is yet a challenging task.

references

1. Li JB, Church GM. Deciphering the functions and regulation of brain-enriched A-to-I RNA editing.

Nat. Neurosci. 2013; 16:1518–1522.

2. Maas S, Kawahara Y, Tamburro KM, et al. A-to-I RNA editing and human disease.

RNA Biol. 2006; 3:1–9

3. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform.

Bioinforma. Oxf. Engl. 2009; 25:1754–1760.

4. Wu TD, Nacu S. Fast and SNP-tolerant detection of complex variants and splicing in short reads.

Bioinforma. Oxf. Engl. 2010; 26:873–881.

5. Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements.

Nat. Methods 2015; 12:357–360.

6. Dobin A, Davis CA, Schlesinger F, et al. STAR: ultrafast universal RNA-seq aligner.

Bioinformatics 2012; bts635.

7. Picardi E, Pesole G. REDItools: high-throughput RNA editing detection made easy.

Bioinforma. Oxf. Engl. 2013; 29:1813–1814.

8. Wang Z, Lian J, Li Q, et al. RES-Scanner: a software package for genome-wide identification of RNA-editing sites.

GigaScience 2016; 5:37

9. John D, Weirick T, Dimmeler S, et al. RNAEditor: easy detection of RNA editing events and the introduction of editing islands.

Brief. Bioinform. 2016.

10. Zhang Q, Xiao X. Genome sequence-independent identification of RNA editing sites.

Nat. Methods 2015; 12:347–350.

11. Piechotta M, Wyler E, Ohler U, et al. JACUSA: site-specific identification of RNA editing events from replicate sequencing data.

BMC Bioinformatics 2017; 18:7.

Authors:

first name	last name	email	country	organization	corresponding?
Maria Angela	Diroma	mariangeladiroma@gmail.com	Italy	University of Bari	
Loredana	Ciaccia	loredana.ciaccia89@gmail.com	Italy	University of Bari	
Graziano	Pesole	graziano.pesole@uniba.it	Italy	University of Bari & IBIOM-CNR	
Ernesto	Picardi	ernesto.picardi@uniba.it	Italy	University of Bari & IBIOM-CNR	✓

dnaasm - new tool to assemble repetitive DNA regions

Keywords: NGS, de-novo DNA assembler, web application

Abstract: We propose a modification of the algorithm for DNA de-novo assembly, which uses the relative frequency of reads to properly reconstruct repetitive sequences (tandem repeats). The main advantage of our approach is that tandem repeats, which are longer than the insert size of paired-end tags, can also be properly reconstructed (other genome assemblers fail in such cases). What is more, tandem repeats could also be restored, if only single-read sequencing data is available.

The application was developed in client-server architecture, where web-browser is used to communicate with end-user and algorithms are implemented in C++ and Python.

Our data structures allow to build and handle graph up to $8 * 10^9$ vertices (e.g. for human genome) in 256 GB RAM, therefore our solution is faster than others.

Availability: http://staff.elka.pw.edu.pl/~rnowak2

Authors:

first name	last name	email	country	organization	corresponding?
Robert	Nowak	rbmnowak@gmail.com	Poland	Warsaw University of Technology, Institute of Computer Science	✓
Wiktor	Kusmirek	wkusmire@mion.elka.pw.edu.pl	Poland	Warsaw University of Technology	

Transcriptome-wide splicing quantification in single cells

Keywords: Single-cell RNA-seq, Isoform estimate, Differential splicing

Abstract: Single cell RNA-seq (scRNA-seq) has revolutionised our understanding of transcriptome variability, with profound implications both fundamental and translational. While scRNA-seq provides a comprehensive measurement of stochasticity in transcription, the limitations of the technology have prevented its application to dissect variability in RNA processing events such as splicing. Here we present BRIE (Bayesian Regression for Isoform Estimation), a Bayesian hierarchical model which resolves these problems by learning an informative prior distribution from sequence features. We show that BRIE yields reproducible estimates of exon inclusion ratios in single cells and provides an effective tool for differential isoform quantification between scRNA-seq data sets. BRIE therefore expands the scope of scRNA-seq experiments to probe the stochasticity of RNA-processing.

Availability: http://homepages.inf.ed.ac.uk/s1333321/

Authors:

first name	last name	email	country	organization	corresponding?
Yuanhua	Huang	huangyh094@gmail.com	United Kingdom	School of Informatics, University of Edinburgh	
Guido	Sanguinetti	G.Sanguinetti@ed.ac.uk	United Kingdom	School of Informatics, University of Edinburgh	✓

Getting the viruses out of metagenomes: new tools for accurate viral profiling and reconstruction of uncharacterized bacteriophages

Keywords: metagenomics, microbiome, bacteriophages, virome

Abstract: Viruses are key players of the human microbiome and their interaction with human-associated bacteria and micro-eukaryotes shape the human-microbiome symbiosis. However, the majority of the viral realm is still unexplored despite recent advancements in metagenomics that allowed us to study the human microbiota at high resolution. Indeed, characterizing viral genomes from metagenomes poses intriguing and unaddressed challenges for computational biologists. These include, for instance, the lack of universal viral markers and the low number of reference genomes that hampers the identification of novel viruses. As a result, currently little is known about specific variants of already known phages and a large fraction of the virome is still transparent to the available metagenomic profiling pipelines.

We have developed a new computational metagenomic pipeline that i) identifies and compares at the single nucleotide variant (SNV) level known viruses and ii) detects novel viruses whose genome is not available as a reference. The framework uses a combination of assembly-free genome reconstruction techniques and a machine-learning based approach for discovering bacteriophage-specific features in assembled contigs. The method is currently being finalized and will be ready to be presented at ISMB/ECCB 2017. We first applied the pipeline to the metagenomes of 29 longitudinally sampled cystic fibrosis (CF) patients identifying a panel of phages and following their variants in time with phylogenetic resolution (Figure 1A). We found that the SNV rates observed for phages between timepoints were substantially higher than those of their bacterial hosts (Figure 1B). We then applied the method to 87 CF metagenomes identifying 8 potential new phages, currently under further investigation. These newly discovered phages showed circular dsDNA genomes, some features of known viral proteins, but low identity with any known virus. Finally, we analysed the gut and skin metagenomes of 25 mothers and their infants and highlighted some vertical mother-to-infant phage transmission events (e.g. Propionibacterium phages, Figure 1C).

We thus provide here computational tools for profiling viruses from raw metagenomic samples and describe our preliminary results that are showing how it is possible to identify, genetically characterise, track, and phylogenetically model viruses in the human microbiome.

Availability: http://segatalab.cibio.unitn.it

Authors:

first name	last name	email	country	organization	corresponding?
Moreno	Zolfo	moreno.zolfo@unitn.it	Italy	CIBIO, University of Trento	✓
Federica	Pinto		Italy	CIBIO, University of Trento	
Federica	Armanini		Italy	CIBIO, University of Trento	
Francesco	Asnicar		Italy	CIBIO, University of Trento	
Francesco	Beghini		Italy	CIBIO, University of Trento	
Pamela	Ferretti		Italy	CIBIO, University of Trento	
Edoardo	Pasolli		Italy	CIBIO, University of Trento	
Serena	Manara		Italy	CIBIO, University of Trento	
Adrian	Tett		Italy	CIBIO, University of Trento	
Nicola	Segata	nicola.segata@unitn.it	Italy	CIBIO, University of Trento	✓

MetaCortex: Assembling variation in metagenomics

Keywords: Metagenomics, variation, assembly, viral

Abstract: Assembling long and accurate contiguous sequence from metagenomic samples that represent species of variable abundance is a challenging problem and is relatively unsolved compared to the problem of genomic assembly. However, it also presents the opportunity to observe systems and populations in more realistic environs, as well as study the variation within closely situated species.

Current state of the art genomic and metagenomic assemblers do not attempt to capture this variation, instead sacrificing information content to produce assemblies which emphasise contiguity over complexity. In general, to simplify the problem of assembly, the coverage (or occurrence) of k-mers is relied upon to discard less common (and therefore less probable) sections in the assembly. Similarly, genomic features such as SNPs and indels are collapsed in order to simplify assembly graph traversal and deliver longer contigs. When considering a single genome, this is a safe assumption to make, and will reduce the overall errors in the dataset. However, in metagenomic data distinct but related subspecies or strains are likely present in same data, and often at vastly differing abundances.

We have developed a new assembler, MetaCortex, to address these issues. It introduces a number of new algorithms to adapt the existing genomic assembler Cortex to the problem of metagenomic assembly with a particular focus on capturing variation. The software initially screens the data for likely contaminants (such as probable host DNA); a de Bruijn graph is built and this is separated into subgraphs based on measures of graph connectivity and localised coverage. For each subgraph, MetaCortex produces a consensus contig, but crucially this stage seeks to preserve local variation which can then be output with contigs in GFA and FASTG format. We limit the production of chimeric contigs by the assembler by maintaining a similar coverage level within a contig.

Authors:

first name	last name	email	country	organization	corresponding?
Martin	Ayling	Martin.Ayling@earlham.ac.uk	United Kingdom	Earlham Institute	✓
Richard	Leggett	Richard.Leggett@earlham.ac.uk	United Kingdom	Earlham Institute	

Dysregulation of Human Endogenous Retroviruses in Primary CD4 T-Cells following Vorinostat treatment

Keywords: Vorinostat, Human endogenous retroviruses, RNA-Seq, Droplet Digital PCR, LTR12, Primary CD4 T-cells

Abstract: Treatment strategies for infection with the Human Immunodeficiency Virus (HIV) have focused primarily upon suppressing viral replication with anti-retroviral therapy (ART). This has drastically increased the lifespan of HIV infected individuals. However, ART has proven ineffective at eradicating the latent HIV reservoir, a pool of long-lived, persistently infected cells that can replicate HIV following the interruption of ART. Strategies for a cure for HIV have focused upon the destruction of this latently infected pool of cells. A method that has shown promise for a cure for HIV is the "Shock and Kill" strategy, where HIV is activated to induce virus expression in these latently infected cells, and then the activated cell either dies or is eliminated by the immune system with the possible assistance of adjuvant therapy. A group of compounds known as histone deacetylase inhibitors (HDACis), such as the drug Vorinostat (a.k.a. SAHA), have been shown to activate HIV in latently infected cells, and have been utilized in the "Shock" portion of this strategy.

The method by which HDACis activate HIV could conceivably upregulate additional human endogenous retroviruses (HERVs) present in the human genome and lead to unwanted or potentially dangerous off-target effects. HERVs comprise nearly 8% of the human genome and arise from retroviral infections of germ line cells and subsequent incorporation into the host genome. Older HERVs entered the mammalian genome as early as 55 million years ago, and as recently as 100 thousand years ago. Several HERVs can actively produce retroviral like particles and previous studies have demonstrated that upregulation of HERV sequences is associated with various diseases such as schizophrenia, cancer, and multiple sclerosis. This potential increased risk of disease and genome instability associated with HERV upregulation advocate for a large scale analysis of HERV upregulation upon treatment with HDACis.

We present the first large scale analysis of HERVs dysregulated upon treatment with the HDACi SAHA and show persistent dysregulation of several HERV species. We further demonstrate that upregulation of one HERV family, LTR12, is seen not only at high dosages of SAHA, but also in lower doses. Additionally, to address the complexities of the study of HERVs such as their imperfect replication across multiple locations in the genome, we present an overall analysis pipeline that utilizes RNA-Seq in an initial step to identify dysregulated HERV elements and droplet-digital (dd) PCR to assess dysregulation at various dosages. This methodology is ideal for primer and probe design for HERV sequences, which may demonstrate different levels of expression across the HERV element. Researchers are encouraged to utilize this method for HERV dysregulation as the RNA-Seq screening methodology can detect transcript expression from any location in the genome, regardless of whether that region has been extensively studied previously. Furthermore, ddPCR is an inexpensive method to expand upon RNA-Seq results for primer and probe design for additional experiments as was done in this work.

Authors:

first name	**last name**	**email**	**country**	**organization**	**corresponding?**
Cory	White	c.h.white@soton.ac.uk	United Kingdom	University of Southampton	✓
Nadejda	Beliakova-Bethell	nbeliakovabethell@ucsd.edu	USA	University of California, San Diego	
Steven	Lada	slada@vapop.ucsd.edu	USA	University of California, San Diego	
Michael	Breen	michaels.breen@mssm.edu	USA	Icahn School of Medicine at Mt. Sinai	
Gkikas	Magiorkinis	gkikas.magiorkinis@zoo.ox.ac.uk	United Kingdom	University of Oxford	
Douglas	Richman	drichman@ucsd.edu	USA	University of California, San Diego	
John	Frater	john.frater@ndm.ox.ac.uk	United Kingdom	University of Oxford	
John	Holloway	J.W.Holloway@soton.ac.uk	United Kingdom	University of Southampton	
Christopher	Woelk	c.h.woelk@soton.ac.uk	United Kingdom	University of Southampton	✓

The SuperTranscriptome: an alternative reference for RNA-seq analysis

Keywords: RNA-seq, transcriptome, assembly, non-model organisms

Abstract:

BACKGROUND

Numerous methods have been developed to analyze RNA sequencing data for a variety of purposes including examining expression at the gene level, inferring transcript abundances, detecting differential isoform usage or even identifying variation in expressed sequence. However, many methods rely on the availability of a reference genome, making them unsuitable for use with data generated from non-model organisms. Instead, for non-model organisms an experiment specific transcriptome can be built from RNA-seq data through de novo transcriptome assembly. However, with the exception of a few methods, such as transcript quantification, many analytical approaches for RNA-Seq are designed to work with a reference genome rather than transcriptome. For example visualization of read coverage across a gene is impossible using a reference transcriptome. At best, reads may be mapped and visualized against a representative transcript from each gene, such as the longest isoform.

RESULTS

Here we present an alternative representation for each gene [1], which we refer to as a superTranscript, designed to enable a variety of RNA-Seq analyses in non-model organisms. SuperTranscripts contain the sequence of all exons of a gene without redundancy. They can be constructed from any set of transcripts including de novo assemblies and we have developed a python program to build them called Lace (available from https://github.com/Oshlack/Lace/wiki). Lace works by building a splice graph for each gene, then topologically sorting the graph using Kahn's algorithm [2]. Building superTranscripts is a simple post-assembly step that promises to unlock numerous analytical approaches for non-model organisms. Although superTranscripts do not necessarily represent any true biological molecule, they provide a practical replacement for a reference genome. We show that reads can be aligned to the superTranscriptome using a splice aware aligner and subsequently visualized using standard tools. In addition, quantification of alternative isoforms can be performed with existing software by counting the reads that overlap superTranscript features. We also show that Lace assembled superTranscripts can be used to accurately call variants. We also demonstrate applications of superTranscripts to model organisms. Specifically, we combined a reference and de novo assembled transcriptome into a compact superTranscriptome using chicken RNA-seq data and found conserved coding sequence in hundreds of genes that was missed in the current chicken reference genome.

CONCLUSIONS

SuperTranscripts are a set of sequences, one for each expressed gene, containing all exons without redundancy and Lace is software to construct them. Lace and superTranscripts can potentially be applied in a broad range of scenarios, and we present just a few examples. SuperTranscripts have the power to transform how studies of non-model organisms are performed as a multitude of the standard analytical tools and techniques can now be applied across all species.

REFERENCES

1. Davidson N.M., Hawkins A.D.K and Oshlack A., SuperTranscript: a data driven reference for analysis and visualisation of transcriptomes. Biorxiv pre-print, http://biorxiv.org/content/early/2017/04/11/077750

2. Kahn A.B., Topological sorting of large networks. Commun ACM 1962, 5:558–562.

Availability: http://oshlacklab.com/

Authors:

first name	last name	email	country	organization	corresponding?
Nadia	Davidson	nadia.davidson@mcri.edu.au	Australia	Murdoch Childrens Research Institute	✓
Anthony	Hawkins	adkhawkins@gmail.com	Australia	Murdoch Childrens Research Institute	
Alicia	Oshlack	alicia.oshlack@mcri.edu.au	Australia	Murdoch Childrens Research Institute and The University of Melbourne	✓

Improved VCF normalization for accurate VCF comparison

Keywords: Variant calling, Variant call format, VCF, Variant calling evaluation, VCF normalisation

Abstract: Motivation: The Variant Call Format (VCF) is widely used to store data about genetic variation. Variant calling workflows detect potential variants in large numbers of short sequence reads generated by DNA sequencing and report them in VCF format. To evaluate the accuracy of variant callers, it is critical to correctly compare their output against a reference VCF file containing a gold standard set of variants. However, comparing VCF files is a complicated task as an individual genomic variant can be represented in several different ways and is therefore not necessarily reported in a unique way by different software.

Results: We introduce a VCF normalization method called Best Alignment Normalisation (BAN) that results in more accurate VCF file comparison. BAN applies all the variations in a VCF file to the reference genome to create a sample genome, and then recalls the variants by aligning this sample genome back with the reference genome. Since the purpose of BAN is to get an accurate result at the time of VCF comparison, we define a better normalization method as the one resulting in less disagreement between the outputs of different VCF comparators.

Availability and Implementation: The BAN Linux bash script along with required software are publicly available on https://sites.google.com/site/banadf16

Availability: http://www.cse.unsw.edu.au/~ignjat

Authors:

first name	last name	email	country	organization	corresponding?
Arash	Bayat	a.bayat@unsw.edu.au	Australia	University of New South Wales	✓
Bruno	Gaeta	bgaeta@unsw.edu.au	Australia	UNSW	✓
Aleksandar	Ignjatovic	ignjat@cse.unsw.edu.au	Australia	UNSW	
Sri	Parameswaran	sri.parameswaran@unsw.edu.au	Australia	UNSW in Sydney	✓

Differential Sample Proportion For Indexing High-Throughput Sequencing Experiments

Keywords: high throughput sequencing, hidden Markov model, barcoding

Abstract: Next-generation sequencing library preparation can be costly and labour intensive. Classically, the per sample sequencing cost can be reduced by pooling samples on the same sequencing lane. In this case, the sequencing library preparation includes a step where short sequences, known as barcodes or indexes, are added to the input DNA to act as sample identifiers. After sequencing occur, barcodes permit to assign short-read sequences to the sample they belong to bioinformatically. However, the cost of the library increases with the number of samples, i.e. barcodes, used in the preparation. We propose to reduce the number of required barcodes by using different known proportions of DNA material as sample identifiers. As a result, each sample is characterised by a specific expected depth of coverage. We have developed a hidden Markov model that takes into account these expected proportions of coverage to reconstruct the sample input sequences using a reference sequence. In silico simulation experiments indicate that sequence coverage can be efficiently used to index the short-reads and that we can successfully reassemble the input sample sequences. Additionally, we validate the method by sequencing pools of mitochondrial DNA amplicons extracted from western and eastern grey kangaroo, Macropus fuliginosus and giganteus respectively, and recovering the original sequences.

Authors:

first name	last name	email	country	organization	corresponding?
Louis	Ranjard	louis.ranjard@gmail.com	Australia	The Australian National University	✓
Thomas	Wong	thomas.wong@anu.edu.au	Australia	The Australian National University	
Allen	Rodrigo	allen.rodrigo@anu.edu.au	Australia	The Australian National University	

Abundance-based reconstitution of microbial pan-genomes from whole-metagenome shotgun sequencing data

Keywords: Metagenomics, Microbiome Analysis, Clustering

Abstract: Background

Recently, there has been a growing interest in studying gene content variability within strains of host-associated microbial species. Indeed, the accessory genes of a species provide functional advantages (e.g. pathogenicity) which impact the health of the host. Advent of shotgun metagenomics where whole-community DNA is randomly sequenced and development of dedicated bioinformatic tools allowed culture-free characterization of individual strain gene content. However, these tools are currently limited by the extent of available reference genomes. For instance, it is estimated that 50% of the species present in the gut microbiota of western individuals lack reference genomes and many have only one or a few representatives.

Results

We introduce MSPminer, a computationally efficient software tool that reconstitutes Metagenomic Species Pan-genomes (MSPs) by binning co-abundant genes obtained from de novo assembly and integration of reference genomes. It can empirically distinguish species core genes from accessory and shared genes based on the survey of thousands of samples. The method relies on a new robust score that is only limited by the number of reads mapped on a gene and its prevalence across samples.

MSPminer was applied to the largest gene abundance table available to date made up of 9.9M genes quantified among 1267 stool samples. This dataset was processed in 3 hours on a regular single node server and identified 1335 MSPs with at least 200 core genes. Many unknown microbes were discovered as only 255 MSPs (19.1%) were annotated at the species level after comparison against RefSeq and KEGG. As reported by others, Escherichia coli and Bacteroides bacteria were the most variable MSPs with an number of accessory genes ranging from 6,720 to 11,764. Finally, many accessory genes were identified for some species with few reference genomes including Prevotella copri (2874 new genes) and Bacteroides thetaiotaomicron (3280 new genes).

Conclusions:

MSPminer successfully reconstructs the genes repertoire of unknown species and enriches that of known species with new representatives. Through the growing efforts in sequencing metagenomics, each new sample offers the opportunity to increase our knowledge of microbial gene content variability, at a pace far beyond the pace of culture-based methods. Moreover, by distinguishing core and modular genes, the proposed method allows more sensible downstream analyses such as estimation of species abundance or discovery of strain-specific biomarker genes.

Availability: https://github.com/fplaza

Authors:

first name	last name	email	country	organization	corresponding?
Florian	Plaza Oñate	florian.plaza@gmail.com	France	Enterome	✓
Alessandra	Cervino	acervino@enterome.com	France	Enterome	
Frédéric	Magoulès	frederic.magoules@centralesupelec.fr	France	CentraleSupélec	
S. Dusko	Ehrlich	dusko.ehrlich@inra.fr	France	INRA MetaGénoPolis	
Matthieu	Pichaud	matthieu.pichaud@gmail.com	France	Enterome	✓

EMBALMER enables mathematically optimal genome database search for big data

Keywords: DNA alignment, short-read, optimal, aligner, read mapping, metagenomics, RefSeq, NGS, high-throughput

Abstract: One of the fundamental tasks in analyzing next-generation sequencing data is genome database search, in which DNA sequences are compared to known reference genomes for identification or annotation. Although algorithms exist for optimal database search with perfect sensitivity and specificity, these have largely been abandoned for next-generation sequencing (NGS) data in favor of faster algorithms that sacrifice alignment quality under precision, accuracy, sensitivity, and recall. Here we introduce EMBALMER, a mathematically optimal high-throughput DNA short-read aligner that enables provably optimal alignment with speed up to one million times faster than the fastest optimal alignment algorithms by relying on several key novel optimizations. Moreover, EMBALMER guarantees to find all equally good matches in the database and can interpolate conservative taxonomic annotation for sequences that match multiple genomes. EMBALMER also losslessly computes the minimal set of matching references for a given set of input sequences and has several other useful modes of operation. Benchmarks on a single compute node at the Minnesota Supercomputing Institute (Mesabi 32-core Ivy Bridge) show alignment speed of roughly 5,000 NGS metagenomic shotgun reads per second against the entire RefSeq 88,000-genome taxonomy-annotated microbial CDS database at a 98% identity threshold and peak RAM usage of 140GB on a 16-million query test set, including all-ties taxonomy interpolation and minimal reference set computation. Performance exceeds 1 million reads per minute on a 5,000-genome "representative microbe" database. EMBALMER is able to scale runtime sub-linearly with input data size by leveraging redundant information contained in the input data; the larger the data set, the faster the runtime per sequence. EMBALMER is open-source and provided as a single, dependency-free binary for all major operating systems. Because it runs up to six orders of magnitude faster than the current best optimal DNA aligners, EMBALMER will support a paradigm shift in next-generation DNA sequencing analysis by allowing, for the first time, complete removal of any error or approximation in genome database search with big data.

Availability: http://knightslab.org

Authors:

first name	last name	email	country	organization	corresponding?
Gabriel	Al-Ghalith	algh0022@umn.edu	USA	University of MN - TC	
Dan	Knights	dknights@umn.edu	USA	University of MN - TC	✓

Unraveling Cortical Development Using Population and Single-cell RNA-Seq Data

Keywords: High-throughput sequencing, Single cell RNA sequencing, RNA-Seq, Machine Learning, Bioinformatics, Next Generation Sequencing, population, fate decision-making, Brain development, Cortical development, Single cell dynamics

Abstract: The brain, as part of the central nervous system, is the most complex organ in the mammalian body and the mechanisms that regulate its development are poorly understood. During brain development, neural stem cells (NSCs) generate thousands of different neuronal subtypes that are organized in precise and functionally distinct layers of the cerebral cortex. This process is a prerequisite for normal brain functions and any deviation from the standard developmental path can lead to debilitating brain disorders.

To unravel these mechanisms, we study changes in the expression of transcription factors and signaling components in NSCs and progenitor populations (NeuroStemX, SystemsX.ch). To this end, we use population and single-cell RNA sequencing of each population at daily intervals during mouse cortical development obtaining a data set containing more than 100 population samples and more than 1000 single cells. This enabled us to identify a set of novel genes that characterizes NSCs and progenitor cells at the population and single cell level at distinct stages of brain development. Using machine-learning methods, we identified a continuous differentiation path and, from this, determined different transcriptional states. Remarkably, we can show that the single cells follow a similar differentiation path to that predicted from transcriptional analysis at the population level. In addition, the single cells can be divided into subpopulations that emerge over time. In summary, our gene expression data of different cell types at the population and single-cell level for daily intervals during neurogenesis combined with the appropriate data analyses give an unprecedented insight into the complex process of stem cell patterning and fate decision-making in early brain development.

Authors:

first name	last name	email	country	organization	corresponding?
Zahra	Karimaddini	zahra.karimaddini@bsse.ethz.ch	Switzerland	ETH Zürich	✓
Tanzila	Mukhtar		Switzerland	University of Basel	
Verdon	Taylor		Switzerland	University of Basel	
Dagmar	Iber		Switzerland	ETH Zürich	

Single-cell doublets detection and Cell Cycle stages prediction in Fluidigm C1 system

Keywords: Single-cell, Cell Cycle Prediction, Gaussian Mixture Models, Doublets Classification

Abstract: Background

The Fluidigm C1 is a single-cell analysis system uses a simplified single-cell isolation and cell processing based on Integrated Fluidic Circuits (IFCs). This system relies on a microfluidic device to achieve maximum efficiency in cell capture. However, it has been shown that this technology is affected by a no negligible 30% rate of cell doublets, which although addressed and improved in recent release of the IFCs, is still a significant issue in the downstream data analysis.

The C1 standard workflows is developed on the assumption that only one cell is present in one capture site of the IFCs, this is ensured by a manual curation of the images of the captured cells. This process needs to happen in a short temporal window to avoid alteration of the genetic material and bias in gene expression quantification. This time constraint limits the resolution of the images which makes harder to analyse them both manually and with automated methods. Moreover the dimensionality of the data associated to the images is low and this compromises classification performance both with supervised and unsupervised methods.

Results

This problem was previously addressed using a fully supervised approach based on gene expression markers by [1] with routinely image checking of the IFCs. However, the majority of doublets come in the form of stacked doublets, which are difficult to identify without high-magnification imaging and not detectable with gene expression markers. Moreover, it is important for the accuracy of the subsequent single-cell RNA-seq analysis that each sample is of a single cell. Hence, we propose here a Mixture of Gaussians based model (MOG) [2] based on the initial IFCs images data with two objectives: (a) to build an unsupervised classifier of doublets using data from fluorescent staining of cells with Hoechst 33342; (b) to predict cell cycle stages by validating true labels against intensities of Hoechst 33342 staining and image features

To achieve this, we learn a MOG from the images IFCs data using an automated microscope Jason Fucci of fluorescently labels red and green embryonic stem cells stained with Hoechst 33342. We optimised the number of the MOG components and the number of needed data features by validation with true labels. We trained the MOG which is a fully unsupervised method and compared with a semi-supervised and a fully supervised. Results show that the ideal unsupervised approach is able to give very reliable results for doublets (99%), G1 phase (100%) and S phase (98%) and seems to have still problems with G2 phase (60%). Same problem is reported for the fully supervised (99%, 100%, 64%, 98% respectively) approach. We show that adding prior knowledge to the MOG and therefore adding more complexity to the model, we are able to improve the G2 discrimination and obtain a 86% success. The G2 phase being very close to the M phase is subject to a high variability in the labelling, this could be addressed by introducing the M phase label.

Conclusions

In this study we show that it is possible to automatically classify doublets and give a characterisation of the cell cycle stages distribution of the population of single cells, using mixture of Gaussians with prior knowledge on captured images of cell depositions. Our approach gives the great advantage of characterising cells before the RNA-seq assay and therefore gives great interpretation power to the following RNA-seq data analysis. Future improvements of this approach are

based on optimising prior selection and data features extracted from the IFCs images.

References: [1] Illicic T. et al. (2016) Classification of low quality cells from single-cell RNA-seq data. Genome Biology. DOI: 10.1186/s13059-016-0888-1; [2] Hensman J. et al. (2012). Fast variational inference in the conjugate exponential family Adv. Neural Inf. Process. Syst., 2012

Authors:

first name	last name	email	country	organization	corresponding?
Luisa	Cutillo	l.cutillo@sheffield.ac.uk	United Kingdom	University of Sheffield	
Max	Zwießele	m.zwiessele@sheffield.ac.uk	United Kingdom	University of Sheffield	
Paul	Gokhale	p.gokhale@sheffield.ac.uk	United Kingdom	University of Sheffield	
Marcelo	Rivolta	m.n.rivolta@sheffield.ac.uk	United Kingdom	University of Sheffield	
Marta	Milo	m.milo@sheffield.ac.uk	United Kingdom	University of Sheffield	✓

Estimation on the number of haplotypes from next-generation sequencing data

Keywords: haplotype reconstruction, high throughput sequencing, number of haplotypes

Abstract: Next-generation sequencing (NGS) provides a good opportunity to perform deep resolution on haplotype sequences. However, given a mixture of NGS reads from different haplotypes, estimation on the number of haplotypes is still a big challenge. Regarding the situation that all the haplotypes have the same frequency in the pool of sequenced DNA, we propose a novel statistical method to predict the number of haplotypes. Results obtained from simulated data sets showed that the method is promising. On average, the accuracy of the method was 100% when the actual number of haplotypes was 25, 98% when the number of haplotypes was 100, and 93.5% when the number of haplotypes was 1000.

On the other hand, when all the haplotypes have different frequencies in the pool of sequenced DNA, we propose another algorithm to predict the number of haplotypes. According to the experiments using simulated data sets, the accuracy reached 93% when the number of haplotypes was 15, and 90% when the number of haplotypes was 20. We also extended the algorithm to reconstruct each of the haplotypes. On average, the accuracy of each haplotype reconstruction (i.e. the number of correct bases on the resulting haplotype / the length of the real haplotype) was 98.6% when the number of haplotypes was 15, and 97.4% when the number of haplotypes was 20, according to the experiments using mixture of error-free reads simulated from a set of 20K-long haplotypes.

Availability: https://github.com/LouisRanjard

Authors:

first name	**last name**	**email**	**country**	**organization**	**corresponding?**
Thomas	Wong	thomas.wong@anu.edu.au	Australia	The Australian National University	✓
Louis	Ranjard	louis.ranjard@gmail.com	Australia	The Australian National University	
Allen	Rodrigo	allen.rodrigo@anu.edu.au	Australia	The Australian National University	

YAMP : Yet Another Metagenomic Pipeline

Keywords: Metagenomics, Analysis Workflow, Quality Control, Next Generation Sequencing

Abstract: Thanks to the increased cost-effectiveness of high-throughput technologies, the number of studies focusing on microorganisms and on their connections with human health and diseases has surged, and, consequently, a plethora of approaches and software has been made available for their study, making it difficult to select the best methods and tools.

Here we present Yet Another Metagenomic Pipeline (YAMP) that, starting from the raw sequencing data and having a strong focus on quality control (QC), allows, within hours, the data processing up to the functional annotation. Specifically, the QC (performed by means of several tools from the BBmap suite [1]), allows de-duplication, trimming, and decontamination of metagenomics sequences, and each of these steps is accompanied by the visualisation of the data quality. The QC is followed by multiple steps aiming at characterising the taxonomic and functional diversity of the microbial community. Namely, taxonomic binning and profiling is performed by means of MetaPhlAn2 [2], which uses clade-specific markers to both detect the organisms present in a microbiome sample and to estimate their relative abundance. The functional capabilities of the microbiome community are currently assessed by the HUMAnN2 pipeline [3] which first stratifies the community in known and unclassified organisms using the MetaPhlAn2 results and the ChocoPhlAn pan-genome database, and then combines these results with those obtained through an organism-agnostic search on the UniRef proteomic database. The next YAMP release, currently under development, will also support MOCAT2 [4] and an optimised version of the HUMAnN2 pipeline. QIIME [5] is used to evaluate multiple diversity measures.

YAMP is constructed on Nextflow [6], a framework based on the dataflow programming model, which allows writing workflows that are highly parallel, easily portable (including on distributed systems), and very flexible and customisable, characteristics which have been inherited by YAMP. Users can decide the flow of their analyses, for instance limiting them to the QC or using already QC-ed sequences. New modules can be added easily and the existing ones can be customised – even though we have already provided default parameters deriving from our own experience. YAMP is accompanied by a Docker container [7], that saves the users from the hassle of installing the required software, increasing, at the same time, the reproducibility of the YAMP results.

YAMP si available at https://sites.google.com/site/populationgenomics/yamp

References

1. https://sourceforge.net/projects/bbmap

2. Truong, D.T., et al. Metaphlan2 for enhanced metagenomic taxonomic profiling. Nature methods 12(10), 902–903 (2015)

3. https://bitbucket.org/biobakery/humann2

4. Kultima, J.R., et al. MOCAT2: A metagenomic assembly, annotation and profiling framework.

Bioinformatics 32(16), 2520–2523 (2016).

5. Caporaso, J.G., et al. QIIME allows analysis of high-throughput community sequencing data. Nature Methods 7(5), 335–336 (2010)

6. Di Tommaso, P., et al. Nextflow enables reproducible computational workflows. Nature Biotechnology 35, 316–319 (2017)

7. https://www.docker.com

Authors:

first name	**last name**	**email**	**country**	**organization**	**corresponding?**
Alessia	Visconti	alessia.visconti@kcl.ac.uk	United Kingdom	King's College London	✓
Tiphaine	Martin	tiphaine.martin@kcl.ac.uk	United Kingdom	King's College London	✓
Mario	Falchi	mario.falchi@kcl.ac.uk	United Kingdom	King's College London	✓

The etiology of somatic mutations in cancer: a probabilistic approach.

Keywords: cancer etiology, genome sequencing, somatic mutations

Abstract: Cancers are caused by mutations that are inherited, induced by environmental factors, or the result of random mistakes made during normal DNA replication. Unlike environmentally-induced (E) mutations, replicative (R) mutations are expected to have approximately the same distribution in all humans, regardless of their environment. To test the implications of this expectation, we correlated the number of normal stem cell divisions with cancer risk among 17 major cancer types in countries with widely different environments. We found that there were striking correlations between normal stem cell divisions and the incidence of these 17 cancer types in every one of the 69 countries that could be assessed (median correlation of 0.80). That R mutations play a major role in these correlations was supported by the results of an independent approach, based solely on cancer genome sequencing and epidemiologic data, suggesting that R mutations are responsible for 2/3 of the mutations that occur in human cancers. These results accentuate the importance of early detection and intervention to reduce deaths from the many cancers arising from unavoidable R mutations.

Tomasetti C*, Li L, Vogelstein B. Stem cell divisions, somatic mutations, cancer etiology, and cancer prevention. Science 2017, 355(6331):1330-1334.

Availability: http://www.cristiantomasetti.com

Authors:

first name	last name	email	country	organization	corresponding?
Cristian	Tomasetti	ctomasetti@jhu.edu	USA	Johns Hopkins University	✓
Lu	Li	lli48@jhu.edu	USA	Johns Hopkins Bloomberg School of Public Health	
Bert	Vogelstein	vogelbe@jhmi.edu	USA	Johns Hopkins Kimmel Cancer Center	

Statistically Synergizing Information from Inherited and Tumor Genomes to Infer the Impact of DNA Variants in Cancer

Keywords: cancer, whole-exome sequencing, susceptibility variants, statistical analysis

Abstract: Since the advent of whole-genome assays, there has been substantial work in the computational research community to develop methods that associate specific DNA variants with human disease. Most recently, these efforts have focused on rare variants, with most methods aggregating a gene's potentially deleterious alleles across disease cases, and comparing with those observed in healthy controls. These methods, termed burden tests or collapsing tests, are largely generic, and may be applied to large case cohorts from any disease, including cancer. However, cancer is unique among human disease in that clues may be gleaned from the matched tumor genome. For instance, genes that are important for malignancy onset and progression are often those that acquire mutations somatically, or that are subject to somatic duplication or deletion events. Therefore, these genes make attractive candidates in the search for important germline mutations. This approach is well-grounded in the literature, and dates back at least to Knudson's "two-hit" hypothesis. Here we describe statistical methods to incorporate information from somatic mutational and copy number aberration data in the search for germline risk variants in cancer. We apply the methods to a large cohort of paired normal-tumor genomes from patients with leukemia and related blood-based malignancies.

Authors:

first name	last name	email	country	organization	corresponding?
Thomas	Laframboise	txl80@case.edu	USA	Case Western Reserve University	✓
Samuel	Li		USA	Case Western Reserve University	

Sensitive protein sequence searching for the analysis of massive data sets

Keywords: Metagenomics, Protein search, Algorithm

Abstract: Sequencing costs have dropped much faster than Moore's law in the past decade, and sensitive sequence searching has become the main bottleneck in the analysis of large metagenomic datasets. We developed the parallelized, open-source software MMseqs2 (mmseqs.org), which improves on current search tools over the full range of speed-sensitivity trade-off, achieving sensitivities better than PSI-BLAST at ¿400 times its speed. MMseqs2 offers great potential to better exploit large-scale (meta)genomic data.

Availability: http://www.soeding.genzentrum.lmu.de/

Authors:

first name	**last name**	**email**	**country**	**organization**	**corresponding?**
Martin	Steinegger	martin.steinegger@mpibpc.mpg.de	Korea, Democratic People's Republic of	Max-Planck-Institute	
Johannes	Soeding	soeding@mpibpc.mpg.de	Germany	Max-Planck-Institute	✓

AmpTaxi toolkit for targeted gene sequencing analysis

Keywords: targeted sequencing, Bioinformatics toolkit, evolutionary analysis

Abstract: The importance of easy-to-use computational tools is evident since high-throughput sequencing technology is available to biology laboratories. Although many platforms and application programming interfaces are available, efforts are needed for targeted sequencing data analyses. Here, we developed a toolkit, AmpTaxi, to address questions about identifying targeted genes, such as duplicated family genes, and their evolutionary relationship from raw sequencing datasets for classification. Additionally, due to the amount of genes in question, we developed a program to parsing evolutionary trees in newick formats. Functionalities include query by bootstrap values or evolutionary distances. Several related workflows are established and users may apply the same procedures on their own data. The tool suite aims to provide straightforward and streamline utilities for Bioinformatics studies. We take the advantage of the popular cross-platform interface Galaxy. Thus, users are free from installation and compiling problems and can store, share and repeat analysis steps in one place. The programs are ready to use at our Galaxy instance, gxy.kuas.edu.tw.

Authors:

first name	last name	email	country	organization	corresponding?
Tzi-Yuan	Wang	tziyuan@gmail.com	Taiwan	Biodiversity Research Center, Academia Sinica	
Wen-Yu	Chung	wychung@kuas.edu.tw	Taiwan	National Kaohsiung University of Applied Sciences	✓

RNA-sequencing for prediction of the conventional breast cancer biomarkers ER, PgR, HER2, Ki67, and NHG

Keywords: breast cancer, RNA-seq, transcriptomics, gene expression profiling, machine learning

Abstract: We aimed to develop RNA-sequencing-based classifiers for the key breast cancer histopathological biomarkers — estrogen receptor (ER), progesterone receptor (PgR), human epidermal growth factor receptor 2 (ERBB2/HER2), Ki67, and Nottingham histological grade (NHG) — which are routinely used for determining prognosis and treatment in the clinic. To obtain reliable training labels we performed a multi-rater histopathological biomarker evaluation on a training cohort of 405 tumor samples. Using the resulting consensus-labels and RNA-seq-derived tumor gene expression data as input, we trained single-gene classifiers (SGC) and multi-gene nearest shrunken centroid classifiers (MGC). We assessed the performance of the resulting classifiers by comparing their predictions to the clinical biomarker status in an independent prospective population-based series of 3273 primary breast cancer cases from the SCAN-B study (ClinicalTrials.gov identifier NCT02306096; Saal et al, Genome Medicine 2015), and by analyzing the overall survival of the patients. The results show that concordance between histopathological evaluations was high for ER, PgR, and HER2, but only moderate for Ki67 and NHG. Within the 3273-cohort the concordance between our biomarker predictions and clinical histopathology was similar to the concordance baseline established in the multi-rater biomarker evaluation. Survival analysis showed that our predictions add clinical value to histopathology by identifying patients that could potentially benefit from additional treatment and patients with poor prognosis.

Availability: http://www.brueffer.io

Authors:

first name	last name	email	country	organization	corresponding?
Christian	Brueffer	christian.brueffer@med.lu.se	Sweden	Lund University	✓
Johan	Vallon-Christersson		Sweden	Lund University	
Dorthe	Grabau		Sweden	Skåne University Hospital	
Anna	Ehinger		Sweden	Blekinge County Hospital	
Jari	Häkkinen		Sweden	Lund University	
Cecilia	Hegardt		Sweden	Lund University	
Janne	Malina		Sweden	Skåne University Hospital	
Yilun	Chen		Sweden	Lund University	
Pär-Ola	Bendahl		Sweden	Lund University	
Jonas	Manjer		Sweden	Skåne University Hospital	
Martin	Malmberg		Sweden	Skåne University Hospital	
Christer	Larsson		Sweden	Lund University	
Niklas	Loman		Sweden	Skåne University Hospital	
Lisa	Rydén		Sweden	Skåne University Hospital	
Åke	Borg		Sweden	Lund University	
Lao H.	Saal	lao.saal@med.lu.se	Sweden	Lund University	

DEScan: Differential enrichment scan of epigenomic high-throughput sequencing data that leverages biological variation to ensure reproducibility

Keywords: epigenomics, differential enrichment, reproducibility

Abstract: A common goal in epigenomic sequencing studies is to identify differences between conditions, i.e differential enrichment. Strategies to do so fall in two general categories: peak and window based. While window based strategies risk testing too many regions in which there is no signal, peak based strategies can introduce biases if peak calling is not done properly. Accuracy of peak location is therefore paramount and can have great impact on the reliability of the study results. Current peak calling algorithms do not explore the issue of peak reproducibility and its effects on subsequent statistical testing. Here we introduce DEScan, an R based integrated peak and differential caller, specifically designed to accommodate epigenomic signal of variable width and leverage biological reproducibility. DEScan has two modules: a novel peak calling module and a differential enrichment module that relies on existing Bioconductor packages. The first module integrates peak calling on individual samples using a variable-width window (without the need of a separate background sample) with evaluation of reproducibility of peak location across replicates to produce a count matrix for statistical testing. The differential enrichment module implements normalization and statistical testing in a manner analogous to RNA-seq data analysis, with a data normalization approach tailored to epigenomic data. Using two differential HTS datasets generated in-house: chromatin accessibility (Sono-seq, 4x2 replicates) and histone acetylation (H3K9ac ChIP-seq, 8x2 replicates) we illustrate the relationship between significance of peak calling within a sample and reproducibility across replicates. We show that more than 3 replicates should be considered for the number of peak locations to remain stable at a given z-score for peak calling. We then evaluate the impact of normalization methods in differential epigenomic analysis. Our results caution against using a normalization factors derived from total counts, pointing to a fundamental difference between transcriptomic and epigenomic data.

Availability: https://medicine.wsu.edu/facultyandstaff/lucia-peixoto/

Authors:

first name	last name	email	country	organization	corresponding?
John	Koberstein	john.koberstein@wsu.edu	USA	Washington State University	
Bruce	Gomes	bruce.gomes@wsu.edu	USA	Washington State University	
Shane	Poplawski	sgpoplawski@gmail.com	USA	Ibis Biosciences	
Nancy	Zhang	nzh@wharton.upenn.edu	USA	University of Pennsylvania	✓
Ted	Abel	ted-abel@uiowa.edu	USA	University of Iowa	✓
Lucia	Peixoto	lucia.peixoto@wsu.edu	USA	Washington State University	✓

In-silico read normalization using set multi-cover optimization

Keywords: Read Normalization, Sequence Assembly, RNA-Seq, Metagenomics, Set Mulcover Problem, De Bruijn graph

Abstract: De Bruijn graphs are a common assembly data structure for large sequencing datasets. But, with the advances in deep sequencing technologies, assembling a high coverage dataset has become computationally challenging in terms of memory and runtime requirements. Read normalization, a lossy read filtering approach, which removes redundancy in large datasets, is widely applied to reduce resource requirements. Although efficient, current normalization algorithms provide no guarantee to preserve important k-mers that might come from lowly expressed or unique regions of interest. Loss of such k-mers also affects the assembly as they form connections between different regions of the graph.

In this work, normalization is phrased as a set multicover problem on k-mer distribution in the complete dataset and a linear time heuristic algorithm named ORNA is proposed. ORNA normalizes the input dataset to the minimum number of reads required to retain all k-mers a certain number of times. With the aim of maintaining relative k-mer abundance, which is important for RNA-seq and metagenomics data, ORNA uses weighted threshold for individual k-mer. Hence, all the connections and relative coverage information from the original dataset is preserved. ORNA was tested on various RNA-seq and metagenomic datasets with different coverage values. The reduction was compared against the current state-of-art normalization algorithms namely Diginorm and Trinity's In Silico Normalization (TIS). These algorithms base their decision upon mean k-mer coverage of a read. The reduction was done over several parametrizations of the three algorithms. The performance was evaluated by assembling the reduced datasets using de-Bruijn graph based methods - Oases and TransABySS. Universally, it was found that for a similar percent of reduction, ORNA reduced datasets were able to assemble 5-10% more annotated transcripts as compared to Diginorm and TIS. In addition, ORNA reduced datasets better maintained gene expression estimates compared to the original dataset, as evident by higher correlation values. Similar results were obtained for metagenomic assemblies. ORNA supports multithreading and is memory efficient since it uses cascading bloom filters to store k-mer information. It is found to be the fastest method with comparable memory requirements to the others. Due to ORNA's unique ability to retain available k-mer connections, a new application is introduced where joint normalization of many datasets can be used to assemble regions of low coverage in individual datasets.

Finally, ORNA is a general purpose normalization algorithm that is fast and significantly reduces datasets with little loss on assembly quality. It is shown that in combination with read error correction ORNA may even improve assembly performance compared to running the full dataset albeit with significantly reduced assembly resource requirements. ORNA is freely available at

https://github.com/SchulzLab/ORNA

Availability: http://hgsb.mpi-inf.mpg.de

Authors:

first name	last name	email	country	organization	corresponding?
Dilip Ariyur	Durai	ddurai@mmci.uni-saarland.de	Germany	MPII Saarbrücken, MMCI Saarbrücken	
Marcel	Schulz	mschulz@mmci.uni-saarland.de	Germany	MMCI Saarbrücken, MPII Saarbrücken	✓

Shotgun metagenomic DNA sequencing of hypersaline semi-arid desert reveals the coexistence of archaeal, bacterial and viral community

Keywords: next generation sequencing, semi-arid desert, community, MG-RAST, MetaVir

Abstract: The next generation sequencing (NGS) technologies open the new avenue to analyze the massive microbial communities in the environment. The semi-arid hypersaline desert harbors the unique diversity of halophilic and haloalkaliphilic bacteria and archaea. Viruses are also paly the key role in the host interaction. However, the coexistence and distribution viruses and their microbial host in the semi-arid desert are yet not well explored by NGS. This is the first study to decode the whole microbial through Torrent PGM platform; which shed new light on microbial ecology, diversity, virus–host interactions and the discovery of novel viruses and bacterial species. Isolated and electrophoretic purify metagenomic DNA of desert called Little Rann of Kutch was subjected to whole-genome shotgun sequencing performed using the 318 Chip and 300-bp chemistry Ion Torrent PGM platform. Taxonomic composition of the metavirome was done by genome length normalization with a threshold of 103 on e-value, as generated by MetaVir; While the prokaryotic structure was depicted from the MG-RAST server. Total 2487 virus reads were obtained, among them 70% reads that produced a significant database hit were recognized as belonging to dsDNA viruses with no RNA stage, 21% ssDNA viruses and 4% were unclassified phages. More than half taxonomic hits were comprised the three families i.e. Siphoviridae family (26%) followed by Myoviridae (15%) and inoviridae. In prokaryotes, 43 phyla together with the unclassified category at phyla level were recorded. Alteromonadaceae (51.6%) and Halobacteriaceae (12.6%) were a major dominating family in the bacterial and archaeal domain. Finally, at species level total 1802 species were reported. Putative functions of the assembled dataset were predicted using MGRAST server. Nearly 19% hits were, directly and indirectly, related to the viruses, 18.2% fell in poorly characterized group suggest the possibilities of gaining novel gene from the metagenome. 4% gene belonged to various stress responses. Most abundant viruses belong to bacterial and archaeal phages suggest the coexistence of viruses and prokaryotic communities. The various unclassified and unassigned genes suggest the possibilities of auxiliary metabolic genes and new species of prokaryotes and viruses.

Authors:

first name	last name	email	country	organization	corresponding?
Pravin	Dudhagara	dudhagarapr@gmail.com	India	Department of Biosciences, VNSG University, surat	✓
Rajesh	Patel	raj252000@gmail.com	India	Department of Life Sciences, HNG University, patan	
Anjana	Ghelani	ghelanianjana@gmail.com	India	HNG University	

MinHash techniques for scalable compositional analyses

Keywords: database search, high throughput sequencing, Oxford Nanopore, approximate alignment, minhash, single-molecule sequencing

Abstract: The rapid growth of genomic and metagenomic data has begun to outpace traditional methods for sequence analysis. We previously addressed the problem of pairwise mutation distance estimation with Mash, which uses MinHashing to reduce large sequences to small, representative sketches. While Mash enabled clustering and search operations that were infeasible with traditional alignment, it was insensitive to the individual components of a sequence, such as specific genes in a genome or individual species in a metagenome. Here we extend Mash to estimate the containment of target sequences within a greater set, and we demonstrate several use cases for this operation, including antibiotic resistance detection in an Oxford Nanopore sequencing dataset and searching the Sequencing Read Archive (SRA) for a target organism. In combination with MashMap, which uses localized MinHashing to perform approximate alignments, we demonstrate that the Mash toolkit can be used to efficiently compute all significant alignments and coverage profiles of datasets while scaling to large databases, including the NCBI RefSeq database containing ¿800Gbp of sequence.

Availability: http://www.cbcb.umd.edu/~sergek/

Authors:

first name	last name	email	country	organization	corresponding?
Brian	Ondov	brian.ondov@nih.gov	USA	NHGRI/NIH	
Alexander	Dilthey	alexander.dilthey@nih.gov	USA	NHGRI/NIH	
Chirag	Jain	cjain@gatech.edu	USA	Georgia Institute of Technology	
Sergey	Koren	sergek@umd.edu	USA	NHGRI/NIH	✓
Adam	Phillippy	adam.phillippy@nih.gov	USA	NHGRI/NIH	

Nanopore Sequencing Technology and Tools: Computational Analysis of the Current State, Bottlenecks, and Future Directions

Keywords: Nanopore Sequencing, Single-Molecule DNA Sequencing, MinION, High-Throughput Sequence Analysis, Basecalling, Genome Assembly, Polishing, Performance Analysis, Bottleneck Analysis

Abstract: Nanopore sequencing, a promising single-molecule DNA sequencing technology, exhibits many attractive qualities and, in time, could potentially surpass current sequencing technologies. Nanopore sequencing promises higher throughput, lower cost, and increased read length, and it does not require a prior amplification step. Nanopore sequencers rely solely on the electrochemical structure of the different nucleotides for identification, and measure the ionic current change as long strands of DNA (ssDNA) pass through the nanoscale protein pores.

Biological nanopores for DNA sequencing were first proposed in the 1990s, but were only made commercially available in May 2014 by Oxford Nanopore Technologies (ONT). The first commercial nanopore sequencing device, MinION, is an inexpensive, pocket-sized, high-throughput sequencing apparatus that produces real-time data using the R7 nanopore chemistry. These properties enable new potential applications of genome sequencing, such as rapid surveillance of Ebola, Zika or other epidemics, near-patient testing, and other applications that require real-time data analysis. This technology is capable of generating very long reads (50,000bp) with minimal sample preparation. Despite all these advantageous characteristics, it has one major drawback: high error rates. In May 2016, ONT released a new version of MinION that uses a nanopore chemistry called R9. Although R9 improves data accuracy over R7, the error rate remains high. To take advantage of the real-time data produced by MinION, the tools used for nanopore sequence analysis must be fast and must overcome high error rates.

Our goal in this work is to comprehensively analyze current publicly available tools for nanopore sequence analysis, with a focus on understanding the advantages, disadvantages, and bottlenecks of them. It is important to understand where the current tools do not perform well in order to develop better tools. To this end, we analyze the multiple steps and tools in the nanopore genome analysis pipeline; and also provide some guidelines for determining the appropriate tools for each step of the pipeline and the corresponding parameters of them.

The first step, basecalling, translates the raw signal output of MinION into nucleotides to generate DNA sequences. Metrichor is the cloud-based basecaller of ONT; while Nanocall and Nanonet are publicly available nanopore basecallers. Overlap-layout-consensus (OLC) algorithms are used for nanopore sequencing reads since they perform better with longer error-prone reads. The second pipeline step finds read-to-read overlaps. Minimap and GraphMap are the commonly used tools for this step. After finding the overlaps, OLC-based assembly algorithms generate an overlap graph, where each node is a read and each edge is an overlap connecting them. The third pipeline step, genome assembly, traverses this graph, producing the layout of the reads and then constructing the draft assembly. Canu and Miniasm are the commonly used error-prone long-read assemblers. In order to increase the accuracy of the assembly, further polishing may be required. The first step of polishing is mapping the raw basecalled reads to the generated draft assembly from the previous step. The most commonly used long read mapper is BWA-MEM. After aligning the basecalled reads to the draft assembly, the final polishing of the assembly can be performed with Nanopolish.

We analyze the aforementioned nanopore sequencing tools with the goals of determining their bottlenecks and finding improvements to these tools. First, we compare the performance of the chosen tools for each step in terms of accuracy and speed. After the basecalling, read overlap finding, and assembly steps, the generated draft assemblies are compared with their reference genome; and the coverage of and identity with the reference genome are used to gauge their accuracy. For two of the draft assemblies, read mapping and polishing steps are further applied and the generated polished sequences are compared similarly. The execution time of each tool is recorded in order to compare the performance of the tools. Second, we analyze the first two steps of the pipeline in detail in order to assess the scalability of these tools. The performance of each basecaller and each read overlap finder as we vary the thread count is analyzed; wall clock time, peak memory usage, and parallel speedup are the metrics used for comparison. We present our key results in this work, and we expect future work to examine other stages of the pipeline and provide end-to-end results and analyses.

Authors:

first name	last name	email	country	organization	corresponding?
Damla	Senol	dsenol@andrew.cmu.edu	USA	Carnegie Mellon University	✓
Jeremie	Kim	jeremiek@andrew.cmu.edu	USA	Carnegie Mellon University	
Saugata	Ghose	ghose@cmu.edu	USA	Carnegie Mellon University	✓
Can	Alkan	calkan@cs.bilkent.edu.tr	Turkey	Bilkent University	
Onur	Mutlu	onur.mutlu@inf.ethz.ch	Switzerland	ETH Zurich	✓

mint: Integrating DNA methylation and hydroxymethylation

Keywords: DNA methylation, DNA hydroxymethylation, Bisulfite sequencing, Pipeline, High-throughput sequencing analysis

Abstract: DNA methylation (5mC) plays roles in many biological processes, including: mammalian development, oncogenesis, and response to environmental exposures. DNA hydroxymethylation (5hmC) is also an informative epigenetic mark having roles in regulation, development, and cancer that are distinct from 5mC. Two classes of technologies are used to detect DNA methylation and hydroxymethylation: bisulfite-conversion (BS) followed by deep sequencing, and immunoprecipitation to the particular mark followed by deep sequencing. Bisulfite-conversion methods are the gold-standard technologies for assessing DNA methylation, but cannot distinguish 5mC and 5hmC because both marks are resistant to bisulfite-conversion. Consequently, additional experiments are needed to tell the difference between the two marks, and in silico methods are required to analyze, integrate, and interpret these data.

We developed the Methylation INTegration (mint) pipeline to enable the comprehensive analysis of bisulfite-conversion (BS) and immunoprecipitation (IP) based methylation and hydroxymethylation assays. The mint pipeline supports three different platform configurations: a ‘hybrid’ of platform types where BS-based platforms, measuring 5mC and 5hmC together, are combined with IP-based platforms, measuring 5mC or 5hmC alone; a combination of IP-based platforms measuring 5mC and 5hmC separately; or data from any one of the platforms alone.

The mint pipeline performs standard analysis steps including QC, read trimming, alignment, and methylation quantification (BS-based) or peak calling (IP-based). Differential methylation can be determined for simple group comparisons or with general multi-factor designs with covariates. If data for different methylation marks is present, then mint will integrate signal from both marks into CpG- or region-level classifications across the genome. The mint pipeline creates a customized UCSC Genome Browser track hub which collates the data tracks to enable easy and fast visual exploration. To help interpret each step of the analysis (i.e. sample-wise methylation, group differences, and integrated classifications) we annotate corresponding tracks to genomic annotations using our recently developed Bioconductor package, annotatr. Default annotations include: genic features, CpG island features, enhancers, and lncRNA for human and mouse genomes. Alternatively, custom annotations enable annotation to any organism. Additional summary visualizations are output which map categorical or numerical data associated with the tracks onto the genomic annotations.

The mint pipeline is implemented as a command line tool (https://github.com/sartorlab/mint/) and as a Galaxy tool (https://github.com/sartorlab/mint_galaxy/). The command line tool is implemented in Snakemake so that analyses are flexible, reproducible, and restartable. The mint pipeline facilitates complex, comprehensive analyses of genome-wide methylation and hydroxymethylation data, enabling new biological discoveries.

Authors:

first name	**last name**	**email**	**country**	**organization**	**corresponding?**
Raymond	Cavalcante	rcavalca@umich.edu	USA	University of Michigan	✓
Yanxiao	Zhang	shz254@ucsd.edu	USA	Ludwig Institute, University of California, San Diego	
Yongseok	Park	yongpark@pitt.edu	USA	University of Pittsburgh	
Snehal	Patil	snehal@med.umich.edu	USA	University of Michigan	
Maureen	Sartor	sartorma@umich.edu	USA	University of Michigan	

Advancing the functional interpretation of high-throughput gene regulatory data

Keywords: ChIP-seq, genomic regions, gene set enrichment analysis, pathway analysis

Abstract: Background and Motivation: Various high-throughput sequencing approaches are used to study gene regulation by transcription factors, histone modifications, DNA methylation, and combinations thereof in various biological contexts. We previously developed two methods and software applications, ChIP-Enrich and Broad-Enrich, aimed at enabling improved pathway interpretation of these data. Functional, or gene set, enrichment (GSE) testing is one of the most common downstream tests performed after gene regulatory sequencing experiments, illustrating its critical importance and central role in enabling biomedical researchers to interpret their results. Yet the test is also one of the most misused, and often performed or interpreted in a cursory manner. This is because current software does not guide users to choose the appropriate approach, and few programs correct for confounding variables, adequately control for type 1 error, or utilize additional information in the experimental data or about the reference genome. We have made several recent advances by developing more sophisticated tests for enrichment and improving definitions of regulatory regions. Results from different methods across a wide range of ENCODE ChIP-seq data reveal new information about how certain cellular processes are regulated by different classes of DNA-binding proteins.

Methods: I will present several GSE methods for testing sets of genomic regions with various properties: (1) 1000s to 10000s of narrow genomic regions; (2) larger sets of narrow regions; and (3) large sets of broad genomic regions, each potentially regulating genes from promoters, enhancers, introns, CpG islands/shores, etc., and with or without considering the strength of binding. Each of these methods uses a generalized linear regression model with an embedded cubic spline to empirically adjust for the length of each gene's modeled regulatory region. We show appropriate control of the type 1 error rates under various conditions using permuted ChIP-seq data, and compare the power of the methods for datasets with different properties, which can be used to guide researchers on the optimal approach for their data. I will also show how each method can be used to understand regulation at promoters, enhancers, or from within genic regions. For enhancers, we have tested ¿400 combinations of ways to define enhancers, refine their sizes, and link them to their target genes.

Results and Conclusions: We are improving our chipenrich Bioconductor package and associated web tools to guide users in choosing the optimal GSE approach for their data and biological hypotheses. Comparing the above methods on a large set of ENCODE datasets across transcription factors and cell types reveals a broad overview of the different ways that cells regulate pathways and processes. We find that how regulation occurs depends more on the biological process than on the transcription factor. While certain processes require only a single binding event, the number of binding events is important for others. Additionally, we find that certain processes are regulated by enhancers often separated by multiple genes. Enrichment of these pathways is often missed when using the naïve approach of assigning peaks to the nearest transcription start site. Finally, we have updated our gene set databases, including data from the Comparative Toxicogenomics Database (CTD) that will help understand effects of environmental exposures. Alternative methods take a one-size-fits-all approach to GSE with genomic regions, resulting in an overabundance of false positives, loss of signal, oversimplified interpretations of the results, or all of the above. Our methods allow new perspectives on the cellular regulatory mechanisms in a biological system.

Authors:

first name	last name	email	country	organization	corresponding?
Maureen	Sartor	sartorma@umich.edu	USA	University of Michigan	✓
Christopher	Lee	leetaiyi@umich.edu	USA	University of Michigan	
Raymond	Cavalcante	rcavalca@umich.edu	USA	University of Michigan	
Shuze	Wang	shuzwang@umich.edu	USA	University of Michigan	
Heming	Yao	hemingy@umich.edu	USA	University of Michigan	
Peter	Orchard	porchard@umich.edu	USA	University of Michigan	
Tingting	Qin	qinting@med.umich.edu	USA	University of Michigan	
Tsung-Yeh	Tsai	ttsai@umich.edu	USA	University of Michigan	

Algorithms for Structural Variation Discovery Using Hybrid Sequencing Technologies

Keywords: structural variation, PacBio, Illumina, split reads, read pairs

Abstract: Structural variation (SV) is defined as genomic variation that affects more than 50 base pairs. Recent studies found that in the human genome, there exist thousands of SVs on average which affect around 15-20 million base pairs. As a result of this, they show a high probability to cause functional effects. Additionally, various studies associated several SVs to human disease. SV detection studies were made possible by the introduction of high-throughput DNA sequencing (HTS). While these technologies produce large amounts of sequencing data in a cost effective and fast manner, they still suffer from various disadvantages. Second-generation sequencing technologies, such as Illumina, create short reads (75-150 bp) with low cost and low error rates. On the other hand, third-generation sequencing technologies, such as Pacific Biosciences, produce long reads (average 10Kb) with high cost and high error rate as they use single molecule real time sequencing. For accurate SV detection, long reads with low error rate are desired. Even though PacBio data with high coverage ($> 40X$) may be more reliable for this purpose, high costs associated with this technology diminish its practicality for large number of samples. On the other hand, short Illumina reads cannot span over repeats and duplications where most SVs are known to occur. Another problem with SV discovery is that the accurate detection of breakpoints is difficult in the homologous segments and the repeated sections of the DNA. Thus, the coupling low coverage (i.e. lower cost) PacBio and high coverage Illumina data would, in theory, complement the strengths of these two technologies with each other and, correct for the biases.

The aim of this study is to detect large deletions and inversions in human genome with low cost and high accuracy. This is achieved by, first, broadly defining inversions and deletions using low coverage PacBio sequencing data and then, comparing the SV signals with those detected using Illumina data.

We downloaded PacBio data set generated from the genome of NA12878 from the Genome in a Bottle project. The Illumina data set from the same genome was generated as part of the Platinum Genomes collection. Briefly, we searched for split read signature in the PacBio alignments that signal large deletions ($> 5Kbp$), and we then tested whether there exists any Illumina read pair signature in the periphery of the PacBio-detected regions. PacBio reads were previously aligned using BLASR, and Illuma reads were aligned using BWA-MEM. We used an approximation to the quasi clique detection problem for clustering SV sequence signals. We provide the preliminary results below for deletions only, which we will further improve through incorporating local assembly and split read mapping of Illumina reads, and extend to detect inversions.

Authors:

first name	last name	email	country	organization	corresponding?
Ezgi	Ebren	ezgiebren@gmail.com	Turkey	Bilkent University, Department of Computer Engineering	
Ayse Berceste	Dincer	ayse.dincer@ug.bilkent.edu.tr	Turkey	Bilkent University, Department of Computer Engineering	
Can	Alkan	calkan@gmail.com	Turkey	Bilkent University, Department of Computer Engineering	✓

Rescue the missing variants-lessons learned from large sequencing projects

Keywords: variant calling, next generation sequencing, GATK best practices, large sample size

Abstract: Identifying novel disease variants through next generation sequencing (NGS) has been a fruitful practice in medical research in recent years, leading to the discoveries of new disease mechanisms as well as therapeutic strategies. The GATK best practices have since been established to provide general recommendations on core processing steps required to go from raw reads to final variant call sets. However, with the sample size drastically increasing in today's sequencing experiments, many default variant calling strategies and the choice of tools call for a closer examination.

Our study utilized the whole exome sequencing data provided by the Alzheimer's Disease Sequencing Project (ADSP) to test for different variant calling strategies and tools involved in the variant discovery workflow in the context of sample sizes. We first investigated the impact of using different sequence aligners on variant callsets while keeping the default GATK settings of the variant calling and QC steps identical. We selected 1952 samples to align by both BWA and NovoAlign, and compared the variant callsets in 50, 100, 200, 500, 1000 and 1952 samples. We discovered that the percentage of variants unique to aligner increased dramatically with increasing sample sizes. At sample size of 1952, the unique variants generated by BWA and NovoAlign account for more than 20% of total called variants. These unique variants have good variant quality metrics: 80% have Genotype Quality (GQ) score of 60 or above, and their distribution of B allele concentration (BAC) centers around 0.5 and 1, consistent with what is expected of diploid genomes. What's more, over 96% of the unique variants have population B allele frequency (BAF) of less than 0.01, indicating that these variants are rare in the population. All these metrics suggest that these unique variants are important to be included in downstream variant analysis. In addition to aligner comparison, we also evaluated single-sample genotyping versus the default multi-sample joint genotyping strategy in 50, 100, 500, 2000, and 5000 samples. Our data showed that, with increasing sample sizes, the single-sample genotyping strategy added increasing percentage of unique variants. At sample size of 5000, single-sample genotyping added 58,884 variants, accounting for 5.55% of total variants called by both strategies. 7331 of these unique variants passed Variant Quality Score Recalibration (VQSR) and had GQ of 60 or above in at least 5 samples.

Our study identified a large number of good-quality variants from the ADSP exome data that were missed by using one aligner or using multi-sample joint genotyping strategy alone. Our findings revealed the relationships between bioinformatics pipelines and biomedical research results, and suggested that alternative variant calling strategies may be beneficial for optimal variant discovery in face of today's large sequencing scale.

Availability: http://www.mayoclinic.org

Authors:

first name	**last name**	**email**	**country**	**organization**	**corresponding?**
Yingxue	Ren	ren.yingxue@mayo.edu	USA	Mayo Clinic	✓
Joseph	Reddy	reddy.joseph@mayo.edu	USA	Mayo Clinic	
Vivekananda	Sarangi	Sarangi.Vivekananda@mayo.edu	USA	Mayo Clinic	
Shulan	Tian	Tian.Shulan@mayo.edu	USA	Mayo Clinic	
Jason	Sinnwell	Sinnwell.Jason@mayo.edu	USA	Mayo Clinic	
Nilufer	Ertekin-Taner	Taner.Nilufer@mayo.edu	USA	Mayo Clinic	
Owen	Ross	Ross.Owen@mayo.edu	USA	Mayo Clinic	
Rosa	Rademakers	Rademakers.Rosa@mayo.edu	USA	Mayo Clinic	
Shannon	McDonnell	McDonnell.Shannon@mayo.edu	USA	Mayo Clinic	
Joanna	Biernacka	Biernacka.Joanna@mayo.edu	USA	Mayo Clinic	
Minerva	Carrasquillo	Carrasquillo.Minerva@mayo.edu	USA	Mayo Clinic	
Liudmila	Mainzer	lmainzer@life.illinois.edu	USA	University of Illinois at Urbana-Champaign	
Yan	Asmann	Asmann.Yan@mayo.edu	USA	Mayo Clinic	

Conta: methods for detecting trace amounts of contamination

Keywords: next generation sequencing, high throughput sequencing, targeted sequencing, liquid biopsy, early cancer detection, circulating tumor DNA, cfDNA, contamination detection, single nucleotide polymorphisms, SNP, minor allele frequency, maximum likelihood, linear regression, cross validation

Abstract: Next Generation Sequencing-based assays of circulating tumor DNA must achieve high sensitivity and specificity in order to detect cancer early. Early cancer detection and liquid biopsy both require highly sensitive methods to spot low tumor burden as well as highly specific methods to reduce false positive calls. Contaminating DNA from adjacent samples can compromise specificity, because rare SNPs from the contaminant may look like low level mutations, resulting in false positive calls. One method to detect contamination involves examining the allele frequency signature of single nucleotide polymorphisms (SNPs). In the case of contamination, the observed signature is then a blend of the SNPs from the host and the contaminating sample. At lower levels of contamination, the mix signal becomes harder to distinguish from that of the background noise signature.

To distinguish noise from contamination, first, we model the observed variant frequencies as a linear combination of population minor allele frequency (MAF) and a background noise model, and solve the resulting regression problem. Second, we calculate the probability of contamination from both the MAF of SNPs and the host genotype, and use this as a prior probability in a likelihood model. The likelihood model calculates the probability of observing the data at a given contamination level. Third, we further inform the likelihood model with priors from known genotypes of other samples that were processed in the same batch, and check if the source of contamination can be identified.

To train our methods, we first built a noise model from 94 healthy clean cfDNA samples, where contaminated samples were removed from the model. Then, we took another set of 85 clean samples and by each time selecting two of them randomly, generated 1500 Poisson mixtures ranging from 0.01% to 1%. We used these clean and mixture samples in a cross validation setting to choose optimal thresholds and estimate the test error. At the target limit of detection of 0.05%, the sensitivity of our method was 96.1% (CI=[87.61%,100%]) and specificity was 100% (CI=[94.4%,100%]).

Next, we tested our methods on healthy and cancer samples with the trained thresholds. We detected in vitro cases of contamination events down to 0.01%. Preventing contaminations at this level requires clean and robust workflows in the lab. Dual indexing of the samples across batches may alleviate post-indexing contaminations, but robust detection methods are still necessary to detect and possibly clean index-swapping and residual sample mixing events that happened before indexing.

Availability: http://www.grailbio.com/

Authors:

first name	**last name**	**email**	**country**	**organization**	**corresponding?**
Onur	Sakarya	osakarya@grailbio.com	USA	Grail Bio	✓
John	Lamping	jlamping@grailbio.com	USA	Grail Bio	
Alexander	Blocker	ablocker@grailbio.com	USA	Grail Bio	
Earl	Hubbell	ehubbell@grailbio.com	USA	Grail Bio	
Catalin	Barbacioru	cbarbacioru@grailbio.com	USA	Grail Bio	
Franz	Och	och@grailbio.com	USA	Grail Bio	✓

Cleaner Expression Signatures by Alignment-Free Degradation Assessment of RNA-seq Data

Keywords: Quality Control, RNA-seq, degradation

Abstract: In the age of Big Data where genomic sequencing on a large scale is becoming a new standard in medical and biological research, automatic quality assessment becomes a crucial component of the process. Whereas small projects often have a uniform design and a manageable structure allowing for a manual per-sample analysis of quality, large scale studies tend to be much more heterogeneous and complex. With the increase of study sizes such manual quality investigation becomes increasingly unfeasible. Thus the design of sensitive and efficient automated quality control processes for RNA-seq data becomes crucial to avoid conclusions that are the result of technical artifacts. In this study, we show that single, commonly used quality criteria like the RIN-score alone are not sufficient to ensure RNA sample quality. Another limitation is that degradation based quality criteria typically require read data to be aligned to a reference genome. In the presence of a systematic problem, this rather time consuming step will delay potential measures to address the data quality potentially diminishing or even destroying valuable samples in the process. We developed a new method and provide an efficient new tool, that estimates RNA sample degradation by computing the 5'/3' bias of the first and last constitutive exon of long transcript genes, using an alignment free approach. Our analysis shows that this strategy is robust, provides complementary quality information to RIN-scores and allows the accurate identification of degraded samples. Our tool utilizes a low dimensional projection of the read count ratios observed across a candidate gene set and determines their distance to a reference set of non-degraded samples. We performed a Pan-cancer analysis of all currently available TCGA expression data sets and demonstrate that samples passing the standard RIN-score assessment would not pass a degradation analysis based on 5'/3' bias. Further, we show how degraded samples can lead to patterns that are strong enough to lead to the false discovery of new cancer subtypes. Allowing the assessment of a sample on raw read data, without requiring the computationally expensive alignment step, will allow researchers to detect and address potential quality issues further upstream of rather complex analysis pipelines.

Availability: http://www.raetschlab.org/members/akahles/

Authors:

first name	last name	email	country	organization	corresponding?
Kjong-Van	Lehmann	kjong.lehmann@inf.ethz.ch	Switzerland	ETH Zurich	✓
Andre	Kahles	andre.kahles@ratschlab.org	Switzerland	ETH Zurich	
Niki	Schulz	schultz@cbio.mskcc.org	USA	MSKCC	
Chris	Sander	chris@sanderlab.org	USA	Dana Farber Cancer Institute	
Gunnar	Ratsch	gunnar.ratsch@ratschlab.org	Switzerland	ETH Zurich	✓

Enhancing BLAST+ speed and user experience with CrocoBLAST

Keywords: sequence alignment, BLAST, parallelization

Abstract: NCBI BLAST+ is one of the most cited computational tools in science, being included in pipelines for sequence annotation, genome mapping, genome assembly, and direct sequence similarity comparison. However, NCBI BLAST+ is not well optimized for large calculations or for computers with a large number of cores, which constitutes a significant setback when running alignments involving large datasets, especially in the context of Next Generation Sequencing.

To address these limitations, we have developed CrocoBLAST, a tool for dramatically speeding up BLAST+ execution on any computer, so that alignments that would take days or even weeks with NCBI BLAST+ can be run overnight with CrocoBLAST. CrocoBLAST scales well on low-end computers, workstations, servers, and cluster nodes, under different operational systems and storage architectures. This high efficiency is achieved by an automated procedure of verifying the optimal internal parameters for each BLAST+ calculation requested, and dynamically re-allocating computational resources while the job is running.

In addition to the improvement in computational efficiency of BLAST+ jobs, CrocoBLAST provides features critical for large data analyses, including:

- real-time information regarding calculation progress and remaining run time;
- access to partial alignment results;
- queueing, pausing, and resuming BLAST+ calculations without information loss;
- results identical to those of BLAST+;
- compatibility with any BLAST+ version;
- internal queuing system for planning and managing BLAST+ jobs in real time.

CrocoBLAST accepts both FASTA and FASTQ as input files. Furthermore, CrocoBLAST keeps an internal index of up-to-date databases of nucleotide and protein sequences in BLAST-ready format (e.g., downloaded from NCBI), which facilitates running additional alignments against a single database. Finally, CrocoBLAST was developed with an update function, that, once executed, automatically verifies if there is a newer version available and updates itself without any complicated installation process.

CrocoBLAST is implemented in C, while the easy-to-use graphical user interface is implemented in Java. CrocoBLAST is supported under Linux and Windows, and can be run under Mac OS X in a Linux virtual machine. CrocoBLAST is freely available online, with ample documentation (webchem.ncbr.muni.cz/Platform/App/CrocoBLAST). No installation or user registration is required.

We here describe the principles underlying CrocoBLAST and further exemplify its use on case studies involving NGS data.

Authors:

first name	last name	email	country	organization	corresponding?
Ravi Jose	Tristao Ramos	souoravi@gmail.com	Czech Republic	NCBR / CEITEC - MU	✓
Crina-Maria	Ionescu	ionescu@ncbr.muni.cz	Czech Republic	NCBR / CEITEC - MU	
Jaroslav	Koča	jkoca@ceitec.muni.cz	Czech Republic	NCBR / CEITEC - MU	

Advancing parasite transcriptomics with spliced-leader sequencing experimental and computational workflows

Keywords: RNA-seq, Transcriptomics, Next Generation Sequencing, Trypanosomatids

Abstract: BACKGROUND

The Trypanosomatida family contains many human pathogenic, parasitic species including Leishmania donovani (visceral leishmaniasis), Trypanosoma gambiense/rhodesiense (sleeping sickness) and Trypanosoma cruzi (chagas disease). Transcriptome studies of these parasites are essential for fundamental insights in their development, pathogenicity and drug resistance. However, in most patient tissue samples, host RNA is much more abundantly present than parasite RNA, imposing a complicated and time consuming parasite isolation step prior to sequencing. Interestingly, mature mRNA of Trypanosomatida differs from the host's by starting with a fixed 39 nucleotide sequence or spliced-leader (SL), added to pre-mRNA by a process called 'trans-splicing'[1].

RESULTS

We exploited the presence of a SL on each parasite mRNA and developed an RNA-seq protocol (SL-seq) to specifically amplify and sequence SL-containing RNA out of a pool of host cell RNA. SL-Seq first converts SL-containing mRNA to cDNA using a SL-specific primer. Amplification is carried out with overhang-extension PCR, which adds additional motives and indexes, allowing multiplexing hundreds of samples on a single sequencing lane. In addition, we developed and a new bio-informatic pipeline that can deal with the intricacies of the technology, specific to the method. It uses existing RNAseq tools (including Samtools, TopHat, HTseq and DESeq) and new tools that were developed in Python. One of the main differences with a conventional RNA-seq method that had to be addressed is that most SL-Seq sequencing reads map in the 5'-UTRs. However, since 5'-UTRs have a variable length for the same gene in Trypansomatids, they are left unannotated in all Trypanosomatid reference genomes. The SL-seq pipeline is developed to associate the reads automatically with the closest upstream gene, without the need for UTR annotation. Other modifications include the trimming of the SL sequence artifacts from the reads and differential splice site usage detection. We verified the validity and performance of SL-seq and its bio-informatic pipeline by comparing the results with those obtained with the Illumina Stranded mRNA kit (ILL-seq) on an identical pool of RNA. A strong correlation was observed between the expression values obtained with both methods ($p < 2e - 16andR^2 = 0.8$) and also the differentially expressed genes and enriched GO categories were largely identical. We also successfully sequenced Leishmania transcriptomes directly from infected THP-1 cells, without prior isolation of the parasites. With ILL-seq only 1.6% of the ILL-seq data was Leishmania mRNA, while this was 65.0% using the SL-seq protocol, indicating SL-seq resulted in a 42 fold enrichment of parasite mRNA.

CONCLUSIONS

We developed an RNA-seq method and corresponding bio-informatic analysis toolkit that can sequence and analyze Trypanosomatid transcriptomes, directly from infected tissues with unprecedented resolution. SL-seq could also be useful for other SL-containing organisms including nematodes, trematodes and primitive chordates.

REFERENCES

1. Martinez-Calvillo S, Nguyen D, Stuart K, Myler PJ: Transcription initiation and termination on Leishmania major chromosome 3. Eukaryotic cell 2004, 3(2):506-517.

Authors:

first name	last name	email	country	organization	corresponding?
Bart	Cuypers	bart.cuypers@uantwerpen.be	Belgium	Univeristy Of Antwerp	✓
Malgorzata	Domagalksa		Belgium	Institute of Tropical Medicine, Antwerp	
Geraldine	de Muylder		Belgium	Institute of Tropical Medicine, Antwerp	
Pieter	Meysman	pieter.meysman@uantwerpen.be	Belgium	University of Antwerp	✓
Manu	Vanaerschot		USA	Columbia University in the City of New York	
Hideo	Imamura		Belgium	Institute of Tropical Medicine	
Franck	Dumetz		Belgium	Institute of Tropical Medicine	
Thomas-Wolf	Verdonckt		Belgium	Institute of Tropical Medicine	
Peter	J. Myler		USA	Center for Infectious Disease Research	
Gowthaman	Ramasamy		USA	Center for Infectious Disease Research	
Kris	Laukens	kris.laukens@uantwerpen.be	Belgium	University of Antwerp	✓
Jean-Claude	Dujardin		Belgium	Institute of Tropical Medicine, Antwerp	

Prophyle: a phylogeny-based metagenomic classifier using Burrows-Wheeler Transform

Keywords: metagenomics, metagenomic classification, taxonomic classification, phylogenetic trees, Burrows-Wheeler transform, BWT-index, k-mer-based methods

Abstract: Metagenomics is a powerful approach to study genetic content of environmental samples and it has been strongly promoted by NGS technologies. The metagenomic classification problem is to assign each sequence of the metagenome to a corresponding taxonomic unit, or to classify it as “novel”.

To cope with increasingly large metagenomic projects, researchers resort to alignment-free methods. The most popular tool – Kraken [2] – performs metagenomic classification of NGS reads based on the analysis of shared k-mers between an input read and each genome from a pre-compiled database. Kraken provides an extremely rapid read classification, but its index suffers from two major limitations. First, its enormous memory consumption, due to a large hash table, does not allow one to perform classification other than on high-performance clusters. This prohibits the use of Kraken when computational resources are limited. For instance, point-of-care sequencing and real-time disease surveillance projects often have to rely on data analysis on laptops with little memory. Second, every k-mer in the Kraken's index is represented through its lowest common ancestor, which can result in an inaccurate classification, especially when many k-mers are present in multiple branches of the tree as it is common, e.g., in phylogenetic trees for a single species.

In our talk, we present Prophyle [1], a metagenomic classifier based on BWT-index. Prophyle uses a classification algorithm similar to Kraken but with an indexing strategy based on a bottom-up propagation of k-mers in the tree, assembling contigs at each node and matching using a standard full-text search. The obtained index occupies only a fraction of RAM compared to Kraken – 13 GB instead of 120 GB for index construction and 14 GB instead of 75 GB for index querying. The resulting index is also more expressive as we can, for every queried k-mer, retrieve a list of all genomes in which the k-mer occurs. Overall, Prophyle provides an index for resource-frugal metagenomic classification, which is accurate even with single-species phylogenetic trees. Prophyle is available at http://github.com/karel-brinda/prophyle, released under the MIT license.

[1] Břinda, K., Salikhov, K., Pignotti, S., & Kucherov, G. ProPhyle: accurate and resource-frugal phylogeny-based metagenomic classification. To appear.

[2] Wood, D. E., & Salzberg, S. L. (2014). Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biology, 15(3), R46

Availability: http://brinda.cz

Authors:

first name	last name	email	country	organization	corresponding?
Karel	Brinda	karel.brinda@gmail.com	USA	Center for Communicable Disease Dynamics, Department of Epidemiology , Harvard T.H. Chan School of Public Health, Boston MA 02115, USA	✓
Kamil	Salikhov	salikhov.kamil@gmail.com	Russia	Mechanics and Mathematics Department, Lomonosov Moscow State University, Russia	
Simone	Pignotti	pignottisimone@gmail.com	France	LIGM/CNRS Université Paris-Est, 77454 Marne-la-Vallée, France	
Gregory	Kucherov	gregory.kucherov@univ-mlv.fr	France	LIGM/CNRS Université Paris-Est, 77454 Marne-la-Vallée, France	

Top-down de novo assembly of 10x Genomics Chromium data

Keywords: Chromium, 10x Genomics, Linked reads, De novo assembly, Sequence overlap

Abstract: 10X Genomic Chromium sequencing provides long range information between Illumina sequence read pairs. Each read pair is linked to a group of reads by a barcode within a fragment size typically between 10-100 kbp. This powerful technology has been used in phasing haplotypes, but is difficult to use for assembly and scaffolding due to the low relative coverage of each barcode and the absence of positional information (ie. read order, orientation, and position within each molecule).

Chromium-barcoded reads provide information that can be used to localize read sequencing information for de novo assembly. For example, the Supernova software from the vendor utilizes de Bruijn graphs to first assemble the data and incorporates the barcode information in later steps. Alternatively, one can localize and assemble the linked reads into contiguous synthetic molecules using a top-down approach.

The Chromium linked reads represent each molecule with below 1x coverage, so they cannot be used directly to generate synthetic long reads. Here, we propose a method to generate sufficient coverage for synthetic molecules, by finding the overlapping barcodes given a target barcode. We demonstrate that these collections can then be used to generate local assemblies. We note that the process may collect off target sequences, though these can be dealt with, for instance, by filtering repetitive sequences. Because longer sequences have a higher specificity, we anticipate these local assemblies to be of use in a number of downstream applications, such as whole genome assemblies using overlap-layout-consensus approaches, reconstruction of transcript isoforms, and strain typing in metagenomics studies.

Authors:

first name	last name	email	country	organization	corresponding?
Justin	Chu	cjustin@bcgsc.ca	Canada	Canada's Michael Smith Genome Sciences Centre, BCCA	✓
Ben	Vandervalk	benv@bcgsc.ca	Canada	Michael Smith Genome Sciences Centre	
Shaun	Jackman	sjackman@bcgsc.ca	Canada	BC Cancer Agency Genome Sciences Centre	
Hamid	Mohamadi	hmohamadi@bcgsc.ca	Canada	BC Cancer Agency	
Inanc	Birol	ibirol@bcgsc.ca	Canada	BC Genome Sciences Centre	

De novo repeat identification in whole genome shotgun sequencing data

Keywords: repeat filtering, repeat identification, sequence analysis

Abstract: Repetitive sequence elements such as microsatellites, transposons, gene families, and segmental duplications are abundant in eukaryotic genomes. They often induce many local alignments, complicating sequence assembly, comparisons between genomes and analysis of large-scale duplications and rearrangements. Hence, identification and classification of repeats is a fundamental step in many genomics applications and their downstream analysis tools. Most available tools for the analysis of repeats rely on reference genomes and are based on similarity search, which is a computationally costly process. On the other hand, de novo computational identification and classification of such elements is a challenging problem.

In this work, we propose an efficient de novo approach for repetitive sequence identification and classification based on statistical analysis of the k-mer content profile of the input DNA sequencing data. In the proposed approach, we first obtain the k-mer coverage histograms of genomics datasets using our recently published ntCard algorithm, an efficient streaming algorithm for estimating the k-mer coverage histograms. Using the estimated k-mer coverage histograms, we fit a mixture model to estimate the relative abundances of repeated contents. To make the model fitting more robust, we exclude the erroneous k-mers that are likely caused by sequencing errors. This is performed by fitting a Weibull distribution to the k-mer coverage histogram. After excluding the effect of erroneous k-mers, we fit a mixture model on the true k-mer coverage histogram, and then identify the threshold measures for the most repeated k-mers in the input dataset. In the last step, we go through the input dataset a second time to determine the most repeated k-mers, and consequently the repeated regions, using the measures obtained from the statistical analysis of the k-mer profile.

Authors:

first name	last name	email	country	organization	corresponding?
Hamid	Mohamadi	hmohamadi@bcgsc.ca	Canada	BC Cancer Agency Genome Sciences Centre	✓
Stewart Austin	Hammond	shammond@bcgsc.ca	Canada	BC Cancer Agency Genome Sciences Centre	
Shaun	Jackman	sjackman@bcgsc.ca	Canada	BC Cancer Agency Genome Sciences Centre	
Ben	Vandervalk	benv@bcgsc.ca	Canada	BC Cancer Agency Genome Sciences Centre	
Justin	Chu	cjustin@bcgsc.ca	Canada	BC Cancer Agency Genome Sciences Centre	
Inanc	Birol	ibirol@bcgsc.ca	Canada	BC Cancer Agency Genome Sciences Centre	

Development and validation of a somatic copy-number alteration calling algorithm for targeted resequencing of actionable cancer genes

Keywords: next-generation sequencing, cancer, copy number alterations, CNV, tumor profiling, targeted re-sequencing, clinical diagnostics, precision medicine, oncology

Abstract: Next-generation, deep sequencing of gene panels is being adopted as diagnostic tests to identify actionable mutations in samples from cancer patients. DNA extracted from patient samples is enriched for cancer genes and sequenced to identify cancer somatic mutations. Somatic mutations with potential implications for the patient's treatment and prognosis include SNVs, indels, and rearrangements that result in copy number alterations (CNAs). Targeted sequencing assays involve positive selection of gene exons, typically involving PCR at some stage in the laboratory protocol. PCR creates biases associated with GC-content and length that can skew the representation of the original DNA fragments within the sequenced library. Combined with the small counts of reads in some of the targeted exons, these biases often result in reduced sensitivity to the identification of CNAs. Here we present a CNA calling algorithm optimized for targeted resequencing data which delivers high sensitivity and specificity. To reduce the impact of sequence-context biases, the method compares the data of a test sample to that of a diploid negative control run alongside it, or to a model carefully constructed from a panel of normal samples. Depth of coverage data is corrected for G+C content, a weighted median normalization is applied and log ratio versus the control or model are calculated. Further, the effect of low read counts in short exons is ameliorated by calling CNA segments across targeted exons going beyond, if needed, a single gene boundary. Analogous to the program VEGAWES, we use a 1-dimensional form of the Mumford-Shah energy function to determine if adjacent segments should be merged, considering intra-segment spacing. Finally, calls are annotated with a segment quality score (SQS). The program reports CNV segment-level calls in VCF 4.1 file format, and additionally in a gene-level report based on the list of the exons targeted by the assay. We benchmarked and validated this algorithm with a simulated CNV dataset (generated by SCNVSim), as well as with sequencing data obtained from DNA of cancer cell lines with orthogonally verified CNAs, using an oligonucleotide-selective sequencing (OS-Seq) enrichment assay for a 130-cancer gene panel. We analyzed our calls with those from the CNVkit and VEGAWES programs using Receiver Operating Characteristic (ROC) curve analysis. Our results show that our CNA caller outperforms these other programs in terms of ROC area under the curve. We identified several novel unannotated CNA calls in some of the analyzed cell lines which we confirmed orthogonally using ddPCR. A distinctive feature of our method is its ability to filter raw calls to a specified sensitivity and specificity trade-off by selecting empirical SQS thresholds based on the analysis of the reference datasets. Filtering was evaluated using synthetic materials spiked into the background of a well-characterized germline (NA24385) and sequenced with the OS-Seq gene panel, enabling systematic assessment of the algorithm's performance. We show that we can detect hemizygous deletions and amplifications from 0.6 to more than 150 excess copies over the diploid background, with minimal or no false positives across the 130 genes in the panel. We have implemented this algorithm as part of an analysis pipeline for tumor profiling of cancer patients in support of precision medicine workflows.

Authors:

first name	**last name**	**email**	**country**	**organization**	**corresponding?**
Francisco	De La Vega	delavefm@gmail.com	USA	Stanford University	✓
Sean	Irvine	sea@realtimegenomics.com	New Zealand	Real Time Genomics Ltd.	
Kurt	Gaastra	kurt@realtimegenomics.com	New Zealand	Real Time Genomics Ltd.	
Ryan	Koehler	ryan@tomabio.com	USA	TOMA Biosciences Inc.	
Yannick	Pouiliot	yannkc@tomabio.com	USA	TOMA Biosciences Inc.	
Daniel	Mendoza	daniel@tomabio.com	USA	TOMA Biosciences Inc.	
Philip	Cavales	philip@tomabio.com	USA	TOMA Biosciences Inc.	
Anna	Vilborg	anna@tomabio.com	USA	TOMA Biosciences Inc.	
Len	Trigg	len@realtimegenomics.com	New Zealand	Real Time Genomics Ltd.	

Single-cell RNA sequencing identifies novel roles and interacting partners of APE1 in Panceatic Ductal Adenocarcinoma Cells

Keywords: scRNA-seq, transcriptomics, single-cell, cancer, Pancreatic Cancer

Abstract: Apurinic/apyrimidinic endonuclease/redox factor-1 (APE1/Ref-1 or APE1) is a multifunctional protein involved in repairing DNA damage via endonuclease activity in base excision repair and in redox signaling control of transcription factors such as HIF-1 (hypoxia inducible factor-1) STAT3, NFB, and others. APE1 is overexpressed in several cancers, including in pancreatic ductal adenocarcinoma (PDAC). APE1 overexpression in cancer is associated with resistance to radiation and chemotherapy as well as increased tumor cell migration, proliferation, and survival. Deciphering APE1's role in cell survival, hypoxia signaling, and resistance to chemotherapy is complicated by the fact that APE1 is essential for cell viability and by its dual functionality in DNA repair and in redox regulation of transcription factors. Due to an inability to generate a stable APE1-knockout cell line and the incomplete, transient nature of APE1 siRNA knockdown, the use of bulk RNA-seq would lead to difficulty in conclusively defining a comprehensive list of genes regulated by APE1. In this study, single-cell RNA sequencing was utilized to compare the transcriptomes of siAPE1 and scrambled control cells under normal oxygen conditions and under hypoxia. Low passage patient-derived PDAC cells were used to investigate and characterize APE1 function under these conditions. Cell cycle-related genes were identified and used to determine a correction factor for the expression of other genes and fit a latent variable model accounting for treatment and control covariates using the R package scLVM. The R package BPSC was then utilized to test for differential expression, which models the gene expression counts using the beta Poisson distribution. Overall under normal O2, 1,950 genes were differentially expressed between the siAPE1 knockdown and control cells using a false discovery rate cutoff of 5%. Additional analyses were performed to fully take advantage of the power of single-cell sequencing, including an analysis using a statistical model that split all cells into three categories: scrambled control, cells transfected with siAPE1 that retained some expression of APE1 transcript (siAPE1-non zero), and cells transfected with siAPE1 that gave undetectable APE1 (siAPE1-zero) which identified 2,837 differentially expressed genes. A pathway analysis identified numerous pathways influenced by APE1 knockdown including mTOR, mitochondrial dysfunction, and the apoptosis signaling pathway. Biological validation was performed, including qRT-PCR validation of several genes in PDAC cells and in various other patient-derived tumor cells. Using data from TCGA, the clinical relevance of the DEGs was assessed by fitting a Cox proportional hazards model. Of the DEGs, 16% of the genes overlapping between this single-cell sequencing study and previous bulk RNA-Seq datasets included in TCGA were found to be clinically relevant to pancreatic cancer. Thus the current scRNA-seq study both contains overlap with genes already found to be clinically relevant and also provides new putative APE1 interacting partners as well as potentially novel mechanisms through which APE1 acts in the cell. This study has identified novel roles for APE1 in the cell and has utilized the power of single-cell to identify well-established as well as new, putative partners in the APE1 interactome. This study paves the way for future experiments aimed at identifying novel combination therapies for PDAC.

Authors:

first name	**last name**	**email**	**country**	**organization**	**corresponding?**
Nadia	Atallah	natallah@purdue.edu	USA	Purdue University, Purdue University Center for Cancer Research	✓
Emery	Goossens	emerygoossens@gmail.com	USA	Purdue University, Department of Statistics	
Fenil	Shah	fshah@iupui.edu	USA	Indiana University School of Medicine, Department of Pediatrics, Wells Center for Pediatric Research	
Mark	Kelley	mkelley@iu.edu	USA	Indiana University School of Medicine, Department of Pediatrics, Wells Center for Pediatric Research, Department of Pharmacology and Toxicology	
Melissa	Fishel	mfishel@iu.edu	USA	Indiana University School of Medicine, Department of Pediatrics, Wells Center for Pediatric Research, Department of Pharmacology and Toxicology	

Classification and visualization of metagenomics data for infectious disease diagnosis, and its application to corneal infections

Keywords: Metagenomics, Infectious Disease, Visualization, Pathogen

Abstract: The cause of a majority of clinically relevant inflammations remains undiagnosed. Infectious causes are typically diagnosed using methods that are time-consuming, have limited sensitivity, or do not give results for unexpected pathogens. High-throughput sequencing has the potential to revolutionize the clinical diagnosis of infectious diseases. Every pathogen has a genome, and can thus be uniquely identified if its genome is known. Metagenomic sequencing generates millions of short reads from the RNA or DNA in a sample, which are matched against the genomes of thousands of different microbes. Often less than 1 permille of the reads stem from the pathogen, and sensitive computational methods are necessary to correctly identify them.

We developed Centrifuge (www.ccb.jhu.edu/software/centrifuge) as novel metagenomics classifier using an optimized indexing scheme based on the FM index. Centrifuge can search the vast nr/nt sequence seq that BLAST uses, while being over 1000x faster than megaBLAST. However, the interpretation of the results is usually not straightforward. The samples are often dominated by host, background and contaminant sequences, and span the taxonomical tree. We present Pavian, a novel method to visualize, rank pathogens and compare metagenomics results in an interactive interface (https://github.com/fbreitwieser/pavian). Pavian incorporates virus-host data to allow filtering for human-associated species. Implemented in R using the Shiny framework and interactive Javascript/D3 graphs, it can be run on Windows, Linux and MacOS.

We demonstrate the functionality of Centrifuge and Pavian on 20 samples from corneal infections that were also analyzed in a microbiology laboratory. Generating 20 to 46 million reads per sample, on average 1.7% represented microbial sequences. Using z-score based ranking we successfully identified fungal, bacterial, viral and amoebal pathogens in the patients.

In conclusion, sequencing together with new classification, analysis and visualization methods can help identify a wide range of pathogens in a single test, showing a potential future of clinical diagnosis of infectious diseases. Both Centrifuge and Pavian are freely available under GPL v3.0.

Authors:

first name	last name	email	country	organization	corresponding?
Florian P	Breitwieser	florian.bw@gmail.com	USA	Johns Hopkins University	✓
Zhigang	Li	zli74@jhu.edu	USA	Johns Hopkins University School of Medicine	
Jen	Lu	jlu26@jhu.edu	USA	Johns Hopkins University	
Daehwan	Kim	infphilo@gmail.com	USA	Johns Hopkins University	
Li	Song	lsong10@jhu.edu	USA	Johns Hopkins University	
Charles G	Eberhart	cerberha@jhu.edu	USA	Johns Hopkins University	
Steven L	Salzberg	salzberg@jhu.edu	USA	Johns Hopkins University	

A novel approach using area under the Profile of Shannon Difference to detect clonal heterogeneity differences for single cell data

Keywords: single cell, heterogeneity, clonality

Abstract: Background: Tumor heterogeneity plays a critical role in tumor aggression and the development of drug resistance. Characterizing clonality and quantifying cellular heterogeneity using single cell technology is the first step to enable us to gain insights into cancer progression and develop effective therapeutic strategies.

Methods: We proposed a cluster-based profile analytical approach to analyze the single cell data and quantify the cellular heterogeneity. This is achieved by first quantifying cellular heterogeneity using Shannon Profile (SP) at different clonal resolutions for each population and then characterizing heterogeneity differences using the Profile of Shannon Difference (PSD) between two populations. The PSD is the profile of the differences between two SPs from two populations of interests. We propose to use a novel D statistic, the area under the PSD, to quantify the heterogeneity difference between two populations. Multivariate adaptive regression splines (MARS) model was used to detect the change points in PSD to determine the number of phenotypic clones. In addition to individual comparisons, a combined score, Generalized Fisher Product Score (GF) was developed to prioritize biomarkers for further investigating heterogeneity. We have implemented this algorithm in a software package, SinCHet (Single Cell Heterogeneity) in MATLAB with a GUI, which can analyze both continuous and discrete omics data.

Results: As proof of principle, we applied the proposed algorithm to published single-cell gene expression and methylation datasets. The results from expression data showed that the heterogeneity is significantly higher in the samples with EGFR-mutation than that in EGFR wild type lung cancer tumors (D=-63.8, p¡0.001). Nine clones were identified. The dominant clone from each group identified by SinCHet was in general agreement with each of the two clusters identified in the original paper. Additional clonal heterogeneity was characterized by SinCHet, with 7 additional clones identified. In addition to identifying the same reported epithelial cell markers (e.g., MUC1, SFTPC and KRT7), using the GF score, we were able to identify novel markers such as CD44, MT2A within subpopulations.

Conclusions: Our proposed novel D statistics, area under the PSD, enables the quantitative comparison of clonal heterogeneities between populations and prioritize clonal-specific and/or population-specific biomarkers. It provides unique insights into emerging or disappearing clones between populations or states. As the single-cell technologies become affordable and widely used, this algorithm could be easily applied in cancer research. We are currently planning an adaptive clinical trial and SinCHet will be used to characterize clonal evolutions in patient samples during the course of treatment to assist the decision of adaptive treatment strategies. The SinCHet software is freely available for non-profit academic use.

Availability: http://labpages2.moffitt.org/chen/

Authors:

first name	**last name**	**email**	**country**	**organization**	**corresponding?**
Yian	Chen	ann.chen@moffitt.org	USA	Moffitt Cancer Center	✓
Jiannong	Li	Jiannong.Li@moffitt.org	USA	Moffitt Cancer Center	
Inna	Smalley	Inna.Smalley@moffitt.org	USA	Moffitt Cancer Center	
Michael	Schell	Michael.Schell@moffitt.org	USA	Moffitt Cancer Center	
Keiran	Smalley	Keiran.Smalley@moffitt.org	USA	Moffitt Cancer Center	

Tibanna: a cloud-based workflow automation system for the 4D Nucleome project

Keywords: Cloud, Amazon Web Services, Docker, Common Workflow Language, workflow, pipeline, AWS lambda, AWS step functions

Abstract: We introduce a cloud-based genomic workflow management system called Tibanna. There is an increasing demand for processing large-scale genomic data using cloud computing. Our goals are to accommodate flexible and automated handling of massive genomic data of heterogeneous types, to improve reproducibility by utilizing standardized workflows, and to facilitate an effective use of the elastic nature of the cloud platform. To meet this end, we built an integrated system with adaptable components, tailored for the 4D Nucleome (4DN) Data Coordination and Integration Center (DCIC).

Tibanna adopts Amazon Web Services (AWS) as a main platform and consists of an upstream scheduler utilizing AWS's step functions ('Tibanna scheduler') and a set of 'minions', or lambda functions that performs specific tasks in a serverless environment. The lambda functions use three different utility components; AWSF, SBG pipe and annotator.

Tibanna AWSF and Tibanna SBG pipe are workflow executors. AWSF, an independent Autonomous Workflow machine Submission and monitoring Facility, launches an autonomous EC2 instance that executes a specified workflow, reports logs and self-terminates (AWSEM, Autonomous Workflow Step Executor Machine). AWSF serves as a stand-alone tool as well as a lambda utility. The SBG pipe is a connector to the proprietary Seven Bridges Genomics (SBG) platform and controller for file import/export. A workflow may run either on the SBG platform (SBG pipe) or on the given account's EC2 instance (AWSF).

Tibanna annotator creates and updates metadata for workflow runs, output files and quality metrics. Tibanna annotator is designed to interact with the 4DN-DCIC metadata database via the web portal's REST API module. The annotator holds the components that are specific to the metadata database, while the rest of Tibanna is agnostic to the 4DN infrastructure. This modular architecture ensures Tibanna can be connected to third party metadata systems.

Tibanna handles workflows described in Common Workflow Language and is naturally Docker-friendly. Along with data processing pipelines for 4DN, the general utility of Tibanna has been demonstrated by running the Genome Analysis Toolkit (GATK) pipeline. Tibanna is open-source and is available on http://github.com/4dn-dcic/tibanna.

Authors:

first name	last name	email	country	organization	corresponding?
Soohyun	Lee	duplexa@gmail.com	USA	Harvard Medical School	✓
Jeremy	Johnson	Jeremy_Johnson@hms.harvard.edu	USA	Harvard Medical School	✓
Burak H.	Alver	Burak_Alver@hms.harvard.edu	USA	Harvard Medical School	✓
Peter J.	Park	peter_park@hms.harvard.edu	USA	Harvard Medical School	✓

The spectrum of sequence and copy-number variants in 80,000 patients: Implications for test development and validation

Keywords: Genetic testing, Variant detection, Copy number, Indel, Long read, Sensitivity

Abstract: Many medically important genes are located in technically challenging regions of the genome, and complex but highly medically relevant classes of mutation are well-known. The overall impact of these facts on diagnostic yield and thus on appropriate technologies for routine clinical genome/exome sequencing has not yet been thoroughly described.

We examined over 80,000 patients undergoing germline clinical testing for physician-specified gene panels underlying a hereditary cancer, cardiovascular, neurological or pediatric condition. Sensitive methods using NGS, long read sequencing, MLPA, and arrays were used to detect and confirm a broad spectrum of variants.

Of 12,489 pathogenic, potentially actionable findings, approximately 10% belong to a technically challenging class not easily addressed by short-read NGS methods. Approximately 3% are CNVs affecting only a single exon, 2% are either large indels or complex variants, and 5% were in low-complexity, highly conserved or extreme-GC regions. This general observation was consistent across clinical areas, although specifics varied. Very large triplet repeat expansions and cytogenetic abnormalities are not included and would increase this total.

In summary, technically challenging variants are a substantial fraction of the pathogenic findings in routine clinical genetic testing. However, published validation studies often omit these variants, and benign SNPs dominate most sensitivity calculations. This may, in part, be due to difficulty obtaining positive controls. We have thus developed synthetic controls containing a diverse set of challenging mutations in commonly tested genes. These have been tested in multiple laboratories using different protocols and will be available to the community by the time of the meeting.

Authors:

first name	**last name**	**email**	**country**	**organization**	**corresponding?**
Stephen	Lincoln	steve.lincoln@me.com	USA	Invitae	✓
Rebecca	Truty		USA	Invitae	
Justin	Zook		USA	National Institute for Standards and technology	
Mark	Salit		USA	National Institute for Standards and technology	
Russell	Garlick		USA	SeraCare	
Matthew	Ferber		USA	Mayo Clinic	
Brian	Shirts		USA	University of Washington	
Swaroop	Aradhya		USA	Invitae	
Robert	Nussbaum		USA	Invitae and University of California, San Francisco	

Improved substitution and indel error correction using nearest-neighbor based error rate estimation

Keywords: high throughput sequencing, error correction, homopolymer repeat errors, substitution and indel errors

Abstract: Correction of errors is often a necessary step in the analysis of high throughput sequence data. We have previously developed the general-purpose error correction software Pollux, which is highly effective at correcting substitution, insertion and deletion errors, including homopolymer repeat errors. The error correction software is effective at identifying errors in Illumina and Ion torrent data, and can be applied to single- or mixed genome data sets while remaining sensitive to low coverage areas of sequencing projects. Using published data sets, we demonstrate an accuracy of error correction greater than 94% for Illumina data and 88% for Ion Torrent data. Here we present the updated version of the software, Pollux 2.0, which further increases error correction rates while greatly reducing the occurrence of introduced errors (false positives). The new version of the software implements a new nearest-neighbor based algorithm to estimate the rate of substitution, insertion, deletion, and homopolymer repeat errors directly from kmer data. The algorithm is used to set data-dependent thresholds for the correction of the different error types, and is able to distinguish sequencing errors from low-frequency variants in data sets. The new version of the software also implements a more efficient memory management system, reducing memory requirements by up to six-fold. Here, a compressed hash table is used to store kmer count data, where the kmer is represented by an index and remainder using an XOR-based hash function. The index is dual purpose, and is used to locate kmer data within the hash table and also to represent approximately half of each kmer sequence. A user-specified memory limit is also implemented so that if available memory is limited, the hash table is written to disk and a new hash table constructed for additional reads. The final representation of kmers in memory is further reduced by omitting low frequency kmers and blank entries in the hash tables, providing a dense representation of kmer frequencies and permitting large data sets to be analyzed on a single workstation. The introduced software is highly effective at correcting errors across platforms, and provides general-purpose error correction that may be used in applications with or without assembly.

Availability: http://www.biology.uwaterloo.ca/people/mcconkey/

Authors:

first name	last name	email	country	organization	corresponding?
Brendan	Mcconkey	mcconkey@uwaterloo.ca	Canada	University of Waterloo	✓

Insilico Identification of Protein-Coding and Non-Coding Regions in Next-generation Technology Transcriptome Sequence Data: A Machine Learning Approach

Keywords: Protein-Coding, Non-Coding, Next-generation Technology, Transcriptome Sequence Data, Machine Learning

Abstract: With the rapid increase in the volume of sequence data and multi-species transcripts generated using next-generation sequencing technologies, designing algorithms to process these data in an efficient manner and gaining biological insight is becoming a significantly growing challenge.

But there is still no known effective method to discriminate between non-coding and protein-coding regions in transcriptomes because RiboNucleic Acid (RNA) types show similar features to each other.

We describe a few techniques that discriminate between non-coding and protein-coding regions in transcripts, but all of these involve slow computational speed for small datasets or multi-threading in large datasets just to achieve small difference in computational speed, and thus risking a high execution time of the tool.

To solve this problem, we propose a fast, accurate and robust alignment-free predictor based on multiple feature groups using Logistic Regression, for the discrimination of protein-coding regions in multispecies transcriptome sequence data, where the predictive performance is influenced by Open Reading Frame(ORF)-Related and ORF-unrelated features used in the model rather than the training datasets, thereby achieving a relatively high performance and computational speed in processing small and large datasets of full-length and partial-length protein-coding and non-coding transcripts derived from transcriptome sequencing.

We describe a series of experiments on each of the datasets for human, mouse and fly species, with a goal that, our predictor generally performs better than competing techniques.

We expect this new approach to dramatically reduce the computational cost of identifying non-coding and protein-coding regions in transcripts, and hence make transcriptome analysis easier.

Authors:

first name	last name	email	country	organization	corresponding?
Olaitan	Awe	laitanawe@gmail.com	Nigeria	University of Ibadan	✓
Angela	Makolo		Nigeria	University of Ibadan	
Segun	Fatumo		United Kingdom	Wellcome Trust Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge	

NeatSeq-Flow: A Lightweight Software For Efficient Execution Of High Throughput Sequencing Workflows

Keywords: High throuput sequencing, Workflow, Computer cluster

Abstract: High Throughput Sequencing (HTS) experiments produce large sequence datasets, requiring multi-step analysis using open-source, commercial and/or in-house tools. Different workflows are required for different HTS applications, often with shared steps. Bioinformaticians typically write shell scripts to run individual tools, and need to administer intermediate files which accumulate during the process.

However, this approach has many drawbacks: Hand-written scripts are time-consuming and error prone and documentation and reproduction of an analysis can be daunting. Therefore, there is need for a modular and expandable system that will manage the execution of a variety of possible workflows from a single interface.

NeatSeq-Flow is a lightweight python software for computer clusters. The user defines the samples and related files; chooses the required modules (analysis steps), their order and dependencies; and specifies modules' parameters. NeatSeq-Flow then creates a hierarchy of shell scripts for modules' execution and structured directories for storing outputs. The scripts can be executed either as a whole, or one module or even one sample at a time, enabling full control of the workflow execution. NeatSeq-Flow efficiently exploits the computer cluster capabilities for managing module dependencies and parallelization.

NeatSeq-Flow provides analysis documentation, version control and time & memory usage reports. Adding new modules to NeatSeq-Flow can be achieved by anyone with basic knowledge of python, based on existing templates.

NeatSeq-Flow is routinely used on a high performance cluster at our Bioinformatics Core Facility in many types of analyses, such as DNA assembly and annotation, RNA-seq, ChIP-seq and others.

Authors:

first name	last name	email	country	organization	corresponding?
Menachem	Sklarz	sklarz@bgu.ac.il	Israel	The National Institute for Biotechnology in the Negev	✓
Michal	Gordon	gordonmi@bgu.ac.il	Israel	The National Institute for Biotechnology in the Negev	
Vered	Chalifa-Caspi	veredcc@bgu.ac.il	Israel	The National Institute for Biotechnology in the Negev	

HISEA: HIerarchical SEed Aligner for PacBio data

Keywords: PacBio sequecning, Genome assembly, Lond read aligner

Abstract: The next generation sequencing (NGS) techniques have been around for over a decade. Many of their fundamental applications rely on the ability to compute good genome assemblies. As the technology evolves, the assembly algorithms and tools have to continuously adjust and improve. The currently dominant technology of Illumina produces reads that are too short to bridge many repeats, setting limits on what can be successfully assembled. The emerging SMRT (Single Molecule, Real-Time) sequencing technique from Pacific Biosciences produces uniform coverage and long reads of length up to sixty thousand base pairs, enabling significantly better genome assemblies. However, SMRT reads are much more expensive and have a much higher error rate than Illumina's – around 10-15% – mostly due to indels. New algorithms are very much needed to take advantage of the long reads while mitigating the effect of high error rate and lowering the required coverage.

An essential step in assembling SMRT data is the detection of alignments, or overlaps, between reads. High error rate and very long reads make this a much more challenging problem than for Illumina data. We present a new read aligner, HISEA (HIerarchical SEed Aligner) for SMRT sequencing data. Our algorithm has the best alignment detection sensitivity among all programs for SMRT data, significantly higher than the current best. The currently best assembler for SMRT data is the Canu program which uses the MHAP aligner in its pipeline. We have incorporated our new HISEA aligner in the Canu pipeline and bench-marked it against the best pipeline for multiple data-sets at two relevant coverage levels: 30x and 50x. Our assemblies are better than those using MHAP for both coverage levels. Moreover, Canu+HISEA assemblies for 30x coverage are comparable with Canu+MHAP assemblies for 50x coverage, while being faster and cheaper.

Availability: http://www.csd.uwo.ca/faculty/ilie/

Authors:

first name	**last name**	**email**	**country**	**organization**	**corresponding?**
Nilesh	Khiste	nkhiste@uwo.ca	Canada	University of Western Ontario	✓
Lucian	Ilie	ilie@uwo.ca	Canada	University of Western Ontario	✓

SwiftCluster: A unified parallel computational framework for large scale genetic variation analysis in heterogeneous cloud environments

Keywords: cloud computing, parallel computing, genetic variation analysis, next generation sequencing, variant detection, genomics, bioinformatics

Abstract: Large scale genetic variant analysis plays a critical role in elucidating the causes of various human diseases. However, the tera- and peta-byte scale footprint for sequence data imposes significant technical challenges and financial burdens for data management and analysis, including the tasks of collection, storage, transfer, sharing, and privacy protection. Cloud Computing has recently emerged as a compelling paradigm for large scale data management and big data analysis due to the availability, scalability, and cost effectiveness of computing resources across the global. However, it is still a challenge, and often big headache, for researchers themselves to either set up appropriate instance clusters in a popular cloud (Amazon cloud, Google cloud, or Microsoft Cloud), or transfer existing analysis pipeline to a cloud for parallel computing with all the necessary software environments even before a meaningful analysis tasks can be performed. Here, we present an easy-to-use parallel computational framework, SwiftCluster, for the large scale genetic variation analysis using next-generation sequencing data in heterogeneous cloud environments. This framework not only includes a customized machine image with preconfigured tools for read alignment and variant calling (https://github.com/ngs-swift/Introduction), but also enables users to launch and manage grid-engine cluster with multiple instance nodes so that the time consuming analysis can be expedited by distributing vast computing jobs into multiple nodes. Moreover, this tool has a unified command interface allowing users to work on multiple heterogeneous cloud platforms such as AWS and GCLOUD. Performance of this framework has been evaluated in cloud using dbGaP study phs000710.v1.p1, and security handling in cloud environment when dealing with control-accessed sequence data has also been addressed.

Authors:

first name	last name	email	country	organization	corresponding?
Chunlin	Xiao	xiao2@mail.nih.gov	USA	NIH	✓

MINTIA : Metagenomic INsertT Bioinformatic Annotation in Galaxy

Keywords: metagenomic, activity-based screening, DNA assembly, gene annotation

Abstract: Functional metagenomics is used to understand who is doing what in microbial ecosystems . DNA sequencing can be priorized by activity-based screening of libraries obtained by cloning and expressing metagenomic DNA fragments in an heterologous host. When large insert libraries are used, allowing a direct access to the functions encoded by entire metogenomic loci sizing several dozens of kbp, NGS is required to identify the genes which are responsible for the screened function. The pipeline presented here allows biologists to easily assemble, clean and annotate their NGS sequences. It has been set up in Galaxy as two tools which can easily be chained in a pipeline. The first one produces the cleaned assemblies and their metrics from the sequencing reads. It provides users with a table containing links to the different files corresponding to the assembly and vector cleaning steps as well as an interactive graph showing contig depth and length for all metagenomic inserts. It enables a quick assembly validation. Another output of this module is a compressed file including metagenomic insert sequences after assembly and cleaning, in fasta format. The second tool generates an annotation table including Metagene ORF finding and blast annotation against the nr, Swissprot and COGs databases which will help biologists to make functional and taxonomic annotation. The table includes links to the alignments files enabling a precise analysis. The poster presents the results for simulated and real read sets.

Authors:

first name	last name	email	country	organization	corresponding?
Christophe	Klopp	christophe.klopp@inra.fr	France	INRA	✓
Laguerre	Sandrine	laguerre@insa-toulouse.fr	France	INSA	
Maman	Sarah	sarah.maman@inra.fr	France	INRA	
Legoueix	Sabrina	sabrina.legoueix@inra.fr	France	INRA	
Laville	Elisabeth	laville@insa-toulouse.fr	France	INSA	
Potocki-Véronèse	Gabrielle	veronese@insa-toulouse.fr	France	INSA	

From RNA-seq to gene expression: comparative evaluation of multiple pipelines and differential expression methods on experimental human data

Keywords: RNA-seq data analysis, Algorithms, Differential gene expression

Abstract: Background: Many algorithms and pipelines have been developed to analyse RNA-seq data. Nevertheless, there is much debate about which of these approaches provide the best results.

Material and Methods: KMS12-BM and JJN-3 multiple myeloma cell lines were used to test amiloride and TG003 drugs. Control and treatment samples were sequenced by triplicate. Several pipelines combining 3 trimming methods, 5 algorithms for alignment, 5 counting methods and 6 normalization approaches were tested. Precision and accuracy of the 108 resulting pipelines were determined by analysing the MAD ranks and qPCR correlation for the most stable expressed genes, respectively. 9 differential expression programs were compared under 5 testing approaches.

Objective: To assess the performance of the most widely used RNA-seq algorithms.

Results: We observed that most of the pipelines provided comparable results in precision, except those based on eXpress, which performed worse regarding this parameter. With respect to the accuracy, pipelines based on the counting algorithm HTSeq reached the best results, being the preferred alignment methods RUM and HISAT2. When evaluating the differential expression of genes (DEG) we found a high degree of overlap in methods such as EdgeR, Limma and DESeq2. We also noticed that the detection power of these methods was dependent of the similarities between the 2 compared groups.

Conclusion: To our knowledge, the best analysis approach included the HTSeq method, being RUM and HISAT2 the most recommendable aligners. EdgeR, Limma and DESeq2 were the methods which performed similarly well for DEG detection.

Funding: “Fundación Española de Hematología y Hemoterapia”

Authors:

first name	last name	email	country	organization	corresponding?
Luis Antonio	Corchete	lacorsan@usal.es	Spain	Hospital Universitario de Salamanca	✓
Elizabeta A.	Rojas	personitha@gmail.com	Spain	Instituto de Investigación Biomédica de Salamanca	
Diego	Alonso	diego.alonso@usal.es	Spain	Centro de Investigación del Cáncer de Salamanca	
Javier	De Las Rivas	jrivas@usal.es	Spain	Centro de Investigación del Cáncer de Salamanca	
Norma C.	Gutiérrez	normagu@usal.es	Spain	Hospital Universitario de Salamanca	✓
Francisco Javier	Burguillo	burgui@usal.es	Spain	Facultad de Farmacia, Universidad de Salamanca	✓

ImmunExplorer Online: Implementation of a Web Service for Analyzing Clonality and Diversity of Next-Generation Sequencing Data

Keywords: next-generation sequencing, immunoinformatics, clonality, diversity, web service

Abstract: For analyzing the immune system, a very precious and essential factor for humans' health, we need computational methods to properly analyze the huge amount of generated immunosequencing data in the World Wide Web.

We here present a new web service called ImmunExplorer Online that enables users to upload their NGS or preprocessed IMGT/HighV-QUEST data, to perform clonality analysis, diversity calculations, and additionally provide important statistics based on the functional and non-functional sequences. Additionally, users can create and manage their own projects and results can be visualized and downloaded.

The basis of these analyses forms the freely available software framework ImmunExplorer (IMEX) which enables the analysis of raw next-generation sequencing data and preprocessed IMGT/HighV-QUEST data. Several features, such as the calculation of the clonality, the diversity, primer efficiency analyses, and the prediction of the status of the adaptive immune repertoire using machine learning algorithms are implemented in IMEX. Moreover, various analyses about the V-(D)-J rearranged regions, genes and alleles, and statistics about preprocessed IMGT/HighV-QUEST data can be determined.

Using ImmunExplorer Online users can run a full pipeline to profile the human adaptive immune system starting from raw NGS data to health state prediction. The profiling of the immune system and the analysis of the interaction of its key players is one of the most addressed research topics in the field of immunoinformatics and therefore this web service shall help to gain detailed insights in medical research, transplantation medicine, and in the diagnosis and treatment of diseases.

Authors:

first name	**last name**	**email**	**country**	**organization**	**corresponding?**
Marina	Pavic	S1410458013@students.fh-hagenberg.at	Austria	University of Applied Sciences Upper Austria, Hagenberg Campus	
Kimberly	Bouchal	S1410458002@students.fh-hagenberg.at	Austria	University of Applied Sciences Upper Austria, Hagenberg Campus	
Sara	Fregnan	S1510458009@students.fh-hagenberg.at	Austria	University of Applied Sciences Upper Austria, Hagenberg Campus	
Florian	Hametner	S1510458012@students.fh-hagenberg.at	Austria	University of Applied Sciences Upper Austria, Hagenberg Campus	
Marcel	Luckeneder	S1520458042@students.fh-hagenberg.at	Austria	University of Applied Sciences Upper Austria, Hagenberg Campus	
Simon	Weinzinger	S1510458038@students.fh-hagenberg.at	Austria	University of Applied Sciences Upper Austria, Hagenberg Campus	
Johannes	Weinberger	weinberger.johannes@gmail.com	Austria	Österreichisches Rotes Kreuz, Wien	
Stephan	Winkler	stephan.winkler@fh-hagenberg.at	Austria	University of Applied Sciences Upper Austria, Hagenberg Campus	
Susanne	Schaller	susanne.schaller@fh-hagenberg.at	Austria	University of Applied Sciences Upper Austria, Hagenberg Campus	✓

Feasibility of sample size calculation for RNA-seq differential expression studies

Keywords: sample size calculation, power, RNA-seq, study design, replicates, tool evaluation

Abstract: Sample size calculation is crucial to ensure sufficient statistical power for detecting existing effects, but rarely performed for RNA-seq studies.

To evaluate feasibility and provide guidance, we performed a systematic search and evaluation of open sources tools for RNA-Seq sample size estimation. We used simulations based on real data to examine which tools performs well for different levels of fold changes between conditions, different sequencing depth and variable numbers of differentially expressed genes. Furthermore we examined the effect of the pilot replicate number on the results and if real pilot data are necessary for reliable results at all. In addition we looked at the actual false discovery rate correction.

The six evaluated tools provided widely different answers for human data, which were affected by fold change, the demanded power values and the used data. While all tools failed for small fold changes, some tools can at least be recommended when closely matching pilot data are available and relatively large fold changes are expected.

Authors:

first name	last name	email	country	organization	corresponding?
Alicia	Poplawski	alpoplaw@uni-mainz.de	Germany	UNIVERSITY MEDICAL CENTER of the Johannes Gutenberg University Mainz	✓
Harald	Binder	binderh@imbi.uni-freiburg.de	Germany	University of Freiburg	

RNA sequencing approach to investigate transcripts involved in platinum resistance in high-grade serous epithelial ovarian cancer

Keywords: biomarkers for platinum resistance, high-grade serous ovarian carcinoma, RNA sequencing

Abstract: The most common histological subtype of epithelial ovarian cancer, the high-grade serous ovarian carcinoma (HGS-EOC), shows five-years survival rate less than 30%, despite an initial response to platinum agents, the patients become progressively resistant and die becoming incurable.

The conventional array-based approaches, drawn on known transcript structures, failed to identify biomarkers for platinum resistance. Thus, with the aim to discover new mutations or transcript variants associated with the mechanism of resistance, we sequenced the transcriptomes of multiple HGS-EOC biopsies: 14 biopsies from platinum-sensitive patients, 14 biopsies from platinum-resistant patients and 16 matched longitudinal biopsies (tumor specimens collected from insurgence throughout the progression of the disease) sensitive at the time of first biopsy and resistant at the time of the last biopsy.

The samples and analyses, here presented, are part of an ongoing project, called VIOLeTS, that aims, using novel methodological and computational approaches coupled to next generation sequencing of DNA and RNA, to study the transcriptional and genomics changes occurring in the HGS-EOC transcriptomes leading to relapse and resistance, allowing the study of the tumor evolution during therapies. After the first-year project, the transcriptome reconstruction of the 14 sensitive and 14 resistant tumors highlights 1371 transcripts differentially expressed between resistant and sensitive samples: 125 known transcripts, 686 potentially novel isoforms of known transcripts and, the remaining, if validated, suggest novel intergenic transcripts and anti-sense transcripts. Interestingly, a very small part of the collected transcriptional alterations can be ascribed to coding-genes, suggesting a prominent non-coding role in HGS-EOC platinum resistance.

Availability: http://romualdi.bio.unipd.it/

Authors:

first name	last name	email	country	organization	corresponding?
Enrica	Calura	enrica.calura@unipd.it	Italy	University of Padova	✓
Luca	Beltrame		Italy	Mario Negri Institute, Milan, Italy	
Paolo	Martini		Italy	University of Padova	
Gabriele	Sales		Italy	University of Padova	
Antonella	Ravaggi		Italy	University of Brescia, Brescia, Italy	
Eliana	Bignotti		Italy	University of Brescia, Brescia, Italy	
Maurizio	D'Incalci		Italy	Mario Negri Institute, Milan, Italy	
Sergio	Marchini		Italy	Mario Negri Institute, Milan, Italy	
Chiara	Romualdi		Italy	University of Padova	

Tracing fine-scale temporal evolution of yeast populations using de novo meta-assembly

Keywords: Meta-assembly, Temporal evolution, Genome variation

Abstract: Abnormal variations are frequent in clonal genome evolution of cancers. Such aberrational variations often function as a driver in cancer cell growth. Understanding fundamental evolutionary dynamics underlying these variations in tumor metastasis still remains understudied owing to their genetic complexity.

Recently, whole genome sequencing empowers to determine genome variations in short-term evolution of cell populations. This approach has been applied to evolving

populations of unicellular organisms including yeast. It is substantial progress in evolutionary genomics to examine sequence changes at such fine-scale resolution. These studies, however, have been limited to observing only point mutations and small insertions and deletions relying on a given reference sequence due to the incomplete fragmented construction of individual de novo genome assemblies.

We herein design a new meta-assembly approach for building the sequence assembly of each population at different time points and use time-series analysis for identifying novel genome-wide variations. We improve the continuity and accuracy of the genome assembly and determine the evolutionary patterns of variations in big data of yeast (Saccharomyces cerevisiae) W303 strain genomes from 40 populations at 12 time points.

Availability: http://www.stanford.edu/~gsong

Authors:

first name	last name	email	country	organization	corresponding?
Giltae	Song	gsong@stanford.edu	South Korea	Pusan National University	✓
Jongin	Lee	jongin333@gmail.com	South Korea	Konkuk University	
Jaebum	Kim	jaebum.kim@gmail.com	South Korea	Konkuk University	✓

Mining for Viral Genomes within Sequencing of Complex Communities

Keywords: Virus, Bacteriophage, Metagenomics, Genomics

Abstract: Bacteriophages (phages), or viruses that infect bacteria, are the most abundant group of viruses on Earth and play a critical role in structuring bacterial communities. Despite their prevalence and ubiquity, only a small number of phage genomes have been sequenced and characterized. Viral communities inhabiting niches from across the globe have now been sequenced, uniformly discovering a wealth of sequence data with no recognizable homology to extant sequence collections. Nevertheless, novel viral species genomes have been successfully excavated directly from complex community metagenomes. Discovery of such viral genomes often relies heavily on manual curation and prior studies have employed a variety of different criteria when sifting through sequencing data.

To provide an automated solution for identifying viral genomes from complex sequence data sets, we developed the tool virMine. Synthetic metagenome data sets were created and examined to assess the performance of this new tool, testing its sensitivity to both variation in the abundance of viral relative to non-viral sequences and sequence divergence from previously characterized taxa. VirMine was next used to mine viral metagenomic data sets from several studies of: (1) the gut virome, (2) the urinary virome, and (3) the freshwater virome. Numerous complete and largely-complete phage genome sequences resembling previously characterized phage species were extracted from each dataset without manual intervention. Furthermore, novel putative phage genomes were identified warranting further investigation and confirmation in the lab. The virMine tool provides a robust and expedient means for viral genome sequence discovery from complex community sequence data.

Authors:

first name	last name	email	country	organization	corresponding?
Andrea	Garretto	agarretto@luc.edu	USA	Loyola University	✓
Catherine	Putonti	cputonti@luc.edu	USA	Loyola University	✓

Comparison of Sequencing Platforms to Measure Phylogenetic Diversity of Gut Microbiota

Keywords: Next Generation Sequencing, Gut Microbiota, Phylogenetic Diversity

Abstract: Standardization of measurement of the human gut microbiota is the first step to combine individual's microbial profile to the other types of healthcare data. The microbial profile is affected by various factors, such as fecal sampling condition, sequencing platform, targeted sequencing region, etc. In this study we focused to evaluate commercial sequencing platforms and targeted 16S region. With assist of bioinformatics tools, the 454 FLX pyrosequencer has been successfully applied to obtain microbial community profiling. Illumina MiSeq platform has increasingly outpaced the 454 systems, mainly due to its much higher throughput in contrast of the limitation of the short read length. Pacific Biosciences (PacBio) Single Molecule, Real-Time (SMRT) DNA sequencing system has become available for microbial phylogenetic profiling with the ability of full-length 16S sequencing. For this purpose, we generated fecal sequences from 170 Korean subjects using the GS FLX+ (V1-4), Illumina MiSeq (V1–3, V3–4 and V4), and PacBio (V1–9) systems. We compared the phylogenetic resolution and abnormality of the simulation study of public 16S rRNA gene database. The information generated from this study will become a valuable source of construction of the standard protocol for Korean gut microbiome analysis.

Authors:

first name	last name	email	country	organization	corresponding?
Won-Hyong	Chung	whchung@kfri.re.kr	South Korea	Korea Food Research Institute	✓

Abseq: Quality Control Pipeline for the Construction of Antibody Libraries

Keywords: Antibody library, Immune repertoire, Next-generation sequencing, TCR and BCR analysis, Rep-Seq

Abstract: During the past two decades, a large number of antibody libraries have been constructed to meet the needs of drug discovery and diagnostic processes. The advent of next-generation sequencing (NGS) technology has enabled scientists to rigorously assess library size, quality, diversity and robustness at different stages of the construction process. The currently available bioinformatic tools mainly focus on the analysis of clonotypes of T cell receptors. We propose a new software pipeline, Abseq, designed to facilitate a high-throughput analysis of NGS reads of the variable domain of an antibody chain. The Abseq pipeline includes all the essential analysis steps from merging paired-end reads, annotating V-(D)-J rearrangement, estimating the abundance levels of germline genes and families, visualising the alignment quality of germline genes (including the filtering of low quality sequences), predicting frame shifts and identifying functional clones, and finally calculating spectratypes and clonotypes to estimate the diversity of the library. Importantly, Abseq also contains functionality to facilitate the selection of the best combination of restriction enzymes for the construction of library vectors. We illustrate the pipeline capabilities by applying it to a naïve IgM repertoire extracted from peripheral blood lymphocyte (PBL) of pooled human donors. The results show that the abundance of germline genes is inline with the natural distribution that is reported in the literature. The integrity of frameworks, complementarity-determining regions and secretion signals has been examined through comprehensive motif analysis. Overall, the results confirm that the repertoire is not pathological and can be used for library construction.

Availability: http://ww.csl.com.au

Authors:

first name	last name	email	country	organization	corresponding?
Monther	Alhamdoosh	m.hamdoosh@gmail.com	Australia	CSL Limited	✓
Chao-Guang	Chen	ChaoGuang.Chen@csl.com.au	Australia	CSL Limited	
Milica	Ng	Milica.Ng@csl.com.au	Australia	CSL Limited	✓
Georgina	Sansome	Georgina.Sansome@csl.com.au	Australia	CSL Limited	✓
Irene	Kiess	Irene.Kiess@csl.com.au	Australia	CSL Limited	✓
Tobias	Sargeant	tobias.sargeant@gmail.com	Australia	CSL Limited	✓
Michael	Wilson	Michael.Wilson@csl.com.au	Australia	CSL Limited	✓
Con	Panousis	Kosta.Panousis@csl.com.au	Australia	CSL Limited	✓

Profiling immunoglobulin repertoires by RNA Sequencing across 8555 samples from 53 GTEx tissues

Keywords: immunoglobulin, RNA-Seq, GTEx, immune response

Abstract: Assay-based approaches provide a detailed view of the adaptive immune system by profiling immunoglobulin (Ig) receptor repertoires. However, these methods carry a high cost and lack the scale of standard RNA sequencing (RNA-Seq). Here we report the development of ImReP, a novel computational method for rapid and accurate profiling of the immunoglobulin repertoire from regular RNA-Seq data. ImReP can also accurately assemble the complementary determining regions 3 (CDR3s), the most variable regions of Ig receptors. We applied our novel method to 8,555 samples across 53 tissues from 544 individuals in the Genotype-Tissue Expression (GTEx v6) project. ImReP is able to efficiently extract Ig-derived reads from RNA-Seq data. Using ImReP, we have created a systematic atlas of Ig sequences across a broad range of tissue types, most of which have not been studied for Ig receptor repertoires. We also compared the GTEx tissues to track the flow of Ig clonotypes across immune-related tissues, including secondary lymphoid organs and organs encompassing mucosal, exocrine, and endocrine sites, and we examined the compositional similarities of clonal populations between these tissues. The Atlas of Immune Immunoglobulin repertoires (The AIR), is freely available at https://smangul1.github.io/TheAIR/ , is one of the largest collection of CDR3 sequences and tissue types. We anticipate this recourse will enhance future immunology studies and advance development of therapies for human diseases. ImReP is freely available at https://sergheimangul.wordpress.com/imrep/ .

Availability: http://cs.ucla.edu/~serghei/

Authors:

first name	last name	email	country	organization	corresponding?
Serghei	Mangul	smangul@ucla.edu	USA	UCLA	✓
Igor	Mandric	mandric.igor@gmail.com	USA	Georgia State University	
Harry Taegyun	Yang	Harry2416@gmail.com	USA	UCLA Department of Computer Science, Zarlab	
Nicolas	Strauli	Nicolas.Strauli@ucsf.edu	USA	UCSF	
Sagiv	Shifman	sagiv.shifman@gmail.com	Israel	The Hebrew University of Jerusalem	
Noah	Zaitlen	noahaz@gmail.com	USA	University of California San Francisco	
Maura	Rossetti	mrossetti@mednet.ucla.edu	USA	UCLA	
Mark	Ansel	Mark.Ansel@ucsf.edu	USA	University of California San Francisco	
Eleazar	Eskin	eeskin@cs.ucla.edu	USA	University of California, Los Angeles	

AMPS: A pipeline for screening archaeological remains for pathogen DNA

Keywords: Metagenomics, Pathogen Screening, aDNA

Abstract: High-throughput DNA sequencing (HTS) enables metagenomic studies to be performed at large-scale. Such analyses are not restricted to present day environmental or clinical samples but can also be applied to molecular data from archaeological remains (ancient DNA) in order to provide insights into the relationship between hosts and bacteria through time. Here we present AMPS (Ancient Metagenomic Pathogen Screening), an automated bacterial pathogen screening pipeline for ancient DNA sequence data that provides straightforward and reproducible information on species identification and authentication of its ancient origin. AMPS consists of a customized version of (1) MALT (Megan ALignment Tool) (Herbig, et al. 2016), (2) RMAExtractor, a Java tool that evaluates a series of authenticity criteria for a list of target species, and (3) customizable post-processing scripts to identify, filter, and visualize candidate hits from the RMAExtractor output.

We evaluated AMPS with DNA sequences obtained from archaeological samples known to be positive for specific pathogens, as well as simulated ancient DNA data from 33 bacterial pathogens of interest spiked into diverse metagenomics backgrounds (soil, archaeological bone, dentine, and dental calculus). AMPS successfully confirmed all experimental samples. AMPS further correctly identified all simulated target pathogens that were present with at least 500 reads in the metagenomic library. In addition, we used these data to assess and compensate for biases resulting from the reference database contents and structure.

AMPS provides a versatile and fast pipeline for high-throughput pathogen screening of archaeological material that aids in the identification of candidate samples for further analysis.

Authors:

first name	last name	email	country	organization	corresponding?
Ron	Hübler	huebler@shh.mpg.de	Germany	Max Planck Institute for the Science of Human History	✓
Felix M.	Key	key@shh.mpg.de	Germany	Max Planck Institute for the Science of Human History	
Christina	Warinner	warinner@shh.mpg.de	Germany	Max Planck Institute for the Science of Human History	
Kirsten	Bos	bos@shh.mpg.de	Germany	Max Planck Institute for the Science of Human History	
Johannes	Krause	krause@shh.mpg.de	Germany	Max Planck Institute for the Science of Human History	
Alexander	Herbig	herbig@shh.mpg.de	Germany	Max Planck Institute for the Science of Human History	✓

Identification of robust and sub-clonal variants in poly-ploid panel-seq data by stochastic bootstrapping

Keywords: Next-Gen sequencing, Panel sequencing, Personalized medicine, Bootstrapping, Stochastic optimization, Variant calling, Cancer research

Abstract: Panel sequencing of patient-derived cancer samples has become an important tool for clinicians and researchers. Calling variants on panel-seq data however, is error-prone due to a limited amount of raw material, unknown ploidy-settings and purity-rates for commonly hyper-ploid and impure samples. The Perturber method identifies robust or sub-clonal variants in addition to determining the parameters that maximize the likelihood of the observed variant calls. This is achieved by calculating the a posteriori likelihood for a GATK null-hypothesis-run that is bootstrapped by a stochastic grid-search over different algorithms and parameter settings. The method is freely available and has been benchmarked on the 1000 genomes and a corresponding somatic variant calling goldstandard dataset.

Availability: http://www.informatik.hu-berlin.de/forschung/gebiete/wbi/

Authors:

first name	last name	email	country	organization	corresponding?
Raik	Otto	raik.otto@hu-berlin.de	Germany	Humboldt-Universität zu Berlin	✓
Ulf	Leser	leser@informatik.hu-berlin.de	Germany	Institut für Informatik, Humboldt-Universität zu Berlin	✓
Christine	Sers	christine.sers@charite.de	Germany	Laboratory of Molecular Tumorpathology, Institute of Pathology, University Medicine Charite	✓

Development of a bioinformatics pipeline for the routine analysis of viral whole genome sequencing data

Keywords: NGS, bioinformatics, public health, virus, consensus sequence

Abstract: Viral infections can be a major public health threat, sometimes causing massive epidemics. The viral genome is of importance for both characterization and prevention. As the viral genotype can diverge rapidly because of high mutation rates, traditional sequencing approaches such as genome walking are laborious and time consuming, while often only generating fragmented sequences. Alternatively, utilizing whole genome sequencing (WGS), the complete viral genome can be obtained providing unprecedented resolution for the study of viral genomics. The lack of the required bioinformatics expertise for analysing WGS data is however a hurdle preventing its broad adaptation in many national reference centers. We developed a pipeline specifically designed to bridge this gap. Our pipeline is flexible and moreover species-agnostic. It performs automated quality control based on Illumina data (including removal of host DNA when working in a metagenomics context), and then either de novo subtyping or subtyping based on a user-provided set of input reference genomes, to generate the viral consensus sequence contained within a sample. A detailed output report containing intermediary results and quality parameters of importance is also created. A user-friendly interface has been deployed in an in-house Galaxy instance to facilitate access to a broad audience of scientists. Preliminary validation on influenza and mumps data demonstrates that our pipeline is capable of obtaining high-quality viral consensus sequences, providing a solid basis for downstream analyses such as viral genotyping, in silico serotyping, virulence and/or resistance characterization. Our pipeline can easily be adapted to other viral species and will be made publicly available upon its publication.

Availability: http://www.wiv-isp.be

Authors:

first name	last name	email	country	organization	corresponding?
Qiang	Fu	qiang.fu@wiv-isp.be	Belgium	Scientific Institute of Public Health	✓
Bert	Bogaerts	bert.bogaerts@wiv-isp.be	Belgium	Scientific Institute of Public Health	
Raf	Winand	raf.winand@wiv-isp.be	Belgium	Scientific Institute of Public Health	
Julien	Van Braekel	julien.vanbraekel@wiv-isp.be	Belgium	Scientific Institute of Public Health	
Cyril	Barbezange	cyril.barbezange@wiv-isp.be	Belgium	Scientific Institute of Public Health	
Veronik	Hutse	Veronik.Hutse@wiv-isp.be	Belgium	Scientific Institute of Public Health	
Steven	Van Gucht	steven.vangucht@wiv-isp.be	Belgium	Scientific Institute of Public Health	
Sigrid	De Keersmaecker	Sigrid.DeKeersmaecker@wiv-isp.be	Belgium	Scientific Institute of Public Health	✓
Nancy	Roosens	nancy.roosens@wiv-isp.be	Belgium	Scientific Institute of Public Health	✓
Kevin	Vanneste	kevin.vanneste@wiv-isp.be	Belgium	Scientific Institute of Public Health	✓

Automation from whole genome sequencing to comprehensive genome comparison

Keywords: sequence analysis, genome comparison, automation

Abstract: The aim of the project is to develop a workflow that automates the processing, assembly and annotation of whole genome sequencing data. Therefore, the snakemake workflow engine is applied, which automates the use of executable tools and scripts.

The steps include quality control, de novo assembly, gene prediction, gene annotation and gene comparison. Usually they require time for configuration and interpretation of the intermediate data, the workflow combines these tasks in one step. The key features of the presented workflow are the generation of hybrid assemblies from PacBio and Illumina sequencing data, optional filtering of prokaryotic sequences (from non-axcenic eukaryotic cultures) and advanced gene prediction from available RNA sequencing data.

The workflow is applied to five whole genome sequenced Chrysophyceae strains using the technologies of Illumina Hiseq XTen and Pacbio RSII. In the course of evolution, Chrysophyceae frequently reduced their plastids and accordingly their nutritional mode. Hence, the genome comparison reveals details in dependence of nutrition: essential and optional genes, gene density and arrangement, GC content and genome size.

The automated Snakemake workflow incorporates SPAdes to assemble the Illumina and Pacbio reads. Additionally, in non-axenic cultures the software MaxBin2.0 and Kraken separates the contigs in eukaryotic and prokaryotic sequences. If available RNA-Seq data aides the gene prediction process of the programs Tophat2, Augustus, Genemark and Braker. The predicted genes are searched with Diamond against the KEGG database. Finally, gene matches of each species are clustered and compared among the different nutritional modes.

This workflow will simplify future genomic analysis.

Authors:

first name	last name	email	country	organization	corresponding?
Stephan	Majda	stephan.majda@uni-due.de	Germany	university of Duisburg-Essen	✓

CoCo: RNA-seq Read Assignment Correction for Nested Genes and Multimapped Reads

Keywords: RNA, High-throughput sequencing, Read assignment, snoRNA, Small non-coding RNA, Transcriptomics, Nested genes

Abstract: The human genome is complex, it holds nested genes, genes with multiple copies and overlapping genes, the study of which is difficult. Small nucleolar RNA (snoRNA) are one such family, most being intronic, many nested in an intron retention. Furthermore, the latest advances in RNA sequencing enable the study of different types of RNA at once. Existing tools are generally designed for a specific RNA type at the expense of others and do not address these genome particularities, or only one of them.

We developed CoCo (Count Corrector for nested genes and multimapped reads) which modifies an annotation, inserting holes in exons and in retained introns containing nested genes. This annotation is then submitted to an existing tool like Subread's featureCounts. Afterwards, CoCo distributes the counts from multimapped reads, usually coming from duplicated genes, based on the proportion of uniquely mapped reads. This approach prevents from allocating counts to a non-expressed gene, for example in the case of a protein coding gene and its pseudogene.

CoCo salvages over 15% of reads that are usually left out. Using CoCo, we detect 133 more snoRNA species than with traditional methods and 60% of rescued reads come from snoRNA. With the multimapped read distribution, the estimated counts triple for genes with multiple copies like the signal recognition particle 7SL. The correlation between different types of RNA measured by PCR and sequencing is higher using CoCo than traditional methods. Thus, CoCo gives a better portrait of the abundance of most RNA types.

Authors:

first name	last name	email	country	organization	corresponding?
Gabrielle	Deschamps-Francoeur	Gabrielle.Deschamps-Francoeur@USherbrooke.ca	Canada	Université de Sherbrooke	
Vincent	Boivin	Vincent.Boivin2@USherbrooke.ca	Canada	Université de Sherbrooke	
Sherif	Abou-Elela	Sherif.Abou.Elela@USherbrooke.ca	Canada	Université de Sherbrooke	
Michelle	S. Scott	Michelle.Scott@USherbrooke.ca	Canada	Université de Sherbrooke	✓

Systematic comparative study of immunogenetics analysis algorithms

Keywords: Bioinformatics, Algorithms, Immunosequencing, T/B cell receptors, Comparative assessment

Abstract: High-throughput sequencing has broadened the possibility to dissect the immune repertoire at a higher resolution to deepen our understanding of the adaptive immune system. Most significant insights can be gained relative to various states as cancer, autoimmune conditions, infection and aging process. Additionally, it can assist to uncover the underlying mechanisms of immunity in health and disease conditions. With further progress in sequencing technologies higher amount of data is being generated, which requires sophisticated analysis methods. Various tools have been developed so far for immunosequencing (T and B cell receptors) analyses and in this aspect progress is currently underway. The major analysis aim is to unravel the diversity of the immune system and perform composition profiling to obtain clinically relevant information. In T cell receptor analyses the general strategy includes steps to determine gene segments by aligning sequences to the reference set, clonotypes identification, detection of complementarity determining region 3 and their abundance estimation. However, there is still a clear lack of guidelines and consensus about features crucial for reliable data analysis. To provide a practical comparison of the computational methods for immune repertoire analyses, we have conducted an in-depth and systematic comparative study of eight available methods. We have employed numerous in silico and experimental datasets to perform thorough assessment of each approach in view of various analysis factors. Moreover, a clonal plane analysis strategy is used to perform clonality analysis of samples under investigation. In addition, we describe in detail the substantial effects of choice of analysis method on interpretation and outcome. Our study will help researchers in this filed to select an optimal analysis method and in this regard will provide basic evaluation based guidelines.

Authors:

first name	last name	email	country	organization	corresponding?
Saira	Afzal	saira.afzal@nct-heidelberg.de	Germany	German Cancer Research Center (DKFZ)	✓
Irene	Gil-Farina	irene.gil-farina@nct-heidelberg.de	Germany	German Cancer Research Center (DKFZ)	
Richard	Gabriel	richard.gabriel@nct-heidelberg.de	Germany	German Cancer Research Center (DKFZ)	
Shahzad	Ahmad	s.ahmad@erasmusmc.nl	Germany	German Cancer Research Center (DKFZ)	
Christof	von Kalle	christof.kalle@nct-heidelberg.de	Germany	German Cancer Research Center (DKFZ)	
Manfred	Schmidt	manfred.schmidt@nct-heidelberg.de	Germany	German Cancer Research Center (DKFZ)	
Raffaele	Fronza	raffaele.fronza@nct-heidelberg.de	Germany	German Cancer Research Center (DKFZ)	

A computational approach for inferring the sequential acquisition of somatic copy number alterations in genomically unstable cancers

Keywords: Somatic copy number alterations, Whole genome duplication, Computational biology, Colon-rectal cancer

Abstract: Somatic copy number alterations (SCNAs) are pervasive in cancer due to genomic instability that can lead to whole genome duplication (WGD), focal amplifications or deletion, and other chromosomal abnormalities.

WGD is frequently observed in cancer and is typically associated with adverse outcomes suggesting that it plays an essential role in the development of an aggressive tumour phenotype. However, as WGD does not in occur in isolation but in concert with other chromosomal aberrations, the doubling of whole genomic content will convolute with the effect of other SCNAs, resulting in a complex landscape of chromosomal rearrangements that is highly challenging to interpret.

Here we developed a computational approach for parsing complex copy number profiles from multiple tumour samples that can be used to deconvolute the effect of WGD and focal alterations. The method seeks to separate recurrent and tumour-specific SCNA events and can be viewed as a form of dimensionality reduction for structured high-dimensional discrete data. The output from our method provides an estimate of the sequential series of copy number alteration events that occur in the tumours. The problem is modelled as an optimization problem with an quadratic objective function and constraints. The Lagrange dual for the original optimization problem is solved by fixed point method. There are also a handful user-defined parameters in the model, so that the method is quite flexible to account for various forms of user needs.

We demonstrate the utility of the method by analysing 380 CRC sample data from The Cancer Genome Atlas. The method gives better indication of the existence of WGD in the sample than merely average ploidy. The results also showed that there are different causes for the copy number differences among different genome loci. Thus, this method could potentially help researchers understand the evolution of SCNA in cancer.

Availability: http://cwcyau.github.io/

Authors:

first name	last name	email	country	organization	corresponding?
Yun	Feng	yun.feng@jesus.ox.ac.uk	United Kingdom	University of Oxford	✓
Christopher	Yau	c.yau@bham.ac.uk	United Kingdom	University of Birmingham	✓

BCFtools/csq: haplotype-aware variant consequences

Keywords: variant consequence prediction, haplotype-aware calling, bcftools

Abstract: Prediction of functional variant consequences is an important part of sequencing pipelines, allowing the categorization and prioritization of genetic variants for follow up analysis. However, current predictors analyze variants as isolated events, which can lead to incorrect predictions when adjacent variants alter the same codon, or when a frame-shifting indel is followed by a frame-restoring indel. Exploiting known haplotype information when making consequence predictions can resolve these issues. BCFtools/csq is a fast program for haplotype-aware consequence calling which can take into account known phase. Consequence predictions are changed for 501 of 5019 compound variants found in the 81.7M variants in the 1000 Genomes Project data, with an average of 139 compound variants per haplotype. Predictions match existing tools when run in localized mode, but the program is an order of magnitude faster and requires an order of magnitude less memory. The program is freely available for commercial and non-commercial use in the BCFtools package which is available for download from http://samtools.github.io/bcftools

Authors:

first name	last name	email	country	organization	corresponding?
Petr	Danecek	petr.danecek@gmail.com	United Kingdom	Wellcome Trust Sanger Institute	✓
Shane A.	McCarthy	sm15@sanger.ac.uk	United Kingdom	Wellcome Trust Sanger Institute	

Clustering of cells for single cell analysis using Latent Dirichlet Allocation

Keywords: single cell analysis, Latent Dirichlet Allocation, clustering

Abstract: The body of an organism is one system composed of a large number of cells, analyzing the behavior of cells contributes to elucidation of life phenomena and treatment of diseases. In particular, analytical methods such as single cell analysis with RNA-seq have attracted attention. This is due to the fact that the expression levels of cells are different for each cell even though they are of the same cell type, and also the biological function they play differs.

Clustering of cells by their gene expression profiles is common as an informatics method in single cell analysis, and it aims to discover cell heterogeneity. Many of clustering methods reduce the dimensions of expression profiles by using principal component analysis or independent component analysis. However, gene expression profiles obtained by single-cell RNA-seq protocols contain a vast amount of zeros, and this fact makes it difficult to appropriately reduce the dimension.

In this research, we propose a novel clustering method by using Latent Dirichlet Allocation (LDA) as a dimensionality reduction method, which is known to operate also on sparse matrix. Our method allocates genes into some gene sets called topics, and each topic is the result of dimensionality reduction. The topics are presumed by bias of expression levels, and genes that have similar function are assigns into the same topic. We experimented to classify cells against the actual expression profiles and showed that the method was able to classify more accurately than the conventional methods.

Authors:

first name	last name	email	country	organization	corresponding?
Mitsuhiro	Eto	m-etou@ist.osaka-u.ac.jp	Japan	Osaka University	✓
Shigeto	Seno	senoo@ist.osaka-u.ac.jp	Japan	Osaka University	
Yoichi	Takenaka	takenaka@kansai-u.ac.jp	Japan	Kansai University	
Hideo	Matsuda	matsuda@ist.osaka-u.ac.jp	Japan	Osaka University	✓

A Fast CNV Segmentation Algorithm for WGS data

Keywords: Copy Number Variation, Structural Variants, WGS Analysis, Segmentation

Abstract: Copy number variation (CNV) is a type of structural variant that affect a large range of nucleotide sequences (usually more than 1000 bp),which has been proved holding strong correlation with many genomic diseases, such as cancer. Currently, NGS-based CNV profiling are more and more prevailing. It can provide higher-resolution when compared with array-based approach, and also brings in more computational challenges at the same time. In this work, we proposed an efficient algorithm for the task of CNV segmentation on NGS data. Different from previous approaches, we proposed a vector-based bin representation and made use of the distance distribution of adjacent bins for detecting potential breakpoints in bins. Our algorithm runs in linear time approximately and owns scalability for whole genome sequencing (WGS) data. We compared our method with classic methods, such as binary circular segmentation and event-wise testing, in both simulation data and real data based on GIAB NA12878.

Availability: http://bonsai.ims.u-tokyo.ac.jp/~imoto/index.html

Authors:

first name	**last name**	**email**	**country**	**organization**	**corresponding?**
Yao-Zhong	Zhang	yaozhong@ims.u-tokyo.ac.jp	Japan	The University of Tokyo	✓
Rui	Yamaguchi	ruiy@ims.u-tokyo.ac.jp	Japan	Human Genome Center, Institute of Medical Science, University of Tokyo,	
Seiya	Imoto	imoto@ims.u-tokyo.ac.jp	Japan	Human Genome Center, Institute of Medical Science, University of Tokyo,	
Satoru	Miyano	miyano@hgc.jp	Japan	Human Genome Center, the Institute of Medical Science, University of Tokyo	

An Ultra-Fast Correction Algorithm for Hybrid Genome Assembly

Keywords: Genome Assembly, FM-index, High-Throughput Sequencing

Abstract: The 2nd and 3rd generation sequencing are now preferred choices for de novo genome reconstruction. The 2nd generation sequencing offers high throughput in low cost but the read length is inadequate to resolve large repeats. On the other hand, the 3rd generation sequencing can generate much longer reads for spanning large repeats, but the error rate and sequencing cost are much higher. This poster presented a novel algorithm for correcting low-quality long reads using FM-index constructed from high-quality short reads. In particular, long reads are mapped onto FM-index of short reads and correct sequences are generated via FM-index extension without time-dynamic programming alignment. The experimental results indicated that the correction power, accuracy, and speed are better than existing methods. The strength of hybrid assembly can be easily seen when coverage of 3rd generation sequencing is low.

Authors:

first name	**last name**	**email**	**country**	**organization**	**corresponding?**
Yao-Ting	Huang	ythuang@cs.ccu.edu.tw	Taiwan	National Chung Cheng University	✓
Ping-Ye	Chen	t7050415t@gmail.com	Taiwan	National Chung Cheng University	

Improved data-driven likelihood factorizations for transcript abundance estimation

Keywords: RNA-seq, data-driven factorization, Quantification, lightweight methods

Abstract: Motivation: Many methods for transcript-level abundance estimation reduce the computational burden associated with the iterative algorithms they use by adopting an approximate factorization of the likelihood function they optimize. This leads to considerably faster convergence of the optimization procedure, since each round of e.g. the EM algorithm, can execute much more quickly. However, these approximate factorizations of the likelihood function simplify calculations at the expense of discarding certain information that can be useful for accurate transcript abundance

estimation.

Results: We demonstrate that model simplifications (i.e. factorizations of the likelihood function) adopted by certain abundance estimation methods can lead to a diminished ability to accurately estimate the abundances of highly related transcripts. In particular, considering factorizations based on transcript-fragment compatibility alone can result in a loss of accuracy compared to the per-fragment, unsimplified model. However, we show that such shortcomings are not an inherent

limitation of approximately factorizing the underlying likelihood function. By considering the appropriate conditional fragment probabilities, and adopting improved, data-driven factorizations of this likelihood, we demonstrate that such approaches can achieve accuracy nearly indistinguishable from methods that consider the complete (i.e. per-fragment) likelihood, while retaining the computational efficiently of the compatibility-based factorizations.

Availability and Implementation: Our data-driven factorizations are incorporated into a branch of the Salmon transcript quantification tool: https://github.com/COMBINE-lab/salmon/tree/factorizations.

Authors:

first name	last name	email	country	organization	corresponding?
Mohsen	Zakeri	mzakeri@cs.stonybrook.edu	USA	Stony Brook University	
Avi	Srivastava	asrivastava@cs.stonybrook.edu	USA	Stony Brook University	
Fatemeh	Almodaresi	falmodaresit@cs.stonybrook.edu	USA	Stony Brook University	
Rob	Patro	rob.patro@cs.stonybrook.edu	USA	Stony Brook University	✓

NAGPP: Gene annotation pipeline for plant genomes with self-training and core-gene models.

Keywords: gene prediction, protein-coding regions, HMM, genome

Abstract: Many de novo genome assembly projects have been performed using high-throughput sequencers, many genomic sequences are being produced. Gene prediction is one of the most important steps in the process of genome annotation, along with the genetic assembly process. A large number of software tools and pipelines developed with various computing technologies can be used for gene prediction. However, such a pipeline does not accurately predict all or most of the protein coding regions. Also, among currently available gene prediction programs, there is no Hidden Markov Model (HMM) that can automatically perform gene prediction for all life forms. Therefore, species-specific HMMs are required for specific genome annotation.

We present a NAGPP, an automated gene prediction pipeline using a self-training HMM model, core-gene model and transcriptomic data. In this pipeline, the genome sequence and transcript sequence of the target species to be predicted using CEGMA, GlimmerHMM, SNAP, and AUGUSTUS were processed, and then the MAKER2 program was used to analyze protein sequence and the gene structure is unified. NAGPP uses the CEGMA for the plant genome that is currently being performed and generates a HMM that can be used universally without being devided into monocots and dicots, and then produces a species specific HMM. We evaluated this pipeline using the known arabidopsis and rice genomes. It was confirmed that gene structure can be identified by probabilities of 22% and 28% for Arabidopsis and rice, respectively. Because it uses CEGMA and species specific HMM, it shows better prediction results than GlimmerHMM, SNAP and Augustus used in existing MAKER2.

NAGPP provides researchers with a pipeline that can reveal a more accurate gene structure for species that are not precisely cleared of the gene structure through species specific HMMs or for new species. This pipeline concludes that gene structure prediction for new species, as well as for new model species, can yield better results than conventional pipelines used.

Authors:

first name	last name	email	country	organization	corresponding?
Kyooyeol	Lee	kylee@insilicogen.com	South Korea	Insilicogen, Inc	
Jehong	Lee	jhong@insilicogen.com	South Korea	Insilicogen, Inc	
Hyesun	Park	hspark@insilicogen.com	South Korea	Insilicogen, Inc	
Junmo	Kang	jmkang@insilicogen.com	South Korea	Insilicogen, Inc	
Myeonghee	Jung	mhjung@insilicogen.com	South Korea	Insilicogen, Inc	
Byeongchul	Kang	bckang@insilicogen.com	South Korea	Insilicogen, Inc	✓

CONFESS: Single-cell ordering by fluorescence signal estimation and modelling in R

Keywords: single-cell sequencing, transcriptomics, R

Abstract: Modern high-throughput single-cell technologies facilitate the efficient processing of hundreds of individual cells to comprehensively study their morphological and genomic heterogeneity. Fluidigm's C1 Auto Prep system isolates fluorescence-stained cells into specially designed capture sites, generates high-resolution image data and prepares the associated cDNA libraries for mRNA sequencing. Existing methods such as Monocle and Oscope for downstream analysis sort and classify cells using single-cell RNA-seq expression data and do not take advantage of the important information carried by the images themselves. We propose a novel statistical model whose multiple steps are integrated into the Cell OrderiNg (by) FluorEScence Signal (CONFESS) R package. CONFESS performs image analysis and fluorescence signal estimation for data coming from the Fluidigm C1. It collects extensive information on the cell morphology, location and signal that can be used for quality control and phenotype prediction. If applicable, it normalizes and uses the signals for unsupervised cell ordering (pseudotime estimation) and 2-dimensional clustering via scalar projection, change-point analysis and Data Driven Haar Fisz transformation for multivariate data. One could potentially use CONFESS to classify and sort fluorescent cells in various applications (cell cycle, cell differentiation, etc). Here we illustrate the use of CONFESS to trace Fucci-labeled Hela cells in their cell cycle progression. The output can be easily integrated with available single-cell RNA-seq (or other) expression profile packages for subsequent analysis.

Authors:

first name	last name	email	country	organization	corresponding?
Diana	Low	lowdiana@gmail.com	Singapore	Institute of Molecular and Cell Biology, Singapore	
Efthymios	Motakis	mefthymios@gis.a-star.edu.sg	Japan	RIKEN Center for Life Science Technologies, Yokohama, Japan	✓

An integrated process for misassembly detection using Genotyping by sequencing, BAC ends library sequences and gene synteny relations.

Keywords: NGS, BAC ends library, genome assembly, GBS, gene synteny, misassembly

Abstract: As Next Generation Sequencing technology advances, enormous amounts of whole genome sequence information in variety species have been released. However, it is still difficult to assemble the whole genome precisely due to inherent limitations of the short read sequencing technology. In particular, the complexities of plants are incomparable to those of microorganisms or animals because of whole genome duplications, repeat insertions, Numt insertions, etc. In this study, we describe a new methodology for detecting misassembly sequence regions of Brassica rapa with Genotyping-by-sequencing (GBS) followed by MadMapper clustering. The missembly candidate regions were cross-checked with BAC clone paired ends library sequences that have been mapped on the reference genome. The list were further verified with gene synteny relations between Brassica rapa and Arabidopsis thaliana. We conclude that this method will help to detect misassembly regions and be applicable to incomplete assembled reference genomes from a variety of species.

Authors:

first name	last name	email	country	organization	corresponding?
Young-Joon	Ko	suppakoko@gmail.com	South Korea	School of Systems Biomedical Science, Soongsil University, Seoul 06978, Korea.	
Jung Sun	Kim	jsnkim@korea.kr	South Korea	Genomics Division, Department of Agricultural Bioresources, National Academy of Agricultural Science, Rural Development Administration (RDA), Wansan-gu, Jeonju, Korea	
Sangsoo	Kim	sskimb@ssu.ac.kr	South Korea	School of Systems Biomedical Science, Soongsil University, Seoul 06978, Korea.	✓

Understanding non-genetic variations in the clonal population of cancer cells

Keywords: Single cell-based RNA sequencing, Non-genetic variations, Geneset-based analysis, Functional categories

Abstract: Analyzing single cell-based transcriptome profile highlights the heterogeneity of cancer cells. Although genomic instability is the major cause for the cellular variation of transcriptome in the given sample, non-genetic clonal variations of gene expression may also contribute to the differentiation and transformation of cancer cells in the course of anticancer therapies. Here we characterized the varied profile of cancer transcriptome among the homogeneous cancer cell population. Single cell-based RNA sequencing data were retrieved and analyzed for a total of 50 cells of lung cancer cell lines, H358. Varied transcriptome profiles in the clonal population were compared to the lineage-dependent variation of gene expression in diverse lung cancer cell lines. Geneset-based analysis provided new insights on functional categories associated with the non-genetic variation among homogeneous cancer cells. The present approach has applications in dissecting genetic and non-genetic factors in cancer progression.

Authors:

first name	last name	email	country	organization	corresponding?
Hyojung	Kang	kang9069@naver.com	South Korea	Sookmyung Women's University	✓
Sukjoon	Yoon	yoonsj@sookmyung.ac.kr	South Korea	Sookmyung Women's University	✓

Application of next generation sequencing for marker development in next generation plant breeding

Keywords: SNP, NGS, next generation sequencing, Breeding

Abstract: For a long time, the finger-printing methods such as RFLP, AFLP and SSR have been used for plant molecular breeding. Recently, SNP has been known that are related to specific traits and are used as new molecular markers for target trait for molecular marker-assisted selection. The development of next generation sequencing(NGS) technology has made SNPs more powerful than conventional finger-printing techniques. A large mount of SNPs produced by NGS technology enable new molecular breeding .

A Rice is a major food crop in Korea, hundreds to thousands of resequencing data are produced for the study of new varieties, and these data are stored in the National Agricultural Biotechnology Information Center(NABIC). These data are important for new plant molecular breeding studies, we need a new applications for identification of individual SNPs in two different genomes from NGS data.

This pipeline is largely composed of comparisons between individuals for the development of molecular breeding and group analysis for discrimination of origin using manual python scripts, BIOPYTHON, VCF-tools, Primer3 and PLINK. A comparision pipeline is largely composed of three parts: a) specific SNP discovery among individuals b) restriction enzyme cleavage check c) primer sequence construction for SNPs. As a result of the experiment, 30 40% of the SNP candidates were confirmed to be actual SNPs and selected as the marker candidates.

This pipeline has simplified the manual analysis process, which diffcult and complex analysis for plant molecular breeding. Although its accuracy and efficiency are different depending on the accuracy of the sequencing, it may be a tool that will be of great help to breeders.

Authors:

first name	last name	email	country	organization	corresponding?
Kyungyun	Kim	kykim@insilicogen.com	South Korea	Insilicogen	
Kyooyeol	Lee	kylee@insilicogen.com	South Korea	Insilicogen	
Giyong	Lee	jerry@insilicogen.com	South Korea	Insilicogen	
Donghun	Kim	dhkim@insilicogen.com	South Korea	Insilicogen	
Byeongchul	Kang	bckang@insilicogen.com	South Korea	Insilicogen	✓

Large-scale comparative analysis of spider fibroin genes using hybrid sequencing approach

Keywords: Spider, Fibroin gene, Hybrid sequencing, Transcriptome

Abstract: Synthetic spider silks have been explored for potential industrial applications, taking advantage of their immense toughness and renewability to realize protein-based plastic biomaterial as an alternative to those rely on petroleum. On the other hand, complete identification of spider fibroin genes still remains relatively uncharted, due to the many challenges in sequencing these genes. There are up to seven morphologically differentiated silks, and all of these genes are extremely long (¿10kbp) genes, that are almost entirely comprised of tandem iterations of repeat sequences. In order to understand the sequence design principles of spider fibroins by marrying the genotype to phenotype, we are currently conducting a de novo transcriptome study of 1,000 spiders, and we have developed a sequential read extension algorithm using a hybrid of short and long read sequencing technologies to overcome the challenges. Here, we introduce a streamlined feasibility study of various storage and logistic conditions of field samples and their effects on de novo transcriptome assembly results, the sequencing protocols and analysis algorithms, as well as the obtained knowledge about phylogenetically conserved and diverse features of spider silk gene.

Authors:

first name	last name	email	country	organization	corresponding?
Nobuaki	Kono	ciconia@sfc.keio.ac.jp	Japan	Institute for Advanced Biosciences, Keio University	
Masayuki	Fujiwara	fujimasa@ttck.keio.ac.jp	Japan	Institute for Advanced Biosciences, Keio University	
Hiroyuki	Nakamura	nakamura@spiber.jp	Japan	Spiber Inc.	
Rintaro	Ohtoshi	rintaro_ohtoshi@spiber.jp	Japan	Spiber Inc.	
Masaru	Tomita	mt@sfc.keio.ac.jp	Japan	Institute for Advanced Biosciences, Keio University	
Kazuharu	Arakawa	gaou@sfc.keio.ac.jp	Japan	Institute for Advanced Biosciences, Keio University	✓

Prioritizing candidate genes for digenic inheritance in arrhythmogenic cardiomyopathy

Keywords: digenic inheritance, arrhythmogenic cardiomyopathy, PKP2, prioritization

Abstract: Arrhythmogenic cardiomyopathy (ACM) is a genetic disorder, in which the heart muscle is progressively substituted with fibro-fatty tissue, leading to severe ventricular arrhythmias, heart failure and sudden cardiac death. Even though numerous genes are known to be involved in the disease, causal variants cannot be identified in 40% of patients and identified mutations often have low penetrance, suggesting the involvement of unknown genetic or environmental factors. In fact, recent studies have reported mutations in two different genes, implying digenic inheritance as a disease causal mechanism. We applied whole exome sequencing to investigate digenic inheritance in two ACM families in which all affected and some healthy individuals were known to carry mutations in PKP2, the gene most commonly mutated in ACM. We determined all genes that harbor variants in affected but not in healthy PKP2 carriers or vice versa. We identified likely candidates in each family by computationally prioritizing these genes and restricting to known ACM disease genes and genes related to PKP2 through protein interactions, functional relationships or shared biological functions. The top candidate in the first family is FRZB, which is located at the border of a known ACM locus and has been previously associated with other cardiac diseases. TTN, the most likely candidate in the second family, is a known ACM gene which, however, has not yet been reported in a digenic disease causal context. We propose that these variants might impair or modify protein function or structure and may cause ACM in combination with the PKP2 variant.

Authors:

first name	last name	email	country	organization	corresponding?
Eva	König	eva.koenig@eurac.edu	Italy	Eurac Research	
Claudia	Volpato	claudia.volpato@eurac.edu	Italy	Eurac Research	
Benedetta Maria	Motta	BenedettaMaria.Motta@eurac.edu	Italy	Eurac Research	
Hagen	Blankenburg	hagen.blankenburg@eurac.edu	Italy	Eurac Research	
Anne	Picard	anne.picard@eurac.edu	Italy	Eurac Research	
Peter	Pramstaller	peter.pramstaller@eurac.edu	Italy	Eurac Research	
Giulio	Pompilio	gpompilio@ccfm.it	Italy	Centro Cardiologico Monzino-IRCCS	
Viviana	Meraviglia	vivana.Meraviglia@eurac.edu	Italy	Eurac Research	
Fransisco S.	Domingues	francisco.domingues@eurac.edu	Italy	Eurac Research	
Elena	Sommariva	elena.sommariva@cardiologicomonzino.it	Italy	Centro Cardiologico Monzino-IRCCS	
Alessandra	Rossini	alessandra.rossini@eurac.edu	Italy	Eurac Research	✓

IsoPlot: a database and visualization for comparison of alternative splicing events in mosquitoes

Keywords: alternative splicing, mosquitoes, cross-species comparison

Abstract: Mosquitoes are vectors of numerous human pathogens that cause enormous public health problems but the splice isoforms of gene transcripts in these vector species are poorly curated. IsoPlot is a publicly available database with visualization tools for exploration of alternative splicing events, including three major species of mosquitoes, Aedes aegypti, Anopheles gambiae, and Culex quinquefasciatus, and one model insect species of fruit fly Drosophila melanogaster. IsoPlot includes annotated transcripts and 17,037 newly predicted transcripts from massive transcriptome data at different life stages of insects. The web interface is interactive to explore the patterns and abundance of isoforms in different experimental conditions as well as cross-species sequence comparison of orthologous transcripts.

Availability: http://bits.iis.sinica.edu.tw/

Authors:

first name	last name	email	country	organization	corresponding?
I-Man	Ng	nim.1111.ou@gmail.com	Taiwan	Institute of Information Science, Academia Sinica	
Jia-Hsin	Huang	jhhuang@iis.sinica.edu.tw	Taiwan	Institute of Information Science, Academia Sinica	
Huai-Kuang	Tsai	hktsai@iis.sinica.edu.tw	Taiwan	Institute of Information Science, Academia Sinica	✓

RNA-Seq analysis protocol development for Komagataella phaffii

Keywords: RNA-Seq analysis, RNA-Seq, Komagataella phaffii, yeast

Abstract: For obvious reasons many tools in bioinformatics have been developed having the human genome in mind. Nonetheless there are a lot more eukaryotic organisms which are utilized in biotechnology which have a quite distinct genome composition. This fact has to be taken into consideration during protocol development and software selection. RNA sequencing data of the yeast Komagataella phaffii, a versatile host for recombinant protein production[3], was analyzed for differential gene expression using two different approaches. Due to two distinct requirements, the data was analyzed using the recent successor of the tuxedo workflow[2] – HISAT2-StringTie-Ballgown – and the rather new count-number-based workflow kallisto-DESeq2. Where the HISAT2-StringTie-workflow utilizes a genome guided assembly and can be used to identify new genes, alternative transcripts and subsequent gene expression analysis using ballgown, the kallisto-workflow[1] is transcript guided solely for differential expression analysis. To achieve the optimal protocol the parameters needed to be adjusted according the different genome composition like smaller intron sizes. Besides the protocol development, comparison and parameter optimisation, the data were analyzed for alternative transcripts.

[1] Michael I. Love et al. "RNA-Seq workflow: gene-level exploratory analysis [...]" In: F1000Research 4 (2015).
[2] Mihaela Pertea et al. "Transcript-level expression analysis of RNA-seq [...]" In: Nat Protocols 11.9 (2016).
[3] Minoska Valli et al. "Curation of the genome annotation of Pichia pastoris (Komagataella phaffii) CBS7435 [...]" In: FEMS Yeast Research 16.6 (2016)

Authors:

first name	last name	email	country	organization	corresponding?
Nadine E.	Tatto	nadine.tatto@acib.at	Austria	ACIB GmbH (Wien)	✓
Josef	Moser	josef.moser@stud.fh-campuswien.ac.at	Austria	ACIB GmbH (Wien)	
Brigitte	Gasser	brigitte.gasser@boku.ac.at	Austria	University of Natural Resources and Life Sciences, Vienna	
Diethard	Mattanovich	diethard.mattanovich@boku.ac.at	Austria	University of Natural Resources and Life Sciences, Vienna	
Alexandra B.	Graf	alexandra.graf@fh-campuswien.ac.at	Austria	FH CAMPUS Wien	✓

Discovering genetic signatures of extreme physiology using African mole-rats

Keywords: RNAseq, genomics, phylogenetics, naked mole-rat

Abstract: The African mole-rats (Bathyergidae) are a family of subterranean rodents with very unusual physiological traits for mammals. The most famous member of African mole-rats is the naked mole-rat (Heterocephalus glaber), which shows several extraordinary phenotypes like poikilothermy, extreme longevity, cancer resistance and extreme adaptation to low oxygen environments [Park, Reznick et al., Science 2017]. Additionally, the naked mole-rat and some other Bathyergidae species are insensitive to several noxious substances or algogens (e.g. acid, capsaicin, or mustard oil) [Park et al., PLOS Biology 2008].

This study focuses on understanding the sensory phenotypes of at least 8 African mole-rat species, as these closely related species show different patterns of insensitivity to noxious substances. Recently, a sequence motif in the NaV1.7 ion channel of the naked mole-rat was found to be directly connected to its acid insensitivity [Smith et al., Science 2011].

We sequenced poly-A selected mRNA from multiple tissues of 8 African mole-rat species. As there are no annotated genomes available for most of the species, we performed de-novo transcriptome assembly to obtain the protein-coding sequences. We developed a bioinformatic workflow to annotate putatively coding transcripts and exclude contaminating or falsely assembled sequences and chimeras. Using this approach, we were able to identify more than 9,000 unique protein-coding transcripts per species. We also directly compared the protein-coding sequences and transcript levels across species boundaries. Using statistical models correcting for phylogenetic relationships between species we were able to robustly identify differentially expressed genes in the species tree. Maximum likelihood methods for phylogenomics yielded insights in differences in selection pressure along the African mole-rat lineage. This approach allows a multivariate analysis of the relationship between gene expression level, sequence variation and extreme phenotypes across this rodent family.

Authors:

first name	last name	email	country	organization	corresponding?
Ole	Eigenbrod	ole.eigenbrod@mdc-berlin.de	Germany	Max-Delbrueck-Center for Molecular Medicine	✓
Jane	Reznick		Germany	Max-Delbrueck-Center for Molecular Medicine	
Karlien	Debus		Germany	Max-Delbrueck-Center for Molecular Medicine	
Damir	Damir Omerbašić		Germany	Max-Delbrueck-Center for Molecular Medicine	
Thomas J.	Park		USA	Laboratory of Integrative Neuroscience, Department of Biological Sciences, University of Illinois at Chicago	
Gary R.	Lewin		Germany	Max-Delbrueck-Center for Molecular Medicine	

BioSeq-Zip: compression of high throughput sequencing data for fast alignment computation

Keywords: Read alignment, Read mapping, Fast alignment, Read compression, read to tag

Abstract: Today, High-Throughput Next-Generation-Sequencing machines can generate millions of reads of variable length ranging from 50nt to 200nt. These reads are usually annotated for the quality assessment and aligned on the reference database/genome. During the alignment procedures, the probability to align the same sequences thousands of times is very high, thus affecting the overall execution time and resources utilisation.

To reduce alignment execution time and required hardware resources we developed BioSeq-Zip, a tool that compresses the read sequences before the alignment step and restores the expression levels after the mapping. BioSeq-Zip is composed of two modules:

1. Fastq2tagq collapses the reads with the same sequences from one or multiple samples of a dataset and assign to each of them an unambiguous label called tag and a consensus read-quality. Then, this new collection of sequences can be stored in a file and used for the alignment procedure performed with standard tools such as Bowtie, Star, TopHat, isomiR-SEA, Yara etc.

2. ReadEx recovers the expression levels of mapped reads. This module has an interface for each supported alignment tool and can be extended with customizable filtering/processing steps.

We tested BioSeq-Zip on three RNA-Seq samples (235 millions-of-reads and read length from 75nt to 150nt). We aligned both the original and the collapsed tags with Bowtie2, STAR, and TopHat2 on the Human-Transcriptome achieving a reduction in memory requirements of about 71% for Single-End and 60% for Paired-End. Alignment of collapsed tags has influenced the time elapsed by alignment tools up to 75%.

Availability: http://www.polito.it/

Authors:

first name	last name	email	country	organization	corresponding?
Gianvito	Urgese	gianvito.urgese@polito.it	Italy	Politecnico di Torino	✓
Orazio	Scicolone	orazio.scicolone@polito.it	Italy	Politecnico Di Torino	
Elisa	Ficarra	elisa.ficarra@polito.it	Italy	Politecnico di Torino	✓

Exploring Cancer-Specific Pathway Associations from NGS Data

Keywords: pathway association, RNA-seq data, pathway activity, cancer-specific signature

Abstract: The task of discovering cancer-specific genetic markers or signatures is extremely important in uncovering disease mechanisms and predicting the effect of treatments. In particular, since pathway activities and their associations are distinguishable according to the cancer type, exploring notable cancer-specific associations among the activities of pathways would be interesting.

In this study we aim to investigate the activities of pathways found in a specific type of cancer and identify their distinct associations as cancer-specific signatures. To this end, we employ RNA-seq data to define the activity level of pathways for each cancer type and use them to find pathway associations by applying association rule mining approach. Specifically, we find interesting sets of pathways frequently active together appeared in gene expression profiles of specific cancer type. In addition, we visualize pathway activities and their associations as cancer-specific signature.

Authors:

first name	last name	email	country	organization	corresponding?
Hyeonjeong	Lee	dic1224@naver.com	South Korea	Kyungpook National University	
Miyoung	Shin	shinmy@knu.ac.kr	South Korea	Kyungpook National University	✓

Independent component analysis helps classifying melanoma patients

Keywords: independent component analysis, classification, feature selection, cancer, transcriptomics

Abstract: Cancer samples investigated by high-throughput transcriptomics are often highly heterogeneous. Intra-tumour heterogeneity in conjunction with sample composition variability may lead to the observation of an unrealistic averaged combination of abundant transcripts. Lowly abundant cells are masked by such averaging. However, computational methods can help to separate mixed transcriptional signals. Here we investigated the independent component analysis (ICA) as a feature extraction method for sample classification and patient diagnostics.

The method was applied to a TCGA RNA-seq melanoma dataset. First, the stability of ICA deconvolution was improved by performing multiple runs and building a consensus signal and mixture matrices. We optimized the number of independent components by minimising the correlation between component weights. Several gene expression metrics were investigated and FPKM, which showed the most promising results, was chosen for the subsequent analysis. Importantly, unlike PCA, the ICA method resulted in both gene signatures (signals) and clinical predictors (weight coefficients). Using this information, each component was associated to a biological, technical or clinical factor. The weight coefficients showed a strong statistical linkage with clinical data and were used as input features for a support vector machine classifier. We validated the method by leave-one-out cross-validation. This resulted in a 91% accuracy for classifying melanoma subtypes (immune, keratin or MITF-low). Interestingly, several identified components were strong predictors of patient survival. Next, we used the method to successfully classify a new patient. Thus, the proposed data-driven method not only improves patient classification, but also gives prognostic information.

Availability: https://www.lih.lu/page/departments/biomod-bioinformatics-and-modelling-1364

Authors:

first name	last name	email	country	organization	corresponding?
Petr	Nazarov	petr.nazarov@lih.lu	Luxembourg	Luxembourg Institute of Health	✓
Anke	Wienecke-Baldacchino	anke.wienecke@uni.lu	Luxembourg	University of Luxembourg	
Gunnar	Dittmar	gunnar.dittmar@lih.lu	Luxembourg	Luxembourg Institute of Health	
Stephanie	Kreis	stephanie.kreis@uni.lu	Luxembourg	University of Luxembourg	
Francisco	Azuaje	francisco.azuaje@lih.lu	Luxembourg	Luxembourg Institute of Health	

QC Fail - identifying and solving NGS problems

Keywords: NGS, sequencing, artefacts, Illumina, RNA-Seq, ChIP-Seq, troubleshooting, Blog, News

Abstract: Reaching valid conclusions from DNA sequencing requires accurate data and a thorough understanding of that data, but yet we – a bioinformatics facility based at the Babraham Institute, Cambridge UK – all too often see researchers misinterpreting artefacts as genuine results and consequently presenting incorrect findings and wasting time chasing false leads.

To help researchers be aware of the key technical issues affecting their experiments we developed the website QC Fail. When we encounter a problem which may be of interest to the wider Life Sciences community we record our experiences on the QC Fail website.

Each article typically discusses how a problem was identified; whether we determined the underlying cause; what measures should be taken to ameliorate the problem and how this experience should shape the planning of future work. We shall also make available example datasets for each problem, which people can download and analyse themselves, or use in the development of processing or QC tools. In addition, these datasets should prove useful for teaching purposes.

We hope the QC Fail website is an invaluable resource to help other scientists plan, perform and analyse experiments. The stories generally focus on sequencing-related matters, but we intend to build on the site, reporting new technology trends as they develop.

To visit the website please go to: qcfail.com

Availability: http://qcfail.com

Authors:

first name	last name	email	country	organization	corresponding?
Steven	Wingett	steven.wingett@babraham.ac.uk	United Kingdom	The Babraham Institute	✓

The MGX Framework for Metagenome Analysis

Keywords: Metagenomics, Taxonomic classification, High-throughput sequencing, Metagenome, Data visualization, Workflow analysis, Statistics

Abstract: The characterization of microbial communities based on sequencing and analysis of their genetic information has become a popular approach also referred to as metagenomics; in particular, the recent advances in sequencing technologies have enabled researchers to study even the most complex communities consisting of thousands of species.

Metagenome analysis, the assignment of sequences to taxonomic and functional entities, however, remains a tedious task, as large amounts of data need to be processed.

There are a number of approaches that aim to solve this problem addressing particular aspects, however, scientific questions are often too specific to be answered by a general-purpose method.

We developed MGX, an extensible framework for the management and analysis of unassembled metagenome datasets. MGX is a client/server application providing a comprehensive set of predefined analysis pipelines including most recent tools like Kraken (Wood et al, 2014) or Centrifuge (Kim et al, 2016).

MGX allows to include own data sources and/or to devise custom analysis pipelines based on the Conveyor workflow engine (Linke et al, 2011). As all analysis tasks are executed on the server infrastructure, thus no extensive compute resources need to be provided by the user.

The intuitive and easy-to-use graphical user interface is available for all major operating systems (Windows, Linux, Mac OS X) and allows to create interactive as well as high-quality charts based on taxonomic and functional profiling results.

References

Wood DE, Salzberg SL: Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biology, 15:R46. (2014)

Kim D, Song L, Breitwieser FP, Salzberg SL: Centrifuge: rapid and sensitive classification of metagenomic sequences. Genome Research 26: 1721-1729 (2016)

Linke B, Giegerich R, and Goesmann A: Conveyor: a workflow engine for bioinformatic analyses. Bioinformatics 27(7): 903-911 (2011)

Authors:

first name	last name	email	country	organization	corresponding?
Sebastian	Jaenicke	Sebastian.Jaenicke@computational.bio.uni-giessen.de	Germany	Professorship of Systems Biology, Justus Liebig University Giessen	
Patrick	Blumenkamp	patrick.blumenkamp@computational.bio.uni-giessen.de	Germany	Professorship of Systems Biology, Justus Liebig University Giessen	✓
Stefan P.	Albaum	alu@cebitec.uni-bielefeld.de	Germany	Bioinformatics Resource Facility, Center for Biotechnology, Bielefeld University	
Burkhard	Linke	Burkhard.Linke@computational.bio.uni-giessen.de	Germany	Professorship of Systems Biology, Justus Liebig University Giessen	
Jens	Stoye	jens.stoye@uni-bielefeld.de	Germany	Faculty of Technology and Center for Biotechnology, Bielefeld University	
Alexander	Goesmann	Alexander.Goesmann@computational.bio.uni-giessen.de	Germany	Professorship of Systems Biology, Justus Liebig University Giessen	

Bayesian Hierarchical Modelling of Single-cell Methylation Profiles

Keywords: DNA methylation, single cell, epigenetics, machine learning, bayesian methods, generalized linear models (GLMs)

Abstract: New technologies enabling the measurement of DNA methylation at the single cell level are promising to revolutionise our understanding of epigenetic control of gene expression. Yet, intrinsic limitations of the technology result in very sparse coverage of CpG sites (around 20% to 40% coverage), effectively limiting the analysis repertoire to a semi-quantitative level. Here we propose a Bayesian hierarchical method to share information across cells and quantify spatially-varying methylation profiles across genomic regions from single-cell bisulfite sequencing data (scBS-seq). The method clusters individual cells based on genome-wide methylation patterns, enabling the discovery of epigenetic diversities and commonalities among individual cells. The clustering also acts as an effective regularisation method for imputation of methylation on unassayed CpG sites, enabling transfer of information between individual cells. We show that by jointly learning the posterior distribution of all parameters of interest, the proposed model is more robust and allows the sharing of information across cells to improve its imputation accuracy both on simulated and real data sets.

Availability: http://homepages.inf.ed.ac.uk/s1263191/

Authors:

first name	**last name**	**email**	**country**	**organization**	**corresponding?**
Chantriolnt-Andreas	Kapourani	C.A.Kapourani@ed.ac.uk	United Kingdom	University of Edinburgh	✓
Guido	Sanguinetti	gsanguin@inf.ed.ac.uk	United Kingdom	University of Edinburgh	✓

Ra – Rapid de novo genome assembler

Keywords: de novo genome assembly, third generation sequencing platforms, long reads

Abstract: Ra is a novel de novo genome assembler based on the Overlap-Layout-Consensus paradigm tailored for reads produced by third generation sequencing platforms. It integrates previously developed Minimap overlap (Li, 2016), Racon consensus (Vaser et al, 2017) tools and a newly developed layout module Rala into one package. Omitting time consuming error correction in the preprocessing step enables fast genome assembly while keeping high accuracy levels. The achieved results on several read datasets generated by the Pacific Biosciences sequencing platforms are comparable with those of similar de novo assemblers Hinge (Kamath et al, 2017) and Miniasm+Racon in contiguity, accuracy and running time.

Authors:

first name	last name	email	country	organization	corresponding?
Robert	Vaser	robert.vaser@fer.hr	Croatia	University of Zagreb, Faculty of Electrical Engineering and Computing	
Mile	Sikic	mile.sikic@fer.hr	Croatia	University of Zagreb, Faculty of Electrical Engineering and Computing	✓

Rala - Rapid layout module for de novo genome assembly

Keywords: Long reads, Overlap-Layout-Consensus, Layout module, de novo assembly, third generation sequencing

Abstract: Rala is a standalone layout module intended for assembly of raw reads generated by third generation sequencing platforms. It consists of two parts, read preprocessing inspired by HINGE (Kamath et al, 2017), and assembly graph simplification as described in Miniasm (Li, 2016). In preprocessing, coverage graphs are generated from pairwise mappings using Minimap (Li, 2016) and are used to detect chimeric reads as well as reads from repetitive regions. Afterwards, the assembly graph is simplified with transitive reduction, trimming, bubble popping and a heuristic which untangles leftover junctions in the graph. As a side result, we show that the percentage of chimeric reads produced by either the Pacific Biosciences or Oxford Nanopore Technologies platforms is correlated with the read length.

Authors:

first name	last name	email	country	organization	corresponding?
Robert	Vaser	robert.vaser@fer.hr	Croatia	University of Zagreb, Faculty of Electrical Engineering and Computing	
Mile	Sikic	mile.sikic@fer.hr	Croatia	University of Zagreb, Faculty of Electrical Engineering and Computing	✓

Transcriptome-wide identification of ceRNAs in tumor-cell migration and invasion

Keywords: microRNA, long non-coding RNA, circular RNA, competing endogenous RNA, tumor-cell migration

Abstract: Cancer metastasis is a series of stages that drives the movement of tumor cells to a distant location and is the main cause of mortality and morbidity of cancer patients. Despite plenty of remarkable advances in understanding causes and treatments of cancer during past few decades, the molecular mechanisms underlying the invasion and metastasis of cancer cells still remain unclear. On the other hand, many non-coding RNAs were found playing important roles in a diversity of biological processes with recently great advances in high-throughput sequencing technologies, and dysregulation of these non-coding RNAs may cause many acute diseases and cancers. Lots of non-coding RNAs involving in tumor invasion were also identified, such as many long non-coding RNAs (lncRNAs) found in promotion of cancer metastasis. Therefore, a more comprehensively transcriptomic regulatory mechanism of tumour-cell invasion and migration is requred, especially the competing endogenous RNA (ceRNA) network which composed of these mRNAs and non-conding RNAs. In this work, we have performed expression data analysis on microarray and RNA-seq datasets of breast cancer, and integrated multiple -omics data (including PPI, TF-gene, microRNA-mRNA, microRNA-lncRNA, microRNA-circRNA interaction data, etc.) to identify the key regulators and targets during cancer migration. Combining the results from previous step and regulator-target pair informations, we constructed a multi-level regulatory network. Finally, applying network analysis and functional analysis to the network, we've identified a module which is most relevant to cancer migration and invasion, which would be helpful for the prevention and treatment of metastatic breast cancer.

Authors:

first name	last name	email	country	organization	corresponding?
Chia-Chun	Chiu	jiajyun.ciou@gmail.com	Taiwan	LightUp Biotech Co., Ltd.	✓
Yi-Fan	Liou	yifan@gmail.com	Taiwan	LightUp Biotech Co., Ltd.	
Chun-Hui	Yu	itsjeffreyy76@gmail.com	Taiwan	LightUp Biotech Co., Ltd.	
Hsiao-Han	Lin	sallyluckyday@gmail.com	Taiwan	National Cheng Kung University	
Hung-Teng	Liang	htl519@gmail.com	Taiwan	LightUp Biotech Co., Ltd.	✓

A Sequence Simulation Based Approach to Evaluating Metagenomic Sequence Assignment Accuracy

Keywords: metagenomics, accuracy, pipeline evaluation, next-generation sequencing, 16S rRNA gene, taxonomic assignment

Abstract: The revolution in next-generation sequencing (NGS) technologies has enabled a step-change in the way that sequence data is collected and used in Biology including in metagenomics, the sequencing of mixed source nucleic acid samples. These studies have profound implications for human, animal and plant health and disease as well as in diverse areas such as forensic science, environmental pollution monitoring and climate modelling. The increasing quantity of metagenomic sequence data being generated and the diversity of its application areas requires highly optimised and computationally scalable solutions to process and interpret these data. Yet, there is no standardised way of evaluating the accuracy of methods that assign these sequences to taxonomies.

We present a comparative evaluation of metagenomic analysis methods in which we use sequence simulators to generate gold-standard data against which to benchmark the efficacy of the methods. We use our method to develop an approach to estimate errors in taxonomic sequence assignment by perturbing the underlying taxonomic trees used in our simulations. Our method demonstrates the high dependency of taxonomic classification success and accuracy on the information present in the reference database and the methods used for classification. We also present an evaluation of the relative importance of different regions of the 16S rRNA marker gene in taxonomic assignment for meta-genetic studies.

Authors:

first name	**last name**	**email**	**country**	**organization**	**corresponding?**
Alba	Crespi-Boixader	alba.crespi@ed.ac.uk	United Kingdom	School of Informatics, University of Edinburgh	✓
Alex	Mitchell	mitchell@ebi.ac.uk	United Kingdom	European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI)	
Robert	D Finn	rdf@ebi.ac.uk	United Kingdom	European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI)	
David	Longbotom	david.longbottom@moredun.ac.uk	United Kingdom	Moredun Research Institute	
Ian	Simpson	ian.simpson@ed.ac.uk	United Kingdom	Biomathematics & Statistics Scotland	

Gene model complexity evolution

Keywords: gene model, evolution, alternative splicing, complexity metrics

Abstract: Genome-scale expression profiling has become a key tool of functional genomics, critically supporting progress in the post-genomic era. It improves our understanding of living systems at molecular level. The fast development of sequencing technologies have recently led to many updates of genome sequences as well as annotations and has revealed the complexity of the gene models of many species.

In this work, we have analysed extensively human gene model evolution with focusing on alternative splicing events (ASE). In addition to the well defined canonical ones, we have focus on ones which are more complex and do not fit to any canonical category. These in the latest EnsEMBL releases made over 40% of all ASE. We here define a 4 new ASE categories which encounter for about 2/3 of all complex 'ASE'. The remaining 1/3 seems to be a combination of already known and these new 4 ASE types.

In our future work we would like to investigate possible evolutionary origins of these 'complex' ASE. Based on the detailed analysis of the gene model complexity evolution and appearance of alternative splicing events, we would like to take advantage of this knowledge and incorporate it into newly introduced combined metric to assess gene model complexity. Motivation is that the currently available metrics are of limited use as they describe/assess just part of the gene model and they do not reflect the evolutionary sources of the gene model complexity.

Authors:

first name	last name	email	country	organization	corresponding?
Ognyan	Kulev	okulev@fmi.uni-sofia.bg	Bulgaria	Faculty of Mathematics and Informatics, Sofia University "St. Kliment Ohridski"	✓
Milko	Krachunov	milko@3mhz.net	Bulgaria	Faculty of Mathematics and Informatics, Sofia University "St. Kliment Ohridski"	✓
Dimitar	Vassilev	jim6329@gmail.com	Bulgaria	Faculty of Mathematics and Informatics, Sofia University "St. Kliment Ohridski"	✓
Paweł P.	Łabaj	pawel.labaj@boku.ac.at	Austria	Boku University Vienna	✓

TaxMapper: An Analysis Tool, Reference Database and Workflow for Metatranscriptome Analysis of Eukaryotic Microorganisms

Keywords: metatranscriptome, microeukaryotes, community composition, workflow

Abstract: Next generation sequencing (NGS) technologies are increasingly applied to analyse complex microbial ecosystems by mRNA sequencing of whole communities, also known as metatranscriptome sequencing. This approach is currently limited to prokaryotic communities and communities of few eukaryotic species with sequenced genomes. For eukaryotes the analysis is hindered mainly due to inappropriate reference databases to infer the community composition.

In this study, we focus on the development of a tool (TaxMapper) for a reliable mapping to a microeukaryotic reference database and a comprehensive analysis workflow. We focus on the assignment of higher taxonomic groups and therefore collected publicly available genomic and transcriptomic sequences from the databases of NCBI, Marine Microbial Eukaryote Transcriptome Sequencing Project and JGI. 143 references were selected such that the taxa represent the main lineages within each of the seven supergroups of Eukaryotes and possess predominantly complete transcriptomes or genomes. TaxMapper is used to assign taxonomic information to each NGS read by mapping to the database and filtering low quality assignment. Therefore, a logit classifier was trained and tested on sequences in the database, sequences of related taxa to those in the database and randomly generated reads. TaxMapper is part of a metatranscriptome Snakemake workflow developed to perform quality assessment, functional and taxonomic annotation and (multivariate) statistical analysis including environmental data. The workflow is provided and described in detail to empower researchers to easily apply it for metatranscriptome analysis of any environmental sample.

Authors:

first name	last name	email	country	organization	corresponding?
Daniela	Beisser	Daniela.Beisser@uni-due.de	Germany	Biodiversity, University of Duisburg-Essen	✓
Nadine	Graupner	Nadine.Graupner@uni-due.de	Germany	Biodiversity, University of Duisburg-Essen	
Jens	Boenigk	Jens.Boenigk@uni-due.de	Germany	Biodiversity, University of Duisburg-Essen	
Sven	Rahmann	Sven.Rahmann@uni-due.de	Germany	Genome Informatics, University of Duisburg-Essen	

gargammel: simulating Illumina reads from ancient genomes

Keywords: ancient DNA, sequence simulation, sequencing, computer science

Abstract: Sequencing simulators prove to be a useful procedure to test the accuracy of algorithms and their robustness to sequencing errors. Despite being an exceptional tool to infer the history of past populations, ancient DNA data are characterized by a series of idiosyncrasies such as extensive fragmentation, damage and contamination, all of which can influence downstream analyses. We present gargammel, a package to simulate sequencing reads from a set of user-provided reference genomes. This package simulates the entire molecular process from post-mortem DNA fragmentation, DNA damage, experimental sequencing errors, GC-bias as well as potential bacterial and present-day human contamination. We present two case studies to illustrate the capabilities of our software and how it can be used to assess the validity of specific read alignment procedures and inference of past population histories. First, we evaluate the impact of present-day human contamination on admixture analyses for hominin species. Second, we present the impact of microbial contamination on ancient DNA alignments to the human reference genome. The package is publicly available on github: https://grenaud.github.io/gargammel/ and released under the GPL.

Availability: http://grenaud.github.io/

Authors:

first name	last name	email	country	organization	corresponding?
Gabriel	Renaud	gabriel.reno@gmail.com	Denmark	University of Copenhagen	✓
Kristian	Hanghoej	kristianhanghoej@gmail.com	Denmark	University of Copenhagen	
Eske	Willerslev	ewillerslev@snm.ku.dk	Denmark	University of Copenhagen	
Ludovic	Orlando	orlando.ludovic@gmail.com	Denmark	University of Copenhagen	

Locating CNV candidates in WGS data using wavelet-compressed Bayesian HMM

Keywords: HMM, CNV, wavelet, NGS

Abstract: The avalanche of NGS data and the growing demand for Bayesian methods pose huge algorithmic challenges in the case of whole-genome CNV inference. At the same time, fast, accurate and efficient computation is crucial in both clinical and fundamental research settings. Recently, a method based on dynamic wavelet compression was shown to drastically improve computation of full latent state marginals of Bayesian HMM in terms of speed and convergence behavior, but handling the memory requirements due to the sheer size of the input remained challenging. We present a new data structure to alleviate this problem, and demonstrate Bayesian CNV inference on a laptop within minutes using WGS data from divergently selected rat populations.

Availability: https://schlieplab.org/People/JohnWiedenhoeft/

Authors:

first name	last name	email	country	organization	corresponding?
John	Wiedenhoeft	john@wiedenhoeft.biz	Sweden	Chalmers University of Technology	✓
Alex	Cagan		United Kingdom	Wellcome Trust Sanger Institute	
Rimma	Kozhemjakina		Russia	Institute of Cytology and Genetics of the Russian Academy of Sciences	
Rimma	Gulevich		Russia	Institute of Cytology and Genetics of the Russian Academy of Sciences	
Alexander	Schliep		Sweden	Chalmers University of Technology	

Improving long-read mapping with simple lossy sequence transformations

Keywords: Long-read sequencing, Read mapping, Sequence transformation, Indel errors, k-mer counting

Abstract: Long-read sequencing technologies from Pacific Biosciences and Oxford Nanopore have dramatically increased achievable read length, with reads routinely exceeding 10 Kbp. These reads are crucial for resolving ambiguities when mapping and assembling reads from repetitive genomes such as human, with consequences for many applications, from closing gaps of the reference to mapping the structural variation that underlies many human diseases. However, the increased read length comes at the cost of a significantly higher error rate. Despite continuing improvements, read mappers and assemblers designed for short-read technologies such as Illumina, where indels are almost nonexistent, struggle to map these reads accurately or at all.

We describe six very simple ”squash” transformation functions that can be applied to any DNA sequence to produce a smaller sequence, on average one quarter of the input size, and which have a useful “indel-tolerant” property: on average, 75% of deletions and 62.5% of insertions leave the result of the transformation unchanged. The transformations are fast, streamable and in-place. When applied as preprocessing before k-mer lookup, they can be viewed as a form of gapped k-mers, but with sequence-dependent gaps. We show that two of the functions significantly improve the accuracy of the initial k-mer lookup phase of read mapping: for simulated PacBio data, transforming both reference sequence and reads before mapping k-mers yields relative support scores higher than baseline for at least 76% of the reads, suggesting that these transformations have the potential to simultaneously improve both the speed and accuracy of long read mapping.

Authors:

first name	last name	email	country	organization	corresponding?
W. Timothy J.	White	tim.white@bihealth.de	Germany	Berlin Institute of Health	✓
Birte	Kehr	birte.kehr@bihealth.de	Germany	Berlin Institute of Health	✓

Interrogating the Human Antibody Repertoire

Keywords: Next Generation Sequencing, Antibody, Motifs, Human antibody repertoire

Abstract: Antibodies are proteins of the immune system that tag noxious molecules for elimination. They can be adjusted to bind with high affinity and specificity to a target molecule. This property has been extensively exploited in biopharmaceuticals, diagnostics and research agents. Their binding malleability arises from their diversity ($> 10^{10}$ possible sequences). The advent of Next Generation Sequencing (NGS) has made it possible to produce snapshots of this diversity. In this poster we describe our work with a large NGS dataset comprising 13.5m heavy and light chains from $\sim$ 500 individuals. By studying this dataset we aim to establish a set of descriptors that will allow us to formally interrogate the properties of immune repertoires. The descriptors we have explored include length of complementarity determining region (CDRs), gene usages, amino acid distributions. One feature we have identified within the data is that 6-7% of H3 loops in our dataset contain a cysteine pair motif. H3 length correlates with the proportion of cysteine residues inside the loop (Pearson correlation, R2 = 0.89). Two of the most common motifs containing two surrounding cysteines are 5 and 6 amino acids long. Our analysis of amino acid distribution reveals that both motifs have a strong preference for tyrosines in the flanking positions, although the amino acid distributions inside each motif is distinct. This knowledge is one example of how NGS can be used to rationally design antibodies in novel ways.

Authors:

first name	last name	email	country	organization	corresponding?
Aleksandr	Kovaltsuk	aleksandr.kovaltsuk@stats.ox.ac.uk	United Kingdom	University of Oxford	✓
Konrad	Krawczyk	konrad.krawczyk@stats.ox.ac.uk	United Kingdom	University of Oxford	
Sebastian	Kelm		United Kingdom	UCB Pharma	
James	Snowden		United Kingdom	UCB Pharma	
Charlotte	Deane	deane@stats.ox.ac.uk	United Kingdom	University of Oxford	✓

DeepWAS: Directly integrating regulatory information into GWAS using deep learning

Keywords: GWAS, functional data, chromatin features, cell-type specificity, ENCODE, deep learning, multiple regression

Abstract: Genome-wide association studies (GWAS) have been highly successful in identifying genetic variants associated with risk for common diseases. The majority of the phenotype-associated SNPs identified in GWA studies are in non-coding, regulatory regions. Despite this fact, the existing functional studies fall short in exploiting the regulatory impact of variants since most methods either do not go beyond positional overlap of annotated regulatory regions and associated variants or they integrate other types of molecular readouts such as eQTL or tfQTL. While the overlapping approaches cannot assess the actual impact of the variant on regulatory elements, the integration methods need additional data which is not always available. We here describe DeepWAS, a new approach where the phenotype-genotype link is interrogated in a cell line and transcription factor specific manner via multilocus regression models using the regulatory features of variants predicted by the deep learning method DeepSEA. DeepWAS, a method combining classical GWAS with deep learning-based functional variant annotation, has a potential as a powerful tool to uncover disease mechanisms for common disorders, including relevant cell types.

Availability: http://icb.helmholtz-muenchen.de

Authors:

first name	last name	email	country	organization	corresponding?
Gökcen	Eraslan	goekcen.eraslan@helmholtz-muenchen.de	Germany	Helmholtz Zentrum München	
Janine	Arloth	janine.arloth@helmholtz-muenchen.de	Germany	Helmholtz Zentrum München	
Ivan	Kondofersky	ivan.kondofersky@helmholtz-muenchen.de	Germany	Helmholtz Zentrum München	
Fabian J.	Theis	fabian.theis@helmholtz-muenchen.de	Germany	Helmholtz Zentrum München	
Nikola S.	Mueller	nikola.mueller@helmholtz-muenchen.de	Germany	Helmholtz Zentrum München	✓

RNA-Seq read count correction by intron bias estimation

Keywords: RNA-Seq, Intron bias, Correction

Abstract: In RNA-Seq experiments, up to 20 percent of reads mapping to intronic regions can be observed. There are numerous hypotheses on the true origin of these reads. Whether reads mapping to non-coding regions should be considered noise and be ignored as a consequence is a subject of recent debate as RNA-Seq aims at estimating protein abundance.

This study was conducted to determine whether the incorporation of reads mapping to non-coding regions into RNA-Seq transcript abundance estimation improves accuracy of differential expression analysis compared to the standard pipelines.

To adjust transcript counts for so-called non-coding reads, we estimate for each transcript the abundance of non-coding RNA fragments and subtract these from the original transcript abundance. To estimate the reduction rate we test a brute-force method and a binning approach that corrects for 5' to 3' read distribution bias. We determine accuracy of results obtained in differential gene expression analysis using corrected read counts for synthetic data sets mimicking well-known RNA-Seq biases.

Correcting for RNA-Seq reads possibly originating from non-coding elements improves reproducibility between replicates. Accuracy of differential gene expression analysis was significantly improved in synthetic data sets compared to results obtained with HTSeq read counts. Testing our approach on a freely available experimental data set, we could increase the number of detected differentially expressed genes. An R-package will be available soon.

Authors:

first name	last name	email	country	organization	corresponding?
Valentina	Klaus	valentina.klaus@helmholtz-muenchen.de	Germany	Helmholtz Center Munich / Technical University Munich	✓
Dominik	Lutter	dominik.lutter@helmholtz-muenchen.de	Germany	Helmholtz Center Munich	

A robust LASSO regression approach for mutational signature identification in cancer genomics

Keywords: Mutational signature, Mutational processes, LASSO, Cancer genomics

Abstract: Multiple endogenous and exogenous mutational processes drive cancer mutagenesis and leave distinct fingerprints. Noticeably, they have inherent mutational nucleotide context biases. Mutation profiling of cancer sample finds all mutations accumulate over the lifetime, including somatic alterations both before the cancer initiation and during cancer development. In a generative model, over-time multiple latent processes produce mutations, drawing from their corresponding nucleotide context distributions (the "mutational signature"). In cancer sample, mutations from various mutation processes are mixed and observable by sequencing.

Many mutation processes are recognized and linked with known etiologies. Understanding the fundamental underlying processes helps understand cancer initiation and development. A key issue in the field is to detect operative signatures in new cancer samples by leveraging current known signatures derived from large-scale pan-cancer analyses.

Previously published methods use empirical forward selection or iterate all combinations (brute force). Here, we formulate this as a LASSO linear regression problem. By parsimoniously assigning signatures to cancer genome mutation profiles, the solution becomes sparse and biologically interpretable. Additionally, LASSO organically integrates biological priors into the solution by fine-tuning penalties on coefficients. Compared with the current approach of subseting signatures in fitting, our method leaves leeway for noise and allows promoting similarity within sample subgroups, leading to a more reliable and interpretable signature solution. Last, our method can be automatically parameterized based on cross-validation. This objective, robust approach promotes data replicability and fair comparison across research.

Availability: https://scholar.google.com/citations?user=GHyTQc8AAAAJ&amp;hl=en

Authors:

first name	last name	email	country	organization	corresponding?
Shantao	Li	shantao.li@yale.edu	USA	Yale University	✓
Mark	Gerstein	mark@gersteinlab.org	USA	Yale University	✓

The cellular proteome may be less complex than we think

Keywords: Alternative splicing, Proteomics, Variation

Abstract: Alternative splicing is well documented at the transcript level, but reliable large-scale proteomics experiments detect many fewer alternative isoforms than expected. Instead proteomics evidence suggests that the vast majority of coding genes have a single dominant splice isoform, irrespective of cell type. Where a main proteomics isoform can be determined there is almost perfect agreement with two orthogonal sources of reference isoforms, principal isoforms from the APPRIS database and unique CCDS variants, based on the conservation of protein structure and function and cDNA evidence respectively.

When alternative isoforms are detected in proteomics experiments they tend to be highly conserved and are enriched in subtle splice events such as mutually exclusively spliced homologous exons and tiny indels. Only a small fraction of proteomics-supported alternative events disrupt protein functional domain composition. Two thirds of annotated alternative transcripts would disrupt functional domains.

Many annotated alternative splice transcripts have little cross-species conservation. However, it has been suggested that these alternative variants may play an important role in evolutionary innovation. We have analysed the results of human population variation studies and find that this is not the case. Indeed most alternative exons appear to be evolving neutrally in present-day human populations.

While a small number of annotated alternative variants are conserved across species and are translated in detectable quantities, most are evolving neutrally. This strongly suggests that most alternative variants will not generate functionally relevant proteins.

Availability: http://www.cnio.es

Authors:

first name	last name	email	country	organization	corresponding?
Michael	Tress	mtress@cnio.es	Spain	Spanish National Cancer Research Centre	✓
Federico	Abascal	fa8@sanger.ac.uk	United Kingdom	Wellcome Trust Sanger Institute	
Alfonso	Valencia	avalencia@cnio.es	Spain	Spanish National Cancer Research Centre	

SNP discovery in Leishmania infantum transcriptome using RNA sequencing approach

Keywords: RNA sequencing, Transcriptome, Single Nucleotide Polymorphisms, Leishmania infantum, Trivalent antimony, Drug resistance

Abstract: Leishmaniasis refers to a disease complex caused by protozoan parasites of the genus Leishmania. Annually, approximately 300,000 new cases and 20,000 deaths related to visceral leishmaniasis are reported. The treatment of leishmaniasis is problematic due mainly to the high toxicity of pentavalent antimonials and the emergence of parasites resistant to these compounds. In order to investigate if the presence of single nucleotide polymorphisms (SNPs) could be associated with antimony resistance mechanisms, we used transcriptome data obtained by NGS Illumina RNA sequencing from susceptible (LiWTS) and trivalent antimony (SbIII) resistant (LiSbR) L. infantum (MHOM/BR/74/PP75) lines. Considering the availability of Leishmania genomic data and the high synteny observed between all sequenced genomes, for an initial assessment of SNPs, Burrows-Wheeler Aligner (BWA) was used for mapping the reads against the reference genome, L. infantum JPCM5. SAMtools and BCFtools were used for SNP calling and SnpEFF was used for variants identification. In addition, functional annotation was performed using Blast2GO software. The pipeline applied in the analysis process allowed the identification of variant rate of one variant every 3,532 bases in the LiWTS and one variant every 35,716 bases in LiSbR, most of them resulting in missense effects. The functional effect of the polymorphism, variants rate by chromosome and Indels were addressed. For the LiSbR line, the SNPs with high impact are related to proteins having domains of unknown function (DUF proteins), amastin and cysteine peptidase, which play important roles in survival and virulence of the parasites.

Authors:

first name	last name	email	country	organization	corresponding?
Leilane	Gonçalves	leilane1002@hotmail.com	Brazil	Instituto Oswaldo Cruz	✓
Frederico	Guimarães	frederico.guimaraes@cpqrr.fiocruz.br	Brazil	Centro de Pesquisas René Rachou	
Paul	Guimarães	contato@paulanderson.com	Brazil	Centro de Pesquisas René Rachou	
Henrique	Toledo	henrique.toledo@cpqrr.fiocruz.br	Brazil	Centro de Pesquisas René Rachou	
Daniela	Resende	dani.melo.resende@gmail.com	Brazil	Instituto Oswaldo Cruz	
Silvane	Murta	silvane@cpqrr.fiocruz.br	Brazil	Centro de Pesquisas René Rachou	
Jeronimo	Ruiz	jeronimo@cpqrr.fiocruz.br	Brazil	Centro de Pesquisas René Rachou	✓

QDNAseqFLOW: A Computational Analysis Workflow of DNA Copy Number Aberrations from low-coverage whole genome sequencing reads

Keywords: low-coverage whole genome sequencing, DNA copy number, aberrations, cancer, R programming, Bioconductor

Abstract: BACKGROUND: Gains and losses of genetic material, also known as DNA copy number alterations are aberrations that are involved in the development of cancer. Their analysis is therefore critical for research and diagnostics in oncology. DNA sequencing based determination of copy number aberrations is becoming the most cost effective way as compared to microarray based techniques at equal resolution. To obtain copy number calls from low coverage whole genome sequencing reads requires the combined usage of several programs with various steps, followed by more statistical analysis tools for pairwise comparisons, etc. A complete workflow would therefore be useful for many other researchers in the field.

RESULTS: Here we present QDNAseqFLOW, a computational workflow that produces DNA copy number plots along with various summaries and statistics, including the aberration differences found between groups of input samples. Written in the R programming language, it relies on Bioconductor packages QDNAseq, DNAcopy, CGHcall and CGHregions as well as the open-source R packages NoWaves and CGHtest, all of them described in peer-reviewed journal articles.

USAGE: The program is written in the R programming language and can be run without programming skills on Windows, MacOSX and Linux through provided wrapper scripts. The user is guided by simple graphical pop-ups to enter parameters or select file locations, while access to the program code allows users with R programming skills to change advanced paramters. FEATURES and WORKFLOW: (1) Reads obtained from low-coverage (= “shallow”) whole genome sequencing of DNA samples need to be provided as BAM files obtained by alignment to the human reference genome hg19. (2) Copy number plots and -files are created using Bioconductor package QDNAseq. (3) ‘Waves’ in the profiles are smoothed with the R package NoWaves (van de Wiel et al., 2009) and subsequently, aberrated regions are combined with the circular binary segmentation (CBS) algorithm implemented in Bioconductor package DNAcopy and the copy numbers of obtained segments are called using Bioconductor package CGHcall. (3) Summarizing frequency plots and quality statistics for all plots are created. Plots are flagged if their noise and/or number of segments is higher than expected, based on the inter-quartile range of values observed for all samples, and can then be checked and removed by the user from subsequent analysis. (4) If the user provides a grouping for the samples, individual frequency plots, aberration summaries (per chromosome arm) and a differential aberration analysis will be produced. To obtain the latter, Bioconductor package CGHregions is used to slightly adjust the segments in all samples in a way to obtain regions with start and end positions identical in all samples with minimal information loss. Then, with the help of R package CGHtest (van de Wiel et al., 2005), a Wilcoxon-Mann-Whitney two-sample test or Kruskal-Wallis k-sample test is applied to all aberrated regions to calculate which aberration is significantly different between the groups.

CONCLUSIONS: QDNAseqFLOW is a comprehensive workflow for the analysis of copy number aberrations. It will be made available at github.com/NKI-Pathology.

Authors:

first name	**last name**	**email**	**country**	**organization**	**corresponding?**
Christian	Rausch	christian.rausch@gmail.com	Netherlands	Netherlands Cancer Institute	✓
Beatriz	Carvalho	b.carvalho@nki.nl	Netherlands	Netherlands Cancer Institute	
Remond	Fijneman	r.fijneman@nki.nl	Netherlands	Netherlands Cancer Institute	
Gerrit	Meijer	g.meijer@nki.nl	Netherlands	Netherlands Cancer Institute	
Mark	van de Wiel	m.vdwiel@vumc.nl	Netherlands	VU Medical Center and Vrije Universiteit Amsterdam	

Fuzzy clustering and machine learning for high-variation high-error datasets

Keywords: next generation sequencing, metagenomics, data analysis, error detection, machine learning

Abstract: Itroduction

High variation in sequencing data—as present in metagenomics and polyploidy genome analysis—poses many difficulties during the data processing and analysis. Error detection is particularly impeded, as erroneously read bases are hard to discern among correct bases of varying quantities. The problem becomes particularly significant when using modern sequencing technologies like Oxford Nanopore that offer low-cost sequencing with very high error rates.

Rationale and methods

Our work has been focused on the development of an aggregate error detection approach to take the variation into account. It uses the conjunctive summary of two predictors—an analytic error predictor that creates per-position fuzzy clusters of similar sequences, and a machine learning (ML)-based model that is trained to discern errors from the naturally occurring bases. Variants of the approach have been tested on metagenomics and hexaploid wheat.

Deliverables and conclusion

During our study, the aggregate approach showed very promising results particularly on low quality datasets where the error rates were significant. The different ML-based models alone had a precision and recall of over 99.5% on metagenomics, and even higher on wheat. Because of the Bayes rule, the accuracy on metagenomics was insufficient for improving the average already low error rates of Illumina or 454, however it still leaves the approach highly applicable for use in Oxford Nanopore datasets.

Work is ongoing on applying this hybrid approach on metagenomics datasets sequenced using technologies such as Oxford Nanopore, and demonstrating that the consistently high accuracy of the ML model persists.

Authors:

first name	last name	email	country	organization	corresponding?
Milko	Krachunov	mie2013@milko.3mhz.net	Bulgaria	Faculty of Mathematics and Informatics, Sofia University ”St. Kliment Ohridski”	
Pawel	Labaj	pawel.labaj@boku.ac.at	Austria	Boku University Vienna, Austria	
Ognyan	Kulev	okulev@fmi.uni-sofia.bg	Bulgaria	Faculty of Mathematics and Informatics, Sofia University ”St. Kliment Ohridski”	
Dimitar	Vassilev	jim6329@gmail.com	Bulgaria	FMI, Sofia University ”St. Kliment Ohridski”	✓

A custom bioinformatics analysis pipeline for the stratification of admixed populations and prioritisation of rare single nucleotide variants

Keywords: Personalised medicine, Admixed population, PCA, Exome Sequencing

Abstract: Exome sequencing has become an emerging technique in the identification of rare single nucleotide variants (SNVs) that harbor deleterious impact and possible health risk. However, identification of these variants is dependent on minor alleles frequencies (MAF) that are derived from large sequencing projects of known populations. The stratification of individuals within admixed populations is a challenge in personalized medicine. In particular, the lack of knowledge regarding the patient's ethnic background may interfere with the ability to identify bona fide rare SNVs that could be associated with the disease. In this study, we utilized a customized principle component analysis (PCA) tool in order to classify a cohort of whole-exomes data derived from cases originating from the western region of the Kingdom of Saudi Arabia; a region known to be inhabited by an admixed and poorly genetically characterized population. To this end, we performed whole-exome sequencing on thirty individuals from the western region of Saudi Arabia on the SOLiD 5500 XL platform. The generated variant call format files (vcf) were filtered for quality to include the SNVs with more than 10x converge and MQV of ¿=20. Exome data of 5700 individuals from 7 populations of 1000 genome project were downloaded and queried using tabix and VariantAnnoation Bioconductor package. Custom-written codes in R along with SNPRelate package were used to merge the downloaded data with the in-house cohort to plot the PCA from the shared SNVs. The relative relationships between Arabian individuals to the nearby population cluster were measured in two-dimensional plot of the two first principle components. Our approach showed that the proposed tool is promising to prioritise rare SNVs of admixed-ethnicity individuals and hence guiding to use the most relevant MAF values in filtering these variants.

Authors:

first name	last name	email	country	organization	corresponding?
Nada	Salem	nfsalem@kau.edu.sa	Saudi Arabia	King Fahd Medical Research Center, King Abdulaziz University	✓
Ashraf	Dallol		Saudi Arabia	Center of Innovation in Personalised Medicine - King Abdulaziz University	
Adel	Abuzenadeh		Saudi Arabia	Center of Innovation in Personalised Medicine - King Abdulaziz University	
Emmanouil	Dermitzakis		Switzerland	Department of Genetic Medicine and Development - University of Geneva	

Pathway Discordance Analysis Reveals Pseudo-Temporal Dysregulation of the Erythroid Stem Cell Lineage in Acute Erythroid Leukemia

Keywords: Single-cell, Pseudo-temporal, Pathway Analysis

Abstract: Single-cell sequencing technology is rapidly improving the resolution at which cellular heterogeneity in complex tissues is studied, particularly in stem cell biology where rare intermediate cell types are often difficult to capture or isolate.

We have developed a computational method which leverages high-throughput single-cell RNA-seq data to asses which biological pathways undergo dysregulation in the context of neoplastic development and where in differentiation maximum dysregulation occurs.

Our method is built upon our previous computational strategy for aligning single cells along a developmental or pseudo-temporal axis to estimate and describe changes in gene expression as tissues differentiate and mature.

We have leveraged publicly available data from 10X Genomics and the Pathway Commons to estimate pathway activity changes during differentiation of the erythroid lineage in bone marrow captured from two healthy patients and a patient with Acute Erythroid Leukemia.

This Pathway Discordance Analysis has identified biologically discordant pathways previously implicated in myeloproliferative disorders by the scientific literature as well as less-studied pathways which may represent new opportunities for future research.

Authors:

first name	last name	email	country	organization	corresponding?
Geoffrey	Schau	schau@ohsu.edu	USA	Oregon Health and Science University	✓
Devorah	Goldman		USA	Oregon Health and Science University	
Michael	Mooney		USA	Oregon Health and Science University	
Guanming	Wu		USA	Oregon Health and Science University	
Andrew	Adey		USA	Oregon Health and Science University	
William	Fleming		USA	Oregon Health and Science University	

Integrated Theory- and Data-driven Dimensionality Reduction in High-throughput Gene Expression Data Analysis

Keywords: Causal modeling, Latent variables, Graphical models

Abstract: The exponential growth of high dimensional biological data has led to a rapidly increase in demand for automated approaches for knowledge production. Previous studies rely on two major approaches to address this type of challenge, 1) the Theorydriven approach, and 2) the Data-driven approach. The former

constructed future knowledge based on the prior background knowledge that is acquired and the latter formulates scientific knowledge solely based on analyzing the data that is obtained previously. In this work, we argue that using either approach alone suffers from the bias towards past/present knowledge as

they fail to incorporate all of the current knowledge that is available for knowledge production. To address such challenge, we propose a novel two-step analytical workflow that incorporates a new dimensionality reduction paradigm as the first step to handling high-throughput gene expression data analysis and utilizes graphical causal modeling as the second step to handle the automatic extraction of causal relationships. Our results, on real world clinical datasets from The Cancer Genome Atlas (TCGA), show that our approach is capable of wisely selecting genes for learning effective causal networks.

Availability: https://panos.cs.pitt.edu/

Authors:

first name	**last name**	**email**	**country**	**organization**	**corresponding?**
Vineet	Raghu	vineetraghu@gmail.com	USA	University of Pittsburgh	
Xiaoyu	Ge	XIG34@pitt.edu	USA	University of Pittsburgh	
Panos K	Chrysanthis	panos@cs.pitt.edu	USA	University of Pittsburgh	✓
Panayiotis V	Benos	benos@pitt.edu	USA	University of Pittsburgh	✓

Recursive Cluster Elimination (RCE) Based on Ensemble Clustering for Gene Expression Data

Keywords: Gene Expression, Significant Genes, Recursive Cluster Elimination, Ensemble Clustering f

Abstract: Abstract

Background

Advances in technology that resulted in lowering the cost for generating gene expression data from large numbers of samples has led to the development of “Big Data” approaches to analyzing gene expression in basic and biomedical systems. That being said, the data still includes relatively small numbers of samples and tens of thousands of variables/genes. Different techniques have been proposed for searching these gene spaces in order to select the most informative genes that can accurately distinguish one class of subjects/samples from another. We now describe a new approach for selecting those significant clusters of genes using recursive cluster elimination (RCE) based on an ensemble clustering approach called Support Vector Machine RCE-Ensemble Clustering (SVM-RCE-EC) that improves on the traditional SVM-RCE approach. We present our results comparing the performance of SVM-RCE-EC with different methods applied to the same datasets.

Results

SVM-RCE-EC uses an ensemble-clustering method, to identify clusters that are robust. Support Vector Machines (SVMs), with cross validation is first applied to score (rank) those clusters of genes by their contributions to classification accuracy. Recursive cluster elimination (RCE) is then applied to iteratively remove the gene clusters that contribute the least to the classification performance. SVM-RCE-EC searches the cluster space for the most significantly differentially expressed clusters between two classes of samples. Utilization of gene clusters using the ensemble method enhances the accuracy of the classifier as compared to SVM-RCE and other similar methods.

Conclusions

The SVM-RCE-EC outperforms or is comparable to other methods. Additional advantage of SVM-RCE-EC is that the number of clusters is determined based on the ensemble approach thus capturing the real structure of the data, rather than by having the number of clusters defined by the user as is the case with SVM-RCE (k-means). Moreover, we show that the clusters generated by SVM-RCE-EC are more robust.

Availability: The Matlab version of SVM-RCE-EC is available upon request to the first author.

Availability: http://www.wistar.upenn.edu/showe/

Authors:

first name	last name	email	country	organization	corresponding?
Malik	Yousef	malik.yousef@gmail.com	Israel	Zefat College	✓
Waleed	Khalifa	khwalid@hotmail.com	Israel	sakhnin college	
Louise	Showe	lshowe@wistar.org	USA	Wistar Institute	
Loai	Abdallah	loai1984@gmail.com	Israel	Haifa University	✓

Comparison of Two Machine Learning Techniques to Predict Virulence, and Habitat of Non-Typeable Haemophilus influenzae Via Gene Possession

Keywords: genomics, machine learning, gene prediction, bacteriology

Abstract: Whole genome sequencing was performed on 855 different strains of clinically isolated non-typeable Haemophilus influenzae. After gene prediction and gene clustering, a gene presence/absence matrix was produced by homology search. The three goals were to use these gene clusters as features used to predict 1) if the strain was isolated from a sick or healthy patient, 2) if we could predict what body site the strain had been recovered from, and 3) if we could identify an informative subset of genes with high predictive power for either of the previous predictions. Feature selection was done using a combination of variance analysis and a GLM lasso implementation to identify informative features, and reduce/remove redundancy. In predicting if the strain came from a sick or healthy patient an artificial neural network (ANN) implementation achieved an accuracy of 0.7917 which increased to an accuracy of 0.7976 using the lasso selected features. The random forest (RF) implementation initially did slightly better with an accuracy of .8095, which interestingly went down to 0.7917 using only the selected features. In predicting the body site origin, ANN accuracy was 0.4385 percent, and after feature selection 0.4154. RF correctly identified the correct body origin 0.497 percent before feature selection, and at 0.5308 after.

Authors:

first name	last name	email	country	organization	corresponding?
Joshua	Earl	joshearl1@hotmail.com	USA	Drexel University College of Medicine	✓
Garth	Ehrlich	garth.ehrlich@drexelmed.edu	USA	Drexel University College of Medicine	

Predicting somatic mutations for exome sequencing data of patient-derived xenograft (PDX) mouse without patient normal or tumor samples

Keywords: PDX mouse model, preclinical setup, personalized medicine, precision medicine, Exome sequencing data, Somatic mutation calling

Abstract: PDX mouse model is an emerging platform for testing treatment responses in preclinical setup and it provides ample opportunities to realize the personalized and precision medicine. Ideally, trio samples of patient normal, patient tumor, and mouse tumor tissues are required to identify somatic mutations in PDX mouse that are concordant with patient tumor. However, it is often the case that patient tissues are not enough to generate the deep sequencing data, thus making subsequent analysis of somatic calling process difficult and error-prone. Here we developed a computational pipeline to predict somatic mutations for exome sequencing data of PDX mouse in such circumstances. It consists of intricate read mapping and filtering processes to remove mouse-originated mutations and germline mutations. We tested our pipeline for over 60 trio cases of lung cancer assuming either patient normal or tumor data are missing, and demonstrated that we could retrieve most of genuine somatic mutations without too much of false positives. Our results indicate that PDX mouse without patient reference tissues can be utilized effectively.

Availability: http://ercsb.ewha.ac.kr:8080/ErcsbHome/

Authors:

first name	last name	email	country	organization	corresponding?
Sangok	Kim	kimsangok0622@gmail.com	South Korea	Ewha Research Center for Systems Biology	
Jaewon	Kim	kimjae001@gmail.com	South Korea	Ewha Research Center for Systems Biology	
Sanghyuk	Lee	sanghyuk63@gmail.com	South Korea	Ewha Research Center for Systems Biology (ERCSB), Department of Bio-Information Science, Ewha Womans University, Seoul 120-750, KOREA	✓

Web-based applications for large-scale comparative genomics

Keywords: Comparative genomics, Synteny block, Visualization

Abstract: Recent advances in next-generation sequencing technologies and genome assembly algorithms have enabled the accumulation of a huge volume of genome sequences of various species. This trend has provided new opportunities for large-scale comparative genomics together with unprecedented burden on handling large-scale genomic data. Identifying and utilizing synteny blocks, which are conserved genomic regions among multiple species, is the key step for large-scale comparative genomics, such as comparing genomes of multiple species, reconstructing ancestral genomes, and revealing the evolutionary changes of genomes and their functional consequences. However, the construction of the synteny blocks is very challenging, especially for biologists unfamiliar with bioinformatics skills, because it requires the systematic comparison of whole-genome sequences of multiple species. To alleviate these difficulties, we recently developed a web-based application, called Synteny Portal, for constructing, visualizing, and browsing synteny blocks. Synteny Portal can be used to (i) construct synteny blocks among multiple species by using prebuilt alignments, (ii) visualize and download syntenic relationships as high-quality images, such as the Circos plot, (iii) browse synteny blocks with genetic information, and (iv) download the raw data of synteny blocks to use it as input for downstream synteny-based analyses. It also provides an intuitive and easy-to-use web-based interface. In addition, a stand-alone version of Synteny Portal has been being developed, which supports the construction of a user's own Synteny Portal interface. Synteny Portal will play a pivotal role in promoting the use of large-scale comparative genomic approaches. Synteny Portal is freely available at http://bioinfo.konkuk.ac.kr/synteny_portal/.

Authors:

first name	last name	email	country	organization	corresponding?
Jongin	Lee	jongin333@gmail.com	South Korea	Konkuk University	
Daehwan	Lee	leedaehwan90@gmail.com	South Korea	Konkuk University	
Mikang	Sim	bombi8983@gmail.com	South Korea	Konkuk University	
Daehong	Kwon	kwanyi4113@gmail.com	South Korea	Konkuk University	
Juyeon	Kim	kimjy904@gmail.com	South Korea	Konkuk University	
Younhee	Ko	younhee.ko@gmail.com	South Korea	Hankuk University of Foreign Studies	
Jaebum	Kim	jaebum.kim@gmail.com	South Korea	Konkuk University	✓

Nanopore Sequencing Technology and Tools: Computational Analysis of the Current State, Bottlenecks, and Future Directions

Keywords: Nanopore Sequencing, Single-Molecule DNA Sequencing, MinION, High-Throughput Sequence Analysis, Basecalling, Genome Assembly, Polishing, Performance Analysis, Bottleneck Analysis

Abstract: Nanopore sequencing, a promising single-molecule DNA sequencing technology, exhibits many attractive qualities and, in time, could potentially surpass current sequencing technologies. Nanopore sequencing promises higher throughput, lower cost, and increased read length, and it does not require a prior amplification step. Nanopore sequencers rely solely on the electrochemical structure of the different nucleotides for identification, and measure the ionic current change as long strands of DNA (ssDNA) pass through the nanoscale protein pores.

Biological nanopores for DNA sequencing were first proposed in the 1990s, but were only made commercially available in May 2014 by Oxford Nanopore Technologies (ONT). The first commercial nanopore sequencing device, MinION, is an inexpensive, pocket-sized, high-throughput sequencing apparatus that produces real-time data using the R7 nanopore chemistry. These properties enable new potential applications of genome sequencing, such as rapid surveillance of Ebola, Zika or other epidemics, near-patient testing, and other applications that require real-time data analysis. This technology is capable of generating very long reads (50,000bp) with minimal sample preparation. Despite all these advantageous characteristics, it has one major drawback: high error rates. In May 2016, ONT released a new version of MinION that uses a nanopore chemistry called R9. Although R9 improves data accuracy over R7, the error rate remains high. To take advantage of the real-time data produced by MinION, the tools used for nanopore sequence analysis must be fast and must overcome high error rates.

Our goal in this work is to comprehensively analyze current publicly available tools for nanopore sequence analysis, with a focus on understanding the advantages, disadvantages, and bottlenecks of them. It is important to understand where the current tools do not perform well in order to develop better tools. To this end, we analyze the multiple steps and tools in the nanopore genome analysis pipeline; and also provide some guidelines for determining the appropriate tools for each step of the pipeline and the corresponding parameters of them.

The first step, basecalling, translates the raw signal output of MinION into nucleotides to generate DNA sequences. Metrichor is the cloud-based basecaller of ONT; while Nanocall and Nanonet are publicly available nanopore basecallers. Overlap-layout-consensus (OLC) algorithms are used for nanopore sequencing reads since they perform better with longer error-prone reads. The second pipeline step finds read-to-read overlaps. Minimap and GraphMap are the commonly used tools for this step. After finding the overlaps, OLC-based assembly algorithms generate an overlap graph, where each node is a read and each edge is an overlap connecting them. The third pipeline step, genome assembly, traverses this graph, producing the layout of the reads and then constructing the draft assembly. Canu and Miniasm are the commonly used error-prone long-read assemblers. In order to increase the accuracy of the assembly, further polishing may be required. The first step of polishing is mapping the raw basecalled reads to the generated draft assembly from the previous step. The most commonly used long read mapper is BWA-MEM. After aligning the basecalled reads to the draft assembly, the final polishing of the assembly can be performed with Nanopolish.

We analyze the aforementioned nanopore sequencing tools with the goals of determining their bottlenecks and finding improvements to these tools. First, we compare the performance of the chosen tools for each step in terms of accuracy and speed. After the basecalling, read overlap finding, and assembly steps, the generated draft assemblies are compared with their reference genome; and the coverage of and identity with the reference genome are used to gauge their accuracy. For two of the draft assemblies, read mapping and polishing steps are further applied and the generated polished sequences are compared similarly. The execution time of each tool is recorded in order to compare the performance of the tools. Second, we analyze the first two steps of the pipeline in detail in order to assess the scalability of these tools. The performance of each basecaller and each read overlap finder as we vary the thread count is analyzed; wall clock time, peak memory usage, and parallel speedup are the metrics used for comparison. We present our key results in this work, and we expect future work to examine other stages of the pipeline and provide end-to-end results and analyses.

Authors:

first name	last name	email	country	organization	corresponding?
Damla	Senol	dsenol@andrew.cmu.edu	USA	Carnegie Mellon University	✓
Jeremie	Kim	jeremiek@andrew.cmu.edu	USA	Carnegie Mellon University	
Saugata	Ghose	ghose@cmu.edu	USA	Carnegie Mellon University	✓
Can	Alkan	calkan@cs.bilkent.edu.tr	Turkey	Bilkent University	
Onur	Mutlu	onur.mutlu@inf.ethz.ch	Switzerland	ETH Zurich	✓

Real time pathogen identification from metagenomic Illumina datasets

Keywords: Diagnostics, Metagenomics, Viromics, Next Generation Sequencing, Real-time mapping, K-mers, Visualization

Abstract: In the past years, Next Generation Sequencing has been utilized in time critical applications such as pathogen diagnostics with promising results. Yet, long turnaround times had to be accepted to generate sufficient data, as the analysis was performed sequentially after the sequencing was finished. Finally, the interpretation of results can be hindered by various types of contaminations, clinically irrelevant sequences, and the sheer amount and complexity of the data.

We designed and implemented a real-time diagnostics pipeline which allows the detection of pathogens from clinical samples up to five days before the sequencing procedure is even finished. To achieve this, we adapted the core algorithm of HiLive, a real-time read mapper, while enhancing its accuracy for our use case. Furthermore, common contaminations, low-entropy areas, and sequences of widespread, non-pathogenic organisms are automatically marked beforehand using NGS datasets from healthy humans as a baseline. The results are visualized in an interactive taxonomic tree, providing the user with several measures regarding the relevance of each identified potential pathogen.

We applied the pipeline on a human plasma sample spiked with Vaccinia virus, Yellow fever virus, Mumps virus, Rift Valley fever virus, Adenovirus and Mammalian orthoreovirus, which was then sequenced on an Illumina HiSeq. All spiked agents could already be detected after only 12% of the complete sequencing procedure. While we also found a large number of other sequences, these are correctly marked as clinically irrelevant in the resulting visualization, allowing the user to obtain the correct assessment of the situation at first glance.

Availability: http://www.rki.de

Authors:

first name	last name	email	country	organization	corresponding?
Simon H.	Tausch	tauschs@rki.de	Germany	Robert Koch Institute	✓
Jakob	Schulze	schulzeja@rki.de	Germany	Robert Koch Institute	
Andreas	Andrusch	andruscha@rki.de	Germany	Robert Koch Institute	
Tobias P.	Loka	LokaT@rki.de	Germany	Robert Koch Institute	
Jeanette	Klenner	JeanetteKlenner@bundeswehr.org	Germany	Bundeswehr	
Piotr W.	Dabrowski	dabrowskiw@rki.de	Germany	Robert Koch Institute	
Bernhard Y.	Renard	renardB@rki.de	Germany	Robert Koch Institute	
Andreas	Nitsche	NitscheA@rki.de	Germany	Robert Koch Institute	

Identification and characterization of bacteriophages in global sewage samples

Keywords: comparative metagenomics, environmental metagenomics, phage metagenomics

Abstract: Sewage is a major source of both human pathogens and their associated bacteriophages. Phages control bacterial population by predation and can act as natural reservoirs for accessory genes such as antimicrobial resistance genes and virulence factors. However, currently limited knowledge is available about the sequence and functional diversity of such sewage phage communities. We here present a study of the phage communities of 81 sewage samples from 62 different countries around the world.

The samples consist of metagenomic assemblies in which we identified phage contigs by using the MetaPhinder tool. These contigs were subsequently screened for the presence of known virulence and resistance genes with the VirFinder and ResFinder tools. Additionally, we performed host prediction with HostPhinder and taxonomic classification.

Antimicrobial resistance genes were found in the phage population of 52 out of 80 samples and virulence factors in 18 of the samples. Potential hosts were predicted for 12.7 ± 3 % of phage contigs. Among the most common host genera were Escherichia, Caulobacter and Bacillus. Taxonomic classifications were assigned to 0.5% of the phage contigs on average. Among the most common taxonomic assignments is crAssphage which was identified in 74 of the samples.

In conclusion, we found that the phage communities in sewage are extremely diverse and contain many novel sequences.

Authors:

first name	last name	email	country	organization	corresponding?
Henrike	Zschach	henrike@cbs.dtu.dk	Denmark	Technical University of Denmark	✓
Barbara	Lindhard	b.lindhard@live.dk	Denmark	Technical University of Denmark	
Vanessa Isabell	Jurtz	vanessa@cbs.dtu.dk	Denmark	Technical University of Denmark	✓
Frank	Aarestrup	fmaa@food.dtu.dk	Denmark	Technical University of Denmark	
Morten	Nielsen	mniel@cbs.dtu.dk	Denmark	Technical University of Denmark	✓
Ole	Lund	lund@cbs.dtu.dk	Denmark	Center for Biological sequence analysis (CBS)	✓

From miRNAs to isomiRs: isomiRs population profiles in rheumatoid arthritis

Keywords: NGS, miRNA, isomiR, isomiRs population, Rheumatoid arthritis

Abstract: Introduction

Rheumatoid Arthritis (RA) is a chronic autoimmune disease which leads to inflammation of joints in a patient. The cause of RA is still not well understood, but smoking, gender, pregnancy and genetic factors are all known to contribute to the development of RA. MicroRNAs are short non-coding RNAs with length varying from 19 to 26 nucleotides that regulate gene expression by binding to mRNA targets. Microarray experiments have identified several miRNAs that appear to play a functional role in RA patients, but there are few miRNA studies on RA using Next Generation Sequencing. In this work, we investigate changes both in miRNA expression levels, as well as variation in miRNA isoform (or isomiRs) population in miRNA sequence data extracted from three immune cell lines from Norwegian RA patients and healthy controls.

Methods

We collected blood samples and three types of immune cells (CD19, CD4 memory and CD4 naïve) from RA patients at three time points: newly diagnosed; three months treatment after diagnosis; and long-term patients. Samples from newly diagnosed and long-term patients were prepared for small RNA sequencing and the sequence data was submitted to a standard preliminary analysis including QC control and adapter trimming. Prior to mapping, identical reads were collapsed into single sequences and mapped to human reference miRNA hairpin sequences using bowtie with 2 mismatches. We then investigated the variation amongst conditions in the population of the reads (i.e. isoforms or isomiRs) that map to the reference set of human miRNAs (according to MiRBase version 21). To facilitate this, we introduced a comprehensive nomenclature to describe the modifications between a specific isomiR and the reference "parent" miRNA as specified in miRBase.

Results

We identified a set of isomiRs that are differentially expressed amongst the two RA cohorts, with many more isomiRs differentially expressed in CD19 than CD4 memory and CD4 naïve cells, and distinct isomiRs observed in each cell line. Additionally, computational target prediction identified distinct targets sets for each isomiR, which are also distinct from the predicted targets for the parent miRNA. For example, the mature form of hsa-miR-126 is predicted to target to more than 1600 target genes. In contrast, the differentially expressed isomiRs each have a dramatically decreased set of target genes; the shorter isoform (both one nucleotide deleted at both 5' and 3' end) have 5 targets and same length isomiR (one nucleotide extended at 5' end and one nucleotide deleted at 3' end) has 61 targets.

Conclusions

Investigating the additional dimensionality of small RNA NGS data (in the form of isomiR populations) can reveal additional structure that can provide further insight into differences among tested conditions.

Availability: http://eia.udg.edu/~apla/index_en.html

Authors:

first name	**last name**	**email**	**country**	**organization**	**corresponding?**
Xiangfu	Zhong	xiangfu.zhong@studmed.uio.no	Norway	Oslo University Hospital	✓
Fatima	Heinicke	fatima.heinicke@studmed.uio.no	Norway	Department of Medical Genetics, University of Oslo and Oslo University Hospital	
Albert	Pla Planas	a.p.planas@medisin.uio.no	Norway	Universitetet i Oslo	
Benedicte A.	Lie	b.a.lie@medisin.uio.no	Norway	University of Oslo and Oslo University Hospital	
Simon	Rayner	simon.rayner@medisin.uio.no	Norway	Oslo University Hospital	✓

Integrated transcriptomic analysis of skeletal disease

Keywords: Transcriptomics, Network, Clustering, Disease

Abstract: Skeletal diseases, including the complex diseases osteoarthritis and osteoporosis, present a large and growing health care burden with often poor treatment options. There is a critical need for more detailed mechanistic understanding of these diseases to enable the development of rational disease modifying treatments. Transcriptomics analysis has been often used to provide both characterisation of tissue affected by skeletal diseases and to find disease gene candidates in relevant cell types. Despite a large number of existing skeletal disease datasets, these expression profiles are difficult to interrogate and have not been examined in an integrated way.

Using an automated pipeline for reproducible analysis of publicly available transcriptomics data, we have produced a large collection of consistently analysed, annotated expression datasets to allow mining for hidden connections and shared pathogenic mechanisms between different skeletal diseases. Unsupervised clustering at multiple regulatory levels was performed which revealed clusters of similarity at the pathway and transcription factor level which were not readily visible from simple examination of the gene expression signatures, demonstrating the utility of this integrative analysis. Our knowledge base of gene signatures, enriched pathways, active sub-networks and upstream transcription factors provides a resource for querying skeletal disease related datasets, so to contextualise the data with prior knowledge to provide more meaningful biological insight.

Authors:

first name	last name	email	country	organization	corresponding?
Jamie	Soul	jamie.soul@manchester.ac.uk	United Kingdom	University of Manchester	✓
Jean-Marc	Schwartz	Jean-marc.Schwartz@manchester.ac.uk	United Kingdom	University of Manchester	

Exploring variation within the genus Lens with 10x Chromium data

Keywords: 10x Genomics Chromium, genome assembly, structural variation, linked reads, copy number variation

Abstract: The wild lentil species, Lens lamottei and Lens odemensis, are potential sources of novel genetic variation for disease resistance and other desirable traits for Lens culinaris (cultivated lentil) breeding. Populations from crosses with L. odemensis have been made but hybrids with L. lamottei have been very difficult to produce. Understanding the structural differences between the wild and cultivated genomes will give insight into both the evolution within the Lens genus and domestication of cultivated lentil, as well as identify large-scale structural differences that may contribute to the level of success in obtaining viable hybrid offspring.

Short-read assemblies were improved with 10x scaffolding for L. lamottei (3.5Gb total assembly, 4.4Mb N50, 28,638 scaffolds) and L. odemensis (3.7Gb total assembly, 3.9Mb N50, 23,352 scaffolds). The scaffolds were anchored on high-density genetic maps generated from GBS data to create pseudomolecules and additional unanchored scaffolds. The 10x Chromium data (30X read depth) initially used in superscaffolding were then remapped against the alternate genomes to identify both variations in coverage (PAV/CNV) and breakpoints (structural variation) amongst the wild and cultivated lentil species. The long-range information from linked reads was used to confirm interspecific differences in genome organization. Although large-scale rearrangements were expected based on previous cytogenetic studies, this approach allowed for much higher confidence and more fine-grained identification of variation between the genomes beyond simple SNP and indel calling.

Authors:

first name	last name	email	country	organization	corresponding?
Chushin	Koh	kevin.koh@gifs.ca	Canada	Global Institute for Food Security	✓
Larissa	Ramsay	l.ramsay@usask.ca	Canada	University of Saskatchewan	✓
Andrew	Sharpe	andrew.sharpe@gifs.ca	Canada	Global Institute for Food Security	
Kirstin	Bett	k.bett@usask.ca	Canada	University of Saskatchewan	

The genome of N. lovaniensis, the base for a comparative approach to unravel pathogenicity factors of the brain-eating amoeba N. fowleri

Keywords: Naegleria lovaniensis, Naegleria fowleri, long read sequencing, genome de novo assembly, comparative genomics

Abstract: Naegleria fowleri, commonly known as the brain-eating amoeba, is a free-living eukaryote found in soil and fresh warm water sources all over the world. Once entered the nose, N. fowleri follows the olfactory nerves to the brain and causes primary amoebic meningoencephalitis (PAM), a fast progressing and mostly fatal disease of the central nervous system. The mechanisms involved in the pathogenesis are still poorly understood. To gain a better understanding of the relationships within the genus of Naegleria and to investigate pathogenicity factors of N. fowleri, we characterized the genome of its closest non-pathogenic relative N. lovaniensis.

To achieve a nearly complete assembly of the N. lovaniensis genome, long read sequencing was applied followed by assembling of the data using FALCON, a diploid-aware string graph assembler. To unravel the relatedness of Naegleria species, a phylogenetic tree based on maximum likelihood and bootstrapping using RAxML was constructed. Keeping pathogenicity in mind, proteins specific for N. fowleri were defined by clustering of orthologous gene families between different Naegleria species and their function was characterized by functional annotation and GO enrichment analysis.

In this study, we present the 30Mb genome of N. lovaniensis for the first time. Sequencing and de novo assembly of the genome supports the hypothesis of the close relationship to the human pathogen N. fowleri. Thus, knowledge of the N. lovaniensis genome provides the basis for further comparative approaches to unravel pathways involved in the pathogenicity of PAM and to identify structures for possible treatment options.

Authors:

first name	last name	email	country	organization	corresponding?
Nicole	Liechti	nicole.liechti@bioinformatics.unibe.ch	Switzerland	University of Bern	✓
Rémy	Bruggmann	remy.bruggmann@bioinformatics.unibe.ch	Switzerland	University of Bern	
Matthias	Wittwer	matthias.wittwer@babs.admin.ch	Switzerland	Spiez Laboratory	

Detection and localization of mutations in MUC1 gene, causing an Autosomal Dominant Tubulointerstitial Kidney Disease (ADTKD)

Keywords: High-throughput sequencing, NGS, Nanopore, Long reads, Haplotype phasing, Repetitions, Human genetics, Genetic variant

Abstract: MUC1 gene is coding for transmembrane glycoprotein mucin-1 and its coding sequence is GC-rich (82%), containing 25-120 polymorphic tandem repeats (VNTR) of the length of 60bp. Frameshift mutations in MUC1 are leading to the synthesis of abnormal, highly basic, cysteine-rich protein MUC1-fs. MUC1-fs accumulates in the tubular cells of kidneys, causes progressive deterioration of renal functions and leads to a renal failure. The age of kidney failure varies from 17 to 75 years and we hypothesize, that the exact location of the mutation may be related to the age of renal failure.

Using current technologies, it is very difficult to detect the mutations in MUC1 and it seems to be impossible to determine its exact position because of the repetitive and GC-rich sequence, the length of the VNTR and the homopolymer stretch of 7 cytosines within each repeat.

To detect the mutation, we amplified the VNTR region using Long Range PCR and sequenced the amplified region on Illumina HiSeq, followed by a bioinformatic analysis of the raw reads. We tested this method successfully on samples with previously known C insertion and then applied it to samples with unknown mutations. Using this approach, we identified three completely new mutations. To determine haplotype and the exact position of the mutations, we are currently using the MinION sequencer from Oxford Nanopore.

These methods enable genetic diagnostics of ADTKD and could contribute to the understanding of the genetic factors determining the progression and the age of the onset of the kidney failure in ADTKD.

Authors:

first name	last name	email	country	organization	corresponding?
Anna	Pristoupilova	anna.pristoupilova@lf1.cuni.cz	Czech Republic	Institute of Inherited Metabolic Disorders, First Faculty of Medicine, Charles University	✓
Viktor	Stranecky		Czech Republic	Institute of Inherited Metabolic Disorders, First Faculty of Medicine, Charles University	
David	Hoksza		Czech Republic	Department of Software Engineering, Faculty of Mathematics And Physics, Charles University	
Hana	Hartmannova		Czech Republic	Institute of Inherited Metabolic Disorders, First Faculty of Medicine, Charles University	
Alena	Vrbacka		Czech Republic	Institute of Inherited Metabolic Disorders, First Faculty of Medicine, Charles University	
Katerina	Hodanova		Czech Republic	Institute of Inherited Metabolic Disorders, First Faculty of Medicine, Charles University	
Martina	Zivna		Czech Republic	Institute of Inherited Metabolic Disorders, First Faculty of Medicine, Charles University	
Anthony	Bleyer		USA	Wake Forest School of Medicine, Medicine Center Blvd, Winston-Salem	
Stanislav	Kmoch	skmoch@lf1.cuni.cz	Czech Republic	Institute of Inherited Metabolic Disorders, First Faculty of Medicine, Charles University	✓

systemPipeR: NGS workflow and report generation environment

Keywords: High Throughput Sequencing (HitSeq), Analysis software environment, Analysis reports, Ribo-Seq, ChIP-Seq, RNA-Seq, VAR-Seq

Abstract: The analysis of Next-generation sequencing (NGS) data remains a major obstacle to the efficient utilization of the technology. While substantial effort has been invested on the development of software dedicated to the individual analysis steps of NGS experiments, insufficient resources are currently available for integrating the individual software components within the widely used R/Bioconductor environment into automated workflows capable of running the analysis of most types of NGS applications from start-to-finish in a time-efficient and reproducible manner. To address this need, we have developed the R/Bioconductor package systemPipeR. It is an extensible environment for both building and running end-to-end analysis workflows with automated report generation for a wide range of NGS applications. Its unique features include a uniform workflow interface across different NGS applications, automated report generation, and support for running both R and command-line software on local computers and computer clusters. A flexible sample annotation infrastructure efficiently handles complex sample sets and experimental designs. To simplify the analysis of widely used NGS applications, the package provides pre-configured workflows and reporting templates for RNA-Seq, ChIP-Seq, VAR-Seq and Ribo-Seq. Additional workflow templates will be provided in the future. systemPipeR accelerates the extraction of reproducible analysis results from NGS experiments. By combining the capabilities of many R/Bioconductor and command-line tools, it makes efficient use of existing software resources without limiting the user to a set of predefined methods or environments. systemPipeR is freely available for all common operating systems from Bioconductor (http://bioconductor.org/packages/devel/systemPipeR).

Availability: http://girke.bioinformatics.ucr.edu

Authors:

first name	last name	email	country	organization	corresponding?
Tyler	Backman	tbackman@lbl.gov	USA	Lawrence Berkeley National Laboratory	
Thomas	Girke	thomas.girke@ucr.edu	USA	University California, Riverside	✓

ClearUp: sample identity validation in NGS data

Keywords: quality control, software, next generation sequencing, homo sapiens, single nucleotide polymorphism, fingerprint

Abstract: Human DNA sequencing studies are often compromised by mix-ups happening during either sample preparation or data management, bringing a demand of checking for mislabeled samples as a routine quality control step. We present ClearUp, a method and a software package for sample identity validation from BAM files. By selecting selecting a set of common population SNPs shared across input with high enough sequencing quality, ClearUp builds "SNP fingerprints" and uses them to determine relatedness, ancestry and sex. The user friendly web-based interface allows to review and refine the results using an interactive dendrogram and a built-in genome browser. The method works across different types of sequencing data, including WGS, WES, RNA-seq, and targeted sequencing, as soon as the input targets overlap. We demonstrate that SNP fingerprints give enough variation in order to accurately detect mislabeled and related samples. In contrast t similar tools, ClearUp is undemanding in terms of input and does not require any data pre-processing, taking only files in BAM format. The tool is open sourced and available on GitHub at https://github.com/AstraZeneca-NGS/Fingerprinting, and provides both a command line interface and a Flask-driven web-server with a graphical user interface.

Authors:

first name	**last name**	**email**	**country**	**organization**	**corresponding?**
Vladislav	Saveliev	vladislav.sav@gmail.com	Russia	St. Petersburg State University	✓

Epigenetic programming in acute intestinal inflammation

Keywords: Whole methylome analysis, Whole transcriptome analysis, BS-sequencing, Inflammation and pathogenesis of tumors, Activation of enhancers, Epigenetic clock

Abstract: Inflammation has been linked to the pathogenesis of tumors in a substantial fraction of human cancers. Here, we used whole-genome bisulfite and transcriptome sequencing of single-dose DSS treated mice and matched controls to characterize colitis-related DNA methylation and the accompanying gene expression patterns in detail. Our analysis of colitis-related methylomes detected a consistent DSS-dependent hypomethylation for several enhancer segments. As enhancers overlap with lowly methylated regions (LMRs) in whole-genome bisulfite sequencing analyses (Burger, 2013), we identified these active regulatory regions in our datasets. We focused our analysis on the large fraction (46 %) of intragenic LMRs, as it allowed the assignment of individual LMRs to specific genes. This approach identified 373 LMRs that become hypomethylated in DSS-treated mice and showed significantly (q¡0.05) increased expression (181 hypomethylated with reduced expression). We analyzed this set of LMRs for its enrichment of transcription factor binding sites and conclude that AP-1 and not NF-kB plays a key role in regulating the response to acute colitis. Furthermore, we applied the recently developed epigenetic clock for mouse (Stubbs, 2017) to the different methylomes and show that inflammation changes the rate of aging of affected tissues compared to healthy tissue.

Authors:

first name	last name	email	country	organization	corresponding?
Guenter	Raddatz	g.raddatz@dkfz.de	Germany	DKFZ	✓
Frank	Lyko	f.lyko@dkfz.de	Germany	DKFZ	
Yehudit	Bergman	yehuditb@ekmd.huji.ac.il	Israel	Hebrew University Medical School	
Ihab	Ansari	ihabansari@gmail.com	Israel	Hebrew University Medical School	

Targeted single cell sequencing unravels the heterogeneity of T-cell acute lymphoblastic leukemia

Keywords: single cell RNA-seq, targeted single cell sequencing, T-cell acute lymphoblastic leukemia, graph-based algorithm

Abstract: T-cell acute lymphoblastic leukemia (T-ALL) comprises 25% of all ALL cases and primarily affects children. We aimed to investigate the heterogeneity of T-ALL patient samples and identify the order of mutation acquisition during leukemia evolution. We performed targeted DNA sequencing and RNA sequencing on 200-400 single cells of 4 human T-ALL samples.

Whole genome sequencing of the bulk diagnostic samples was used to identify the spectrum of genomic lesions present in the major diagnostic clone(s). We then used targeted sequencing of about 20 genomic lesions and 40 heterozygous SNPs (for quality control) in the single leukemia cells. Cells were discarded from analysis if locus and allelic drop-out exceeded 33.3%. Of the 4 patients analysed, two exhibit one homogeneous leukemic cell population with most cells having all mutations. In the other two patients we observed two distinct subclones with different mutation loads. A graph-based algorithm was developed to determine the order at which mutations were acquired, which showed that most chromosomal translocations were early events in leukemia development, while NOTCH1 mutations were typically late events.

Single-cell RNA-sequencing analysis was performed with 10X genomics platform, and revealed limited heterogeneity on the level of gene expression, with the major discriminating factor being cell cycle effects. In conclusion, our novel graph-based algorithm for single cell sequence data was able to provide new information on the order at which mutations are accumulating during T-ALL development.

Authors:

first name	last name	email	country	organization	corresponding?
Sofie	Demeyer	sofie.demeyer@kuleuven.vib.be	Belgium	KULeuven / VIB	✓
Jolien	De Bie	jolien.debie@kuleuven.vib.be	Belgium	KULeuven / VIB	
Ellen	Geerdens	ellen.geerdens@kuleuven.vib.be	Belgium	KULeuven /VIB	
Jan	Cools	jan.cools@kuleuven.vib.be	Belgium	KULeuven / VIB	

In-silico read normalization using set multi-cover optmization

Keywords: Read Normalization, De bruijn graph based assembly, Set multi-cover optimization, RNA-seq

Abstract: De Bruijn graphs are a common assembly data structure for large sequencing datasets. But with the advances in sequencing technologies, assembling high coverage datasets has become a computational challenge. Read normalization, which removes redundancy in large datasets, is widely applied to reduce resource requirements. Current normalization algorithms, though efficient, provide no guarantee to preserve important k-mers that form connections between regions in the graph. Here, normalization is phrased as a set multi-cover problem on reads and a heuristic algorithm, ORNA, is proposed. ORNA normalizes to the minimum number of reads required to retain all k-mers and their relative kmer abundances from the original dataset. Hence, all connections and coverage information from the original graph are preserved. ORNA was tested on various RNA-seq datasets with different coverage values. It was compared to the current normalization algorithms and was found to be performing better. It is shown that combining read error correction and normalization allows more accurate and resource efficient RNA assemblies compared to the original dataset. Further, an application was proposed in which multiple datasets were combined and normalized to predict novel transcripts that would have been missed otherwise. Finally, ORNA is a general purpose normalization algorithm that is fast and significantly reduces datasets with little loss of assembly quality. ORNA is freely available at https://github.com/SchulzLab/ORNA

Authors:

first name	last name	email	country	organization	corresponding?
Dilip	Durai	ddurai@mmci.uni-saarland.de	Germany	MPII saarland and MMCI saarland	
Marcel	Schulz	mschulz@mmci.uni-saarland.de	Germany	MPII saarland and MMCI saarland	✓

Scale4C: an R package for scale-space transformation applied to 4C-seq data

Keywords: 4C-seq, Next generation sequencing, Data visualization

Abstract: 4C-seq is a method to identify chromosomal contact partners for one chosen position in the genome. Since the 3C-based technique inherently causes a fragment structure of the output, and suffers from technical artifacts like PCR bias, most current 4C-seq algorithms use windows or smoothing techniques to decrease noise, introducing arbitrary window sizes or smoothing parameters. This leads to the general problem of parameter choice for optimal analysis and visualization.

We present the R package Scale4C, which uses Witkin's scale-space filtering approach to create a novel multi-scale 4C-seq near-cis visualization. This representation of the data allows for explorative analysis of candidate interactions, and structural comparison of datasets. During scale-space filtering, the 4C-seq signal is smoothed with Gauss kernels of increasing smoothing factors. Inflection points of the resulting curves are subsequently tracked in a so-called fingerprint map, and singular points of these curves are identified. Focusing on features of the data ('peaks' and 'valleys') and their transitions for multiple smoothing parameters, the package's plot functions can create 2D tesselation maps in scale-space from these singularities. Tesselation maps allow to visually assess prominent features of the 4C-seq signal with a high degree of stability for multiple smoothing factors, and thus potential interactions.

Additional functions of the package include further visualization routines for smoothed data with a chosen smoothing factor and its corresponding inflection points, and plot functions for fingerprint maps with their traced singularities. Data import from bed-files and Basic4Cseq is supported, as well as export of the scale-space tesselation in tabular form.

Authors:

first name	last name	email	country	organization	corresponding?
Carolin	Walter	Carolin.Walter@ukmuenster.de	Germany	Westfälische Wilhelms-Universität Münster	✓
Martin	Dugas	dugas@uni-muenster.de	Germany	Westfälische Wilhelms-Universität Münster	

Investigate ongoing retroviral endogenization in koalas (KoRV)

Keywords: aDNA, endogenous retrovirus, KoRV, PacBio, Illumina, iPCR, targeted sequencing

Abstract: Endogenous retroviruses invade the host genome and get horizontally transmitted from parents to the offspring. The koala retrovirus (KoRV) is currently invading the genome of Phascolarctos cinereus. By investigating KoRV, we can study the endogenization process of an infectious virus in real time. We conducted different studies to examine insertion sites in ancient DNA samples (Cui, P. et al. Comprehensive profiling of retroviral integration sites using target enrichment methods from historical koala samples without an assembled reference genome. PeerJ 4, e1847 (2016).), samples of wild animals and zoo animals, as well as comparing cancer and control tissues. We found rarely shared integration sites and recombination of two viruses, which may result in a reduced prevalence of the virus.

Authors:

first name	last name	email	country	organization	corresponding?
Ulrike	Löber	uloeber@yahoo.de	Germany	Leibniz institute for zoo and wildlife research	✓
David	Alquezar	d.e.alquezar@gmail.com	Australia	na	
Alex	Greenwood	greenwood@izw-berlin.de	Germany	Leibniz institute for zoo and wildlife research	

Epi-MOLAS: Epi-genoMics OnLine Analysis System

Keywords: Methylation, bisulfite sequence, Online analysis, Web application

Abstract: EpiMOLAS is an open access web service for exploring the whole genome epigenetic modification based on bisulfite sequencing (BS Seq) related technologies. It supports outputs of two popular BS Seq mapping programs, BS Seeker (version 2, CGmap) and Bismark (bismark_methylation_extracter). Information of three different C sequence contexts (CpG, CHG, CHH) in two location type (”promoter” and ”gene-body”) are calculated and joined to gene annotation database to build a user-owned data analysis workbench. Besides accessing the methylation indexes of individual genes, the changes on the epigenetic marks can be performed on the built-in quantitative analysis pipeline and joined to various advanced analysis toolkits like functional enrichment analysis on GO terms and KEGG pathways, or views the methylation level by heatmap or by whole genome plot seamlessly. Those gene lists derived from different approaches can be calculated in logical manner via Venn-diagram to generate the new list for further analysis. The data analysis website can be set to open for public or shared by a keyword-controlled way by the website creator.

epiMOLAS is available for submission on http://symbiosis.iis.sinica.edu.tw/epimolas. Present available reference genomes in epiMOLAS includes human (GRCh37/hg19), mouse (GRCm38/mm10) and Arabidopsis (TAIR10).

*Here is a demo site:

http://symbiosis.iis.sinica.edu.tw/epimolas/grch37/

Availability: http://www.iis.sinica.edu.tw/pages/cylin/index_en.html

Authors:

first name	last name	email	country	organization	corresponding?
Shu-Hwa	Chen	Sophia.emily@gmail.com	Taiwan	Institute of Information Science, Academia Sinica	
Sheng-Yao	Su	daniel0523@gmail.com	Taiwan	Institute of Information Science, Academia Sinica	
Yi-Hsun	Lu	lindalu@iis.sinica.edu.tw	Taiwan	Institute of Information Science, Academia Sinica	
Pao-Yang	Chen	paoyang@gate.sinica.edu.tw	Taiwan	Institute of Plant and Microbiology, Acadmia Sinica	
Chung-Yen	Lin	cylin@iis.sinica.edu.tw	Taiwan	Institute of Information Science, Academia Sinica	✓

Algorithms for Structural Variation Discovery Using Hybrid Sequencing Technologies

Keywords: structural variation, PacBio, Illumina, split reads, read pairs

Abstract: Structural variation (SV) is defined as genomic variation that affects more than 50 base pairs [1]. Recent studies found that in the human genome, there exist thousands of SVs on average which affect around 15-20 million base pairs [2]. As a result of this, they show a high probability to cause functional effects [3]. Additionally, various studies associated several SVs to human disease[4].

SV detection studies were made possible by the introduction of high-throughput DNA sequencing (HTS). While these technologies produce large amounts of sequencing data in a cost effective and fast manner, they still suffer from various disadvantages. Second-generation sequencing technologies, such as Illumina, create short reads (75-150 bp) with low cost and low error rates [5]. On the other hand, third-generation sequencing technologies, such as Pacific Biosciences, produce long reads (average 10Kb) with high cost and high error rate as they use single molecule real time sequencing [7].

For accurate SV detection, long reads with low error rate are desired. Even though PacBio data with high coverage (¿40X) may be more reliable for this purpose, high costs associated with this technology diminish its practicality for large number of samples. On the other hand, short Illumina reads cannot span over repeats and duplications where most SVs are known to occur. Another problem with SV discovery is that the accurate detection of breakpoints is difficult in the homologous segments and the repeated sections of the DNA. Thus, the coupling low coverage (i.e. lower cost) PacBio and high coverage Illumina data would, in theory, complement the strengths of these two technologies with each other and, correct for the biases.

The aim of this study is to detect large deletions and inversions in human genome with low cost and high accuracy. This is achieved by, first, broadly defining inversions and deletions using low coverage PacBio sequencing data and then, comparing the SV signals with those detected using Illumina data. We downloaded PacBio data set generated from the genome of NA12878 from the Genome in a Bottle project. The Illumina data set from the same genome was generated as part of the Platinum Genomes collection.

Briefly, we searched for split read signature in the PacBio alignments that signal large deletions (¿5Kbp), and we then tested whether there exists any Illumina read pair signature in the periphery of the PacBio-detected regions. PacBio reads were previously aligned using BLASR, and Illuma reads were aligned using BWA-MEM. We used an approximation to the quasi clique detection problem for clustering SV sequence signals. We provide the preliminary results below for deletions only, which we will further improve through incorporating local assembly and split read mapping of Illumina reads, and extend to detect inversions.

Authors:

first name	last name	email	country	organization	corresponding?
Ezgi	Ebren	ezgiebren@gmail.com	Turkey	Bilkent University, Department of Computer Engineering	
Ayse Berceste	Dincer	ayse.dincer@ug.bilkent.edu.tr	Turkey	Bilkent University, Department of Computer Engineering	
Can	Alkan	calkan@gmail.com	Turkey	Bilkent University, Department of Computer Engineering	✓

Genome Read In-Memory (GRIM) Filter: Fast Location Filtering in DNA Read Mapping using Emerging Memory Technologies

Keywords: Genome Read Mapping, Bioinformatics, 3D-stacked DRAM, Computer Architecture, Hardware Acceleration, Seed Filter Extend, GRIM-Filter, Processing-in-Memory, High Throughput Sequence Analysis

Abstract: Motivation: Location filtering is critical in DNA read mapping, a process where billions of DNA fragments (reads) sampled from a donor are mapped onto a reference genome to identify genomic variants of the donor. Read mappers quickly generate possible mapping locations for each read, extract reference sequences at each of the mapping locations, and then check similarity between each read and its associated reference sequences with a computationally expensive dynamic programming algorithm (alignment) to determine the origin of the read. Location filters come into play before alignment, discarding locations that alignment would have deemed a poor match. The ideal location filter would discard all poor matching locations prior to alignment such that there is no wasted computation on poor alignments.

Results: We propose a novel filtering algorithm, GRIM-Filter, optimized to exploit emerging 3D-stacked memory systems that integrate computation within a stacked logic layer, enabling processing-in-memory (PIM). GRIM-Filter quickly filters locations by 1) introducing a new representation of coarse-grained segments of the reference genome and 2) using massively-parallel in-memory operations to identify read presence within each coarse-grained segment. Our evaluations show that for 5% error acceptance rates, GRIM-Filter eliminates 5.59x-6.41x more false negatives and exhibits end-to-end speedups of 1.81x-3.65x compared to mappers employing the best previous filtering algorithm.

Authors:

first name	last name	email	country	organization	corresponding?
Jeremie	Kim	Jeremiek@andrew.cmu.edu	USA	Carnegie Mellon University	✓
Damla	Senol	dsenol@andrew.cmu.edu	USA	Carnegie Mellon University	
Hongyi	Xin	gohongyi@gmail.com	USA	Carnegie Mellon University	
Donghyuk	Lee	blepus@gmail.com	USA	NVIDIA Research	
Saugata	Ghose	ghose@cmu.edu	USA	Carnegie Mellon University	✓
Mohammed	Alser	mealser@gmail.com	Turkey	Bilkent University	
Hasan	Hassan	hhassan@etu.edu.tr	Turkey	TOBB University of Economics & Technology	
Oguz	Ergin	oergin@etu.edu.tr	Turkey	TOBB University of Economics and Technology	
Can	Alkan	calkan@cs.bilkent.edu.tr	Turkey	Bilkent University	
Onur	Mutlu	omutlu@gmail.com	Switzerland	ETH Zurich	✓