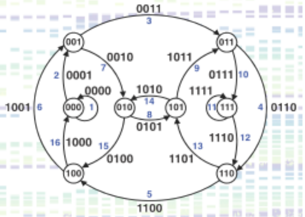
HitSeq 2015

High Throughput Sequencing

Algorithms & Applications

A SIG of IMSB/ECCB 2015. July 10-11, 2015. Dublin, Ireland



Poster Abstracts

How data analysis affects power, reproducibility and biological insight of RNA-seq studies in Neuroscience

Keywords: RNA-seq, normalization, differential expression, reproducibility, neuroscience

Abstract: The sequencing of the full transcriptome (RNA-seq) has become the preferred choice for the measurement of genome-wide gene expression. Despite its widespread use, challenges remain in RNA-seq data analysis. One often-overlooked aspect is normalization. Despite the fact that a variety of factors or "batch effects" can contribute unwanted variation to the data, commonly used RNA-seq normalization methods only correct for sequencing depth. The study of gene expression in the context of brain and behavior is particularly problematic because it is influenced simultaneously by a variety of biological factors in addition to the one of interest. We show that in RNA-seq studies of gene expression in the brain, batch effects often dominate the signal of interest; and that the choice of normalization method affects the power and reproducibility of the results. While commonly used global normalization methods are not able to adequately normalize the data, more recently developed RNA-seq normalization can. We focus on one particular method, RUV, and show that it is able to increase power and biological insight of the results. RUV normalization, available in the open-source Bioconductor package RUVSeq, is applicable to a broad range of studies as well as meta-analysis of publicly available data.

Authors:

first name	last name	email	country	organization	web site	corresponding?
Davide	Risso	davide.risso@berkeley.edu	United States of America	University of California, Berkeley		✓
Lucia	Peixoto	lucia.peixoto@wsu.edu	United States of America	University of Pennsylvania		✓
Shane	Poplawski	shane.poplawski@abbott.com	United States of America	University of Pennsylvania		
Mathieu	Wimmer	mwimmer@mail.med.upenn.edu	United States of America	University of Pennsylvania		
Terry	Speed	terry@wehi.edu.au	Australia	The Walter and Eliza Hall Institute of Medical Research and The University of Melbourne and University of California, Berkeley		
Marcelo	Wood	mwood@uci.edu	United States of America	University of California, Irvine		
Ted	Abel	abele@sas.upenn.edu	United States of America	University of Pennsylvania		✓

ASpli: an integrative R package for the analysis of alternative splicing using RNA-Seq

Keywords: alternative splicing, RNA seq, plants

Abstract: Alternative splicing (AS) is a prevalent mechanism of post transcriptional gene regulation in multicellular eukaryotes. It allows a single gene to increase functional and regulatory diversity, through the synthesis of multiple mRNA isoforms encoding structurally and functionally distinct proteins. AS occurs via 4 main events: intron retention (IR), exon skipping (ES) and alternative use of donor and aceptor splicing sites (Alt 5'ss and Alt 3'ss). The development of novel high-throughput sequencing methods for RNA (RNA-Seq) provided a very powerful mean to study alternative splicing under multiple conditions at unprecedented depth. As long as new studies on post-transcriptional regulation arises, there are an increasing evidence than AS frequency is higher than expected. Despite It has became the new standard for studying gene and transcription expression, the use of RNA-seq for the study of transcripts repertoire in a given condition is not trivial.
Here we introduce a very flexible and easy to use R package named ASpli. We propose a count based integrative method taking into account gene expression, exon and intron differential usage and their relationship with junctions spanning those features. Using an annotated transcriptome we are able to classify subgenic features into alternative or not alternative regions. ASpli is intended to facilitate the analyiss of RNAseq data for the quantification and discovery of AS events and it has been used in many recent publications from our lab. Results of the analysis are presented in a user friendly manner, including plots of the most relevant AS events discovered.

Authors:

first name	last name	email	country	organization	web site	corresponding?
Estefania	Mancini	emancini@leloir.org.ar	Argentina	Fundación Instituto Leloir	http://www.leloir.org.ar	✓
Ariel	Chernomoretz	ariel@df.uba.ar	Argentina	University of Buenos Aires		
Marcelo	Yanovsky	myanovsky@leloir.org.ar	Argentina	IIBBA - Fundación Instituto Leloir	http://www.leloir.org.ar	

RNA-Seq: assessment of transcript level analysis

Keywords: RNA-Seq, transcriptome, alternative splicing, spike-ins

Abstract: One of the major applications of next-generation sequencing (NGS) technologies is RNA-Seq for transcriptome genome wide analysis. Although there are multiple studies evaluating and bench-marking RNA-Seq tools dedicated to gene level analysis there are few evaluation studies performed on the transcript- isoform level. Alternative splicing occurs as a normal phenomenon in eukaryotes, where it greatly increases the biodiversity of proteins that can be encoded by the genome, in humans, ~95% of multi-exonic genes are alternatively spliced. The aim of this study is to assess and compare the ability of the various bioinformatics approaches and tools to assemble, quantify abundance and detect deferentially expressed transcripts using RNA-Seq data in a controlled experiment.

Towards this aim we evaluated many different tools using a differentiating mouse embryonic transcriptome, to which mouse spike-in control transcripts were added. This novel approach was used to assess the accuracy found among the tools as revealed by comparing the observed results versus the mouse controlled spiked-ins expected results. We found that detection of differential expression at the gene level is acceptable, yet on the transcript-isoform level all tools tested were lacking accuracy and precision. In this study we discuss the factors that contribute to the difficulties in transcript level analysis.

Authors:

first name	last name	email	country	organization	web site	corresponding?
Dena	Leshkowitz	dena.leshkowitz@weizmann.ac.il	Israel	Weizmann Institute of Science		✓
Ester	Feldmesser	ester.feldmesser@weizmann.ac.il	Israel	Weizmann Institute of Science		
Gilgi	Friedlander	gilgi.friedlander@weizmann.ac.il	Israel	Weizmann Institute of Science		
Ghil	Jona	ghil.jona@weizmann.ac.il	Israel	Weizmann Institute of Science		
Elena	Ainbinder	elena.ainbinder@weizmann.ac.il	Israel	Weizmann Institute of Science		
Yisrael	Parmet	iparmet@bgu.ac.il	Israel	Ben-Gurion University of the Negev		
Shirley	Horn-Saban	shirley@netium.com	Israel	Galil Genetic Analysis Ltd.		

When is a gene expressed?: RNA-Seq profiles bimodal hypothesis (tests on human stem cells).

Keywords: RNA-Seq, transcriptomics, gene expression profiling, human gene expression, human cell types, mesenchymal stem cells, MSC, functional genomics

Abstract: At present, there is not a clear answer to this simple and fundamental question: When is a gene expressed in a cellular system?, or When is a gene really ON or OFF in a cellular system where RNA expression is measured?. It seems that current high throughput RNA-Seq global transcriptomic profiling should be able to answer this question. However, a clear response cannot be found in popular scientific forums or even in HiTSeq published studies. To address this question, we present and discuss in this work the "bimodal hypothesis" that is derived from the analysis of RNA-Seq expression profiles. This hypothesis proposes that to discriminate between switched-ON and OFF genes in metazoan cells using deep sequencing data, the expression distributions can be described as a "bimodal function" that fits the density of FPKM values derived from RNA-Seq experiments and defines two main classes of gene expression levels: (i) the highly expressed genes with an active function in the cell, which can be considered "switched-ON genes"; and (ii) the poorly expressed genes, considered non functional (non active) genes, i.e. "switched-OFF genes" [Hebenstreit et al. Mol Syst Biol 2011]. We reproduced and analyzed a two-peak distribution over a set of RNA-Seq expression data of human mesenchymal stem cells (MSCs), and applying the ON/OFF switch concept, we clearly identified a bimodal behaviour and extracted a common signature of about 6,000 actively expressed genes (i.e. the ON gene-set for this cell type). This was done fitting the expression distribution to a bimodal curve and using a cutoff of log2(FPKMmeans)=2, that corresponds to 4 FPKM mean signal per gene loci (FPKM = Fragments Per Kilobase of transcript per Million mapped reads). The cutoff of 4 corresponded to 95% of the genes expression distribution considering any value of FPKMΓëÑ1. In fact, Mortazavi et al. [Nat Methods 2008] quantifying transcriptomes by RNA-Seq estimated that 3-4 FPKM corresponded to about one transcript per cell. As a general conclusion for RNA-Seq data, the measures of RNA abundance based on a large coverage of reads detected per locus (i.e. large number of FPKMs) generate distributions of accurate expression values that should resemble RNA absolute concentrations and allow discriminate between ON and OFF expressed genes. Other studies have found the bimodal behaviour of expression data in different transcriptomes of human cancer cell lines [Nagaraj et al. Mol Syst Biol 2011; Frenkel-Morgenstern et al. Genome Res 2012], providing also additional proteomic evidence to the proposed "bimodal hypothesis". More recent publications also support this hypothesis [Piccolo et al. PNAS 2013; Hart et al. BMC Genomics 2013], but there are still problems to identify expressed and functionally active genes [Liao & Weng PNAS 2015] and it would be quite good to have a broad and open discussion about this question within the forum of HiTSeq 2015.

Authors:

first name	last name	email	country	organization	web site	corresponding?
Beatr├¡z	Ros├│n-Burgo	beatriz.roson@usal.es	Spain	Cancer Research Center (CiC-IBMCC, CSIC/USAL/IBSAL)		✓
Katia P	Lopes	katiaplopes@gmail.com	Spain	Cancer Research Center (CiC-IBMCC, CSIC/USAL/IBSAL)		
Javier	De Las Rivas	jrivas@usal.es	Spain	Cancer Research Center (CiC-IBMCC, CSIC/USAL/IBSAL)	http://www.cicancer.org/	✓

Structural Variant Detection across Species Boundaries - Mapping-Based Horizontal Gene Transfer Detection from Sequencing Data

Keywords: Structural Variaion, Breakpoint Detection, Horizontal Gene Transfer, Pathogen Diagnostics

Abstract: One of the major fields in the analysis of NGS data is the detection of structural variants (SVs). The focus of SV detection has primarily been on human sequencing data, most prominently in cancer studies. Horizontal gene transfer (HGT) can be seen as a special case of SV. Through HGT, bacteria among other kingdoms are able to acquire novel genes from other, further related, bacteria or other species which often comes with new functions or properties such as antibiotic resistances [3]. A prominent example is the EHEC outbreak 2011 in Germany. Hence, fast and reliable pathogen identification or detection of antibiotic resistance are of particular interest in clinical diagnostics [1]. Integrated into an analysis pipeline, we use the split-read based SV detection tools Gustaf [4] and LASER [2] to identify possible breakpoints of HGT events. We then create candidate regions based on these breakpoints and incorporate read coverage information and read-pair based evidence to support the most likely candidates. We successfully evaluated our approach on two E.coli datasets where we could detect breakpoints and create HGT candidates even in the presence of multiple splits, complex variants and longer gaps between split parts. Transferring the concepts from SV detection methods to bacteria opens up new ways of diagnostics using NGS data, e.g. to distinguish parallel infections of multiple bacteria from single infections where the bacteria have acquired distinct DNA through HGT. We therefore see great potential in applying SV detection approaches across species boundaries.

References

[1] Allyson L. Byrd, Joseph F. Perez-Rogers, Solaiappan Manimaran, Eduardo Castro-Nallar, Ian Toma, Tim McCaffrey, Marc Siegel, Gary Benson, Keith A. Crandall, and William Evan Johnson. Clinical PathoScope: rapid alignment and filtration for accurate pathogen identification in clinical samples using unassembled sequencing data. BMC Bioinformatics, 15(1):262, 2014.

[2] Tobias Marschall and Alexander Schnhuth. Sensitive long-indel-aware alignment of sequencing reads. March 2013.

[3] Michael Syvanen. Evolutionary implications of horizontal gene transfer. Annu Rev Genet, 46:341-358, 2012.

[4] Kathrin Trappe, Anne-Katrin Emde, Hans-Christian Ehrlich, and Knut Reinert. Gustaf: Detecting and correctly classifying SVs in the NGS twilight zone. Bioinformatics, 30(24):3484-3490, Dec 2014.

Authors:

first name	last name	email	country	organization	web site	corresponding?
Kathrin	Trappe	TrappeK@rki.de	Germany	Robert Koch Institute	http://www.rki.de	
Tobias	Marschall	t.marschall@mpi-inf.mpg.de	Germany	Saarland University and Max Planck Institute for Informatics		
Bernhard	Renard	renardB@rki.de	Germany	Robert Koch Institute	http://www.rki.de	✓

Building bioinformatics pipelines for core facilities using SeqWare

Keywords: genomics, workflow, data management, pipeline, NGS

Abstract: The rapid development of Next Generation Sequencing (NGS) technologies has greatly accelerated the production of biological sequence data while at the same time stimulating development of bioinformatics tools for analysis. Complex, multi-step analyses using several such tools can become very hard to manage, especially when conducted on a larger scale.

The Genome Sequence Informatics team at OICR designs and implements software solutions for automated processing of sequence data generated at our institute. Projects run at OICR typically produce tens or hundreds of terabytes of data and thousands of files, which creates a strong demand for efficient data management. We develop computational pipelines (workflows) using SeqWare platform for parallel analysis of large data sets from sequencing experiments, both for basic and translational research.

SeqWare offers several features to solve the challenges of analysis at scale, including data tracking, automation, and scalability. The modular nature of SeqWare workflows allows them to be used in larger pipelines and evolve independently from each other. The SeqWare engine is built for distributed computing environments for maximum parallelization, but hides the intricacies of this environment from the casual user.

We describe the process of building SeqWare pipelines from a developerΓÇÖs perspective: using version control, overcoming SeqWare limitations by incrementally processing large data sets, rapid prototyping, continuous integration and testing to facilitate development of workflows and improve their stability. We also describe the advantages of centralized configuration files that can be used by multiple workflows, conditional launching workflows from other workflows, and developing reporting tools for better data tracking. We hope that the community finds our solutions for managing complex computational pipelines useful when dealing with similar challenges.

Authors:

first name	last name	email	country	organization	web site	corresponding?
Peter	Ruzanov	pruzanov@oicr.on.ca	Canada	Ontario Institute for Cancer Research	http://www.oicr.on.ca	✓
Michael	Laszloffy	michael.laszloffy@oicr.on.ca	Canada	Ontario Institute for Cancer Research	http://www.oicr.on.ca	
Dillan	Cooke	Dillan.Cooke@oicr.on.ca	Canada	Ontario Institute for Cancer Research	http://www.oicr.on.ca	
Xuemei	Luo	Xuemei.Luo@oicr.on.ca	Canada	Ontario Institute for Cancer Research	http://www.oicr.on.ca	
Morgan	Taschuk	Morgan.Taschuk@oicr.on.ca	Canada	Ontario Institute for Cancer Research	http://www.oicr.on.ca	

GSAseq: gene-set enrichment analysis of RNA-seq data accounting for read number bias

Keywords: RNA-seq, gene set enrichment analysis, GSEA, technical replicates, Read count bias, adjustment of read count bias

Abstract: We present a web-based tool, GSAseq, for gene-set enrichment analysis (GSEA) of RNA-seq count data for eight species. We show that RNA-seq count data from technical replicate samples as modeled by Poisson distribution suffer from serious read count bias. We show that such bias is a new source of false positives in GSEA that cannot be removed by sample permutations. On the other hand, read count data from different samples as modeled by over-dispersed Poisson distribution exhibits no such bias, which indicates GSEA can be safely applied to RNA-seq count data with biological replicates. GSAseq provides a simple algorithm to adjust for the read count bias of Poisson-modeled count data. Analysis of a fly RNA-seq dataset between male and female cells is illustrated.

Authors:

first name	last name	email	country	organization	web site	corresponding?
Sora	Yoon	yoonsora@unist.ac.kr	Korea, The Republic of	School of Life Science, Ulsan National Institutes of Science and Technology, Ulsan		
Seon-Kyu	Kim	seonkyu@kribb.re.kr	Korea, The Republic of	Medical Genomics Research Center, Korea Research Institute of Bioscience and Biotechnology, Daejeon		
Sang-Mun	Chi	smchiks@ks.ac.kr	Korea, The Republic of	School of Computer Science and Engineering, Kyungsung University, Busan		
Seon-Young	Kim	kimsy@kribb.re.kr	Korea, The Republic of	Medical Genomics Research Center, Korea Research Institute of Bioscience and Biotechnology, Daejeon		
Dougu	Nam	dougnam@unist.ac.kr	Korea, The Republic of	Ulsan National Institutes of Science and Technology		✓

AFAB: A flexible alignment benchmark

Keywords: reference mapping, diversity, benchmarking, sequence alignment, tool comparison

Abstract: Most high throughput sequencing (HTS) read mapping tools are designed to find the most probable location(s) of reads in a very similar reference sequence, such as a eukaryotic genome of the same species. However, the growing use of HTS in fields such as metagenomics and viral genotyping often results in read mapping tools being used in cases where significant genuine biological diversity exists between the reference genome and the genome being sequenced.

Testing a mapping tools ability to correctly reconstruct diversity, in the form of SNPs and indels, is hampered by the approach taken by most existing benchmarks. Generally, HTS reads are simulated directly from the reference sequence to produce "gold standard" mapping locations, and biological diversity in the reads, if added at all, is simulated at random. The focus is thus on the correction of sequencing errors rather than on correctness in the presence of genuine diversity.

We present AFAB, a novel, easy-to-use benchmarking tool for HTS read mapping. AFAB simulates reads from a user-selected alignment of a reference sequence to sample target sequences, reproducing the true biological diversity expected in a mapping experiment. This enables assessment of mapping tools on their ability to call SNPs and indels in biologically realistic data, by measuring whether reads generated from the target sequence are mapped to the reference sequence such that they correctly reconstruct the given alignment.

AFAB can be used both as a proof-of-concept for a forthcoming mapping experiment and to refine the settings of a chosen mapping tool to a specific problem. We use AFAB to assess a range of mapping tools in three cases: mapping of reads simulated from a coding subsequence of HIV-1 subtype C to its subtype B orthologue; a simulated metagenomic study of hepatitis C and related viruses; and mapping of reads sequenced from the chimpanzee BRCA-1 gene to the human genome.

Authors:

first name	last name	email	country	organization	web site	corresponding?
Imogen	Wright	imogen@sanbi.ac.za	South Africa	South African National Bioinformatics Institute	http://hiv.sanbi.ac.za	✓
Simon	Travers	simon@sanbi.ac.za	South Africa	South African National Bioinformatics Institute	http://hiv.sanbi.ac.za	

Genotype Calling and Imputation for Large Scale Sequencing Studies.

Keywords: Genotyping, Next Generation Sequencing, Big Data, Large Sample Sizes

Abstract: New sequencing technologies have made it feasible to conduct sequencing studies on cohorts of 10,000 individuals or more. However leveraging these large sample sizes is difficult for most imputation and phasing algorithms, where the computational time required grows roughly quadratically with the number of samples. In order to use large data sets effectively (e.g. to study low frequency variants, provide input to phasing algorithms and do population scale GWAS), we would prefer an algorithm for which the complexity grows linearly with increasing sample size.

Based on the work of Menelaou and Marchini, we propose a method to impute genotypes from a reference panel. This method uses a preprocessing step linear in the number of samples to calculate allele frequencies and the covariance matrix. This is followed by a genotype calling algorithm that assumes a multivariate normal distribution for the genotype values. This step is independent of panel size so the allele frequency and covariance matrices for a panel need to be computed once only and can then be used to genotype individuals in constant time.

The accuracy of this method is similar to existing approaches and the excellent scaling with cohort size makes it practical to use with reference panels consisting of tens of thousands of individuals.

Authors:

first name	last name	email	country	organization	web site	corresponding?
Rudy	Arthur	rarthur@illumina.com	United Kingdom	Illumina		✓
Anthony	Cox	ACox@illumina.com	United Kingdom	Illumina		
Ole	Schulz-Trieglaff	oschulz-trieglaff@illumina.com	United	Illumina		
Jared	O'Connell	joconnell@illumina.com	United	Illumina		

Characterization of fusion genes and the significantly expressed fusion isoforms in breast cancer by hybrid sequencing

Keywords: PacBio sequencing, hybrid sequencing, fusion gene, fusion isoform, fusion splice, IDP-fusion, MCF-7, breast cancer

Abstract: We developed an innovative hybrid sequencing approach, IDP-fusion, to detect fusion genes, determine fusion sites and identify and quantify fusion isoforms. IDP-fusion is the first method to study gene fusion events by integrating Third Generation Sequencing long reads and Second Generation Sequencing short reads. We applied IDP-fusion to PacBio data and Illumina data from the MCF-7 breast cancer cells. Compared with the existing tools, IDP-fusion detects fusion genes at higher precision and a very low false positive rate. The results show that IDP-fusion will be useful for unraveling the complexity of multiple fusion splices and fusion isoforms within tumorigenesis-relevant fusion genes.

Authors:

first name	last name	email	country	organization	web site	corresponding?
Jason L.	Weirather	jason-weirather@uiowa.edu	United States of America	University of Iowa		
Pegah	Tootoonchi Afshar	pegahta@stanford.edu	United States of America	Stanford University		
Tyson A.	Clark	tclark@pacificbiosciences.com	United States of America	Pacific Biosciences		
Elizabeth	Tseng	etseng@pacificbiosciences.com	United States of America	Pacific Biosciences		
Linda S.	Powers	linda-powers@uiowa.edu	United States of America	University of Iowa		
Jason	Underwood	jundy@uw.edu	United States of America	University of Washington		
Joseph	Zabner	Joseph-Zabner@uiowa.edu	United States of America	University of Iowa		
Jonas	Korlach	jkorlach@pacificbiosciences.com	United States of America	Pacific Biosciences		
Wing Hung	Wong		United States of America	Stanford University		
Kin Fai	Au	kinfai-au@uiowa.edu	United States of America	University of Iowa	http://www.healthcare.uiowa.edu/labs/au/	✓

Past their primates - comparative evolutionary genomics of great ape Y chromosomes

Keywords: Y chromosome assembly, comparative genomics of primates, pacbio long read data, flow sorted data enrichment algorithm

Abstract: The male-specific region of Y chromosomes (MSY) of chimpanzee and human have been found to be highly divergent with more than 30% of non-homologous sequences. In contrast, the female genomes of the four sequenced great ape species - human, chimpanzee, gorilla, and orangutan - have diverged from each other by less than 3%. To resolve this dichotomy, the remaining Y chromosomes need to be assembled, which would complement the existing reference sequences for human and chimpanzee euchromatic Y - thus enabling comprehensive great ape MSY comparative analysis.

In this study, we sequenced whole genome amplified flow-sorted DNA from the gorilla and orangutan Y chromosomes with both short-read (Illumina) and long-read (PacBio) technologies. We developed in silico strategies for increasing the amount of available Y chromosomal data, by utilizing differential coverage-based information. The Y-specific sequence data was then assembled with a variety of short read assemblers (SPAdes, DISCOVAR de novo), scaffolding software (SSPACE), hybrid tools (PBJelly, SSPACE-LR) and long-read assemblers (HGAP+Celera, Falcon).

This process led us to obtain novel insights into the optimal recipe for primate Y chromosome assembly, combining Illumina and PacBio technologies. Utilizing the generated draft Y assemblies, we estimate the divergence level, evaluate copy number of ampliconic genes, and detected rearrangements between great ape Y chromosomes. Our results indicate that great ape Y chromosomes are remarkably different in size, repeat content, and gene variation. We also demonstrate the utility of the novel Y chromosome sequences to conservation genetics.

Authors:

first name	last name	email	country	organization	web site	corresponding?
Samarth	Rangavittal	szr165@psu.edu	United States of America	The Pennsylvania State University		✓

Pan-cancer analyses reveal biologically and clinically relevant lincRNAs for tumor diagnosis, subtyping and prognosis

Keywords: lincRNA, pan-cancer, biomarker, diagnosis, prognosis

Abstract: Although an increasing number of long intergenic noncoding RNAs (lincRNAs) have been implicated in cancers, their pan-cancer biomarker application has not yet been reported. To seek a panel of lincRNAs as pan-cancer biomarkers, we have analyzed transcriptomes from over 3300 cancer samples with clinical information. Compared to mRNA, lincRNAs exhibit significantly higher tissue specificities which are then lost in cancer tissues. Moreover, lincRNA clustering results accurately classify tumor subtypes. Using RNA-Seq data from 1240 tumor and adjacent normal samples of 12 different cancer types from The Cancer Genome Atlas (TCGA), we identify six lincRNAs as pan-cancer diagnostic biomarkers (AUC=0.947). These lincRNAs are robustly validated (AUC as high as 0.972) using pan-cancer samples from a total of 15 RNA-Seq and microarray data sets. The expression levels of these six lincRNAs have significant prognostic values in various cancers. In summary, our study highlights the emerging role of lincRNAs as potentially powerful pan-cancer biomarkers and represents a significant leap forward in understanding the biological and clinical functions of lincRNAs in cancers.

Authors:

first name	last name	email	country	organization	web site	corresponding?
Travers	Ching	traversc@gmail.com	United States of America	University of Hawaii Cancer Center		
Lana	Garmire	lgarmire@cc.hawaii.edu	United States of America	University of Hawaii Cancer Center		✓

Integrative analysis of ChIP-seq datasets and identification of regions of differential coverage using echipp

Keywords: echipp, ChIP-seq, histone modifications, transcription factor binding sites, pipeline, visualization

Abstract: We developed the open source R-based software Echipp Easy ChIP-seq analysis Pipeline which translates study goals to tasks for visualization and statistical comparisons of sample groups. Our solution helps researchers with limited knowledge in bioinformatics and statistics to get an overview of their datasets, compare with previously published results, and identify subgroup-specific alterations in the landscape of histone modifications.

Our framework incorporates widely used tools for next-generation sequencing analysis, including, among others, the aligners bowtie and bowtie2, fastqc, a variety of peak calling algorithms, and MEDIPS for differential coverage analysis. Echipp manages the execution and monitoring of these tools on a single machine or in an environment of a computational cluster. Another feature that distinguishes our approach from all other available ChIP-seq processing pipelines, is the generation of comprehensive HTML-based interactive reports on quality control, genomic coverage, intra- and intergroup variability, differential coverage, gene set enrichment analysis, and much more. The process of analysis and report generation is fully automated; it can be run using more than one genome assembly. Input is a sample annotation table and corresponding sequence files. Similarly to a Galaxy workflow, every analysis can be customized using a predefined set of analysis options.

In a recent study, we used echipp to integrate three datasets (56 samples in total) on histone modifications of heart tissue and isolated cardiomyocytes. The generated reports show the degree of interoperability between the studies and give valuable insights in cardiomyocyte-specific epigenetic regulation.

Authors:

first name	last name	email	country	organization	Web site	corresponding?
Yassen	Assenov	y.assenov@dkfz.de	Germany	German Cancer Research Center	http://www.computational-epigenomics.com	✓
Daniel	Finke		Germany	University of Heidelberg		
Johannes	Backs		Germany	University of Heidelberg		
Christoph	Plass	c.plass@dkfz.de	Germany	German Cancer Research Center		

A Set Of Computational Tools For Somatic Mutation Analysis

Keywords: Bioinformatics, Software, Cancer, NGS, Sequence Analysis, Mutation Detection

Abstract: Cancer is caused by somatically acquired mutations accumulating in cells causing uncontrolled growth. We have developed a suite of software tools to detect and annotate these somatic mutations in NGS cancer data. Our team has successfully developed algorithms to detect single base substitutions (CaVEMan), insertions and deletions (cgpPindel, a modified version of Pindel), copy number changes (ASCAT NGS), structural variations (BRASS) and accompanying filters for each algorithm to remove false positives. The resulting high quality datasets can be annotated with gene information and predicted effect at the cDNA and protein level using VAGrENT and Grass.

These software tools and some accompanying utilities for NGS data are freely available on Github (http://github.com/cancerit) under an Open Source License agreement. Users and collaborators are welcomed to help maintain and improve this codebase.

Authors:

first name	last name	email	country	organization	web site	corresponding?
David	Jones	drj@sanger.ac.uk	United Kingdom	Wellcome Trust Sanger Institute		✓
Jon	Teague	jwt@sanger.ac.uk	United Kingdom	Wellcome Trust Sanger Institute		
Adam	Butler	apb@sanger.ac.uk	United Kingdom	Wellcome Trust Sanger Institute		
Keiran	Raine	kr2@sanger.ac.uk	United Kingdom	Wellcome Trust Sanger Institute		
Andrew	Menzies	am3@sanger.ac.uk	United Kingdom	Wellcome Trust Sanger Institute		
Lucy	Stebbings	las@sanger.ac.uk	United Kingdom	Wellcome Trust Sanger Institute		
Matthew	Astley	mca@sanger.ac.uk	United Kingdom	Wellcome Trust Sanger Institute		
Kathryn	Beal	kb3@sanger.ac.uk	United Kingdom	Wellcome Trust Sanger Institute		
Shriram	Bhosle	sb43@sanger.ac.uk	United Kingdom	Wellcome Trust Sanger Institute		
Jilur	Ghori	mg15@sanger.ac.uk	United Kingdom	Wellcome Trust Sanger Institute		
Angela	Matchan	am26@sanger.ac.uk	United Kingdom	Wellcome Trust Sanger Institute		
Rebecca	Shepherd	rpe@sanger.ac.uk	United Kingdom	Wellcome Trust Sanger Institute		
Jorge	Zamora	jz1@sanger.ac.uk	United Kingdom	Wellcome Trust Sanger Institute		
Peter	Van Loo	pvl@sanger.ac.uk	United Kingdom	Wellcome Trust Sanger Institute		
David	Wedge	dw9@sanger.ac.uk	United Kingdom	Wellcome Trust Sanger Institute		
Helen	Davies	hrm@sanger.ac.uk	United Kingdom	Wellcome Trust Sanger Institute		
Patrick	Tarpey	pst@sanger.ac.uk	United Kingdom	Wellcome Trust Sanger Institute		
Serena	Nik-Zainal	snz@sanger.ac.uk	United Kingdom	Wellcome Trust Sanger Institute		
Ultan	McDermott	um1@sanger.ac.uk	United Kingdom	Wellcome Trust Sanger Institute		
Michael	Stratton	mrs@sanger.ac.uk	United Kingdom	Wellcome Trust Sanger Institute		
Peter	Campbell	pc8@sanger.ac.uk	United Kingdom	Wellcome Trust Sanger Institute		

WTSI Imputation Service and PBWT (poster)

Keywords: imputation, pbwt, 1000 Genomes, UK10k, Haplotype Reference Consortium

Abstract: Genotype imputation infers missing genotypes in samples from a reference panel, so can fill in full sequences from genome-wide association study (GWAS) data. It is central to modern genetic association studies, supporting meta-analysis and increasing coverage and power. Reference datasets used in imputation analyses usually come from public resources, such as HapMap or 1,000 Genomes Project. With growing number of large sequencing projects, the accuracy of imputation could be increased by including broader range of SNPs. However, because the sample sets in these projects are often of sensitive nature, the sequenced genotype data cannot easily be made publicly available. Nonetheless, even such confidential data can become a valuable resource for imputation.

Here we present a public service developed at WTSI which allows imputation into a choice of reference panels, including 1,000 Genomes Phase 3, UK10K, and the Haplotype Reference Consortium. The imputation server is powered by the PBWT algorithm, which is extremely fast and efficient, and can be used to impute from hundreds of thousands of reference haplotypes.

Authors:

first name	last name	email	country	organization	web site	corresponding?
Petr	Danecek	pd3@sanger.ac.uk	United Kingdom	Wellcome Trust Sanger Institute		✓
Shane A.	McCarthy	sm15@sanger.ac.uk	United Kingdom	Wellcome Trust Sanger Institute		
Richard	Durbin	rd@sanger.ac.uk	United Kingdom	Wellcome Trust Sanger Institute		

A simple iterative approach for assembling large metagenomic datasets.

Keywords: metagenomic, large datasets, assembling, bioinformatics

Abstract: One of the issues faced in the analysis of metagenomic samples is the sheer size of the resulting datasets. This impacts researchers who are then required to gain access to larger computational resources or face the prospect of sub-optimal assemblies. Faced this issue we implemented a simple iterative assembly approach, which divides the assembly problem into smaller, more manageable components.

Our methodology involves creating random subsamples of the sequencing data for which it is possible to generate an assembly given the available computing resources. The proportion of the entire sequencing data that aligns to the subset assembly is determined and the remaining unaligned sequencing data is used for the creation of the subset for the next round. This process is continued until the proportion of original sequencing data that aligns to the assemblies created, reaches a predetermined level.

We applied this approach to a complex rumen metagenomic dataset consisting of 1Tb of raw sequencing data. Following 6 iterative rounds of assembly an 86% alignment rate was achieved. Examination of the taxonomic and functional composition of the assemblies of each consecutive round revealed some interesting trends. For instance, the proportional representation of each taxon changed in each iteration and in the early rounds were not good representations of the final taxonomic composition of the microbiome. This suggests that metagenomic datasets with low sequencing coverage may be subject to assembly biases for both the taxonomic groups present and the functions they possess.

Authors:

first name	last name	email	country	organization	web site	corresponding?
Thomas	Hitch	th.tomhitch@gmail.com	United Kingdom	Aberystwyth University		✓
Chris	Creevey	chc30@aber.ac.uk	United Kingdom	Aberystwyth University		

RNF: a method and tools to evaluate NGS read mappers

Keywords: NGS, bioinformatics, read mapping, evaluation of mappers, read simulation

Abstract: Aligning reads to a reference sequence is a fundamental step in numerous bioinformatics pipelines. As a consequence, the sensitivity and precision of the mapping tool, applied with certain parameters to certain data, can critically affect the accuracy of produced results (e.g., in variant calling applications). Therefore, there has been an increasing demand of methods for comparing mappers and for measuring effects of their parameters.

Read simulators combined with alignment evaluation tools provide the most straightforward way to evaluate and compare mappers. Simulation of reads is accompanied by information about their positions in the source genome. This information is then used to evaluate alignments produced by the mapper. Finally, reports containing statistics of successful read alignments are created. In default of standards for encoding read origins, every evaluation tool has to be made explicitly compatible with the simulator used to generate reads.

In order to solve this obstacle, we have created a generic format RNF (Read Naming Format) for assigning read names with encoded information about original positions.

Futhermore, we have developed an associated software package RNFtools containing two principal components. MISHMASH applies one of popular read simulating tools (among DWGSIM, ART, MASON, CURESIM etc.) and transforms the generated reads into RNF format. LAVENDER evaluates then a given read mapper using simulated reads in RNF format. A special attention is payed to mapping qualities that serve for parametrization of ROC curves, and to evaluation of the effect of read sample contamination.

Authors:

first name	last name	email	country	organization	web site	corresponding?
Karel	Brinda	karel.brinda@univ-mlv.fr	France	LIGM Université Paris-Est	http://brinda.cz	✓
Valentina	Boeva	valentina.boeva@curie.fr	France	Institut Curie		
Gregory	Kucherov	gregory.kucherov@univ-mlv.fr	France	LIGM/CNRS Université Paris-Est		

Integration of Carbon Copy Chromatin Conformation Capture (5C) and ChIP-Seq profiles reveal a high-resolution spatial genomic proximity network controlling epidermal keratinocyte differentiation

Keywords: 5C, ChIP-Seq, epigenetics, chromatin architecture

Abstract: During development, the execution of distinct cell differentiation programs is accompanied by establishing specific higher-order chromatin arrangements between the genes and their regulatory elements. The Epidermal Differentiation Complex (EDC) locus contains multiple co-regulated genes involved in the epidermal keratinocyte (KC) differentiation. Here we applied a probabilistic approach for the investigation of properties of chromatin architecture. Furthermore, we characterize the high-resolution spatial genomic proximity network of a 5Mb region containing the EDC and its flanking regions in mouse epidermal KCs. This was done by integration of data obtained from the 5C experiments and a set of eighteen ChIP-Seq profiles for histone modifications, chromatin architectural and remodeling proteins. The analysis reveals that a substantial number of the spatial interactions at the EDC overlap with chromatin states involving regulators of gene transcription and chromatin architecture. These include different combinations of transcription factors acting predominantly as a transcription repressors in keratinocytes (Cebpa, Cebpb, Mxi1, Ovol2), co-repressor chromatin re-modelers (Sin3a, Kdm5) and higher order chromatin folding regulators (Ctcf, Rad21, Satb1)

We confirmed by using both 5C and 3D FISH that chromatin at the 5Mb genome locus spanning the EDC and its flanking regions form several topologically associated domains (TADs) with similar borders. Moreover, it showed markedly different infra-domain folding in KCs versus thymocytes (TC), e.g. two adjacent TADs at the EDC central part were more condensed and non-randomly folded in KCs versus TCs.

In summary, our integrative approach allows us to suggest an involvement of the chromatin architecture and remodeling proteins into the spatial interaction network of gene cis-regulatory regions controlling co-ordinated gene expression at the EDC locus. It provides an important platform for further studies of the higher order chromatin folding at KC-specific genomic loci involved in controlling gene expression programmes in skin epithelia in health and disease.

Authors:

first name	last name	email	country	organization	web site	corresponding?
Krzysztof	Poterlowicz	K.Poterlowicz1@ bradford.ac.uk	United	University of Bradford		✓
Joanne	Yarker		United	University of Bradford		
Bryan	Lajoie		United States of America	University of Massachusetts, Medical School		
Igor	Malaschchuk		United Kingdom	University of Bradford		
Andrei	Mardaryev		United Kingdom	University of Bradford		
Andrey	Sharov		United States of America	Boston University		
Job	Dekker		United States of America	University of Massachusetts, Medical School		
Vladimir	Botchkarev		United Kingdom	University of Bradford		
Michael	Fessing		United	University of Bradford		

ICE: an iterative, probabilistic, de novo clustering algorithm for error correcting long transcript sequences with variable error rates

Keywords: error correction, clustering algorithm, transcriptome, long reads, fusion transcripts, cancer cell line, PacBio

Abstract: Third-generation sequencing technologies such as Pacific Biosciences' SMRT® Sequencing produces long reads that enable sequencing full-length transcripts up to 10 kb. While this eliminates the need for transcript assembly, the sequencing reads have variable error rates dominated by indels. Even with circular consensus calling, the intra-molecule consensus can still have high error rates, especially in cases where the transcript is long (8 – 10 kb). High error rates prevent accurate mapping back to the genome and preclude useful analysis when there is no genome. Thus, one solution is to create inter-molecule consensus (clustering of transcript sequences) to achieve high accuracy without using a genome.

I describe a de novo algorithm, ICE, for clustering transcript sequences of variable error rates. The input to the algorithm is a set of transcript sequences of any length from a few hundred bases to thousands of kilobases, with variable error rates from 0 – 15%. Each transcript sequence could either represent a fragment or the intact full-length form of a transcript. The goal of the algorithm is to cluster all sequences that originated from the same transcript isoform so that each cluster represents exactly one unique isoform. Importantly, the clustering must distinguish between alternative splice forms, which may have minor differences, such as a skipped 10-bp exon, that can be difficult to distinguish when indel rates are high. Given the alignment of two transcript sequences, I call a "homology hit" if the alignment gaps (sub/ins/del) can be explained by the presence of lower quality values in the sequenced region. This is implemented using a linear time algorithm to detect the presence of large stretches of indels in the alignment.

The clustering algorithm consists of an initial "seed" clustering phase, followed an iterative reassignment – merging phase. In the initial phase, an undirected acyclic graph is constructed where each node represents a sequence and each connecting edge represents a homology hit. A maximal clique finding algorithm is applied to partition the graph into the initial "seed" clusters. A transcript consensus is generated for each cluster, and each sequence's probability of belonging to a consensus calculated. In the subsequent iterative phase, sequences can be reassigned to different clusters, and different clusters can be merged, until a predefined number of iterations or a local maximum is reached.

I show that the algorithm is capable of improving consensus accuracy of 10 kb transcripts from the initial intra-molecule accuracy of 85% to 99-100% with a coverage of 15 – 20. For 5 kb transcripts, a coverage of 10 is sufficient. I show how the algorithm is applied to whole transcriptome datasets generated using the PacBio® SMRT Sequencing technology including the human MCF-7 breast cancer cell line and the fungus Plicaturopsis crispa as well as targeted sequencing of human genes. I show how obtaining a high-quality, full-length, transcript isoform structure enables one to find interesting biological phenomenon such as cancer fusion events, polycistronic transcription, and anti-sense noncoding RNAs.
To conclude, ICE provides a clustering solution for high error sequencing reads of transcriptome data without using a reference genome.

Authors:

first name	last name	email	country	organization	Web site	corresponding?
Elizabeth	Tseng	magdoll@gmail.com	United States of America	Pacific Biosciences		✓
Tyson	Clark	tclark@pacb.com	United States of America	Pacific Biosciences.com		

Dominant protein isoforms from the APPRIS WebServer and WebServices

Keywords: Alternative splicing, Dominant isoforms, Genome annotation

Abstract: We have developed a web server and web services to support the splice isoform annotations in the APPRIS database (http://appris.bioinfo.cnio.es). The APPRIS database houses annotations of splice isoforms for five different vertebrate genomes and Drosophila, and the APPRIS WebServer and WebServices allow users to extend APPRIS annotations to any other vertebrate species.

APPRIS uses protein structural and functional features and conservation information to annotate splice isoforms and to select a single isoform as the main protein isoform for each protein-coding gene. This main isoform has the most conserved protein features and most evidence of cross-species conservation, while those isoforms with unusual, missing or non-conserved protein features are flagged as alternative. APPRIS principal isoforms have been shown to agree overwhelmingly with the main protein isoform detected in proteomics experiments.

The APPRIS WebServer allows users to annotate splice isoforms for individual genes while the APPRIS WebServices provide researchers with the ability to generate annotations automatically in high throughput mode and to interrogate the annotations in the APPRIS Database in an automatic fashion.

Authors:

first name	last name	email	country	organization	Web site	corresponding?
Jose Manuel	Rodriguez	jmrodriguez@cnio.es	Spain	Spanish National Bioinformatics Institute (INB-CNIO)	http://www.inab.org	✓
Angel	Carro	acarro@cnio.es	Spain	Spanish National Cancer Research Centre (CNIO)		
Alfonso	Valencia	avalencia@cnio.es	Spain	Spanish National Cancer Research Centre (CNIO)	http://www.cnio.es	
Michael	Tress	mtress@cnio.es	Spain	Spanish National Cancer Research Centre (CNIO)		✓

Computational and experimental methods for identifying GxE determinants of gene expression

Keywords: gene environment interaction, eQTL, allele specific expression, gene regulatation response

Abstract: Recent studies have shown that GxE interactions can be detected when studying molecular phenotypes (e.g. infection response eQTLs in immune cells) that are relevant for complex traits (e.g., inmune system related diseases). Despite these relevant examples, the extent to which the environment can modulate genetic effects on quantitative phenotypes is still to be defined.

Here we have devised a high-throughput approach to achieve a comprehensive characterization of GxE interactions in humans. To this end we have investigated the transcriptional response to 50 treatments in 5 different cell types (for a total of 250 cellular environments). We then selected the environmental conditions (cell type/treatment) with relevant changes in gene expression and collected deep sequencing data to fully characterize the transcriptional response to environmental changes and to identify genes showing allele specific expression (ASE). We observe that treatments with similar biochemical properties (e.g. nuclear receptor ligands or metal ions) tend to cluster together in principal component analysis (PCA), which also demonstrates cell-type specific transcriptional changes upon environmental perturbation.

We analyzed allele specific expression (ASE) using a new computational approach QuASAR, that allows for joint genotyping and allele specific analysis on RNA-seq data. Across 56 cellular environments we discovered 9548 instances of ASE (FDR<10%), corresponding to 8923 unique ASE genes. We found that in an individual sample, on average, 0.5% of genes with heterozygous SNPs are ASE genes. We then developed a Bayesian model across treatments and cell types to identify genes regulated through GxE interactions (conditional-ASE). Consistent with previous analyses of condition-specific eQTLs, we observe that the majority of ASE is consistent across conditions. For a given gene, the probability of ASE is negatively correlated with average expression, with a 4.2 fold decrease per 10x increase in FPKMs. On the other hand, we find a 1.3 fold increase in probability of ASE per 2x change in expression in response to treatments. When we consider a Bayes factor measuring evidence in support of GxE interaction, we find 227 control-only ASE and 245 treatment-only ASE genes. We observe also a trend for increasing evidence of conditional-ASE for genes with larger differential expression. These results provide the first characterization of ASE across a large number of environmental exposures and will contribute to the understanding of how GxE interactions have shaped human phenotypes in different environments and underlie variation in complex traits.

Authors:

first name	last name	email	country	organization	web site	corresponding?
Roger	Pique-Regi	rpique@wayne.edu	United States of America	Wayne State University		✓
Gregory	Moyerbrailean	ex7689@wayne.edu	United States of America	Wayne State University		
Christopher	Harvey	cbboipdx@gmail.com	United States of America	Wayne State: Center Molecular Medicine & Genetics		
Omar	Davis	odavis@wayne.edu	United States of America	Wayne State University		
Donovan	Watza	dwatza@med.wayne.edu	United States of America	Wayne State University		
Xiaoquan	Wen	xwen@umich.edu	United States of America	Univ. of Michigan		
Francesca	Luca	fluca@wayne.edu	United States of America	Wayne State University		✓

Data-Parallel De Bruijn Graph Construction For De Novo Short Read Micro-Assembly on the GPU

Keywords: micro-assembly, realignment, de Bruijn graph construction, GPU implementation, data parallel

Abstract: We present a novel data-parallel algorithm for de Bruijn graph construction on the GPU in the context of short read micro-assembly. This task is an important step of many existing variant discovery pipelines designed to identify mutations associated with various genetic diseases (e.g. Mendelian disorders or cancer) from high-throughput sequencing (HTS) datasets. In this work we specifically focus on the popular Genome Analysis Toolkit (GATK) HaplotypeCaller implementation, which was shown to outperform many other existing variant detection algorithms. Since typically the micro-assembly procedure is executed independently on many separate windows of the genome, a coarse-grained parallel implementation can be easily achieved by executing the original (highly) sequential algorithm on each window in parallel. However, this approach suffers from poor load balancing and is not very well suited for the GPU. On the other hand, our algorithm is a fine-grained parallel implementation of the de Bruijn graph construction across any given number of genome windows. It relies on data-parallel primitives, such as (segmented) sort, scan, filter, and reduction by key to efficiently populate the graph data structure in the compressed sparse row (CSR) format. Our preliminary results show a 12x speedup when compared to the sequential HaplotypeCaller implementation, while exactly matching its generated graphs.

Authors:

first name	last name	email	country	organization	web site	corresponding?
Victoria	Popic	viq@stanford.edu	United States of America	Stanford		✓
Jacopo	Pantaleoni		United States of America	NVIDIA		
Nuno	Subtil		United States of America	NVIDIA		
Jonathan M	Cohen		United States of America	NVIDIA		
Mauricio	Carneiro		United States of America	Google		
Serafim	Batzoglou		United States of America	Stanford		
Kunle	Olukotun		United States of America	Stanford		

Assembling long reads without error correction using spaced seeds and local de Brujin graphs

Keywords: Assembly, Long NGS Reads, Spaced Seeds, Local de Brujin Graphs, Sequence Disambiguation, Overlap Layout Consensus, Oxford Nanopore Sequencing, Pacific Biosciences Sequencing

Abstract: With the Pacific Biosciences and Oxford Nanopore sequencing platforms, the era of long read, single molecule sequencing is finally upon us. These platforms promise to resolve issues related to short read lengths and to amplification biases seen in the Illumina sequencing platforms. However, the story of progress is not as linear, as these long reads currently have much higher sequencing errors. Thus, correcting errors in long reads is a common first step before sequence assembly; however read correction methods inherently collapse many variations in reads, such as those in near repeats. This limitation has motivated us to design a novel method of sequence assembly that does not rely on error correction, with the aim of distinguishing similar but distinct sequences during assembly.

Here we present a novel long read assembly method based on the Overlap-Layout-Consensus (OLC) paradigm to assemble long reads with high error rates.

Overlap: We first identify overlapping reads using a new algorithm based on spaced seeds. Spaced seeds are the state-of-the-art for approximate sequence matching, and they have been increasingly used to improve the quality and sensitivity in different applications. These include seed-and-extend based alignment methods, and count-based alignment-free sequence analysis methods. A spaced seed S is a string over the binary alphabet {1, *}, where 1 indicates a position in S where a match must occur, whereas * indicates a position where a mismatch is allowed. After constructing a spaced seed index of the reads, we use a count-based paradigm to detect overlaps. We demonstrate that this results in a fast and accurate detection of overlaps between long reads by tolerating their high error rates.

Overlay: We then estimate the overlap sizes using indexed spaced seed coordinates within reads. We also use this information to assert collinearity of matched spaced seeds to further reduce false overlaps. Using the lengths of overlaps, we then partition our data into components to describe contigs. For each contig, we place reads on a linear coordinate system using linear regression.

Consensus: Finally, we perform consensus calling on the overlayed reads, accounting for any indel error-induced sequence drifts. To perform efficient indel tolerant consensus calling we construct a series of local de Brujin graphs. These graphs are constructed from sequences within a small window size, and are deallocated for the next window once the local sequence is assembled. Designed to weave between errors, these graphs utilize the read of origin and location information for each k-mer, allowing us to disambiguate the graph even when using small k-mers sizes.

We demonstrate the performance of our algorithm using publicly available experimental Oxford Nanopore data on E. coli and S. cerevisiae genomes, and compare it to recent assembly algorithms for this new sequencing platform.

Authors:

first name	last name	email	country	organization	web site	corresponding?
Justin	Chu	cjustin@bcgsc.ca	Canada	Canada's Michael Smith Genome Sciences Centre, BCCA, UBC	http://bcgsc.ca	✓
Hamid	Mohamadi	hmohamadi@bcgsc.ca	Canada	Canada's Michael Smith Genome Sciences Centre,	http://bcgsc.ca	

				BCCA, UBC		
Shaun	Jackman	sjackman@bcgsc.ca	Canada	Canada's Michael Smith Genome Sciences Centre, BCCA, UBC	http://bcgsc.ca	
Ben	Vandervalk	benv@bcgsc.ca	Canada	Canada's Michael Smith Genome Sciences Centre, BCCA	http://bcgsc.ca	
Rene	Warren	rwarren@bcgsc.ca	Canada	Canada's Michael Smith Genome Sciences Centre, BCCA	http://bcgsc.ca	
Inanc	Birol	ibirol@bcgsc.ca	Canada	Canada's Michael Smith Genome Sciences Centre, BCCA, UBC, SFU	http://bcgsc.ca	

An integrative somatic mutation analysis to identify pathways linked with survival outcomes across 19 cancer types

Keywords: Somatic mutation, Clinical genomics, HiTSeq, Patient stratification, Data integration, Network analysis

Abstract:
Motivation:
Identification of altered pathways that are clinically relevant across human cancers is a key challenge in cancer genomics. Precise identification and understanding of these altered pathways may provide novel insights into patient stratification, therapeutic strategies and the development of new drugs. However, a challenge remains in accurately identifying pathways altered by somatic mutations across human cancers, due to the diverse mutation spectrum. We developed an integrative approach to integrate somatic mutation data with gene networks and pathways, in order to identify pathways altered by somatic mutations across cancers.
Results: We applied our approach to The Cancer Genome Atlas (TCGA) dataset of somatic mutations in 4,790 cancer patients with 19 different types of tumors. Our analysis identified cancer-type-specific altered pathways enriched with known cancer-relevant genes and targets of currently available drugs. To investigate the clinical significance of these altered pathways, we performed consensus clustering for patient stratification using member genes in the altered pathways coupled with gene expression datasets from 4,870 patients from TCGA and multiple independent cohorts confirmed that the altered pathways could be used to stratify patients into subgroups with significantly different clinical outcomes. Of particular significance, certain patient subpopulations with poor prognosis were identified because they had specific altered pathways for which there are available targeted therapies. These findings could be used to tailor and intensify therapy in these patients, for whom current therapy is suboptimal.

Authors:

first name	last name	email	country	organization	Web site	corresponding?
Sunho	Park		United States of America	University of Texas Southwestern Medical Center (*)		✓
Seung-Jun	Kim			University of Maryland at Baltimore		
Donghyeon	Yu			*		
Samuel	Pena-Llopis			*		
Jianjiong	Gao			Memorial Sloan-Kettering Cancer Center		
Jin Suk	Park			*		
Beibei	Chen			*		
Jessie	Norris			*		
Xinlei	Wang			Southern Methodist University		
Min	Chen			University of Texas at Dallas		
Jeongsik	Yong			University of Minnesota Twin Cities		
Zabi	Wardak			*		
Kevin	Choe			*		
Michael	Story			*		
Tompthy	Star					
Jae-Ho	Cheong			Yonsei University College of Medicine		
Tae Hyun	Hwang	taehyun.hwang@utsouthwestern.edu		*		

HapCol: Accurate and Memory-efficient Haplotype Assembly from Long Reads

Keywords: haplotype assembly, future-generation sequencing, dynamic programming

Abstract: ABSTRACT

Motivation: Haplotype assembly is the computational problem of reconstructing haplotypes in diploid organisms and is of fundamental importance for characterizing the effects of Single Nucleotide Polymorphisms (SNPs) on the expression of phenotypic traits. Haplotype assembly highly benefits from the advent of "future-generation" sequencing technologies and their capability to produce long reads at increasing coverage. Existing methods are not able to deal with such data in a fully satisfactory way, either because accuracy or performances degrade as read length and sequencing coverage increase, or because they are based on restrictive assumptions.

Results: By exploiting a feature of future-generation technologies – the uniform distribution of sequencing errors – we designed an exact algorithm, called HAPCOL, that is exponential in the maximum number of corrections for each SNP position and that minimizes the overall error correction core. We performed an experimental analysis, comparing HAPCOL with the current state-of-the-art combinatorial methods both on real and simulated data. On a standard benchmark of real data, we show that HAPCOL is competitive with state-of-the-art methods, improving the accuracy and the number of phased positions. Furthermore, experiments on realistically-simulated datasets revealed that HAPCOL requires significantly less computing resources, especially memory. Thanks to its computational efficiency, HAPCOL can overcome the limits of previous approaches, allowing to phase datasets with higher coverage and without the traditional all-heterozygous assumption.

Availability: Our source code is available under the terms of the GPL at http://hapcol.algolab.eu/.

Contact: {pirola,simone.zaccaria}@disco.unimib.it

Authors:

first name	last name	email	country	organization	Web site	corresponding?
Paola	Bonizzoni		Italy	University Milano-Bicocca		✓
Riccardo	Dondi			Univ. Bergamo		
Gunnar	Klau			Centrum Wiskunde & Informatica		
Yori	Pirola			University Milano-Bicocca		
Nadia	Pisanti			University of Pica		
Simone	Zaccaria	simone@disco.unimib.it		University Milano-Bicocca		

Skimdiff: Transcript-level Differential Analysis of RNA-Seq Data

Keywords: Bayesian, Transcription, Bioinformatics, Algorithms, Next-generation sequencing

Abstract: ABSTRACT

Motivation: Differential analysis of gene expression has been widely used in biomedical fields. The advent of RNA-Seq techniques has enabled accurate estimation of transcript abundance and has inspired recent development in new computational methods for transcript level differential analysis. However, these methods require aligning reads to a reference genome or transcriptome first, which is computationally expensive and is prone to alignment errors.

Results: We introduce an efficient method, Skimdiff, for transcript- level differential expression analysis using RNA-Seq data. Skimdiff selects a set of k-mers and uses their read counts directly to model transcript expression distribution and to infer differentially expressed transcripts using a Gibbs sampling algorithm. We demonstrate that this alignment-free method is able to detect differentially expressed transcripts with comparable accuracy to that of the state-of-the-art alignment-based methods but only uses a fraction of the computation time.

Authors:

first name	last name	email	country	organization	web site	corresponding?
Liuli	Chen		United States of America	UCLA		
Jennifer	Zhang		United States of America	UCLA		
Chelsea	Jui-Ting Ju		United States of America	UCLA		
Ruirui	Li		United States of America	UCLA		
Wenchao	Yu		United States of America	UCLA		
Wei	Wang	weiwang@cs.ucla.edu	United States of America	UCLA		✓

Informed kmer selection for de novo transcriptome assembly

Keywords: Sequence analysis, Next-generation sequencing, Algorithms

Abstract: Motivation: De novo transcriptome assembly is an integral part for many RNA-seq workflows. Common applications include sequencing of non-model organisms, cancer or meta transcriptomes. Most of these assemblers use the de Bruijn graph (DBG) as the underl- ying data structure. A fundamental parameter with large influence on assembly quality with DBGs is the exact word length k. As such no single kmer value leads to optimal results. Instead, DBGs over diffe- rent kmer values are build and the assemblies merged to improve sensitivity. However, no studies have investigated thoroughly the problem of automatically learning at which kmer value to stop the assembly. Instead a suboptimal selection of kmer values is often used by practitioners.

Here we investigate in detail the contribution of a single kmer in a multi-kmer approach looking at dozens of assemblies. We find that a comparative clustering approach of related assemblies allows to estimate the importance of an additional kmer assembly. We show that a model fit based approach with model selection works well for predicting the kmer value at which no further assemblies are necessary. We test the approach with different de novo assemblers for datasets with different coverage values and read lengths. Our approach is parameter-free and works completely de novo. Conclusion: We provide an automatic method for limiting the number of kmer values without a significant loss in assembly quality but with savings in assembly time. This is a step forward to making multi-kmer methods more reliable and easier to use.

Authors:

first name	last name	email	country	organization	web site	corresponding?
Dilip Ariyu	Durai		Germany	Cluster of Excellence on Multimodal Computing and Interaction		✓
Marcel	Schulz	mschulz@mmci.uni-saarland.de	Germany	Max Planck Institute for Informatics		

Intensive Deep Learning on GPUs for the Prediction of Anticancer Drug Sensitivity

Keywords: Anticancer drug sensitivity, Deep learning, CCLE, GPU

Abstract:
Motivation: The effectiveness of anticancer drugs varies according to the genetic background of patients. In previous studies of the Cancer Cell Line Encyclopedia, data on single nucleotide mutations, copy number variations, and gene expression profiles were generated, and the sensitivity of anticancer drugs was predicted by using the Elastic net regression method. However, the prediction accuracy achieved when only using the information of single nucleotide mutation profile was not necessarily high.

Methods: Deep learning, a highly precise machine learning technique, was employed to develop a predictor of a multi-layer neural network for anticancer drug sensitivity based on genetic variation data of the cell lines, using CCLE database. To speed up the program necessary for a large-scale calculation experiment, a program of the deep learning was implemented to work on GPU where a parallel calculation is efficiently possible. The most suitable hyper-parameters were searched for by using a grid search.

Results: In a comparison with other regression methods (elastic net and support vector regression), deep learning achieved the superior prediction accuracy. In addition, by performing parallelization using GPUs, the program was shown to be run in linear time, despite the computational complexity of learning and prediction in a neural network.

Authors:

first name	last name	email	country	organization	web site	corresponding?
Motoki	Abe		Japan	Department of Biosciences and Informatics		
Kohei	Harujama		Japan	Department of Biosciences and Informatics		
Yohei	Sugawara		Japan	Department of Biosciences and Informatics		
Kengo	Sato		Japan	Department of Biosciences and Informatics		
Yasobumi	Sakakibara	yasu@bio.keio.ac.jp	Japan	Department of Biosciences and Informatics		✓

Discovery of large genomic inversions using pooled clone sequencing

Keywords: pooled clone sequencing, inversion detection, structural variation

Abstract:
Motivation: There are many different forms of genomic structural variation that can be broadly classified into two groups as copy number variation (CNV) and balanced rearrangements. Although many algorithms are now available in the literature that aim to characterize CNVs, discovery of balanced rearrangements (inversions and translocations) remains an open problem. This is mainly because the breakpoints of such events typically lie within segmental duplications and common repeats, which reduce the mappability of short reads. The 1000 Genomes Project spearheaded the development of several methods to identify inversions, however, they are limited to relatively short inversions, and there are currently no available algorithms to discover large inversions using high throughput sequencing technologies (HTS). Results: Here we propose to use a sequencing method (Kitzman et al., 2011) originally developed to improve haplotype phasing to characterize large genomic inversions. This method, called pooled clone sequencing, merges the advantages of clone based sequencing approaches with the speed and cost efficiency of HTS technologies. Using data generated with pooled clone sequencing method, we developed a novel algorithm, dipSeq, to discover large inversions (>500 Kbp). We show the power of dipSeq first on simulated data, and then apply it to the genome of a HapMap individual (NA12878). We were able to accurately discover all previously known and experimentally validated large inversions in the same genome. We also identified a novel inversion, and confirmed using fluorescent in situ hybridization.
Availability: Implementation of the dipSeq algorithm is available at https://github.com/BilkentCompGen/dipseq

Authors:

first name	last name	email	country	organization	web site	corresponding?
Marzieh	Eslami Rasekh		Turkey	Bilkent University		✓
Giorgia	Chiatante			University of Bari		
Mattia	Miroballo			University of Bari		
Joyce	Tang			Benoraya Research Institue		
Mario	Ventura			University of Bari		
Chris	Amemiya			Benoraya Research Institue		
Evan	Eichler			University of Washington		
Francesca	Antonacci			University of Bari		
Can	Alkan	calkan@cs.bilkent.edu.tr		Bilkent University		

CNVera: an assembly-based tool for contig copy number estimation

Keywords: Genome analysis, Genomics, Next-generation sequencing

Abstract: Motivation: Copy Number Variation (CNV) is an important type of genomic variation but detecting CNV from the current high-throughput short read data remains a challenging computational task. Many software packages have been developed for CNV detection and mostly belong to two categories: the pair-end/split-read mapping based and the read-depth coverage based methods. While these CNV algorithms continue to be refined, they face an intrinsic difficulty in detecting CNV of repetitive regions since reads generated from these regions will map to multiple locations in the reference genome. Using the co-assembly approach, Nijkam et al., 2012 made an advance step in CNV detection by first assembled reads from multiple samples into contigs and reduced the problem of CNV detection to the problem of finding copy number of contigs in an assembly. However, the approach for estimating contig copy number in Nijkam et al., 2012 is still limited, since only the read-depth information in each contig is utilized to infer its copy number.
Results: We present CNVera, an assembly-based algorithm for contig copy number estimation. CNVera utilizes read-depth, read pair information, structure of the assembly graph, and a reference genome to infer the copy number of contigs in an assembly. We demonstrate that CNVera outperforms other existing tools in multiple bacterial and yeast data sets, especially for short and repetitive contigs. Contig copy number estimation is not only useful in CNV detection, but also important in various repeat-resolution, scaffolding and reference assisted assembly algorithms.

Authors:

first name	last name	email	country	organization	web site	corresponding?
Meleshko	Dmitrii			Saint Petersburg University		✓
Bushmanova	Vera			St. Petersburg State University		
Son	Pham	spham@salk.edu		UCSD		

Integration of String and de Buijn Graphs for Genome Assembly

Keywords: Algorithms, Genome, Genome analysis, Graph theory

Abstract: Motivation: String and de Bruijn graphs are two underlying graph models used by most genome assemblers. To date, none of existing assemblers clearly outperform the others across all data sets. We observed that, although string graph can make use of entire reads for resolving repeats, the de Bruijn graphs can naturally assemble through error-prone regions owing to sequencing bias.

Results: We develop a generalized assembler having both advantages of string and de Bruijn graphs. First, the reads are decomposed adaptively only in error-prone regions. Second, each paired-end read is extended into a long read directly using FM-index. The decomposed and extended reads are used to build a generalized assembly graph. In addition, several essential components of an assembler are designed or improved. The developed assembler has been fully parallelized, tested and compared with state-of-the-art assemblers using benchmark data sets. The results indicate this assembler is among the best on short read data sets, and outperforms the others over longer-read experiments. The accuracy is also high in comparison with others.

Authors:

first name	last name	email	country	organization	Web site	corresponding?
Yao-Ting	Huang		Taiwan	National Chung Cheng University, Department of Computer Science and Information Engineering		
Fu-Chen	Liao	ythuang@csie.ccu.edu.tw		National Chung Cheng University, Department of Computer Science and Information Engineering		

HapIso : An accurate method for the haplotype-specific isoforms reconstruction from long single-molecule reads

Keywords: Transcriptome, Alternative splicing, Gene expression, Population genetics

Abstract: ABSTRACT

Motivation: Sequencing of RNA provides the possibility to study an individualΓÇÖs transcriptome landscape and determine allelic expression ratios. Single-molecule protocols generate multi-kilobase reads longer than most transcripts allowing sequencing of complete haplotype isoforms. This allows partitioning the reads into two parental haplotypes. While the read length of the single-molecule protocols is long, the relatively high error rate limits the ability to accurately detect the genetic variants and assemble them into the haplotype-specific isoforms.

Results: In this paper, we present HapIso (Haplotype-specific Isoform Reconstruction), a method able to tolerate the relatively high error-rate of the single-molecule platform and partition the isoform reads into the parental alleles. Phasing the reads according to the allele of origin allows our method to efficiently distinguish between the read errors and the true biological mutations. HapIso uses a k-means clustering algorithm aiming to group the reads into two meaningful clusters maximizing the similarity of the reads within cluster, and minimizing the similarity of the reads from different clusters. Each cluster corresponds to a parental haplotype. We use family pedigree information to evaluate our approach. Experimental validation suggests that HapIso is able to tolerate the relatively high error-rate and accurately discriminate the reads into the paternal and maternal alleles of the isoform transcript. Furthermore, our method is the first method able to reconstruct the the haplotype-specific isoforms from long single-molecule reads.

Availability: The open source Python implementation of HapIso is freely available for download at http://genetics.cs. ucla.edu/hapiso/

Authors:

first name	last name	email	country	organization	web site	corresponding?
Serghei	Mangul	serghei@cs.ucla.edu	United States of America	UCLA		✓
Farhad	Hormozdiari			UCLA		
Elizabeth	Tseng			Pacific Biosciences		
Alexander	Zeliovsky			Georgia State University		
Eskin	Eleazar			UCLA		

Accelerating Long Read Alignment Using Locality Sensitive Hashing

Keywords: Alignment, Next-generation sequencing, HitSeq

Abstract: Summary: As long read technologies become more pervasive, there is a need for methods that can efficiently scale with the computational demands of the generated read datasets. Most current long read approaches use seed-and-extend or traditional hash-based techniques. Here we present TOTORO, a novel long read alignment algorithm based on nearest neighbor detection via locality sensitive hashing (LSH). We evaluate our method against several state-of-the-art aligners on simulated and real read datasets (including PacBio CCS and Illumina TruSeq reads). TOTORO achieves speedups of 2-6x over the existing aligners, while maintaining high accuracy.
Availability: http://viq854.github.com/srx
Contact: viq@stanford.edu

Authors:

first name	last name	email	country	organization	web site	corresponding?
Victoria	Popic	viq@stanford.edu	United States of America	Stanford		✓
Stephen	Miller			Stanford		
Serafim	Batzoglou			Stanford		

Scalable multi whole-genome alignment using recursive exact matching

Keywords: Algorithms, Alignment, Comparative genomics

Abstract: Motivation: The emergence of third generation sequencing technologies has brought near perfect de-novo genome assembly within reach. This clears the way towards reference-free detection of genomic variations.

Approach: In this paper, we introduce a novel concept for aligning whole-genomes which allows the alignment of multiple genomes. Alignments are constructed in a recursive manner, in which alignment decisions are statistically supported. Computational performance is achieved by splitting an initial indexing data structure into a multitude of smaller indices.

Results: We show that our method can be used to detect high resolution structural variations between two human genomes, and that it can be used to obtain a high quality multiple genome alignment of at least nineteen Mycobacterium tuberculosis genomes.

Authors:

first name	last name	email	country	organization	web site	corresponding?
Jasper	Linthorst	Jasper.linthorst@gmail.com	United States of America	VU University Medical Center		✓
Marc	Hulsman			VU University Medical Center		
Henne	Holstege			VU University Medical Center		
Marcel	Reinder			Delft University of Technology		

pacFAST: A sensitive alignment tool for single-molecule sequencing reads

Keywords: Sequence alignment, Next-generation sequencing, PacBio sequencing

Abstract: Motivation: Many recent advances in genomics and precision medicine have been made possible through the application of high throughput sequencing (HTS) to large collections of human genomes. Although HTS technologies have proven their use in cataloging human genome variation, computational analysis of the data they generate is still far from being perfect. The main limitation of Illumina and other popular sequencing technologies is their short read length relative to the lengths of (common) genomic repeats. Newer technologies such as Pacific Biosciences and Oxford Nanopore are producing longer reads, making it theoretically possible to overcome the difficulties imposed by repeat regions. Unfortunately, because of their high sequencing error rate (up to 35%), reads generated by these technologies are very difficult to work with and cannot be used in many of the existing standard downstream analysis pipelines. It is not only difficult to find the correct mapping locations of such reads in a reference genome, but also to establish the correct alignment and differentiate sequencing errors from real sequence variants. In order to overcome these problems, we introduce pacFAST, a novel long- read mapper that is specifically designed to align reads generated by PacBio sequencers to a reference. pacFAST employs various filters and techniques to efficiently align the long reads to the reference genome with high sensitivity.

Results: Our experiments indicate that pacFAST is more sensitive than the available alternatives such as PacBioΓÇÖs own BLASR. pacFAST can correctly map 1% more reads and align 5% more bases that any other mapper available. In its most sensitive mode, pacFAST is on par with BLASR in terms of running time, but can be made to run faster by tuning its parameters to make it less sensitive. Thus pacFAST provides the user the ability to tradeoff between speed and sensitivity, which can come handy in certain applications. Availability: pacFAST is implemented in C and supports multi- threading. The source code of pacFAST is available from https://bitbucket.org/compbio/pacfast

Authors:

first name	last name	email	country	organization	web site	corresponding?
Ehsan	Haghshenas		Canada	Simon Fraser University		
Faraz	Hach	fhach@sfu.edu		Simon Fraser University		✓
Iman	Sarrafi			Simon Fraser University		
S. Cenk	Sahinalp			Simon Fraser University		

Optimal Seed Solver: Optimizing Seed Selection in Read Mapping

Keywords: sequence alignment, Algorithms, Alignment

Abstract: Motivation: Optimizing seed selection is an important problem in read mapping. The number of non-overlapping seeds a mapper selects determines the sensitivity of the mapper while the total frequency of all selected seeds determines the speed of the mapper. Modern seed-and-extend mappers usually select seeds with either an equal and fixed-length scheme or with an inflexible placement scheme, both of which limit the potential of the mapper to select less frequent seeds to speed up the mapping process. Therefore, it is crucial to develop a new algorithm that can adjust both the individual seed length and the seed placement, as well as derive less frequent seeds.

Results: We present the Optimal Seed Solver (OSS), a dynamic programming algorithm that discovers the least frequently-occurring set of x seeds in an L-bp read in O(x x L) operations on average and in O(x x L2) operations in the worst case. We compared OSS against four state-of-the-art seed selection schemes and observed that OSS provides a 3-fold reduction of average seed frequency over the best previous seed selection optimizations.

Availability: We provide an implementation of the Optimal Seed Solver in C at: https://github.com/CMU-SAFARI/Optimal-Seed-Solver

Contact: hxin@cmu.edu, calkan@cs.bilkent.edu.tr, onur@cmu.edu

Authors:

first name	last name	email	country	organization	web site	corresponding?
Xin	Hongyi	hxin@cmu.edu	United States of America	Computer Science Department, Carnegie Mellon University, Pittsburgh		✓
Riachard	Zhu			-“-		
Sunny	Nahar			-“-		
John	Emmons			-“-		
Gennady	Pekhimenko			-“-		
Carl	Kingsford			-“-		
Can	Alkan			Bilkent University		
Onur	Mutlu			Computer Science Department, Carnegie Mellon University, Pittsburgh		

Oral Presentation and Poster Presentation

Efficient Privacy-Preserving String Search and an Application in Genomics

Keywords: query database, oblivious transfer, additive homomorphic encryption

Abstract: Motivation: Personal genomes carry inherent privacy risks and protecting privacy poses major social and technological challenges. We consider the case where a user searches for genetic information (e.g., an allele) on a server that stores a large genomic database and aims to receive allele-associated information. The user would like to keep the query and result private and the server the database.

Approach: We propose a novel approach that combines efficient string data structures such as the Burrows-Wheeler transform with cryptographic techniques based on additive homomorphic encryption. We assume that the sequence data is searchable in efficient iterative query operations over a large indexed dictionary, for instance, from large genome collections and employing the (positional) Burrows-Wheeler transform. We use a technique called oblivious transfer that is based on additive homomorphic encryption to conceal the sequence query and the genomic region of interest in positional queries.

Results: We designed and implemented an efficient algorithm for searching sequences of SNPs in large genome databases. During search, the user can only identify the longest match while the server does not learn which sequence of SNPs the user queries. In an experiment based on 2,184 aligned haploid genomes from the 1,000 Genomes Project, our algorithm was able to perform typical queries within ~2 seconds and ~20 seconds seconds for client and server side, respectively, on a laptop computer. The presented algorithm is at least one order of magnitude faster than an exhaustive baseline algorithm.

Authors:

first name	last name	email	country	organization	web site	corresponding?
Kana	Shimizu	shimizu-kana@aist.go.jp	Japan	Biotechnology Research Institute for Drug Discovery		✓
Koji	Nuida			Information Technology Research Institute,		
Gunnar	Rätsch			MSKCC		

Integrating multiple platform cancer methylomes for precision medicine

Lars Feuerbach[1]*, Yassen Assenov[2]*, Sandra D. Koser[1], Lei Gu[2,3], Clarissa Gerhäuser[2], Dieter Weichenhan[2], ICGC Project on Early Onset Prostate Cancer, Christoph Plass[2], Benedikt Brors[1]

[1] Division of Applied Bioinformatics, German Cancer Research Center (DKFZ), Im Neuenheimer Feld 580, D-69120 Heidelberg, Germany
[2] Division of Epigenomics and Cancer Risk Factors, German Cancer Research Center (DKFZ), Im Neuenheimer Feld 280, D-69120 Heidelberg, Germany
[3] Division of Theoretical Bioinformatics, German Cancer Research Center (DKFZ), Im Neuenheimer Feld 580, D-69120 Heidelberg, Germany

* These authors contributed equally to the work

Abstract

Assessment of cancer methylomes for diagnosis and selection of treatment options in the context of precision medicine has become feasible. Due to the large number of differentially methylated regions (DMRs) in an individual cancer patient, the interpretation of the results is still challenging. To meet this challenge, projects such as the International Cancer Genome Consortium (ICGC) and The Cancer Genome Atlas (TCGA) have produced large reference catalogs of cancer methylomes. The majority of the data is contributed by Infinium HumanMethylation450 BeadChip Kit (450k arrays), which cover approximately 1% of the genomic CpG positions. This resource is complemented by higher resolution approaches such as methyl-CpG immunoprecipitation (MCIp) or whole genome bisulfite sequencing.

We here present a software solution for cross-platform integration and visualization of these reference cancer methylomes. In a pilot study on 195 450k arrays and 19 MCIp-seq prostate cancer datasets shows high concordance for the DMRs covered by both platforms. Furthermore, we characterize the improved resolution of promoter hypermethylation detection by integrative methylome analysis.

Our integrative approach facilitates the comparison of data from different research cohorts as well as the interpretation of individual patient epigenomes in the context of published datasets from matched tumor subtypes with the aim to improve patient stratification and to guide therapeutic decisions.

Assessing telomere length and related genomic features from whole cancer-genome sequencing data

Lars Feuerbach[1]*, Lina Sieverling[1]*, David T.W. Jones[2], Philip Ginsbach[1], Katharina Deeg[3], Karsten Rippe[3], Stefan M. Pfister[2], Benedikt Brors[1]

[1] Divison of Applied Bioinformatics, German Cancer Research Center (DKFZ), Im Neuenheimer Feld 580, D-69120 Heidelberg, Germany
[2] Division of Pediatric Neurooncology, German Cancer Research Center (DKFZ), Im Neuenheimer Feld 580, D-69120 Heidelberg, Germany
[3] Research Group Genome Organization & Function, German Cancer Research Center (DKFZ) & BioQuant, Im Neuenheimer Feld 280, D-69120 Heidelberg, Germany

* These authors contributed equally to the work

The application of whole genome sequencing (WGS) as a diagnosis and therapy selection tool for precision medicine opens up analysis perspectives beyond the coding genome. The analysis of telomere features from WGS data represents such a perspective. To ensure infinite cell divisions, every tumor has to overcome the constant shortening of its telomeres. To avoid apoptosis or growth arrest upon reaching a critical telomere length known as the Hayflick limit, the tumor has to either evolve mechanisms to elongate its telomeres or disable the associated tumor-suppressor checkpoint cascades.

We developed a software package to extract the fingerprints of these escape strategies from whole genome sequencing data of matched tumor and control tissues, by monitoring telomere length and structure. These observations are complemented by assessing the presence of somatic alterations which have been linked to specific telomere maintainance strategies, such as activating point mutations in the promoter of the telomerase reverse transcriptase gene (TERT) or alterations in the ATRX gene. The software produces detailed graphical reports for individual patients as well as whole patient cohorts.

A study on 250 WGS pairs from the ICGC PedBrain project showed a strong correlation of the detected signals to patient age, TERT promoter status, and telomere length distribution as determined experimentally by fluorescent *in situ* hybridization (FISH) and terminal restriction fragment (TRF) analysis.

Detection of Copy Number Variations (CNVs) from NGS Data

Sriharsha V, Anwesha M, Prashanthi D, Shanta Pendkar, and Nita Parekh
Center for Computational Natural Science and Bioinformatics
International Institute of Information Technology, Hyderabad, India

Abstract

Copy-number variations (CNVs) are a form of structural variation that lead to abnormal copies of large genomic regions (> 1Kb) in a cell. The importance of CNVs is recognized by their high prevalence in human genome (~12%) and the observation that approximately half of the reported CNVs overlap with protein-coding genes. This results in gain or loss of gene copies, affecting the expression level of genes in the cell. Tumor genomes usually acquire somatic CNVs during carcinogenesis which may result in the amplification of oncogenes or deletion of tumor suppressor genes. Thus, detection of CNVs play an important role in understanding the molecular mechanisms leading to pathogenesis and in drug response.

Recent advances in computational methods have made CNV detection using whole genome NGS data more feasible. Among various CNV detection methodologies, depth of coverage based methods are known to accurately predict exact copy number and can even detect very large insertions. Here we implemented depth of coverage based algorithm, CNV-TV, proposed by Duan *et al* (2013) [1]. In this approach the detection of CNV is modeled as a change-point detection from the read depth signal and is fitted with a total variation (TV) penalized least squares model. For removing bias due to GC content and mapability issues, deepTools [2] were used. The analysis of the algorithm is performed on simulated data and real whole genome data from a DLBCL (GCB subtype) tumour sample (SRR1236468). The results are compared with three other depth of coverage methods, *viz*., ReadDepth, ControlFreeC and CNVnator [3-5]. Some of the features considered for comparison are coordinates of CNVs detected, length of single copy, copy number (gain /loss) and the sequencing depth of the sample (in case of simulated data). We observe that shorter CNVs (~ 1000) with 1 gain/loss require a coverage of ≥ 20X, while longer CNVs (~ 4000) were detected at even 10X coverage. The algorithm, CNV-TV was able to detect smaller CNVs (~ 500) in real data, while the three other tools failed to detect shorter CNVs for default parameters. Our comparative analysis of the four tools suggest that no single method is absolutely dependable in indentifying all true CNVs with the same accuracy. A significant variation is observed in the number and overlap of CNVs detected by the four tools, indicating the need to use more than one tool for their detection.

References:

1. Duan J, Zhang J, Deng H and Wang Y. (2013). *BMC Bioinformatics*, 14: 150

Characterization of bidirectional expression in total RNA reveals prominent active enhancer elements

Sandra D Koser[1], Naveed Ishaque[1,2,3], Jan-Philipp Mallm[4], Sabrina Schumacher[4], Stephan Wolf[5], Stephan Stilgenbauer[6], Karsten Rippe[4], Daniel Mertens[6,7], Benedikt Brors[1,8,9]

[1] Division of Applied Bioinformatics, German Cancer Research Center (DKFZ), Im Neuenheimer Feld 580, 69120 Heidelberg, Germany. [2] Heidelberg Center for Personalized Oncology, DKFZ-HIPO, DKFZ, Im Neuenheimer Feld 580, 69120 Heidelberg, Germany. [3] Division of Theoretical Bioinformatics, German Cancer Research Center (DKFZ), Im Neuenheimer Feld 580, 69120 Heidelberg, Germany. [4] Research Group of Genome Organization and Function, German Cancer Research Center (DKFZ), Im Neuenheimer Feld 580, 69120 Heidelberg, Germany. [5] Genomics & Proteomics Core Facility, German Cancer Research Center (DKFZ), Im Neuenheimer Feld 580, 69120 Heidelberg, Germany. [6] Department of Internal Medicine III, University of Ulm, Albert-Einstein-Allee 23, 89081 Ulm, Germany. [7] Cooperation Unit Mechanisms of Leukemogenesis, German Cancer Research Center (DKFZ), Im Neuenheimer Feld 580, 69120 Heidelberg, Germany. [8] German Consortium for Translational Cancer Research, Im Neuenheimer Feld 580, 69120 Heidelberg, Germany. [9] National Center for TumorDiseases, Heidelberg, Im Neuenheimer Feld 460, 69120 Heidelberg, Germany.

Bidirectional expression is a genome-wide feature that is associated with several regulatory elements. Bidirectional expression with unbalanced strands of origin is related to promoter expression, and in particular so called “divergent expression” which occurs when a promoter site leads to expression of a protein coding gene in one direction and expression of a noncoding RNA in the other. Balanced bidirectional expression often occurs at enhancer sites leading to eRNAs. However, the phenomenon itself has not yet been analyzed in general way. Here, we present a method to detect bidirectional expression genome-wide in long and short RNA with 11 CLL samples and four healthy B-cell pools each with a special focus on potential enhancer RNAs.

We identify bidirectionally expressed regions by interrogating the genome in overlapping windows of 500bp length for a minimum number of ten reads and at least 40% of the reads originate from either strand. We define filters to account for different characteristics of bidirectional expression. First, to determine whether the read distributions of the strands can be separated clearly, we use Ashman's D to test for bimodality. Second, to avoid promoter elements, all regions overlapping transcription start sites are excluded. Third, we exclude regions that have sense as well as and antisense transcription of genes. The read counts are normalized to the average number of aligned reads in the cohort to increase the comparability between the samples. For each filter type we build a consensus track. A consensus region has to show bidirectional expression in 25% of the samples.

In long RNA about 25 times more regions with bidirectional expression are found than in short RNA. However, the number of potentially active enhancer regions, i.e. loci that remain after applying all filters, is slightly higher for short RNAs. We find that each identified locus has a distinct length. Combining both RNA classes and applying all filters, we get a set of 360 potentially active enhancer regions. 60% of the regions overlap with introns, 25% and 15% are in close proximity to genes or in distant intragenic regions, respectively. 90% of the loci overlap with H3K27ac, a histone mark for open chromatin and active transcription. 80% of all potentially active enhancer regions show specific histone marks for active enhancers, i.e. H3K27ac is present together with either H3K4me1 or H3K9ac. The H3K27ac signal is on average stronger for enhancer elements displaying bidirectional expression than for active enhancer regions found only by segmentation based on histone marks. Also, 25% of the potentially active enhancers overlap with annotated enhancers in the FANTOM enhancer database. Hence, we conclude that we can detect prominent active enhancer elements from total RNA-Seq data.

Semi-de novo assembly of antibody sequences using single cell RNA-seq

Qingming Tang*; Toyota Technological Institute at Chicago
Karlynn Neu*; University of Chicago
Patrick Wilson#; University of Chicago
Aly A Khan#; Toyota Technological Institute at Chicago
*Equal contribution; #Correspondence

Among the various cells of the immune system only the B cell has the capacity to produce antibodies, which can identify and neutralize bacterial or viral pathogens. Antibodies obtain their diverse repertoire through a series of somatic rearrangements and mutations. Recent advances in RNA sequencing offer a high-throughput means of profiling all transcripts expressed in a single B cell. However, the assembly of full-length antibody sequences from single-cell RNA sequencing (scRNA-seq) is a non-trivial problem. For example, template-based methods that rely on genome alignment for transcript assembly are vulnerable to the somatic modifications present in the antibody sequences. Furthermore, *de novo* methods typically require solving a genome-scale assembly task that makes antibody repertoire studies difficult. Thus, the lack of efficient methods for quantifying and assembling antibody sequences is a major roadblock in studying their repertoire with scRNA-seq.

Here, we present a novel semi-*de novo* assembly method to determine the full-length sequence of the heavy and the light chains in a B cell using scRNA-seq. We exploit the constant and the non-complementarity determining regions in the chains to guide the *de novo* assembly of the full-length sequences. To demonstrate the utility of our method, we subjected 20 single B cells from a human donor to scRNA-seq, assembled the full-length heavy and the light chains, and identified both the antibody isotype and the specific segments used in recombination. We experimentally confirmed these results by using single-cell primer based nested PCRs and Sanger sequencing.

Taken together, our approach allows investigators to use scRNA-seq for both dissecting cell-to-cell transcriptional heterogeneity in B cells and also characterizing the antibody repertoire. Our semi-*de novo* method also serves as a principled approach to assemble other diverse genes associate with immunological repertoire using scRNA-seq, such as HLA and TCR genes.

We plan to release an open source Python/C++ software implementation of our method upon publication.

Contact: wilsonp@uchicago.edu, aakhan@uchicago.edu

eXclusivity: a genetic algorithm for mutual exclusive analysis of NGS data

Charles D. Imbusch[1], Benedikt Brors[1]

[1]Divison of Applied Bioinformatics, German Cancer Research Center (DKFZ), Im Neuenheimer Feld 580, D-69120 Heidelberg, Germany

Abstract

Through massive parallel sequencing using NGS it has been made possible to detect various aberrations such as SNVs, INDELs, SVs and CNVs. Especially in cancer genomics inter-tumor heterogeneity increases the problem of distinguishing real driver mutations from passenger mutations. Looking at mutational frequencies alone is not sufficient for identifying genes of interest, it becomes necessary to consider that genes and their products are acting cooperatively in a pathway which can be detected by a mutual exclusivity analysis.

eXclusivity makes use of NGS derived cancer aberrations and will search for de novo affected pathways optionally using the String Protein-Protein interaction database. Internally, a genetic algorithm optimizes a fitness function which includes terms for considering mutual exclusivity, the number of genes in the set, recurrently mutated genes and optionally a term for preferring genes which are closer in the Protein-Protein interaction network using the Dijkstra algorithm. The results are presented to the user as Oncoprints together with their respective p-values. Results on test data sets look as expected, finding possible candidate genes in real projects.

2. Ramirez F,*et al*.(2014). *Nucleic Acids Research,* (W1): W187-W191.

3. Miller CA, Hampton O, Coarfa C, Milosavljevic A. (2011). *PloS*, 6(1): e16327

4. Boeva V, *et al*. (2012). *Bioinformatics*, 28(3): 423.

5. Abyzov A, Urban AE, Snyder M, Gerstein M. (2011). *Genome Res* 21: 974

BioPeer: A Fast and Secure Peer-to-Peer Data Sharing Tool

Cihad Oge[1] F. Tugba Dogan[1] Gizem Goktepe[1] Fatma Koc[1] Cem Sevim[1] Can Alkan[1]

[1] Department of Computer Engineering, Bilkent University, Ankara, Turkey

The high throughput DNA sequencing technologies now enable researchers to answer a wide range of biological questions, however they also impose various computational problems. One of the most urgent issues to address is data sharing among collaborators located in different geographical locations, due to the huge amounts of data generated by these platforms. There are multiple levels of data types to keep and/or share among collaborators. First, the sequence data itself is stored in a text file in FASTQ format, which can be compressed using general-purpose compression tools such as gzip, or specialized FASTQ compressors such as SCALCE [5]. When a reference genome is available, the reads are mapped to this reference genome, and the mapping information is kept in a compressed and indexed file format called BAM [6], or its further compressed version, CRAM [3]. All of these file types use vast amounts of storage space. For example, the sequence data from the genome of one human individual sequenced at high depth (30-fold) using the Illumina platform totals to 480 GB in FASTQ format (134 GB gzip, 76 GB SCALCE), and approximately 110 GB in BAM format. The transfer of this data among two or more collaborators over the FTP protocol would require many days. Moreover, most projects generate sequence data from multiple individuals, increasing the amount of the data to be transferred linearly. One example is the Great Ape Diversity Project, where over 20 TB of data was shared between several laboratories located in the United States, Germany, and Spain.

Large scale projects such as the 1000 Genomes Project [1], and large sequencing centers such as the Wellcome Trust Sanger Institute now use a UDP-based file transfer protocol (*fasp*) developed by Aspera Software (`http://www.asperasoft.com/`). Using this server/client tool, it is possible to transfer large files with up to 600 Mb/s speed. This is a tremendous improvement over the TCP-based FTP protocol, where the throughput is usually lower than 1 Mb/s. Although it provides the best data transfer speed, the main drawback for most of the other researchers to use Aspera is its cost. Most labs cannot afford to install and maintain a dedicated Aspera server, and pay for license costs, especially when their data transfer needs are not constant over time. Therefore, the common method to share data among collaborators is writing the data to external disks and circulating these disks using courier services. Thus there is a need for a user-friendly, peer-to-peer (P2P), open source, and very fast file sharing system that would enable researchers share unpublished data with their collaborators.

	BioPeer	FTP
Encryption	Yes	No
P2P	Yes	No
Auto-Sync	Yes	No
Multicast	Yes	No
Protocol	UDP	TCP
Redmond-London	153 Mb/s	90 Mb/s
London-Ankara	41 Mb/s	41 Mb/s
Nürnberg-Ankara	57 Mb/s	55 Mb/s

Table 1: BioPeer vs. FTP. Throughputs are reported as pairwise speed, and do not reflect speed up gained by BioPeer's ability to multicast. Note that BitTorrent-style data sharing would improve speed almost linear to the number of collaborators within the project.

We developed a new cross-platform desktop application (BioPeer) to address this problem, which is a hybrid of various data transfer approaches. Briefly, BioPeer uses the UDP-based open source UDT protocol [4] for data transfer, and provides a P2P file sharing architecture similar to that of BitTorrent, where large files are transferred in chunks, and synchronized between peers (i.e. collaborators) within the same *project*. Different from other P2P platforms, BioPeer also includes user authentication through the ORCID database (`http://www.orcid.org`) to protect data privacy. In addition, files are encrypted using the AES protocol [2]. BioPeer is implemented in Java 8, and supports Linux, Windows, and OS X platforms.

Sample scenario. Assume Alice has two separate projects P1 and P2, where she collaborates only with Bob in project P1, and with both Bob and Ken in project P2. In both projects there are multiple files she wishes to share with her collaborators.

1. Alice starts BioPeer, and defines two projects P1 and P2. She sets the files associated with each project. Alice then adds Bob as a collaborator in projects P1 and P2, and Ken as a collaborator only in project P2.
2. Bob starts BioPeer, and sees that Alice shared two projects P1 and P2 with him. He starts to download from Alice over UDP.
3. Ken starts BioPeer, and sees that Alice shared one project P2 with him, and that Bob is also in the project and is downloading the data. He starts to download from Alice over UDP as well. Ken also adds new files to project P2, to be synchronized to Alice and Bob's computers.
4. At this point, files in P1 are transferred to only Bob; and files in P2 are transferred to both Bob and Ken. If Bob and Ken has their BioPeer instances running at the same time; then the three BioPeer instances in all three computers synchronize sharing files in project P2, in a similar fashion with BitTorrent. BioPeer enables Bob to download from Ken, and Ken to download from Bob at the same time they are downloading different parts of the data from Alice. Here, the "parts" may be different files within the same project, and/or partitions of larger files.

References

[1] 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422):56–65, Nov 2012.

[2] J. Daemen and V. Rijmen. AES proposal: Rijndael. 1998.

[3] M. H.-Y. Fritz, R. Leinonen, G. Cochrane, and E. Birney. Efficient storage of high throughput DNA sequencing data using reference-based compression. *Genome Res*, 21(5):734–740, May 2011.

[4] Y. Gu and R. L. Grossman. UDT: UDP-based data transfer for high-speed wide area networks. *Computer Networks*, 51(7):1777–1799, 2007.

[5] F. Hach, I. Numanagic, C. Alkan, and S. C. Sahinalp. SCALCE: boosting sequence compression algorithms using locally consistent encoding. *Bioinformatics*, 28(23):3051–3057, Dec 2012.

[6] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, and . G. P. D. P. Subgroup. The sequence alignment/map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, Aug 2009.

The Human Inosinome Atlas

Ernesto Picardi[1,2], Caterina Manzari[2], Francesca Mastropasqua[1], Italia Aiello[1], Anna Maria D'Erchia[1,2] and Graziano Pesole[1,2]

[1] *Dipartimento di Bioscienze, Biotecnologie e Biofarmaceutica, Università di Bari, Bari, Italy,* [2] *Istituto Biomembrane e Bioenergetica del Consiglio Nazionale delle Ricerche, Bari, Italy*

*To whom correspondence should be addressed: graziano.pesole@uniba.it

BACKGROUND

RNA editing is a post-transcriptional molecular phenomenon whereby a genetic message is modified from the corresponding DNA template by means of substitutions, insertions and/or deletions [1]. In human, it mainly involves the deamination of adenosines to inosines by the family of ADAR enzymes acting on double RNA strands [2]. A-to-I RNA editing has a plethora of biological effects depending on the RNA region involved in the modification [3]. Changes in UTRs can lead to altered expression, whereas modifications in coding protein regions can induce amino acid replacements with more or less severe functional consequences [4]. The detection of RNA editing events at genomic scale has been largely facilitated by the advent of NGS technologies. Although the computational identification of A-to-I changes in human is yet a challenging task, several attempts have been done and more than 2,4 million events have been collected [5]. RNA editing is a dynamically regulated process and a comprehensive understanding of its biological roles requires large-scale studies and the creation of a specialized atlas.

RESULTS

To this aim, we have massively sequenced total RNA from six human tissues (brain, lung, liver, kidney, heart and muscle) in three different individuals using a HiSeq2500 Illumina sequencer and according to a stranded protocol to preserve RNA orientation. In addition, we have generated whole genome sequencing data for each donor and whole exome sequencing data for each tissue. NGS data have been analysed according to an improved RNA editing algorithm implemented in our REDItools suite taking into account several error sources. Hyper edited reads have also discovered using a recent approach developed to rescue heavily edited RNA-Seq reads that are generally missed by current methods [6]. Overall we detected 3,041,422 events representing the largest collection of A-to-I RNA editing in human with more than 2 millions of novel positions. Of these, 97% was in repetitive regions and ~90% in Alu elements. Only a limited amount of sites fell in non-repetitive regions (3%), as expected. The number of predicted A-to-I events varied greatly among samples because of sequencing depth variation, stringent filters used to recover editing candidates and tissue specific roles of RNA editing. Nonetheless, brain appeared the most edited tissue with on average 511,733 sites per sample. In contrast, heart and muscle showed a smaller number of editing sites than other tissues with on average 79,976 and 28,620 changes, respectively. Regarding the impact on known human protein-coding genes, we discovered that 13062 loci over 20173 (65%) underwent RNA editing in their exons and/or introns. Very interestingly, we found that edited genes were consistently enriched in genes involved in neurological disorders and cancer. In addition, 74% (1842/2501) of essential genes [7] were in the edited set, confirming the relevant biological role of RNA editing in human.

CONCLUSIONS

Here we present the largest collection of RNA editing events in human tissues. We confirm that RNA editing is pervasive in human and indispensable to perverse the cellular homeostasis. Indeed, edited genes are enriched in genes linked to cancer and neurological diseases. Our collection will facilitate the understanding of RNA editing role in normal as well as pathological conditions.

REFERENCES

1. Gott JM, Emeson RB: **Functions and mechanisms of RNA editing**. *Annu Rev Genet* 2000, **34**:499-531.
2. Hogg M, Paro S, Keegan LP, O'Connell MA: **RNA editing by mammalian ADARs**. *Adv Genet* 2011, **73**:87-120.
3. Maas S: **Gene regulation through RNA editing**. *Discov Med* 2011, **10**(54):379-386.
4. Hood JL, Emeson RB: **Editing of Neurotransmitter Receptor and Ion Channel RNAs in the Nervous System**. *Current topics in microbiology and immunology* 2011.
5. Ramaswami G, Li JB: **RADAR: a rigorously annotated database of A-to-I RNA editing**. *Nucleic acids research* 2014, **42**(Database issue):D109-113.
6. Porath HT, Carmi S, Levanon EY: **A genome-wide map of hyper-edited RNA reveals numerous new sites**. *Nature communications* 2014, **5**:4726.
7. Dickerson JE, Zhu A, Robertson DL, Hentges KE: **Defining the role of essential genes in human disease**. *PloS one* 2011, **6**(11):e27368.

High-performance digital API for NGS data

Eric Schendel Rashid Al Ali and Andrey Ptitsyn
Sidra Medical and Research Center,
Doha, Qatar

Abstract

The recent advances in high-throughput sequencing present a series of challenges similar to those of the early trailblazers of Bioinformatics: the power of latest computers is inadequate to the volume of sequence data. Optimizing the code for most common operations in biological sequence analysis makes sense again. At Sidra we are challenged with high volume of genomic fragments (reads) generated by a park of Illumina X10, HiSeq2500, NextSeq, MiSeq, PacBio PS2 and Ion Proton sequencers. Similar challenges are faced by many other research centers. Optimization of code base is part of our response strategy. We have designed and implemented the first version of HPC library for handling, compression and basic manipulations on nucleotide sequences. The principal feature of our library is compression of the nucleotide symbol data into binary bit fields and making all further operations on long unsigned integer numbers. This approach allows effective low-level implementation in which some functions take as little as four CPU instructions to complete. The library includes functions allowing input from popular formats (FASTA, FASTQ); final results can be restored into a standard format or stored effectively as a binary stream. The nearest plans for development include parallel processing in OpenMP standard and implementation of most common sequence match detection and alignment algorithms. We offer this code to the Bioinformatics developer community and welcome any contributions on the free open source principles.

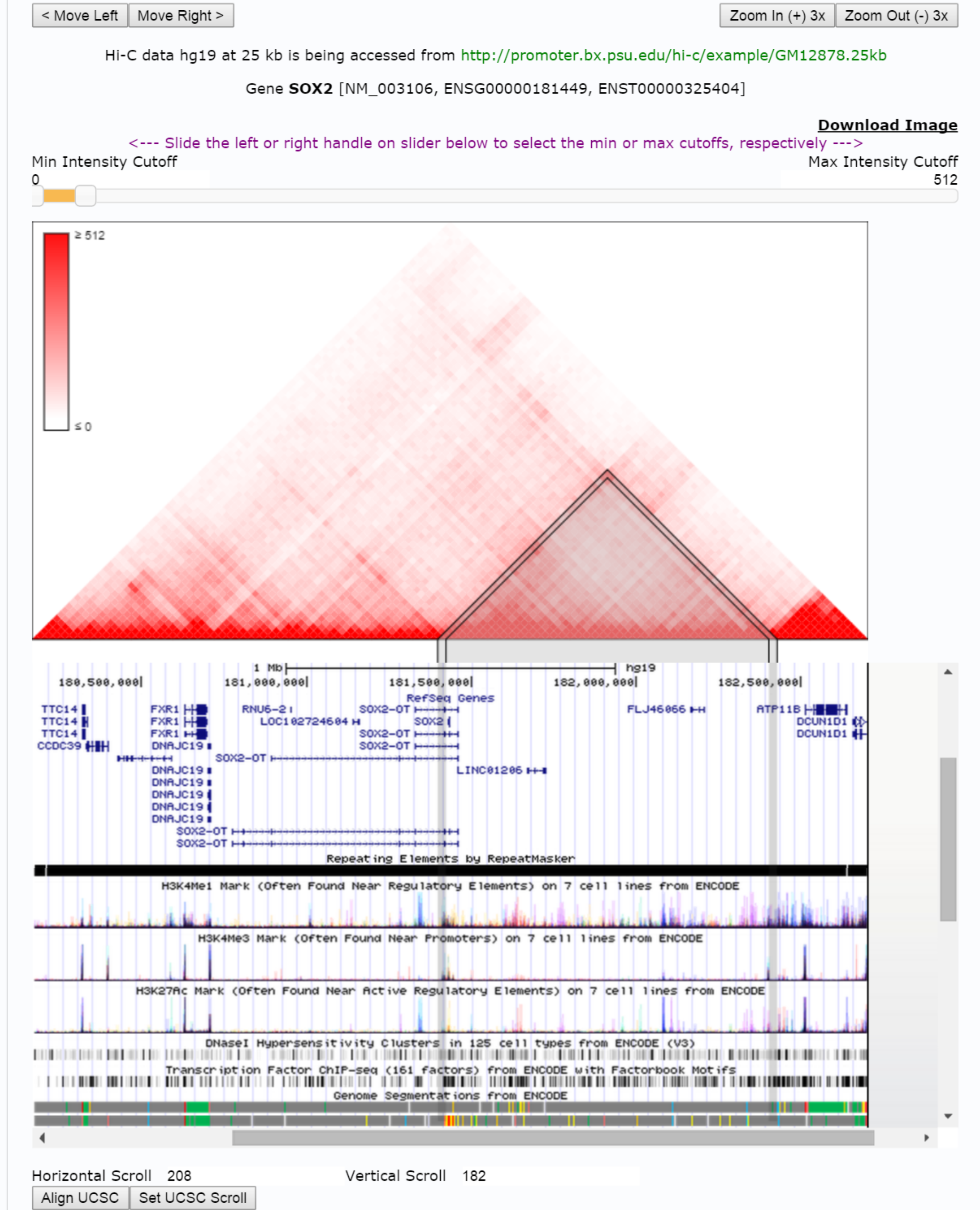
Visualizing three-dimensional organization and long-range interactions of the mammalian genome with the 3D Genome Browser.

Yanli Wang, Gal Yaroslavsky, Tyler Derr, Lijun Zhang, Feng Yue
Department of Biochemistry and Molecular Biology, Penn State College of Medicine, PA, 17033

The mammalian genome subscribes to a complex spatial organization that defines the three-dimensional interactions of potentially distant functional elements that control the regulation of transcription and replication. Recent advancements in sequencing and analysis techniques –specifically Hi-C, or high-throughput chromosome conformation capture– have revealed these interactions genome-wide at unprecedented resolutions. Unfortunately, navigating the Hi-C data remains a daunting feat for many biologists, as its O(n^2) complexity for the already big data intrinsic to mammalian genomes poses a challenge to its analysis (time and memory usage), storage and transfer.

Our laboratory has developed and extended the functionality of the 3D Genome Browser (http://3dgenome.org), a web-based, intuitive and accessible browser of Hi-C data. The browser adopts a gene-centric approach: given the user input of gene symbol or genomic coordinates, it queries the Hi-C intra-chromosomal contact matrix for interactions from the regions in vicinity and display those values as a heatmap. Furthermore, our browser contextualizes the region by directly aligning it to the corresponding region as displayed by the established and familiar University of California Santa Cruz (UCSC) Genomic Browser while retaining its flexibility to customize genome tracks and load personalized UCSC sessions. While our browser contains several existing high-quality Hi-C datasets for a variety of human and mouse tissues for viewing, it also supports the browsing of user-generated Hi-C data with the **"C" Your Data** feature. By converting Hi-C contact matrices into an indexed, binary format file and hosting it on any HTTP accessible server, the 3D Genome Browser could directly query and display the specified region without requiring the upload of entire files onto the server. In addition to the Hi-C heatmap, the contact matrix could also be visualized as virtual 4C, a linear plot detailing the number of interactions between a single genomic site of interest (bait or anchor locus) with other loci. Given the user input of gene or rsid, the virtual 4C plot with the TSS(s) or SNP as anchor locus would facilitate the identification of potential cis-regulatory elements. This feature would be supplemented with the inclusion of DNase I Hypersensitive Site (DHS)-linkage and ChIA-PET data, both currently under development.

With our gene-centric, binary-file browser approach, the 3D Genome Browser improves the accessibility in browsing Hi-C data. With the visualization of the spatial organization and long-range interactions of particular genomic regions along with their genetic and epigenetic context, our browser seeks to drive hypothesis-generation about and enrich the understanding of the intrinsic link between genomic organization and genetic regulation.


< Move Left
Move Right >
Zoom In (+) 3x
Zoom Out (-) 3x
Download Image
≥ 512
≤ 0
1 Mb
hg19
180,500,000
181,000,000
181,500,000
182,000,000
182,500,000
RefSeq Genes
TTC14
TTC14
TTC14
CCDC39
FXR1
FXR1
FXR1
DNAJC19
DNAJC19
DNAJC19
DNAJC19
DNAJC19
DNAJC19
RNU6-2
LOC102724604
SOX2-OT
SOX2-OT
SOX2-OT
SOX2-OT
SOX2-OT
SOX2-OT
SOX2
LINC01206
FLJ46066
ATP11B
DCUN1D1
DCUN1D1
Repeating Elements by RepeatMasker
H3K4Me1 Mark (Often Found Near Regulatory Elements) on 7 cell lines from ENCODE
H3K4Me3 Mark (Often Found Near Promoters) on 7 cell lines from ENCODE
H3K27Ac Mark (Often Found Near Active Regulatory Elements) on 7 cell lines from ENCODE
DNaseI Hypersensitivity Clusters in 125 cell types from ENCODE (V3)
Transcription Factor ChIP-seq (161 factors) from ENCODE with Factorbook Motifs
Genome Segmentations from ENCODE
Align UCSC
Set UCSC Scroll

Hi-C data hg19 at 25 kb is being accessed from http://promoter.bx.psu.edu/hi-c/example/GM12878.25kb

Gene **SOX2** [NM_003106, ENSG00000181449, ENST00000325404]

<--- Slide the left or right handle on slider below to select the min or max cutoffs, respectively --->

Min Intensity Cutoff 0

Max Intensity Cutoff 512

Horizontal Scroll 208

Vertical Scroll 182

The Importance of Mutation Loss in Modelling Evolution and Metastasis in Genomically Unstable Cancers

Andrew McPherson*, Andrew Roth*, Jessica McAlpine, Alexandre Bouchard-Côté, Sohrab P. Shah
*contributed equally

By sequencing primary and metastatic biopsies sampled during initial diagnosis and relapse, researchers aim to reconstruct the evolutionary history of a cancer, allowing for identification of ancestral and descendent drivers in individual cancers, and high level patterns of evolution and metastasis across multiple cancers. Genomically unstable cancers often exhibit high rates of amplification and deletion and significant genomic heterogeneity, confounding existing techniques for inference of evolutionary histories. Many existing methods exclude the possibility that mutations such as single nucleotide variants (SNVs) could be lost by deletion of the encompassing chromosomal segment, an event that is not uncommon in a genomically unstable cancer.

Methods for reconstructing evolutionary histories can be distinguished by their assumptions about mutational processes. Hierarchical clustering of SNV presence/absence aims to group samples with shared ancestry, assuming these groups will have similar presence/absence profiles. Clustering approaches are unable to reconstruct ancestral genotypes, and will misclassify as descendent SNVs that were in fact ancestral and subsequently lost. Recently developed bespoke methods model sample heterogeneity and reconstruct phylogenetic relationships between observed and ancestral genotypes. The predominant assumption is that SNVs occur only once throughout evolution of the tumour (infinite sites) and cannot be subsequently lost.

We have comprehensively profiled 31 samples from 7 High Grade Serous Ovarian Cancer patients using both targeted and whole genome sequencing at bulk and single cell resolution. We predict SNV loss as present in 6 out of 7 patients, affecting between 1% and 10% of SNVs. We have validated, at single cell resolution, an example of SNV loss as the principal event distinguishing tumour clones. We further show that loss unaware, heterogeneous sample models will be confounded by SNV loss, producing incorrect results that are difficult to discern from correct solutions. Many loss unaware methods attempt to mitigate the problem of SNV loss by pre-filtering SNVs. In genomically unstable cancers this may filter interesting events distinguishing tumor clones, and may fail to filter real losses not associated with an observable copy number change.

We contrast existing methods with our own loss aware, homogenous sample approach. Although our approach can be confounded by heterogeneous samples, these situations are comparatively easier to identify and diagnose as large numbers of predicted SNV losses uncorrelated with inferred ancestral copy number changes. Furthermore, the predicted phylogeny is often more interpretable as it is frequently a subtree of the true network relating samples with shared clonal ancestry, whereas for heterogeneous sample models, incorrect phylogenies typically bears little resemblance to the true phylogeny.

We show how our method can be used as part of a larger experimental design that involves selection of targets for deep sequencing, and subsequently single cell sequencing, for comprehensively profiling the clonal composition and phylogeny of a cancer. We use the resulting clonal phylogenies to present examples of mutational and spatial drift contrasted with punctuated evolution of late emergent drivers and subsequent clonal expansions.

GeneTerrain: a visual analytic platform to interpret high-throughput Omics data for clinical genomics applications

Jake Y. Chen[1,2], Peter Li[1], and Zongliang Yue[2]

[1] Medeolinx, LLC & Medeolinx Software, Ltd., Indianapolis, IN 46202 USA

[2] Indiana University School of Informatics and Computing, Indianapolis, IN 46202, USA

Emails: Jake Chen j.chen@medeolinx.com , Peter Li p.li@medeolinx.com, Zongliang Yue zongyue@iupui.edu

Abstract:

There are major challenges for future clinical applications of genomics sequencing or functional profiling assays. One of them is that the genomics-based assays are high-dimensional, often showing hundreds or thousands of mutated/altered genes, making it impractical for even trained clinicians to track the anomalies one at a time. A second one is that the mutations and altered genes (by expression) are highly dependent upon one another, creating a challenge to identify whether observation is a primary "driving forces" that are more upstream or "causal" to the phenotypic conditions observed in clinical settings, or an "effect" that are more downstream or "consequential/peripheral" to the phenotypic conditions observed in clinical settings. A third one is that it's difficult to balance between ease of interpretation of the results for critical clinical decision making and the ease of tracking mechanisms for translational biomedical characterizations of complex diseases.

We developed GeneTerrain as a visual analytic platform that can address the above challenges, therefore bridging the gap between data generated from High-throughput Genomics technology and decisions that clinicians may use. The platform has the following technical characteristics. First, it integrates all hundreds or thousands of mutated/altered gene-centric measurements from different sample group conditions into several geneTerrains, each of which reflects a unique snapshot of the complex biosystem measured for the clinical condition, e.g., ER+ vs ER- breast cancer subtypes. It is represented as a natural heatmap and may be easy for human interpretation after brief training. Second, it integrates molecular network data to make it easy to characterize mutation/alteration of genes from clinical genomics measurements with ranks and modular information critical to prioritize the interpretation of findings. Third, it allows easy comparisons of different sample groups, either through human perceptions or machine learning constructed from limited number of available samples, thus providing powerful clinical decision support. Fourth, for advanced translational medicine professionals, the tool can help connect the visualization to vast amount of available public annotation, making it easy to perform clinical research and mechanistic studies.

We will show a few case studies using cancer functional genomics data to showcase the new platform. The platform has been developed online and is available by-invitation for Academic users interested in beta-testing the new tool.

How Big is That Genome?
Estimating Genome Size and Coverage from k-mer Abundance Spectra

Michal Hozza Tomáš Vinař Broňa Brejová

Faculty of Mathematics, Physics, and Informatics, Comenius University,
Mlynská dolina, 842 48 Bratislava, Slovakia
{hozza,vinar,brejova}@fmph.uniba.sk

Many practical algorithms for sequence alignment, genome assembly and other tasks represent a sequence as a set of k-mers. Here, we address the problems of estimating genome size and sequencing coverage from sequencing reads, without the need for sequence assembly. Our estimates are based on a histogram of k-mer abundance in the input set of sequencing reads and on probabilistic modeling of distribution of k-mer abundance based on parameters related to the coverage, error rate and repeat structure of the genome. Previous works (Li and Waterman, 2003; Williams et al., 2013) concentrated mainly on discovering and compensating for the repeat structure of the genome and neglected modeling of the sequencing errors. Consequently, they require high coverage data ($> 10\times$). Our method provides reliable estimates even at coverage as low as 0.5 or at error rates as high as 10%.

Acknowledgements. This research was funded by VEGA grants 1/1085/12 (BB) and 1/0719/14 (TV).

References

Li, X. and Waterman, M. S. (2003). Estimating the repeat structure and length of DNA sequences using ℓ-tuples. *Genome Research*, 13(8):1916–1922.

Williams, D., Trimble, W. L., Shilts, M., Meyer, F., and Ochman, H. (2013). Rapid quantification of sequence repeats to resolve the size, structure and contents of bacterial genomes. *BMC Genomics*, 14(1):537.

Fishing in Read Collections: Memory Efficient Indexing for Sequence Assembly

Vladimír Boža Jakub Jursa Broňa Brejová Tomáš Vinař

Faculty of Mathematics, Physics, and Informatics, Comenius University,
Mlynská dolina, 842 48 Bratislava, Slovakia
{boza,brejova,vinar}@fmph.uniba.sk

We present a memory efficient index for storing a large set of DNA sequencing reads. The index allows us to quickly retrieve the set of reads containing a certain query k-mer. Instead of the usual approach of treating each read as a separate string (Philippe et al., 2011; Välimäki and Rivals, 2013), we take an advantage of significant overlap between reads and compress the data by aligning the reads to an approximate superstring constructed specifically for this purpose in combination with several succint data structures.

We compare the performance of our data structure, called CR-index, with compressed Gk-arrays (Välimäki and Rivals, 2013). On a data set of 151bp Illumina reads from *E. coli* strain MG1655 (genome length 4.7 Mbp, 184× coverage after removal of low-quality reads, 0.75% error rate). We can index this data set in 142 MB of memory, while compressed Gk-arrays take ≈ 0.8 GB. On the human chromosome 14 (23× coverage of 107 Mbp sequence, 1.5% error rate), the CR-index requires only 571 MB of memory, while compressed Gk-arrays require ≈ 1.7 GB.

Acknowledgements. This research was funded by VEGA grants 1/1085/12 (BB) and 1/0719/14 (TV). We would like to thank authors of SDSL-lite for a functional and well-documented library.

References

Philippe, N., Salson, M., Lecroq, T., Leonard, M., Commes, T., and Rivals, E. (2011). Querying large read collections in main memory: a versatile data structure. *BMC Bioinformatics*, 12(1):242.

Välimäki, N. and Rivals, E. (2013). Scalable and versatile k-mer indexing for high-throughput sequencing data. In *Bioinformatics Research and Applications*, pages 237–248. Springer.