HiTSeq 2016 Poster Abstracts

Accurate Modeling and Correction of GC Content and Gene Length Bias from RNA-seq Data

Keywords: Data analysis, Gene expression, Next-generation sequencing

Abstract: Several approaches exist that deal with GC content and gene length dependent biases in RNA-seq measurements of transcript abundances. However, when computing correction factors, existing approaches do not consider the potential interdependence and interaction of both effects. Additionally, they do not deal satisfactorily with genes that have partially or entirely zero counts. Thirdly, the biases may affect a large fraction of genes such that the assumed global normalization schemes are invalid.

We present a novel method for library bias correction (LBC) with a 2D function that simultaneously depends on GC content and gene length. Our approach is unique by modeling not the absolute biases but the sample-specific deviations from the data-set wide bias. The computed correction factors are precise even in the case of partial zero counts. The presented method is useful for correcting data sets with subsets of deviating samples as well as for the joined analysis of different data sets generated with different biases.

Authors:

first name	last name	email	country	organization	corresponding?
Hubert	Rehrauer	hubert.rehrauer@fgcz.uzh.ch	Switzerland	ETH Zurich	✓
Slavica	Dimitrieva		Switzerland	ETH Zurich	
Ralph	Schlapbach		Switzerland	ETH Zurich	

A Hybrid Genome Assembler for Second- and Third-Generation Sequencing

Keywords: Assembly, Algorithms, Next-generation sequencing, Genome analysis

Abstract: Genome assembly is challenged by the presence of large and complex repeats. The 2^{nd} generation sequencing offers high throughput in low cost but the read length is inadequate to resolve large repeats. On the other hand, the 3^{rd} generation sequencing can generate much longer reads for spanning large repeats, but the error rate and sequencing cost are much higher.

This paper presented several new algorithms for assembling high-quality short reads and low-quality long reads in two stages. In the correction stage, long reads are mapped onto an FM-index constructed from short reads and polished by FM-index extension. In the assembly stage, a novel algorithm (called locality-sensitive backward search) is proposed to efficiently compute inexact overlap among (partially) corrected long reads. The results indicated that the correction power and accuracy is higher than existing methods. The developed assembler is able to assemble high-quality genome (N50 in Mb) using low-coverage PacBio sequencing.

Availability: https://github.com/ythuang0522/StriDe

Authors:

first name	last name	email	country	organization	corresponding?
Yao-Ting	Huang	ythuang@cs.ccu.edu.tw	Taiwan	National Chung Cheng University	✓
Ping-Ye	Chen		Taiwan	National Chung Cheng University	

MICADo - Looking for mutations in PacBio cancer data: an alignment-free method

Keywords: PacBio, Cancer Genomics, Next-generation sequencing, Mutation Calling, de Bruijn graphs

Abstract: Targeted sequencing is commonly used in clinical application of NGS technology since it enables generation of sufficient sequencing depth in the targeted genes of interest and thus ensures the best possible downstream analysis. This notwithstanding, the accurate discovery and annotation of disease causing mutations remains a challenging problem even in such favorable context. The difficulty is particularly salient in the case of third generation sequencing technology, such as PacBio.

We present MICADo, a de Bruijn graph based method that makes possible to distinguish between patient specific mutations and other alterations for targeted sequencing of a cohort of patients. MICADo analyses NGS reads for each sample within the context of the data of the whole cohort in order to capture the differences between specificities of the sample with respect to the cohort. MICADo is particularly suitable for sequencing data from highly heterogeneous samples, especially when it involves high rates of non-uniform sequencing errors. It was validated on PacBio sequencing datasets from several cohorts of patients.

Availability: The source code is available at http://github.com/cbib/MICADo.

Authors:

first name	last name	email	country	organization	corresponding?
Justine	Rudewicz	justinerudewicz@gmail.com	France	University of Bordeaux, CNRS/LaBRI and INSERM	✓
Hayssam	Soueidan		France	University of Bordeaux and CNRS/LaBRI	
Raluca	Uricaru		France	University of Bordeaux and CNRS/LaBRI	
Hervé	Bonnefoi		France	INSERM	
Richard	Iggo		France	INSERM	
Jonas	Bergh		Sweden	Karolinska Institute	
Macha	Nikolski	macha.nikolski@labri.fr	France	University of Bordeaux and CNRS/LaBRI	✓

iMapSplice: a lightweight and personalized RNA-seq alignment approach to improve transcriptome profiling

Keywords: RNA-seq, alignment, transcriptome profiling

Abstract: Genomic variants in both coding and non-coding sequences can have unexpected and functionally important effects on the splicing of gene transcripts. These events, can be measured by RNA sequencing (RNA-seq), but require the accurate alignment of reads across exon splice junctions. Existing alignment algorithms that utilize a standard reference genome as a template may have difficulty in mapping those reads that carry genomic variants. These problems can lead to bias in relative expression abundance of alternative alleles and the failure to detect splice variants created by splice site mutations.

To improve RNA-seq read alignments, we have developed a novel lightweight approach called iMapSplice (Individualized MapSplice) that enables personalized mRNA transcriptional profiling. The algorithm makes use of personal genomic information and performs an unbiased alignment towards genome indices carrying both reference and alternative bases. Importantly, this breaks the limit of dependency on reference genome splice site dinucleotide motifs and enables iMapSplice to discover personal splice junctions created through splice site mutations. We report comparative analyses by applying iMapSplice, MapSplice, and STAR on 4 simulated and 68 real human datasets. Besides general improvements in read alignment and splice junction discovery, iMapSplice greatly alleviates biases in allelic ratios across the genome and unravels many previously uncharacterized splice junctions created by mutations at splice sites, with minimal overhead in computation time and storage.

Availability: iMapSplice is implemented as stand alone C++ code, and can be downloaded via URL: https://github.com/xa6xa6/mps

Authors:

first name	last name	email	country	organization	corresponding?
Xinan	Liu		United States	University of Kentucky	
James N.	MacLeod		United States	University of Kentucky	
Jinze	Liu	liuj@cs.uky.edu	United States	University of Kentucky	✓

MuClone: Detection and classification of somatic mutations through probabilistic integration of clonal population structure

Keywords: Single nuclotide variants, Tumour clones, Mutation detection, Mutation classification, Cancer evolution

Abstract: Accurate detection and classification of somatic single nucleotide variants (SNVs) is important in defining the clonal composition of human cancers. Existing tools are prone to missing low prevalence mutations and methods for classification of mutations into clonal groups across the whole genome are underdeveloped. Increasing interest in deciphering clonal population dynamics over multiple samples in time or anatomic space from the same patient is resulting in whole genome sequence (WGS) data from phylogenetically related samples. We posited that injecting clonal structure information into the inference of mutations from multiple samples would improve mutation detection.

We developed MuClone: a novel statistical framework for simultaneous detection and classification of mutations across multiple samples of a patient from whole genome or exome sequencing data. The key advance lies in incorporating prior knowledge about the cellular prevalences of clones to improve the performance of detecting mutations, particularly low prevalence mutations. We evaluated MuClone through synthetic and real data from spatially sampled ovarian cancers. The results support the hypothesis that clonal information improves the sensitivity without compromising the specificity. In addition, MuClone classifies mutations across whole genomes of multiple samples into biologically meaningful groups that can provide additional phylogenetic insights and permits studying clonal dynamics from WGS data.

Availability: MuClone is available from http://compbio.bccrc.ca/

Authors:

first name	last name	email	country	organization	corresponding?
Fatemeh	Dorri		Canada	University of British Columbia	
Sean	Jewell		United States	University of Washington	
Alexandre	Bouchard-Côté		Canada	University of British Columbia	
Sohrab P.	Shah	sshah@bccrc.ca	Canada	University of British Columbia & British Columbia Cancer Research Center	✓

Fast and Accurate Alignment of Single Molecule Maps

Keywords: Optical Maps, Alignment, Rmap, Restriction Map

Abstract: An optical map is an ordered genome-wide high-resolution restriction map that indicates the positions of one or more short nucleotide sequences. Since optical maps are derived independently of short sequence reads, they are used in *de novo* genome assembly for validating a draft genome (Nature 2013), finding structural variation (Nature Methods 2015), and detecting mis-assembled regions within draft genomes (ISMB 2015). Single molecule maps, referred to as *Rmaps*, is the raw optical mapping data used to construct the genome-wide optical maps which are used for aiding in genome assembly. Currently, there exists very few computational methods to find alignments between the Rmaps - a task that is the first step in assembling the Rmap data into a genome-wide optical maps. and is challenging due to various Rmap specific errors and their frequency.

We present DOPPELGANGER, the first index-based, fully error-tolerant alignment method for optical mapping data. All prior alignment methods use dynamic programming, which is computationally expensive, or have limited error tolerance. We demonstrate a 20x speedup on the plum genome and demonstrate on the Ecoli genome the validity of our alignments. Thus, DOPPELGANGER is the only non-proprietary method that is capable of performing pairwise Rmap alignment for large eukaryote organisms in reasonable time. Lastly, we conclude with other applications of DOPPELGANGER, such as the alignment of long reads with high error rate (e.g. PacBio reads) to a genome-wide optical map.

Availability: The DOPPELGANGER alignment method is available for download at https://github.com/mmuggli/doppelganger.

Authors:

first name	last name	email	country	organization	corresponding?
Martin D.	Muggli	muggli@cs.colostate.edu	United States	Colorado State University	✓
Simon J.	Puglisi		Finland	University of Helsinki	
Christina	Boucher	Christina.Boucher@colostate.edu	United States	Colorado State University	✓

Consensus Representation Estimation of Lineage Expression *CREoLE* Algorithm for scRNA-seq

Keywords: Algorithms, Bioinformatics, Signal processing

Abstract: Single-cell RNA-sequencing (scRNA-seq) is a promising technology widely used to recapitulate gene expression trends through developmental progression of heterogeneous biological tissue. Although several methods have sought to estimate pseudo-temporal gene expression trends, a number of technical limitations presented by scRNA-seq remain, including high expression variability and drop-out measurements, which complicate accurate trend estimation. Consensus Representation Estimation of Lineage Expression (CREoLE) is an efficient and robust algorithm for automatically detecting an underlying branching lineage structure and estimating smooth developmental trends of gene expression by consensus without the need for model fitting or data filtering.

Applied to synthetic data, CREoLE recapitulates underlying gene expression for each gene and across each lineage with an average Pearson correlation coefficients of 0.983 ± 0.009. The impact of simulated technical noise, drop-out measurements, and cell count reduction are evaluated. Our analysis suggests that Creole is robust to additive noise and smaller initial cell populations (as low as 25% initial population). CREoLE correlates with synthetic expression trends with a mean Pearson's correlation coefficient above 0.9 in all cases. Applied to real data, CREoLE accurately identifies lineage structure and computes high-resolution consensus trends that align closely with published findings.

Availability: CREoLE is free and open-source software available from https://github.com/schaugf/creole

Authors:

first name	last name	email	country	organization	corresponding?
Geoffrey F.	Schau	schau@ohsu.edu	United States	Oregon Health and Science University	✓
Shannon	McWeeney		United States	Oregon Health and Science University	
Andrew	Adey	adey@ohsu.edu	United States	Oregon Health and Science University	✓

Organellar Genomes of White Spruce (Picea glauca): Assembly and Annotation

Shaun D Jackman <sjackman@bcgsc.ca>
Genome Sciences Centre, British Columbia Cancer Agency
Vancouver, Canada

The genome sequences of the plastid and mitochondrion of white spruce (Picea glauca) were assembled from whole-genome shotgun sequencing data using ABySS. The sequencing data contained reads from both the nuclear and organellar genomes, and reads of the organellar genomes were abundant in the data as each cell harbors hundreds of mitochondria and plastids. Hence, assembly of the 123-kb plastid and 5.9-Mb mitochondrial genomes were accomplished by analyzing data sets primarily representing low coverage of the nuclear genome. The assembled organellar genomes were annotated for their coding genes, ribosomal RNA, and transfer RNA. Transcript abundances of the mitochondrial genes were quantified in three developmental tissues and five mature tissues using data from RNA-seq experiments. C-to-U RNA editing was observed in the majority of mitochondrial genes, and in four genes, editing events were noted to modify ACG codons to create cryptic AUG start codons. The informatics methodology presented in this study should prove useful to assemble organellar genomes of other plant species using whole-genome shotgun sequencing data.

Chloroplast genomes of gymnosperms, including conifers, are well studied, but little is known about the mitochondria of gymnosperms. In fact, only a single gymnosperm mitochondrion is found in NCBI GenBank. This nearest related mitochondrial sequence is of the Prince Sago palm (Cycas taitungensis) native to Taiwan, which diverged from the white spruce over a hundred million years ago. No conifer mitochondrion genomes are to be found in GenBank at all, until now.

Roughly one percent of the whole genome sequencing reads of white spruce are from its two organellar genomes: the chloroplast and mitochondrion. We assembled these reads using ABySS and found the mitochondrion genome to be nearly six megabases, which is unusually large for a mitochondrial genome. Although many genes typical of mitochondria were found in the genome, most open reading frames had no similarity to any known gene.

White spruce is an economically important species to the forestry industry of Canada. Insights into the conifer mitochondrial genome will provide relevant new information to reconstruct the evolution of this organelle genome relative to other plant lineages, and to identify which genes of a conifer are uniquely inherited through the mitochondria. As the mitochondrial genome is inherited maternally, and the plastid genome is inherited paternally, having a complete genome sequence for both organelles would enable classifying the maternal and paternal species of hybrid seed lots and determining the maternal and paternal lineage of saplings in breeding experiments.

AuPairWise: biologically focused RNA-seq quality control using co-expression

Ballouz, S. and Gillis, J.

Stanley Institute for Cognitive Genomics, Cold Spring Harbor Laboratory, 500 Sunnyside Boulevard Woodbury, NY 11797, USA.

A principal claim for RNA-sequencing has been greater replicability, typically measured in sample-sample correlations of gene expression levels. Replicability of transcript abundances in this way will provide misleading estimates of the replicability of conditional variation, which is what is of interest in most expression analyses. Heuristics which implicitly address this problem have emerged in quality control measures to obtain 'good' differential expression results. However, these methods involve strict filters such as discarding low expressing genes or using technical replicates to remove discordant transcripts, and are costly or simply ad hoc.

Instead, we show that gene-level replicability is a more useful metric, and demonstrate that it can be modeled in a co-expression framework, using known co-expressing gene pairs as pseudo-replicates instead of true replicates. We use this as a quality control metric: by modelling the effects of noise that perturbs a gene's expression, we can then measure the aggregate effect of this perturbation on these co-expressing gene-pairs or 'housekeeping interactions'. We find that perturbing expression by only 5% within its usual range of values is readily detectable (AUROC~0.73), suggesting this test is extraordinarily sensitive. In addition to making the software readily available (github.com/sarbal/AuPairWise), we have adapted the test to optimize RNA-seq alignment with the STAR aligner tool. Our findings suggest that more stringent parameters at the read mapping stage (e.g., minimum alignment scores) would have a modestly positive impact, making the post-hoc filtering done for high-expressing or high fold-changes a more intuitive part of direct quality control.

Indrani Datta, MS, Biostatistician, 3138746229, idatta1@hfhs.org

Co-Expression of Long non-coding RNAs with Epigenetically regulated genes in TCGA Glioma subtypes

Indrani Datta[1,2,3], Laila M. Poisson[1,2,3]

Center for Bioinformatics[1], Public Health Sciences[2], Hermelin Brain Tumor Center[3], Henry Ford Health System, Detroit, Michigan

Background- In recent years RNA-seq deep sequencing technology has emerged as a revolutionary tool to precisely measure transcriptome profiling in eukaryotic genomes. Beyond protein coding RNAs, long non-coding RNAs (lncRNAs) have become recognized as a gene regulators as well as prognostic markers in cancer. In this study, we initiated an *in-silico* analysis of co-expression of lncRNAs with epigenetically regulated genes (EReg) in TCGA Glioblastoma multiform (GBM) and Lower Grade Glioma (LGG) RNA-seq data.

Method- Open-source RNA-seq data set which is manufactured with Illumina HiSeq platform from TCGA GBM and LGG cohort were integrated to capture highly correlated bio-molecules, in our case, lncRNAs and ERegs. A set of 12382 differentially regulated lncRNAs transcripts were identified across various cancers including GBM & LGG samples (372) were derived from Chinnaiyan *et.al* identified by the Tuxedo suite (i.e, Tophat, Cufflink) which perform many aspects of complete RNA-seq analysis in *ab initio* assembly mode. A set of 809 EReg transcripts were obtained from Cecceralli *et al*. which categorizes 7 distinct glioma subtypes in IDHmutant (codal=69,G-CIMP-high=104,G-CIMP-low=8) and IDHwildtype (Classic-like=54,LGm6-GBM=12,Mesenchymal-like=69,PA-like=15) by unsupervised clustering of Illumina methylation 27k and 450k array probes. The expression estimates of these EReg transcripts were generated by Mapsplice/RSEM workflow constructed by broad Institute, were downloaded from GDAC. Expression estimates of lncRNAs were in FPKM and ERegs were in estimated transcript fraction, as these two measures were generated by two different algorithms/workflow, so they were made compatible by converting to transcripts per million (TPM). Following data processing and QC on this integrated data, 315 samples and 12991 (lncRNA=12195 and EReg=796 transcripts) molecules were carried forward for analysis with Weighted Correlation network analysis. Following detection of networks which consists of lncRNAs and ERegs, association of these co-expression networks to glioma subtypes were analyzed with Anova. EReg genes from significantly associated modules were further analyzed by Ingenuity's IPA to delineate biological association as majority of lncRNAs have unknown functions, so "guilt by association" mechanism was used for retrieving functional relevance to these lncRNAs by EReg genes.

Results: There were 27 lncRNA-EReg gene modules were detected. Among these, 2 modules were significantly associated 2 glioma subtypes (IDHWt = PA-Like and IDHWt = LGm6-GBM) at pvalue < 0.05. After multiple testing corrections, both of these modules remain as significant at FDR level < 0.05. EReg genes which were extracted from module associated with LGm6-GBM are working together in cell-To-cell Signaling and Interaction, cellular Growth and Proliferation while EReg genes associated with PA-Like glioma subtype were working together in cell cycle, cellular development, cellular growth and proliferation biological functions. So it can be assumed that lncRNA transcripts which were co-expressed with ERegs transcripts from above mentioned modules will participate in these cellular functions.

Conclusion- This study demonstrates the application of existing bioinformatics algorithms to analyze open source RNA-seq data to capture gene-lncRNA association in respect to glioma subtypes.

ntHash: recursive nucleotide hashing

Hamid Mohamadi, Justin Chu, Benjamin P Vandervalk and Inanc Birol
Canada's Michael Smith Genome Sciences Centre,
British Columbia Cancer Agency, Vancouver, BC, V5Z 4S6, Canada

Background
In bioinformatics, there are many applications that rely on cataloguing or counting DNA/RNA sequences for indexing, querying, and similarity search. These include sequence alignment, genome and transcriptome assembly, RNA-seq expression quantification, and error correction. An efficient way of performing such operations is through the use of hash-based data structures, such as hash tables or Bloom filters. Therefore, improving the performance of hashing algorithms would have a broad impact for a wide range of bioinformatics tools.

Results
Here, we present ntHash, a fast hash method for computing hash values for all possible sub-sequences of length *k* (*k*-mers) in a DNA sequence. The algorithm calculates hash values for consecutive *k*-mers in a given sequence using a recursive approach, in which the hash value of the current *k*-mer is derived from the hash value of the previous *k*-mer. In this work, we have implemented a cyclic polynomial rolling hash function, and adapted it to nucleotide hashing. Particularly, we made use the reduced alphabet of DNA sequences, and handled the reverse complementation efficiently. The proposed method also provides a fast way for calculating multiple hash values for a given *k*-mer without repeating the whole hashing procedure for each value. This functionality would be very useful for certain bioinformatics applications, such as those that utilize the Bloom filter data structure.

Conclusions
Experimental results demonstrate substantial speed improvement over conventional approaches, while retaining near-ideal hash value distribution. Comparison of run time of proposed method with the state-of-the-art general-purpose hash functions demonstrates that ntHash performs over 20x faster than the closest competitor, *cityhash*, the leading algorithm developed by Google.

NanoSim: nanopore sequence read simulator based on statistical characterization

Chen Yang, Justin Chu, René L Warren, Inanç Birol

Background:
The MinION sequencing platform from Oxford Nanopore Technologies (ONT) is still a pre-commercial technology, yet it is generating substantial excitement in the field for its features – longer read lengths and single-molecule sequencing in particular. As groups start developing bioinformatics tools for this new platform, a method to model and simulate the properties of the sequencing data will be valuable to test alternative approaches and to establish performance metrics. Here, we introduce NanoSim, a fast and lightweight read simulator that captures the technology-specific characteristics of ONT data with robust statistical models.

Results:
The first step of NanoSim is read characterization, which provides a comprehensive alignment-based analysis, and generates a set of read profiles serving as the input to the next step, the simulation stage. The simulation tool uses the model built in the previous step to produce in silico reads for a given reference genome. NanoSim is built on our observation that patterns of correct base calls and errors (mismatches and indels) can be described by statistical mixture models. Further, the structures of these models are consistent across chemistries and organisms (*E. coli* and *S. cerevisiae*). NanoSim generates synthetic ONT reads with empirical profiles derived from reference datasets, or using runtime parameters. Empirical profiles include read lengths and alignment fractions (the ratio of alignment lengths after unaligned portions of reads are soft-clipped from their flanks to read lengths). The lengths of intervals between errors (stretches of correct bases) and error types are modeled by Markov chains, and the lengths of errors are drawn from mixed statistical models.

Conclusion:
In this work, we demonstrate the performance of NanoSim on publicly available datasets generated using R7 and R7.3 chemistries and different sequencing kits. NanoSim mimics ONT reads well, true to the major features of the emerging ONT sequencing platform, in terms of read length and error modes. The independent profiling module grants users the freedom to characterize their own ONT datasets, which is expected to perform consistently upon the improvement of nanopore sequencing technology, as the shapes of the error models hold among different datasets. NanoSim

will immediately benefit the development of scalable NGS technologies for the long nanopore reads, including genome assembly, mutation detection, and even metagenomic analysis software. The scalability of NanoSim to human-size genome will benefit the development of scalable NGS technologies for long nanopore reads. Moreover, a mixture of in silico genomes simulating a microbiome will be helpful for benchmarking algorithms with applica-tion in metagenomics, including functional gene prediction, species detection, comparative metagenomics, clinical diagnosis. As such, we expect NanoSim to have an enabling role in the field.

ISMB, HitSeq 2016 Abstract for Oral / Poster Presentation

Template-Based Decomposition of ChIP-exo Profile Reveals Alternative Binding Configuration Repertoire of Transcription Factors

Hee-Woong Lim*# and Kyoung-Jae Won*

The Institute for Diabetes, Obesity, and Metabolism, Department of Genetics, Perelman School of Medicine at the University of Pennsylvania, 3400 Civic Center Blvd, Philadelphia, PA 19103

*To whom correspondence should be addressed.

#Speaker

Contact: heewlim@mail.med.upenn.edu; wonk@mail.med.upenn.edu

Background

ChIP-exo is a next generation sequencing technique to identify genomewide transcription factor (TF) binding sites in single-nucleotide resolution. Improved from the predecessor ChIP-seq, ChIP-exo utilizes an exonuclease to trim off extra 5-prime ends of ChIP DNAs until the enzyme meets exact binding sites. Thanks to its high resolution sensitivity, ChIP-exo is becoming an increasingly popular method for delineating previously unseen landscapes of protein-DNA interaction. Most of the computational analysis pipelines for ChIP-exo data so far depend solely on DNA motif sequences enriched around ChIP-exo peaks to identify exact binding sites in high resolution. However, depending on motif information alone is subject to false positive or false negative errors frequently (**Figure1a**), especially because the presence of a specific motif does not guarantee target protein binding at the locus or a protein can bind to a suboptimal motif in cooperation with other factors.

Result

To overcome this limitation, we propose a template-based decomposition framework for ChIP-exo data analysis. Our framework consists of two major parts: 1) optimized motif mining from ChIP-exo peak-pairs having frequent distances and 2) genomewide template scan of ChIP-exo signal pattern derived from the motifs (**Figure1b**). The basic idea is utilizing ChIP-exo signal at each candidate binding sites in addition to the motifs. After motif mining within ChIP-exo peak-pairs having frequent distances, a template is prepared by aggregating ChIP-exo signal at high quality motif loci. Then all the random signals from false-positive binding sites are averaged-out and only real ChIP-exo signal pattern will be preserved. This template is used for genomewide scan to identify real binding sites corresponding to the motif.

We applied our method to interrogate glucocorticoid receptor (GR) binding in vivo, which is an essential factor for life. From the motif search, we obtained a canonical GR dimer motif (GRE) from the motif search. First, as a control analysis, we selected high quality motif loci with $p < 10^{-4}$ (**Figure1c**). Then we prepared a ChIP-exo template from the GRE motif and did genomewide scan with it. Remarkably, more than 30% of the high-quality motif loci were false positive (**Figure1d**) even though they were located within strong ChIP-seq peaks (**Figure1e**). We also identified substantial number of dimeric binding at suboptimal motif loci (**Figure1d**) that were missed when using motif information only.

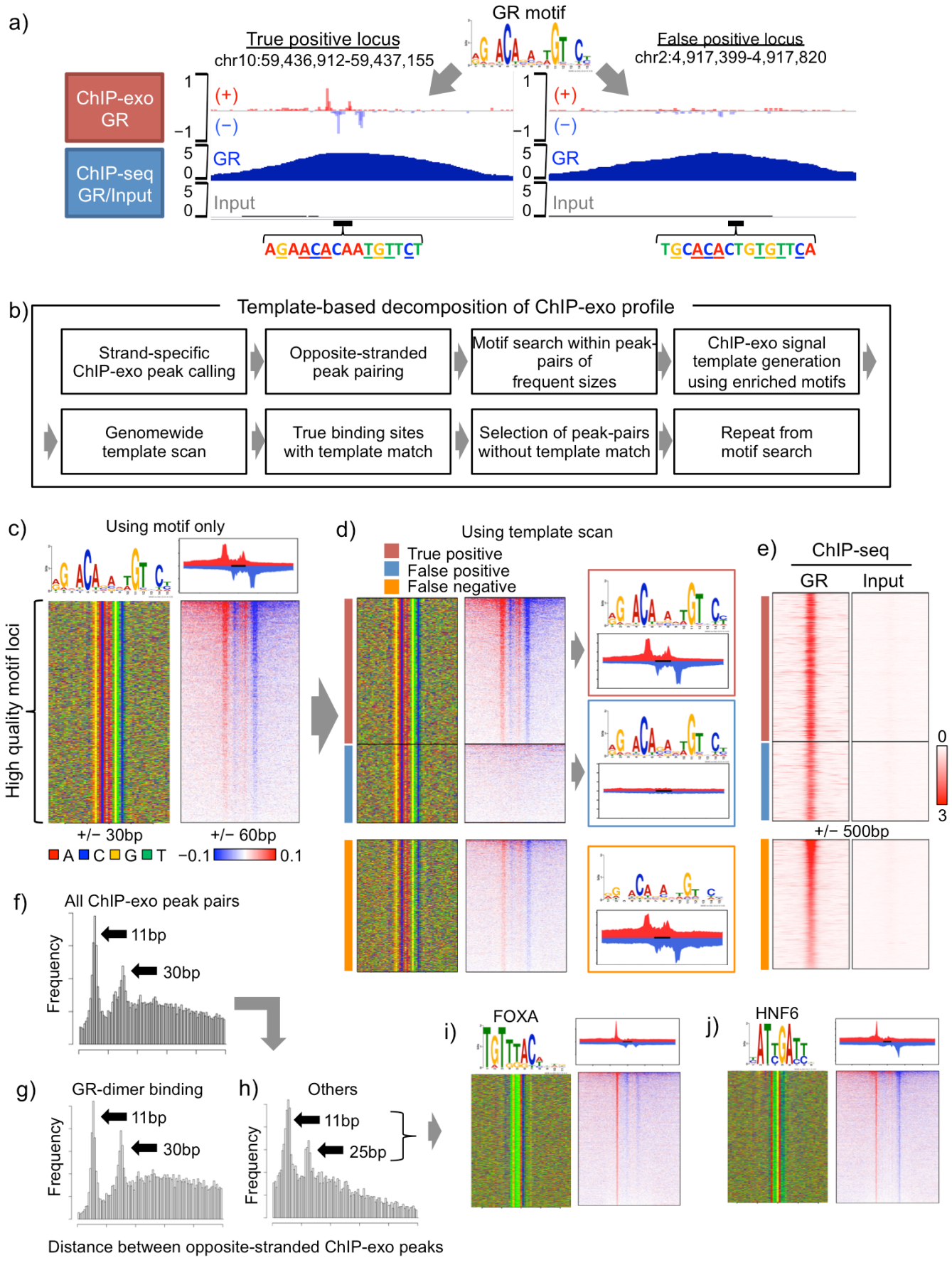
Then we selected ChIP-exo peak-pairs that do not match with the GRE ChIP-exo template (**Figure1f-h**) for the next round analysis. From the second motif search, we obtained HNF6 and FOXA motifs, which were not significantly enriched at the first round. Subsequent template scans revealed thousands of ChIP-exo signature of HNF6 and FOXA binding (**Figure1i-j**), which suggests alternative configurations of GR-binding mediated by these lineage factors.

We also successfully applied our method to other ChIP-exo datasets, such as estrogen receptor (ER) in a breast cancer cell line, SOX2 in mouse embryonic stem cell, etc. The results will be presented at the conference.

Conclusion

Here, we proposed a novel framework for ChIP-exo data analysis based on a template scan for better accuracy and robust identification of alternative binding configuration. There have been many biochemical works to investigate genomic binding of TFs such as PBM or SELEX. However, our method provide more realistic platform to study endogenous binding of TFs in vivo, which will extend our understanding of various repertoire of TF-DNA interactions for gene regulations.

Figure 1


a)
GR motif
True positive locus
chr10:59,436,912-59,437,155
False positive locus
chr2:4,917,399-4,917,820
ChIP-exo GR
ChIP-seq GR/Input
(+)
(−)
GR
Input
AGAACACAATGTTCT
TGCACACTGTGTTCA
b)
Template-based decomposition of ChIP-exo profile
Strand-specific ChIP-exo peak calling
Opposite-stranded peak pairing
Motif search within peak-pairs of frequent sizes
ChIP-exo signal template generation using enriched motifs
Genomewide template scan
True binding sites with template match
Selection of peak-pairs without template match
Repeat from motif search
c)
Using motif only
High quality motif loci
+/− 30bp
+/− 60bp
A C G T
−0.1
0.1
d)
Using template scan
True positive
False positive
False negative
e)
ChIP-seq
GR
Input
+/− 500bp
0
3
f)
All ChIP-exo peak pairs
Frequency
11bp
30bp
g)
GR-dimer binding
11bp
30bp
h)
Others
11bp
25bp
Distance between opposite-stranded ChIP-exo peaks
i)
FOXA
j)
HNF6

(Submission for poster session)

Hybrid genome assembly of Ogye (*Gallus gallus domesticus*) using short and long reads and annotations of noncoding genes.

Kyoungwoo Nam[1], Jang-il Sohn[1], Hyosun Hong[1] and Jin-Wu Nam[1,2,*]

[1]Department of Life Science, College of Natural Sciences, Hanyang University, Seoul 133-791, Republic of Korea

[2]Research Institute for Natural Sciences, Hanyang University, Seoul 133-791, Republic of Korea

Background: Because of ongoing decrease in cost of high-throughput sequencing (HTS), studies for genome assembly and noncoding gene annotations have been getting popular for not only model organisms but also non-model organisms. Ogye, a Korean traditional Gallus gallus breed, is well known for its unique phenotypical characteristics of black leather, skin, fascia, and sclera, and also have strong immune resistance against some specific diseases, such as Marek's disease and avian influenza, in the Korean poultry industry.

Results: To study of the phenotypical characteristics of Ogye in genome level, we first sequenced Illumina (60X paired-end and 170X mate-pair) and PacBio libraries (11X) of Ogye genome, and assembled a draft genome using our hybrid genome assembly pipeline, which consists of ALLPATHS-LG, SSPACE-LongRead, OPERA-LG, PBJelly, LoRDEC, etc. The resulting draft genome of Ogye displayed a high quality of N50 (133 Kbp for contig and 21.2 Mbp for scaffold), and the scaffold N50 length of which is better than that of *Gallus gallus* (Galgal4.0). We also constructed noncoding transcriptome maps on the draft genome and profiled their expression across 20 different tissues including the skin, fascia, and eye by sequencing RNA-seq and small RNA-seq. As a result, we found 23 microRNA (miRNA) and 316 long intervening ncRNAs (lincRNAs) specifically expressed in the black tissues.

Conclusions: We expect that our genomic and transcriptomic resources could provide insights of the genomic evolution during *Gallus gallus subspeciation* and of the medical implication for the viral infection and immune-related diseases.

* Corresponding author: jwnam@hanyang.ac.kr

Keywords : Genome assembly, lncRNA, and miRNA

In Silico Simulation of Low Allele Fraction Gene Rearrangement Detection with Deep Targeted DNA Sequencing

Onur Sakarya[1], Hyunsung John Kim[1], Roger Jiang[1], Tom Chien[1], Payal Shah[1], Hui Xu[1], Chenlu Hou[1], Byoungsok Jung[1], Xiaoyu Chen[2], Han-Yu Chuang[2], and Catalin Barbacioru[1]
[1] GRAIL Inc., Redwood City, CA, 94063
[2] Illumina Inc. San Diego, CA, 92122

Background:
Gene rearrangements are prominent somatic mutations driving cancers. Recent studies identified an increasing number of recurrent gene rearrangements in solid tumors. For example, more than 5% of patients with non-small cell lung cancer (NSCLC) harbor a rearrangement of ALK, ROS1, or RET genes each with multiple partners. Catalogue of Somatic Mutations in Cancer (COSMIC) database provides a curated list of gene rearrangements (*1*). Most of COSMIC gene rearrangement cases are based on RNA sequencing and report the fused exon coordinates of partner genes. However, at DNA level, most breakpoints happen on introns more of than on exons. Furthermore, most introns are in proximity of homologous, low complexity and repeat sequences. Thus, it is more challenging to detect gene rearrangements at DNA level.

We tackled the problem of estimating sensitivity of targeted DNA gene rearrangement detection as a function of breakpoint location within expected introns. We generated random breakpoint templates and simulated artificial reads from these templates in a titration setting. Simulated reads were titrated at low levels to background canonical intron reads to imitate circulating cell-free (cfDNA) setting. We processed the reads through our rearrangement calling pipeline to evaluate the sensitivity and specificity. We also tested our method on real sequencing data from titrated cell lines, a cell line mix and cfDNA from metastatic NSCLC patients whose tissue biopsy confirmed certain oncogenic gene rearrangements.

Results:
We simulated rearrangement breakpoints from 215 COSMIC rearrangements spanning 360,495 base pairs in 87 introns of 28 genes. We required a breakpoint to be at least 500bp apart from an existing breakpoint that was already in the simulation pool. Rearrangements were titrated from 0.2 to 5% within 3000 fragments covering each breakpoint. We repeated the simulation 100 times for each titration level, each time permuting the order of rearranged genes. For each fragment, we simulated 150bp paired-end reads from 167bp long (σ=50) fragments using HiSeq 2500 error profile with ART 2.3.7 (*1*). A custom pyflow (*2*) pipeline was used to map reads with bwa 0.7.10-r789 (*3*) and call breakpoints with Manta-0.29.3 software (*4*). Manta is a two step structural variant detection algorithm based on construction of a breakpoint graph followed by local assembly of individual regions, contig alignment, scoring and calling.

Sensitivity of rearrangement detection was above 99% at 1 to 5% allele fraction (AF), 98% at 0.5% AF, and 73% at 0.2% AF. In general, precise location of the breakpoint was more difficult to detect at lower AF due to lower number of reads going into the assembly process. We

required three paired-end reads as the threshold evidence to initiate the assembly process, which gave an approximate location in the absence of single reads spanning the breakpoint. Improvements to preciseness of the calls were demonstrated with improvements to assembly process. False discovery rate was 0.3% overall for all titration levels.

As real test cases, we deep-sequenced 8 plasma cfDNA samples with known tumor gene rearrangements (EML4>ALK, KIF5B>ALK and CD74>ROS1). We called the associated breakpoint from plasma cfDNA in all cases. Detected breakpoint AF ranged from 0.4 to 12%. There were no false positive breakpoints detected. Two of the cases were biological replicates, i.e. two tubes of whole blood from the same patient. Translocations usually create two reciprocal breakpoints and in one of the replicate cases, both samples had reciprocal calls. In the other replicate case, one sample had a reciprocal call and its replicate sample did not, suggesting reciprocal events may exist in the absence of a reciprocal call. We also sequenced and called gene rearrangements from individual titrated cell lines HCC78 (SLC34A2> ROS1) and H2228 (EML4>ALK) at different input titration levels and Horizon HD753 Structural Variant mix (CCDC6>RET and SLC34A2>ROS1) at AF in the range of 2 to 4%.

Conclusions:

We simulated majority of gene rearrangement breakpoints documented in COSMIC and demonstrated performance of a structural variant calling pipeline at cfDNA setting to achieve high sensitivity and low false discovery rate. We further investigated the performance of oncogenic rearrangement calls from patient plasma cfDNA samples and their localization and reciprocity.

References:

1. Forbes et al. COSMIC: exploring the world's knowledge of somatic mutations in human cancer. Nucleic Acids Research (2015) 43 (D1): D805-D811.
2. Huang, Weichun et al.ART: A next-Generation Sequencing Read Simulator. Bioinformatics (2012) 28 (4): 593–594.
3. Pyflow – a lightweight parallel task engine. https://github.com/Illumina/pyflow
4. Li, Heng, and Richard Durbin. Fast and Accurate Short Read Alignment with Burrows–Wheeler Transform. *Bioinformatics* (2009) 25 (14): 1754–1760.
5. Chen, Xiaoyu et al. Manta: Rapid detection of structural variants and indels for clinical sequencing applications. Bioinformatics (2016) 32 (8): 1220-1222.

Allele-specific expression from single-cell RNA-Seq data

Kwangbom Choi, Narayanan Raghupathy, Steven C. Munger, Gary A. Churchill
The Jackson Laboratory, Bar Harbor, Maine 04609, U.S.A.

Background

In diploid cells, two allelic copies of a gene can differ in the timing and the level of their expression as determined by genetic, environmental, and stochastic factors. With high-throughput sequencing (HTS) technologies, we can now resolve how alleles are preferentially expressed in tissue samples and in individual cells. This information not only reveals which alleles are preferentially expressed, but enables us to decode how gene expression is regulated. This insight is fundamental for understanding the genetic architecture underlying normal phenotypic diversity and disease.

Although the application of single-cell RNA-Seq methods (scRNA-Seq) adds valuable information about the dynamics of allelic expression that is lost in whole tissue samples, it poses multiple technological challenges. High level of sampling noise is common, often accompanied by relatively low depth of coverage (below 10 million reads per cell). Due to inefficient, non-random reverse transcription process, alleles may drop out from measurements. Accurate quantification of allele-specific expression (ASE) from scRNA-Seq data requires allelic variation to distinguish reads from different alleles, but many reads will not overlap polymorphic sites and their origins are ambiguous.

We propose an empirical Bayes model in which we disambiguate the origins of multiply-aligning reads (or multireads), quantify ASE in individual cells using all the aligned reads, refine ASE in each cell referring to other cells in similar expression state, and classify genes by summarizing how cells behave across the population on their transcription events.

Results

Relying solely on uniquely-aligning reads was not a viable option for quantifying ASE from scRNA-Seq data. In many cases, over 80% of reads had to be filtered out just because they aligned to multiple locations of diploid transcriptome. Discarding multireads increases the variability of ASE across genes and results in false discoveries, for example, hundreds of false monoallelic expression patterns. Our model also found strong evidence on coordinated expression of alleles in many genes: gene-specific odds of joint expression of two alleles suggests positive correlation. We were able to classify genes into seven categories with respect to allelic expression state of cell population: maternal monoallelic, paternal monoallelic, biallelic expression, and four combinatorial mixture of those three base classes, including mutually exclusive allelic expression. The classifier helped us, for example, to identify genes dynamically change their expression state along the progress of tissue development.

Conclusions

Our model overcome data sparsity in each individual cell by combining information from other cells of a kind, and control overdispersion by allowing cellular heterogeneity on allele proportion via a hierarchical model. We also provide a heterogeneous model to describe sub-populations of cells resulting from random mono-allelic expression in scRNA-Seq data, which thus enable us to derive shrinkage estimation on allele specificity out of cells in diverse expression states.

Long read- and DNA methylation-based binning of metagenomic contigs and single molecules

Background

Whole-metagenome shotgun sequencing is a comprehensive approach for characterizing the population structure and genetic architecture of complex microbial communities. However, significant challenges arise in the analysis of metagenomic sequences, often stemming from the presence of bacterial species and strains with high sequence similarity, complex DNA repeats, and widely varying relative abundance. Short-read metagenomic assemblies usually result in many thousands of short-to-medium length contigs that subsequently must be annotated using existing reference sequences or segregated by taxa through the process of binning.

Supervised binning methods require existing references to train classification algorithms, while unsupervised (reference-free) methods do not rely on any training data and therefore have the potential to identify novel species. Most reference-free binning methods attempt to cluster contigs from a metagenomic assembly, either using sequence composition alone or using coverage covariance statistics[1]. Sequence composition-based approaches, however, are limited by the fragmented nature of the *de novo* assemblies and often fail to distinguish between genomes with high sequence homology. Coverage covariance-based methods can provide additional power to separate similar genomes, but require the sequencing of many related samples, which is often not feasible, especially in clinical settings when in-depth analysis of few or a single microbiome is desired.

Single-molecule, real-time (SMRT) sequencing has the potential to address many of these challenges, but its applicability in metagenomics has not been extensively explored. Here, we present multiple novel methods for binning metagenomic sequences that not only leverage the long read lengths of SMRT sequencing, but also, for the first time, utilize the DNA methylation signatures inferred from these reads to resolve assembled contigs and single reads into clusters representing distinct biological entities. The diversity of methylation systems found in the bacterial world[2,3], often observed between closely related species and strains, suggests that DNA methylation profiles can be leveraged as an epigenetic feature to separate taxa with high sequence homology. Using single-molecule methylation detection[4] and metagenomic sequencing data from several synthetic microbiome communities and infant gut microbiota, we demonstrate that the proposed methods can segregate both assembled contigs and low abundance reads at a high resolution.

Results

The proposed contig- and read-binning framework relies on the use of two types of features: sequence composition as assessed by k-mer frequencies and DNA methylation profiles (Figure 1a). To evaluate the power of these binning methods, we build upon dimensionality reduction methods that have been shown to provide relatively effective segregation of metagenomic contigs using the t-distributed stochastic neighbor embedding (t-SNE) algorithm.

While previous methods used sequence composition and t-SNE to bin assembled contigs, we incorporate the DNA methylation profiles and further extend the application to unaligned SMRT reads. The inclusion of DNA methylation profiles for binning purposes significantly increases the ability to segregate contigs from closely related species and strains in an infant microbiome sample. By extending the binning methods to unaligned SMRT reads (Figure 1b), we can remove the requirement for a successful *de novo* assembly and thus highlight taxon-specific clusters for very low-abundance members of a metagenomic mixture that would otherwise fail to generate contigs. Finally, we can improve the quality of multi-strain metagenomic assemblies by first binning the reads based on DNA methylation profiles. The binned reads are then assembled separately, generating strain-specific assemblies without the contig fragmentation and chimerism that occurs when multiple strains are assembled together.

Conclusions

The methods discussed here are empowered by emerging sequencing technologies that are not subject to the limitations of short read lengths or biases in amplification and sequencing of GC-rich regions, allowing a more accurate and comprehensive reconstruction of microbial genomes in a metagenomic sample. Furthermore, this work represents the first time that DNA methylation

signatures have been used to improve metagenomic binning and assembly, an advance that facilitates the separation of high-homology species and strains, both at the level of assembled contigs and raw reads.

In addition to helping examine the taxonomic diversity in metagenomic sequencing data, the method will also be useful for associating plasmids to their bacterial hosts, as the plasmids inherit the same methylation signature as their host despite possible differences in sequence composition. This is an important step in understanding the full genomic potential of a certain species in a sample. In addition, bacteriophages that are often responsible for the transmission of antibiotic resistance carry methylation signatures indicative of their most recent host organism. Therefore, methylation profiles can be used to track the transmission of bacteriophages and associated antibiotic resistance elements among species and strains.

The sensitivity and effectiveness of the approaches described here, and interest in long read metagenomic sequencing generally, will only increase as third generation sequencing technologies continue to mature, generating larger yields and longer reads. These methods will serve as a framework for the development of further approaches that take advantage of the unique features of existing and emerging third generation sequencing technologies for characterization of metagenomic communities.

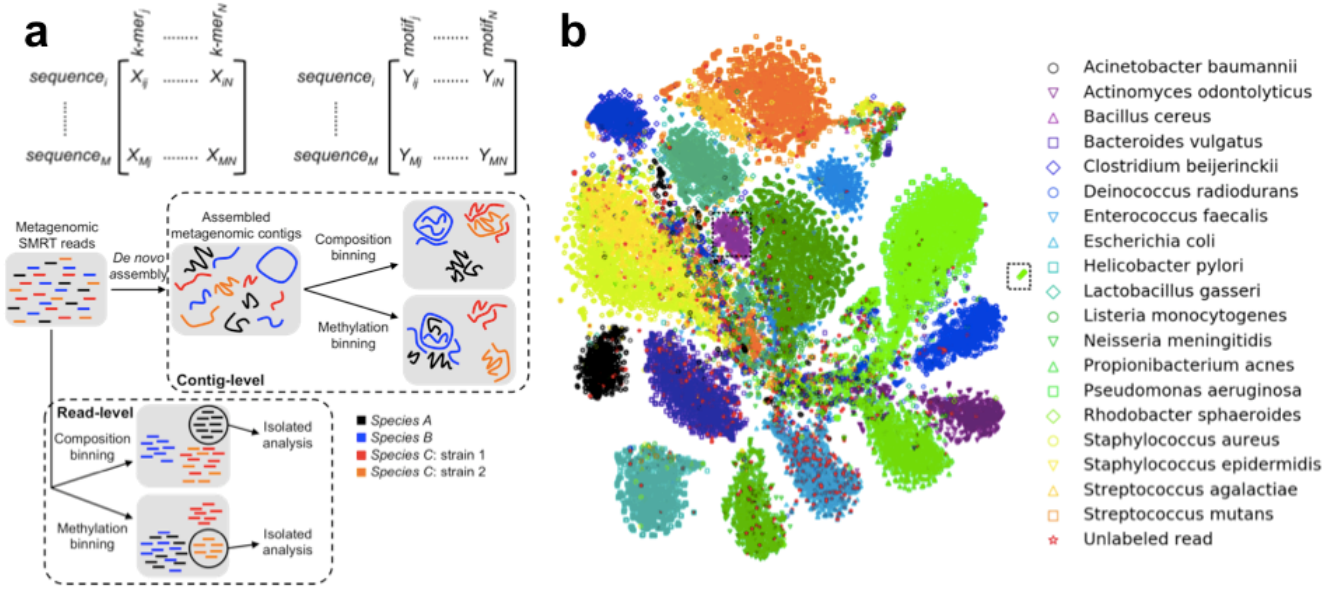


Figure 1: (a) Schematic of metagenomic SMRT binning approaches, where composition-based binning helps separate divergent species, while methylation-based binning helps untangle highly similar genomes. The two binning approaches can also be combined to leverage the strengths of each. (b) Composition-based binning of unaligned reads (length>15kb) from a Human Microbiome mock community containing 20 members. Clusters of reads belonging to two low-abundance species, *Bacillus cereus* and *Rhodobacter sphaeroides*, are highlighted.

1. Nielsen, H. B. *et al.* Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. *Nat. Biotechnol.* **32,** (2014).
2. Fang, G. *et al.* Genome-wide mapping of methylated adenine residues in pathogenic Escherichia coli using single-molecule real-time sequencing. *Nat. Biotechnol.* **30,** 1232–9 (2012).
3. Blow, M. J. *et al.* The Epigenomic Landscape of Prokaryotes. *PLOS Genet.* **12,** e1005854 (2016).
4. Beaulaurier, J. *et al.* Single molecule-level detection and long read-based phasing of epigenetic variations in bacterial methylomes. *Nat. Commun.* **6,** 7438 (2015).

Compact universal *k*-mer hitting sets

Yaron Orenstein[1], David Pellow[2], Guillaume Marçais[3], Ron Shamir[2], and Carl Kingsford[3]

[1] Computer Science and Artificial Intelligence Laboratory, MIT, Cambridge, MA, USA
[2] Blavatnik School of Computer Science, Tel-Aviv University, Tel-Aviv, Israel
[3] School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA
yaronore@mit.edu, dpellow@tau.ac.il, gmarcais@cs.cmu.edu, rshamir@tau.ac.il, carlk@cs.cmu.edu

1 Background

We consider the following problem involving covering strings by selecting short k-mer substrings:

Problem 1. Given integers k and L, find a smallest set U_{kL} of k-mers such that any string of length L or longer must contain at least one k-mer from U_{kL}.

The set U_{kL} is called a *universal* set of hitting k-mers, and we call each k-mer in the set *universal*. Such a set has a number of applications in speeding up genomic analyses since it can often be used in places where minimizers have been used in the past: hashing for read overlapping, sparse suffix arrays and Bloom filters to speed up sequence search.

A universal set U_{kL} has a number of advantages over minimizers for these applications. First, the set of minimizers for a given collection of reads may be as dense as the complete set of k-mers, whereas we show below that U_{kL} is often smaller by a factor of k. Second, for any k and L, the set of universal k-mers needs to be computed only once and not recomputed for every dataset. Third, the hash buckets, sparse suffix arrays, and Bloom filters created for different datasets will contain a comparable set of k-mers if they are sampled according to U_{kL}. The universal set of k-mers also has the advantage over dataset-specific sets because one does not need to look at all the reads before deciding on the k-mers to use, and one does not need to build a dataset-specific de Bruijn graph to select covering k-mers.

The problem is also of theoretical interest as it can be rephrased as an equivalent problem on the complete (original) de Bruijn graph:

Problem 2. Given a de Bruijn graph D_k of order k and an integer L, find a smallest set of vertices U_{kL} such that any path in D_k of length $L-k$ passes through at least one vertex of U_{kL}.

We show that the problem of finding a minimum-size k-mer set that hits every string in a given set of L-long strings is NP-hard, further motivating the need for a universal k-mer set. We provide a heuristic called DOCKS that is based on the combination of three ideas. First, we use a decycling algorithm due to Mykkeltveit to convert a complete de Bruijn graph into a directed acyclic graph (DAG) by removing a minimum number of k-mers. We then supply a novel dynamic program to score remaining k-mers by the number of remaining length-ℓ paths that they hit. Finally, we use that dynamic program in a greedy heuristic to select the additional k-mers and produce a small universal set $\hat{U}_{kL}$, which we show empirically to often be close to the optimal size. Our use of a greedy heuristic is motivated by providing a proof that finding a small ℓ-path cover in a graph G is NP-hard even when G is a DAG. We report on the size of the universal k-mer hitting set produced by DOCKS and demonstrate on two datasets that we can better cover sequences with a smaller set of k-mers than is possible using minimizers. Our results also provide a starting point for additional theoretical investigation of these path coverings of de Bruijn graphs.

2 Results

2.1 DOCKS algorithm

To get the algorithm, we combine the two steps. First, we find a minimum-size decycling set in a complete de Bruijn graph of order k and remove it from the graph, turning it into a DAG. Then, we repeatedly remove a vertex v with the largest hitting number $T_\ell(v)$ (the number of ℓ-long paths the vertex participates in) until there are no ℓ-long paths, where $\ell = L-k$, recomputing $T_\ell(u)$ for all remaining u after each removal. This hitting number can be computed efficiently using dynamic programming. This is summarized below (Algorithm DOCKS).

The running time is polynomial in L and $|\Sigma|^k$. Finding the decycling set takes $O(|\Sigma|^k)$, as the size of the set is $\Theta(|\Sigma|^k/k)$ and the running time for finding each k-mer is $O(k)$. In the second phase, each iteration calculates the hitting number of all vertices using dynamic programming in time $O(|\Sigma|^k L)$. The number of iterations is $1+p$, the number of vertices removed. Thus, the total running time is dominated by steps 4–8 and is $O((1+p)|\Sigma|^k L)$.

Algorithm 1 DOCKS: Find a small k-mer set hitting all L-long sequences

1: Generate a complete de Bruijn graph G of order k, set $\ell = L - k$.
2: Find a decycling vertex set X using Mykkeltveit's algorithm.
3: Remove all vertices in X from graph G, resulting in G'.
4: **while** there are still paths of length ℓ **do**
5: Calculate the number of starting and ending i-long paths at each vertex, for $0 \leq i \leq \ell$.
6: Calculate the hitting number for each vertex.
7: Remove a vertex with maximum hitting number from G', and add it to set X.
8: **end while**
9: Output set X.

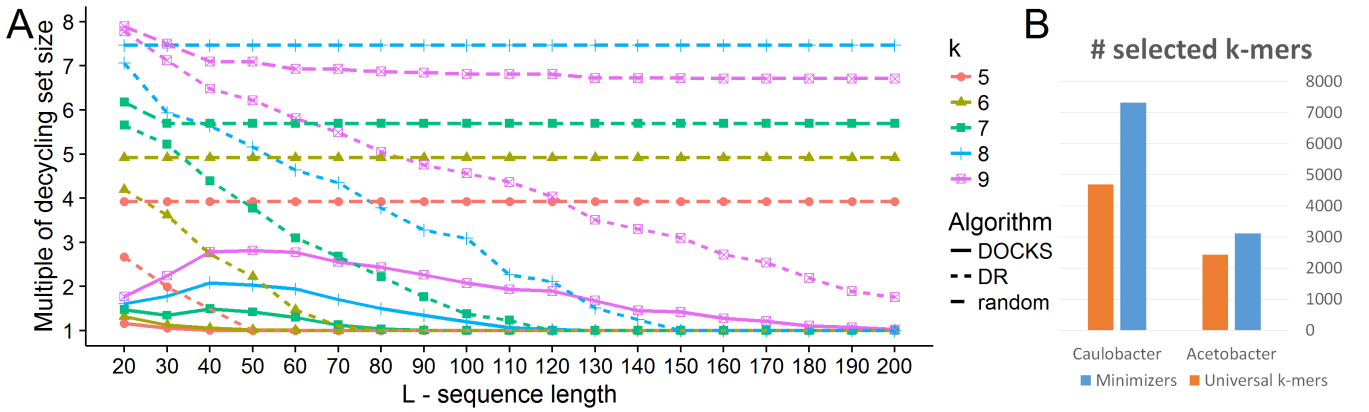


Fig. 1. Performance of DOCKS. A) For different combinations of k and L we ran DOCKS and two random procedures over the DNA alphabet. The results are shown in comparison to the size of the decycling set. When the ratio is 1, all the sequences avoiding the decycling set were of length shorter than L. DR: decycling+random. B) Comparison of the number of selected minimizers and universal k-mers, for $k = 8, L = 100$ in bacterial genomes.

2.2 Computational results

We implemented and ran DOCKS over a range of k and L: $5 \leq k \leq 9$ with $20 \leq L \leq 200$, in increments of 10. These are typical values used for minimizers of longer k-mers and read lengths of short read sequences. We also implemented two random procedures that we compare to as baselines. One, termed "random", removes random vertices until no $\ell = L - k$ paths remains. The second, termed "decycling+random" (DR), first removes a minimum-size decycling set and then randomly removes vertices until no path of length $\ell = L - k$ exists. The results are summarized in Figure 1A. Our method outputs a set of k-mers that is much smaller than both random procedures.

In Figure 1B, we compare the size of the universal hitting k-mers and the minimizers in two bacterial genomes. *Acetobacter tropicalis* (RefSeq NZ_CP011120) has a genome of 2.8 Mbp and a GC content of 47.8%. *Caulobacter vibriodes* (RefSeq NC_002696) is larger at 4.0 Mbp and has a higher GC content of 67.2%. For each genome, we computed the number of minimizers using $k = 8$ and a window length of 100. Also, for each window of 100 bases we found a k-mer from the set $\hat{U_{kL}}$ for $k = 8, L = 100$, computed by DOCKS. Each such window is guaranteed to contain at least one universal k-mer, and usually more than one. In each window, we select only one of the universal k-mers, the smallest one in lexicographic order. Using universal hitting k-mers instead of minimizers gives a smaller set of selected k-mers.

3 Conclusion

In this work, we presented the DOCKS algorithm, which generates a compact set of k-mers that together hit all L-long DNA sequences. DOCKS's good performance can be attributed to its two components. It first optimally removes a minimum-size set that hits all infinite sequences, which takes care of most L-long sequences. It then greedily removes vertices that hit remaining L-long sequences. Its feasibility stems from the first step, which runs in time $O(k)$ times the size of the output, and the second step, which uses dynamic programming to bound the running time to be quadratic in the output size times L.

We demonstrated the ability of DOCKS to generate compact sets of k-mers that hit all L-long sequences. These k-mer sets can be generated once for any desired value of k and L and then used easily for many different purposes. For example, there is a set of only 700 6-mers out of a total of 4096 that hits every sequence longer than 70 bases — a typical read length for many sequencing experiments — enabling efficient binning of reads. These sets of k-mers could improve many of the applications that use minimizers, as we showed that they are both smaller and more evenly distributed across typical sequences.

DOCKS provides the first practical solution to the identification of universal sets of k-mers. The software is freely available on `acgt.cs.tau.ac.il/docks/`, as are universal sets of k-mers over a range of values of L and k.

This work is under review at WABI 2016.

BASIC: BCR assembly from single cells

Stefan Canzar[1,†], Karlynn E. Neu[2,†], Patrick C. Wilson[2], and Aly A. Khan[1,*]

[1] Toyota Technological Institute at Chicago, Chicago IL 60637, USA
[2] Committee on Immunology, The Knapp Center of Lupus and Immunology Research, The University of Chicago, Chicago IL 60637, USA

Background B cells form an important component of the adaptive immune system. They possess the remarkable capacity to recognize antigens through the B-cell receptor (BCR, Figure 1A), which is generated through a series of somatic rearrangements and mutations [1][2]. Recent advances in single cell RNA-sequencing (scRNA-seq) offer a high-throughput means of profiling all transcripts expressed in a single B cell. However, the assembly of full-length BCR sequences from scRNA-seq is a non-trivial problem that neither current reference-based assembly methods nor *de novo* assembly methods address. Thus, the lack of efficient methods for assembling BCR sequences is a major roadblock in studying B-cell biology at a single cell level.

Results Here, we present a novel semi-*de novo* assembly method to determine the full-length sequence of the BCR in single B cells from scRNA-seq data, called BASIC (BCR assembly from single cells). Briefly, BASIC performs semi-*de novo* assembly in two stages (Figure 1B). First, BASIC uses known variable and constant regions in both chains to identify anchor sequences. Second, BASIC performs *de novo* assembly to stitch together the anchor sequences. To demonstrate the utility and accuracy of our method, we subjected single B cells from a human donor to scRNA-seq, assembled the full-length heavy and the light chains, and experimentally confirmed these results by using single cell primer based nested PCRs and Sanger sequencing. Importantly, errors in Sanger sequencing, where specific nucleotides are unresolved and reported as N, were resolved by BASIC and match known germline sequences. Furthermore, we compared our method with a state-of-the-art *de novo* transcript assembly program and report better accuracy for BCR assembly with BASIC (see Figure 1C for an example). In sum, BASIC correctly assembles full-length BCR sequences and demonstrated better performance when compared to a state-of-the-art *de novo* transcript assembly method.

Conclusion BASIC enables investigators to assemble BCR sequences from scRNA-seq data and study B-cell repertoire. We experimentally validated sequences assembled by BASIC, and show it to be robust to potential noise associated with different PCR pre-amplification cycles. The algorithm underlying

[†] Equal contribution
[*] Corresponding author: aakhan@ttic.edu

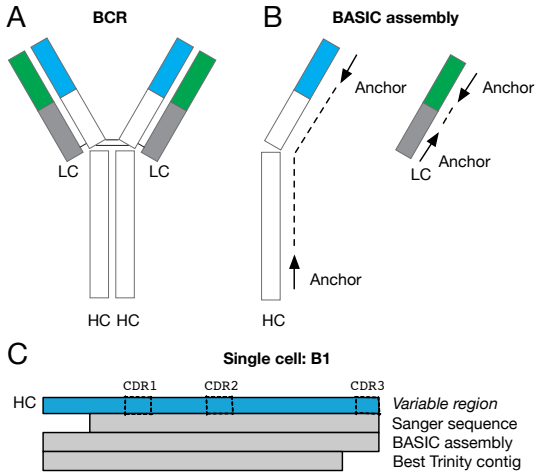


Fig. 1. A) The BCR is a large 'Y' shaped protein complex composed of two identical heavy chains (HC) and two identical light chains (LC). The variable regions are colored blue and green in the paired chains. The complementarity determining regions (CDRs) are those parts of the variable regions that participate in the binding of antigens. B) Anchors are stitched together to assemble the HC and LC. C) Illustration of the HC variable region sequence for single cell B1 along with: the Sanger sequence, the BASIC assembled sequence, and the best contig reported by Trinity. Note the absence of CDR3 from the Trinity contig, and the BASIC assembly extending past the 5' PCR primer site used in the Sanger sequence.

BASIC also serves as a principled approach to assemble other diverse genes associated with immunological repertoire using scRNA-seq, such as HLA and TCR genes. BASIC is available at: http://ttic.uchicago.edu/∼aakhan/BASIC

References

1. Davies, D.R., Padlan, E.A., Segal, D.: Three-dimensional structure of immunoglobulins. Annual review of biochemistry 44(1), 639–667 (1975)
2. Edelman, G.M.: Dissociation of γ-globulin. Journal of the American Chemical Society 81(12), 3155–3156 (1959)

R-loop biology: from a few gene cases to genome scale

Vladimir A. Kuznetsov*, Piroon Jenjaroenpun, Thidathip Wongsurawat

Bioinformatics Institute/A*STAR, Singapore

*Corresponding author

R-loops, which are triple-stranded RNA-DNA hybrid structures, can often occur in the human genome and play crucial roles in many normal biological processes. Such RNA-DNA hybrids could initiate mutations, DNA breaks, genome instability and diseases. However, until 2011 only a few cases of the R-loop formation have been experimentally documented, indicating the roles of R-looping in gene functions. The R-loops, involving in transcription through switch recombination regions at immunoglobulin heavy chain loci in a genome of mammalian B cells, were the well-studied examples. In 2011, we have developed our data-driven quantitative model of RLFS (QmRLFS), which easily demonstrated strong co-localization of predicted RLFS with most genic regions in the human genome and the genome regions associated with open chromatin, promoters and others gene expression control signals, transcript isoforms, splicing, triggering mutation and DNA break loci, fragile and critical disease regions (Wongsurawat et al, 2011). We found that many oncogenes, tumor suppressors and neurodegenerative diseases could be prone to significant R-loop formation. These predictions have been confirmed with several experimental systems and methods including DRIP-qPCR (Yeo et al, 2013, Ginno at al, 2012), DRIP-seq methods (Ginno at al, 2012, Ginno at al, 2013).

The accurate computational prediction (83-92%; Jenjaroenpun et al, 2015) and experimental genome mapping of RLFSs has opened up intriguing possibilities for the studies of RNA-DNA interactome complexity in vivo and R-loop's use targets for diagnostics and treatment of many diseases. Here we review the current knowledge about the mechanisms controlling R-loop formation, methods of experimental R-loop detection, and computational models of R-loop forming sequences at genic and genome-wide scales. Finally, we discuss the observed and putative relationships of R-loops with several basic biological mechanisms, evolution of RLFS motifs and medical conditions including that of cancer, autoimmune and neurodegenerative diseases.

Tracing noisy biological progression and gene network rewiring between cell metastable states in static single-cell transcriptomes

Pablo Cordero and Joshua M. Stuart

UC Santa Cruz Genomics Institute, University of California, Santa Cruz, California, USA

Background

Understanding the dynamics of gene expression and regulatory networks as a cell undergoes biological processes is crucial for dissecting the molecular mechanistic underpinnings of complex biological processes such as differentiation and oncogenesis. Static transcriptome measurements of cell populations at the single-cell level have recently emerged as promising tools to dissect these dynamics by inferring the underlying progression. However, it remains unclear how to adequately elucidate cell types from these noisy data, simultaneously pinpoint their regulatory networks, and how gene expression is rewired during cell state transitions.

Results

To address the above challenges, we propose a strategy, Single Cell Inference of Morphing and Interdependent Trajectories and their Associated Regulatory networks (SCIMITAR), for inferring gene expression network dynamics throughout biological progression from static, single-cell, transcriptomes. SCIMITAR's approach is top-down. First, we focus on detecting recurrent, metastable transcriptional states and any connections between them supported by transitioning cells. Second, we give a detailed, full probabilistic description of each path in the metastable state graph, explicitly accounting for heteroscedastic noise in the data and detecting gene-to-gene expression correlations at each point in the progression. To achieve this, we extend Gaussian mixtures with discrete components to a smooth, continuous mixture of ‘morphing’ distributions. The inferred model allows tracking the rewiring of gene regulatory networks between metastable states and can elicit predictions on data that it wasn’t trained on, such as mapping new samples, including experimental replicates, to the model. Further, the probabilistic nature of SCIMITAR transition models allows for evaluating the shape of the multivariate gene expression distribution as a function of biological progression, which we show can be used to pinpoint stable and transitional cell states.

We tested whether SCIMITAR could yield insights in the developmental trajectory of human fetal neurons by analyzing recent, publicly-available, single-cell transcriptomic measurements, focusing on 578 expressed transcription factors. SCIMITAR pinpointed factors that were expressed in various ways across three metastable states: some went up at the beginning of the transition (in replicating neurons), others were expressed only in the middle of the quiescent state or in the end. A likelihood ratio test designed for the SCIMITAR model revealed 35 genes that significantly varied throughout the progression but that were missed by standard differential expression between cells grouped in

supervised and unsupervised ways. The genes revealed by SCIMITAR involved in the Jak-STAT pathway that presented a coordinated expression pattern in the middle of the developmental trajectory. Further, the SCIMITAR model pinpointed a previously unidentified transitional state between fetal replicating and fetal quiescent neurons. Finally, SCIMITAR also revealed regulatory network rewiring events as gene co-expression degrees changed through the progression, unveiling coordinated regulation of MAP kinases, morphogenesis, and STAT factors throughout the progression as well as potential master regulators.

Conclusions

Static, single-cell transcriptomic measurements hold great promise for revealing the cell state dynamics of a multitude of biological processes. Inferring biological progression from these data requires computational methods that can model the individual cells as an evolving gene expression distribution, a feat that can only be achieved by fully embracing the heteroscedacity of the data. Our proposed method, SCIMITAR, leverages this heteroscedacity to track gene expression rewirings and cell state switches in a continuous model of biological progression. These rich models allowed dissecting the progression dynamics in the transition between human fetal replicating and quiescent neurons, revealing Jak-STAT related genes missed by traditional, population-based differential expression and rise and fall of co-expression networks enriched with diverse kinases and developmental factors. We expect SCIMITAR to be widely used to dissect these gene expression and network progressions from static, single-cell measurements that are now becoming a standard technique to tackle complex cell state processes.

SKE: Ultra fast simultaneous K-mer counting for multiple values of k

Eric Pauley[1], Raunaq Malhotra[1], Guillaume Rizk[3], Paul Medvedev[1], Rayan Chikhi[2], Raj Acharya[1]
[1]Department of Computer Science and Engineering, The Pennsylvania State University, University Park, PA 16802
[2]CNRS Bioinformatics, University of Lille 1, France
[3]IRISA, Rennes, France

K-mer counting is an essential pre-processing step in a number of bioinformatics applications, including de novo assembly using De Bruijn graphs, repeats detection, and multiple sequence alignment. For a given collection of strings or reads, k-mer counting involves computing the number of times every unique substring of length k occurs in all the strings. Additionally, as DNA is double stranded, the counts of a k-mer and its reverse complement are combined together and reported as counts for the canonical k-mer (lexicographically smaller of the two k-mers). However, when working with RNA-seq data, non-canonical counts can be valuable too. It is now possible to obtain a large number of reads (millions to billions) in a single next-generation sequencing run which necessitates the availability of fast and efficient k-mer counting tools [1,2,3,4,5]. Most applications perform k-mer counting for a number of distinct values of k. For example, De Bruijn graph assemblers try a number of k-values for de novo assembly. Currently, a k-mer counting tool such as KMC2[2] or DSK[1] is run multiple times for computing k-mer counts for multiple values of k, which requires accessing the large dataset of reads every time. However, given counts of a k-mer, it is possible to derive the counts of n-mers, where n<k, from the counts of k-mers with minimal disk-IO overhead.

We propose Suffix K-mer Extrapolation, SKE, an algorithm which takes existing non-canonical k-mer counts of size k and computes counts of any n-mers where n<k. Non-canonical counts for length k are obtained using an existing tool such as DSK. The non-canonical counts for an n-mer can be computed from k-mer counts by truncating all k-mers to size n, and combining the counts of same n-mers. The only additional counts that are missed come from suffixes of sizes less than k from each read. We compute the counts for suffixes of lengths less than k separately and combine them with the non-canonical k-mer counts.

The suffix of length (k-1) base pairs (referred as (k-1)-read suffix) is extracted from each read and stored in partitions on disk. Each partition is iterated and n-mers are extracted starting at each index of the (k-1)-read suffix and ending at the final base pair. These n-mers are sorted, combined, and written to disk. A second step of algorithm takes existing non-canonical k-mer counts and the counts of n-mers sorted in the partitions, merges the sorted partitions from both files into a continuous stream of k-mers, then truncates every k-mer down to each requested size n-mer, combining duplicate counts before saving. By counting n-mers from (k-1)-read suffixes in addition to k-mer counts, we obtain correct counts for any length. Whereas existing solutions such as DSK [1] or KMC2 [2] have computational complexity $O(x*bp)$ where x is the number of n-mer sizes to compute and bp is the number of base pairs in the input reads, we achieve complexity $O(bp + x*d)$ where d is the number of distinct k-mers in the largest size counted. Because a significant portion of the time is spent merging (k-1)-read suffix counts with k-mer counts, which takes time linearly proportional to the number of sequences, SKE becomes more efficient as read length increases and error rate goes down.

We have benchmarked our algorithm on reads of both E.coli DNA (read length 151, 1.2GB FASTA, Ecoli_DH10B_110721) and human DNA (read length 250, 250GB FASTA, HG002_NA24385). SKE performed counting faster than KMC2 on human and E.coli reads. For E.coli, all tests were performed on a quad-core server node with 2GB RAM and disk capable of 102MB/s sequential write. Because SKE uses partitioning similar to DSK it benefits greatly from faster disk IO. Counts were performed for k-mer lengths 8-31 using DSK, KMC2, and SKE. SKE took 322s to perform counts for all k-values, including

the initial 31-length counting time using DSK. KMC2 took 657s, and DSK took 1139s. The growth rate for all algorithms is linear with number of k-values being counted, though SKE has a higher constant time requirement. Human counting was performed on an equivalent machine with memory limit changed to 8GB. For DSK and KMC2 counts were performed for lengths 8,15,20,25,31, and were used to interpolate the times required for lengths 8-31. These times were then added to get the total estimated time. SKE took 3217 min including 31-length DSK counting to perform all counts from 8-31, KMC2 took 3481 min, and DSK took 21880 min. SKE shows substantial improvement in counting speed over existing solutions and demonstrates the utility of algorithms targeted at multi k-value counting.

The source for SKE can be found at https://github.com/ericpauley/ske

The current work was supported by NSF grants: 1421908,1533797,1356529,1439057, and 1453527.

References

[1]Rizk G, Lavenier D, Chikhi R: DSK: k-mer counting with very low memory usage. Bioinformatics. 2013, 29 (5): 652-653. 10.1093/bioinformatics/btt020.

[2]Deorowicz, Sebastian, et al. "KMC 2: Fast and resource-frugal k-mer counting." *Bioinformatics* 31.10 (2015): 1569-1576.

[3]Marçais, Guillaume, and Carl Kingsford. "A fast, lock-free approach for efficient parallel counting of occurrences of k-mers." *Bioinformatics* 27.6 (2011): 764-770.

[4]Melsted, Pall, and Jonathan K. Pritchard. "Efficient counting of k-mers in DNA sequences using a bloom filter." *BMC bioinformatics* 12.1 (2011): 1.

[5]Deorowicz, Sebastian, Agnieszka Debudaj-Grabysz, and Szymon Grabowski. "Disk-based k-mer counting on a PC." *BMC bioinformatics* 14.1 (2013): 1.

Abstract for HiTSeq 2016

Title: Individual genome interpretation in newborns with rare disorders

Aashish N. Adhikari[1], Jay P. Patel[2], Alice Y. Chan[3], Divya Punwani[3], Haopeng Wang[3], Antonia Kwan[3], Theresa A. Kadlecek[3], Morton J.Cowan[3], Marianne Mollenauer[3], John Kuriyan[1], Shu Man Fu[4], Yangyun Zou[1], Yaqiong Wang[1], Uma Sunderam[5], Prisni Rath[5], Sadhna Rana[5], Ajithavalli Chellappan[5], Kunal Kundu[5], Dae Lee[6], Flavia Chen[3], Brad Dispensa[3], Mark Kvale[3], Richard Lao[3], Dedeepya Vaka[3], Brandon Zerbe[3], Arend Mulder[7], Frans H J Claas[7], Joseph A Church[8], Arthur Weiss[3], Richard A. Gatti[9], Robert Nussbaum[3], Robert Currier[10], Joseph Sheih[3], Renata Gallagher[3], Sean Mooney[6], Neil Risch[3], Barbara Koenig[3], Pui Kwok[3], Jennifer M. Puck[3], Rajgopal Srinivasan[5] , Steven E. Brenner[1*]

[1]University of California, Berkeley, CA, USA; [2]Children's Hospital of Los Angeles, Los Angeles, CA, USA; [3]University of California, San Francisco, CA, USA; [4]University of Virginia School of Medicine, Charlottesville, VA, USA; [5]Innovation Labs, Tata Consultancy Services Hyderabad, AP, India; [6]University of Washington, Washington, WA; [7]Leiden University Medical Centre, Leiden, The Netherlands; [8]University of Southern California, Los Angeles, CA, USA; [9]University of California Los Angeles, CA, USA; [10]California Department of Public Health, CA, USA;
*Corresponding author email address: brenner@compbio.berkeley.edu

Background

High-throughput sequencing technologies are being increasingly integrated into clinical settings, aiding the detection and diagnosis of disease. However, our ability to reliably interpret genomic data lags behind the ability of the sequencing technologies to generate them. Here, we present an analysis protocol we developed for individual genome interpretation which we applied to exomes from newborns with undiagnosed primary immune disorders. Using multiple callers with multisample calling and an integrated variant annotation, variant filtering, and gene prioritization pipeline, we were able to diagnose several cases of elusive immunodeficiencies.

Results

In two unrelated infant immunodeficient girls with no diagnoses, we discovered compound heterozygous variants in the *ATM* gene for both the infants offering a very early diagnosis of Ataxia Telangiectasia (AT). In addition to avoiding diagnostic odyssey, this allowed for avoidance of undue irradiation and live vaccinations, and for appropriate counseling of the parents regarding their carrier status. In another case, the affected siblings had early onset bullous phempigoid, a chronic autoimmune disorder. Our analysis revealed compound heterozygous mutations in ZAP70, a gene associated with profound primary immunodeficiency, the opposite phenotype. Cellular immunological studies indicated that one variant was hypomorphic and the other was hyperactive. These combined to yield a novel presentation, adding to the existing phenotype repertoire of ZAP70 in humans. We also discovered pathogenic variants in PRKDC occurring after the stop codon encoded in the reference genome; we correctly

identified that the reference genome had a rare pathogenic variant with frameshift leading to a premature stop codon. In a normal reference, the mutations observed in this case led to nonsynonymous changes. Our protocol has been similarly revealing in other SCID and CID cases including Nijmegen Breakage Syndrome, which highlight unique features of the analysis framework that facilitate genetic discovery.

Conclusions

With a diagnostic rate of ~50% in cases involving family trios, these early diagnosis using exome sequencing help provide crucial information to offer prompt appropriate treatment, family genetic counseling, and avoidance of diagnostic odyssey. We have also begun exploring how exome sequencing could potentially augment public health newborn screening of a large number of rare disorders in newborns, currently performed using tandem mass spectrometry (MS/MS) technologies. In collaboration with the California Department of Public Health under an IRB-approved protocol, we aim to evaluate the current ability to predict disease status from exome sequences using de-identified archived dry blood spot samples of all California newborns confirmed to have metabolic disorders for a period of 8.5 years since the introduction of MS/MS screening, as well as samples that were false positives on the MS/MS screening.

CLIA-certified cancer gene panel-based machine learning method to predict sensitivity of anticancer drugs for precision oncology

CLIA certified molecular/genetic panel testing of formalin-fixed, paraffin embedded (FFPE) material including studies of small biopsies offers the potential to identify individualized treatments that target specific genetic alterations such as EGFR mutation. However, molecularly-guided therapy is only available for the minority of lung cancer patients carrying such alterations for targeted drugs (e.g., ~15% of lung adenocarcinoma); thus the selection of chemotherapy or other treatment for the majority of non-small-cell lung cancer (NSCLC) patients without such alterations is still limited. In addition, despite the early success of targeted therapy in NSCLC patient care, patients treated with targeted drugs often developed resistance to these treatments. Thus, it is critical to build a predictive model based on information of genetic panel testing to predict sensitivity/resistance of drug for individualized treatment stratification. To tackle this challenge, we develop a novel machine learning approach called Robust Bayesian Matrix Factorization (RBMF) to integrate genetic information on a large panel of non-small cell lung cancer (NSCLC) lines (e.g., Single Nucleotide Variants (SNVs) found on targeted gene panel or whole exome sequences) with large-scale drug/chemical compound screening profiles on these same NSCLC lines to (a) discover a genetic variation-based predictive biomarker(s) and (b) use this to predict response of drugs in other NSCLC lines (and ultimately in patients). The RBMF method leverages information across multiple related drug/chemical compound screening profiles that have similar mechanisms of actions/targets as well as samples (e.g., NSCLC lines) with similar genetic variant profiles (i.e., exploring clusters of drugs and samples), thus can be robust against noise from each data/drug screening experiment and more accurate to predict sensitivity of drugs.

In experiments with our institutional drug/chemical screening profiles and SNVs present in known cancer-related genes and/or a commercially available genetic panel such as *FoundationOne* in NSCLC cell lines, the RBMF method showed better prediction performance compared to the state-of-the-art methods. Moreover, the RBMF method identified novel mutation-drug sensitive/resistant associations that can serve as a predictive biomarker to stratify patients. Independent validations with Genomics of Drug Sensitivity in Cancer and Cancer Cell Line Encyclopedia datasets demonstrated that the RBMF consistently outperformed current state-of-the-art methods for sensitivity prediction for well-known cancer drugs.

Taken together, our proposed method demonstrated the clinical utility of the use of genetic panels to predict drug response in NSCLC lines. Furthermore, the novel mutation-drug sensitive/resistant association discovered by the RBMF method could provide unprecedented opportunities to develop a clinical assay as a predictive biomarker, which could individualize treatments based on the genetic information of cancer patients.

HiTSeq 2016 Oral Presentations

TwoPaCo: An efficient algorithm to build the compacted de Bruijn graph from many complete genomes

Keywords: Algorithms, graph theory, comparative genomics, parallel computing, de Bruijn graph

Abstract: De Bruijn graphs have been proposed as a data structure to facilitate the analysis of related whole genome sequences, in both a population and comparative genomic settings. However, current approaches do not scale well to many genomes of large size (such as mammalian genomes).

In this paper, we present TwoPaCo, a simple and scalable low memory algorithm for the direct construction of the compacted de Bruijn graph from a set of complete genomes. We demonstrate that it can construct the graph for 100 simulated human genomes in less then a day and eight real primates in less than two hours, on a typical shared-memory machine. We believe that this progress will enable novel biological analyses of hundreds of mammalian-sized genomes.

Availability: Our code and data is available for download from github.com/medvedevgroup/TwoPaCo

Authors:

first name	last name	email	country	organization	corresponding?
Ilia	Minkin	ium125@psu.edu	United States	Pennsylvania State University	
Son	Pham		United States	Salk Institute for Biological Studies	
Paul	Medvedev	pashadag@cse.psu.edu	United States	Pennsylvania State University	✓

Rail-RNA: Scalable analysis of RNA-seq splicing and coverage

Keywords: Sequence alignment, RNA

Abstract: RNA sequencing (RNA-seq) experiments now span hundreds to thousands of samples. Current spliced alignment software is designed to analyze each sample separately. Consequently, no information is gained from analyzing multiple samples together, and it requires extra work to obtain analysis products that incorporate data from across samples.

We describe Rail-RNA, a cloud-enabled spliced aligner that analyzes many samples at once. Rail-RNA eliminates redundant work across samples, making it more efficient as samples are added. For many samples, Rail-RNA is more accurate than annotation-assisted aligners. We use Rail-RNA to align 667 RNA-seq samples from the GEUVADIS project on Amazon Web Services in under 16 hours for US$0.91 per sample. Rail-RNA outputs alignments in SAM/BAM format; but it also outputs (1) base-level coverage bigWigs for each sample; (2) coverage bigWigs encoding normalized mean and median coverages at each base across samples analyzed; and (3) exon-exon splice junctions and indels (features) in columnar formats that juxtapose coverages in samples in which a given feature is found. Supplementary outputs are ready for use with downstream packages for reproducible statistical analysis. We use Rail-RNA to identify expressed regions in the GEUVADIS samples and show that both annotated and unannotated (novel) expressed regions exhibit consistent patterns of variation across populations and with respect to known confounders.

Availability: Rail-RNA is open-source software available at http://rail.bio.

Authors:

first name	last name	email	country	organization	corresponding?
Abhinav	Nellore	anellore@gmail.com	United States	Johns Hopkins University	✓
Leonardo	Collado-Torres		United States	Johns Hopkins University	
Andrew E.	Jaffe		United States	Johns Hopkins University	
José	Alquicira-Hernández		United States & Mexico	Johns Hopkins University & National Autonomous University of Mexico	
Christopher	Wilks		United States	Johns Hopkins University	
Jacob	Pritt		United States	Johns Hopkins University	
James	Morton		United States	University of California San Diego	
Jeffrey T.	Leek		United States	Johns Hopkins University	
Ben	Langmead	langmea@cs.jhu.edu	United States	Johns Hopkins University	✓

popSTR: population-scale detection of STR variants

Keywords: Bioinformatics, Mathematical modeling, Motif finding, Sequence analysis, Statistics, Algorithms

Abstract: Microsatellites, also known as short tandem repeats (STRs), are tracts of repetitive DNA sequences containing motifs ranging from 2-6 bases. The human reference genome contains approximately 1 million microsatellites, covering almost 1% of the genome (Gymrek et al., 2016). Microsatellite analysis has a wide range of applications, including medical genetics, forensics and construction of genetic genealogy. However, microsatellite variations are rarely considered in whole-genome sequencing studies, in large due to a lack of tools capable of analyzing them (Duitama et al., 2014).

Here we present a microsatellite genotyper which is both faster and more accurate than other methods previously presented. There are two main ingredients to our improvements. First we reduce the amount of sequencing data necessary for creating microsatellite profiles by using previously aligned sequencing data. Second, we use population information to train microsatellite and individual specific error profiles. By comparing our genotyping results to genotypes generated by capillary electrophoresis we show that our error rates are 50% lower than those of lobSTR, another program specifically developed to determine microsatellite genotypes.

Availability: Source code is available on Github: https://github.com/snaedis88/popSTR.git

Authors:

first name	last name	email	country	organization	corresponding?
Snædís	Kristmundsdóttir	snaedis.kristmundsdottir@decode.is	Iceland	deCODE Genetics / Amgen	✓
Brynja D.	Sigurpálsdóttir		Iceland	Rejkyavik University	
Birte	Kehr		Iceland	deCODE Genetics / Amgen	
Bjarni V.	Halldórsson	bjarni.halldorsson@decode.is	Iceland	deCODE Genetics / Amgen and Rejkyavik University	✓

Genotyping of Inversions and Tandem Duplications

Keywords: Next-generation sequencing, Genotyping, Structural variant, Duplication, Inversion

Abstract: Next Generation Sequencing (NGS) has enabled studying structural genomic variants (SVs) such as duplications and inversions in large cohorts. SVs have been shown to play important roles in multiple diseases, including cancer. As costs for NGS continue to decline and variant databases become ever more complete, the relevance of genotyping also SVs from NGS data increases steadily, which is in stark contrast to the lack of tools to do so.

We introduce a novel statistical approach, called DIGTYPER (Duplication and Inversion GenoTYPER), which computes genotype likelihoods for a given inversion or duplication and reports the maximum likelihood genotype. In contrast to purely coverage-based approaches, DIGTYPER uses breakpoint-spanning read pairs as well as split alignments for genotyping, enabling typing also of small events. We tested our approach on simulated and on real data and compared the genotype predictions to those made by DELLY, which discovers SVs and computes genotypes. DIGTYPER compares favorable especially for duplications (of all lengths) and for shorter inversions (up to 300 bp). In contrast to DELLY, our approach can genotype SVs from data bases without having to rediscover them.

Availability: https://bitbucket.org/jana_ebler/digtyper.git

Authors:

first name	last name	email	country	organization	corresponding?
Jana	Ebler		Germany	Saarland University	
Alexander	Schönhuth		The Netherlands	Centrum Wiskunde & Informatica	
Tobias	Marschall	t.marschall@mpi-inf.mpg.de	Germany	Saarland University and Max Planck Institute for Informatics	✓

2016 HitSeq Abstract, Orellana. *et al.*

EDGAR: Full-length RNA transcript identification by hybrid sequencing and best edit-distance graph alignment of a single molecule read

Christian F. Orellana, Jacob E. Bogerd, Nathaniel Moorman, Paul Armistead, Corbin D. Jones, Jan F. Prins – UNC Chapel Hill

Background. Ideally, we would characterize the RNA transcriptome by sequencing full length RNA molecules harvested from a cell. Single molecule sequencing could identify novel transcripts produced in virus-infected cells, show novel splicing as a result of disease, and identify linked SNPs in transcripts isoforms. However, current single molecule sequencing technologies from Pacific Biosciences (PacBio) and Oxford Nanopore Technologies operate at the limits of detection and are fundamentally susceptible to noise, resulting in missed or repeated nucleotides as well as errors in nucleotide identity, reaching a 10% error rate or more [1]. By circularizing the cDNA copied from the RNA, a PacBio sequencer has the opportunity to read a single molecule multiple times (subreads) and correct the separate observations with each other to create a circular consensus sequence (CCS). However this comes at the cost of limited length RNA transcripts as the overall number of nucleotides that can be sequenced before the observation fails is in the 1-10 kb regime. On the other hand, short read bulk sequencing is highly accurate but cannot reliably determine full length transcripts because we end up sequencing fragments of many different molecules and attempt to piece them together to infer the identity of the original full length transcripts, and this problem is fundamentally underdetermined [2].

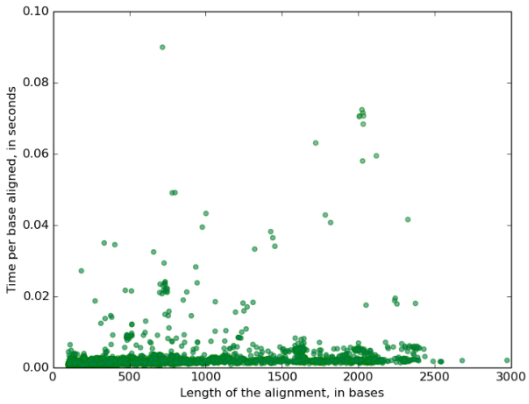
Hybrid long and short read sequencing of two aliquots of an RNA sample has been proposed as a way to combine the accuracy of short read sequencing with the full-length transcript identity of single molecule sequencing. A pioneering method introduced by Au et al. [3] corrects noisy long reads by replacing the local sequence by short reads matching closely to a given interval. In both CCS and hybrid sequencing methods, one shortcoming is correction without a broader context such as the reference genome. In the hybrid sequencing approach, a diploid cell with heterozygous SNPs may be difficult to "correct" because short reads may observe both SNPs. We propose an alternative approach to hybrid sequencing that addresses both these issues by finding the smallest edit-distance alignment of a noisy long read to a path in a directed graph constructed from short reads, either by assembly, or by (spliced) alignment to a reference genome, to form an accurate account of possible transcripts as paths in the graph. Only some of the paths will correspond to actual transcripts. We will describe our approach using the latter form of the graph, termed a splice graph.

Method. A splice graph G is a weighted, directed, multigraph in which nodes represent genomic coordinates in a reference genome and edges represent possible connections between those coordinates: exonic edges represent transcribed sequences, and splice edges join disparate exons. Additional edges are included to represent observed insertions, deletions, and SNPs. Given a single molecule observed as a noisy sequence S, identification of the transcript corresponding to S is reduced to finding that path P in G with smallest edit distance to S. The problem appears to have high complexity, since the total number of paths through the splice graph can be exponential in the number of edges in the graph, and infinite in the presence of cycles. In addition, the traditional minimum edit distance algorithm has cost proportional to the product of the lengths of the sequences being compared, compounding the cost. We developed the EDGAR algorithm to solve this problem with expected cost *linear* in the length of S, using three main strategies. (1) We perform a rapid search of approximate matches between S and the exonic edges of the graph. This search yields "seeds" that serve as starting points for alignments. The algorithm explores paths through the graph in both directions starting from a seed. (2) We define a local bound (r, n) on the number of errors permitted – in any window of length n, at most r edits are permitted. This captures our view that on a sufficiently large scale (ten to hundreds of nts) the errors are distributed uniformly. Thus, if the wrong path is being explored, the local error bound will be exceeded in distance n with high probability, and the path will be discarded, limiting the exponential growth of paths explored. The local error bound also enables a linear time alignment algorithm since the number of cells within a fixed distance r in a traditional dynamic programming tableau is linear in the length of S. (3) When multiple paths starting from one seed reach a given node in G, having aligned an identical subsequence of S, then all paths other than the minimum cost among this set can be deleted, thereby choosing early among paths that differ in a small feature like a SNP.

Results. We tested our method on a hESC hybrid dataset of Illumina and PacBio reads [4] using synthetic and experimental long read data. We generated splice graph G_1 from the short read alignments in chr1 without calling SNPs (thus exonic edges reflect the reference sequence) and generated 1000 random paths through the graph, adding errors to the sequences associated with each path with 9% probability at each nt. The error could be a replacement, insertion, or a deletion of a random nucleotide in the ratio learned from alignments of PacBio reads. This resulted in a synthetic dataset in which each read was at least 1000 nt long with an average length of 1486, for which we know the original paths sampled. Using $(r, n) = (20,100)$, EDGAR aligned 955 of these reads fully to the right path, 43 aligned partially (i.e. exceeded the error threshold at some point), and 2 aligned with one wrong exon at the end. Using $(r, n) = (15,100)$ EDGAR aligned 723 reads fully, 271 reads partially, and 6 reads with at least one incorrect exon at

the end of the path, or in one case skipping a 3 nt early 5'-end splice site of an exon. This establishes the accuracy of the method. Next we used 994 long read CCS from [4] that were corrected using the method of Au, *et al*. [3] which had full length alignment to chr1 on graph G_1' (which includes edges for observed SNPs) using $(r, n) = (20,100)$ and consider these alignments as ground truth (we note these alignments deviated from the exact path by an average of 0.8% suggesting there is value in using the graph for context). We aligned uncorrected CCS and individual subreads to G_1' and report above the number of alignments that match the path of the corrected CCS reads, the average read length, and the number of SNPs included in the alignment and called the same way as in the corrected CCS. The rest of the subreads align to a path that is different from the corrected CCS in at least one exon. We see that SNP calls are reasonably accurate even in subreads. They can be further improved by considering multiple subreads from the same molecule, or by using additional information from the PacBio quality scores.

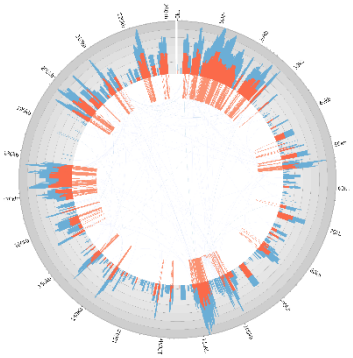
	corrected CCS	uncorrected CCS	subreads
#reads/full/part align	994/994/0 (*)	994/926/52	909/225/559
avg read length	2078	2155	1920
snps incl/matched	16689/16689 = 100%	15837/16665 = 95%	7929/8909 = 89%

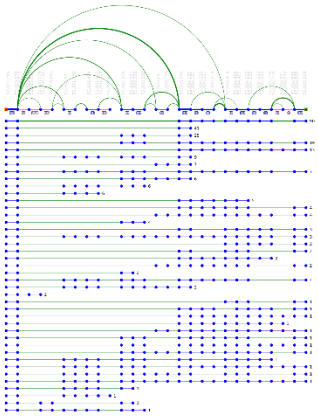


We measured the run time for aligning the long circular consensus sequences provided by Au, *et al*. The figure to the left shows the time in seconds per base aligned, which remains fairly constant for a wide range of read lengths, and varies by less than a factor of 10 over the experiments. The expected time is near the bottom of the range shown.

Applications. We tested our method on two datasets generated at UNC:

(1) Viral transcriptome. The short read alignments to the Human Herpes Virus (HHV5) genome during human cell infection present a high number of cryptic splices (likely due to repeated regions). However, when we aligned the long reads to the splice graph, many of the splices presented by the short reads were not used by any full-length transcript. The figure to the right shows the viral genome in a circular plot. The bars around the circle compare short read coverage (in blue) with long read coverage (in red). The lines in the middle of the circle represent splices confirmed by PacBio long reads (in red) and the ones present only in the short read alignments (in blue). We used 5' and 3' linkers in the protocol to identify full length transcripts which were detected as part of the alignment process.





(2) Novel transcripts in a human cancer cell line. We used our method to analyze the transcriptomes of 10 genes of interest in a human cell line, in order to find novel transcripts. The results of this investigation are under review for publication. The figure to the left shows the full length transcripts found in one of the genes of interest.

Conclusions. The fundamental difference of EDGAR compared with long read correction is the use of context of the short reads represented in the underlying directed graph, identifying a limited set of splices or variants available to the transcript being aligned. Additionally, when used with a splice graph generated from short read alignments to the genome, our method yields an alignment of the long read to the genome, as opposed to just a correction. In this case, long read correction and transcript identification are both achieved simultaneously.

[1] Chaisson, Mark J., and Glenn Tesler. "Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory." BMC bioinformatics 13.1 (2012): 238.

[2] V. Lacroix, M. Sammeth, R. Guigo, A. Bergeron, "Exact transcriptome reconstruction from short sequence reads", WABI 2008 LNCS 5251:50-63, 2008.

[3] Au KF, Underwood JG, Lee L, Wong WH (2012) Improving PacBio Long Read Accuracy by Short Read Alignment. PLoS ONE 7(10): e46679. doi:10.1371/journal.pone.0046679

[4] Au KF, Sebastiano V, Afshar PT, et al. Characterization of the human ESC transcriptome by hybrid sequencing. Proceedings of the National Academy of Sciences of the United States of America. 2013;110(50):E4821-E4830. doi:10.1073/pnas.1320101110.

scphaser: haplotype inference using single-cell RNA-seq data

Daniel Edsgärd[1], Björn Reinius[1] and Rickard Sandberg[1,2]

[1]Department of Cell and Molecular Biology, Karolinska Institutet, 171 77 Stockholm, Sweden and [2]Ludwig Institute for Cancer Research, Box 240, 171 77 Stockholm, Sweden

Abstract

Determination of haplotypes is important for correctly modelling the phenotypic consequences of genetic variation in diploid organisms, including *cis*-regulatory control and compound heterozygosity. Single cell RNA-seq (scRNA-seq) data is exceptionally suited for phasing genetic variants, since both transcriptional bursts and technical bottlenecks cause pronounced allelic fluctuations in individual single cells. Here we present scphaser, an R package that phases alleles at heterozygous variants to reconstruct haplotypes within transcribed regions of the genome using scRNA-seq data. The devised method efficiently and accurately reconstructed the known haplotype for ≥93% of phasable genes in both human and mouse. It also enables phasing of rare and *de novo* variants and variants far apart within genes, which is hard to attain with population-based computational inference. scphaser is implemented as an R package. Tutorial and code are available at https://github.com/edsgard/scphaser* (*Private repository, access available upon request)

Background

The haplotype phase, the sequence of alleles present on the same nucleic acid molecule, such as the maternal or paternal copy of a chromosome, is of importance to elucidate relationships between DNA sequence and phenotype. Major efforts have been made using expression-quantitative-trait-loci studies to identify cis-regulatory variants that affect gene expression. Making use of allele-specific expression (ASE) increases the power of such studies; however, state-of-the-art ASE-based methods to identify cis-regulatory variants require or depend on phased alleles within genes to reach their full potential (Kumasaka et al., 2016; van de Geijn et al., 2015). Phase information is also important for associating clinical outcomes to genetic variation, e.g. to identify cases of compound heterozygosity where risk alleles at different loci do not co-occur on the same DNA molecule but affect both homologous copies of a gene. Such analysis may be especially important to elucidate the impact of mutations in cancer, Mendelian disease and personalized medicine.

Several approaches exist to determine the haplotypes, including direct experimental phasing of a single individual, such as physical separation of the chromosomes, dilution to single-haplotype concentration equivalents, barcoding schemes and long-read sequencing, as well as computational approaches including population phasing using genome reference panels, transmission between related individuals, or utilizing the presence of multiple variants in overlapping reads (S. R. Browning and B. L. Browning, 2011). However, the direct experimental phasing techniques are relatively laborious and the computational methods depend on either DNA data or sequencing read length.

RNA-sequencing allows quantification of the number of transcribed copies from each of the two alleles of a diploid genome; however, short read lengths preclude direct observation of the haplotype sequence. Studies to date evaluated ASE in tissues or cell populations, where the ASE from individual cells is averaged out and it is difficult to obtain gene-based estimates from data at independent heterozygous loci. Instead, scRNA-seq has several unexplored advantages, such as frequent monoallelic or skewed allelic expression (Figure 1A), due to stochastic bursting of gene expression and technical losses of RNA and cDNA molecules (Reinius and Sandberg, 2015). Here, we leverage the pronounced allelic fluctuations in scRNA-seq data to infer the haplotypes of the transcribed parts of a genome (Figure 1B).

Results

We assessed the performance of scphaser on two datasets where the phase was known. This included full-length single-cell RNA-seq data of 336 fibroblast cells from a mouse F1 cross of two inbred strains for which the genomes are known (CAST/EiJ × C57BL/6J, reciprocal cross) and 28 single cells from the human individual NA12878 where phase was inferred via transmission between the sequenced genomes of the family-trio (Marinov, et al., 2014). Using default settings of scphaser 95.1% and 97.5% of variants were correctly phased in the mouse and human dataset, respectively (Figure 1C). At a gene-level 93.6% and 94.9% of genes had all variants correctly phased. Originally, there were 12,247 and 6,065 RefSeq genes with at least two exonic heterozygous variants in the mouse and human dataset, respectively, and 11,512 and 534 genes were phasable (336 vs 28 sequenced cells). In a human dataset with

163 single cells sequenced from an individual (Borel, et al., 2015) we found that 3,155 RefSeq genes were phasable (Figure 1D).

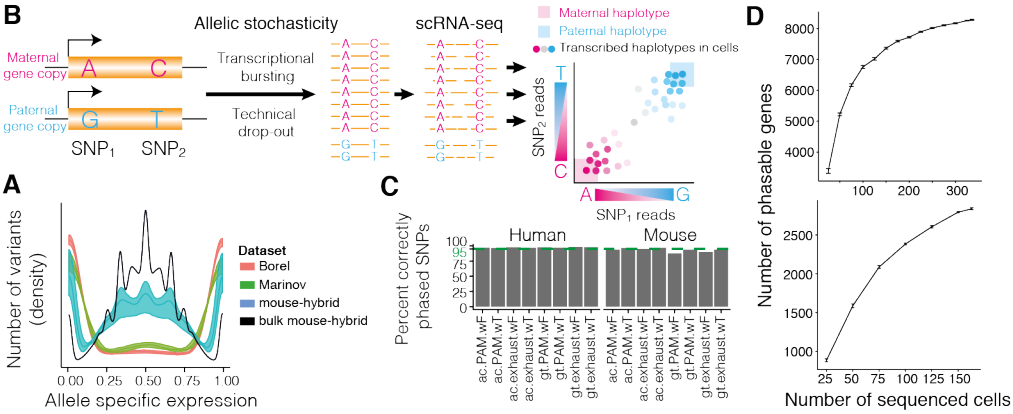


Figure 1. Concept and performance of scphaser. (A) Number of genes against ASE in two human and a mouse scRNA-seq dataset. Line indicates mean and band the inter-quartile range across cells. (B) Stochastic allelic expression bursting and technical drop-out events often cause monoallelic or allele-biased expression in scRNA-seq data. ASE observations in several individual cells can reveal phase of a transcribed sequence, since alleles originating from the same parental copy are co-expressed. (C) Fraction correctly phased SNPs for eight implemented phasing approaches with respect to a human and mouse dataset. X-axis labels denote the input, method and weighing settings for the phasing (Methods). (D) Number of phasable RefSeq genes against number of sequenced cells in a mouse-hybrid (upper) and human (Borel et al.) dataset (lower).

Conclusions

We conclude that phasing by leveraging the imbalanced ASE frequently observed in single-cell RNA-seq data is both accurate and fast. Using RNA instead of DNA enables phasing of variants that are far apart from each other within a gene due to introns. As data from only a single individual is needed we can also phase rare and *de novo* variants. Phasing capacity is facilitated by data from full-length scRNA-seq methods. The more cells that are sequenced the likelihood increase that there are a number of cells where an imbalance is present in the ASE for a particular gene in that individual. The retrieved gene phase information has important applications in functional and clinical genomics, such as empowering cis-regulatory variation studies and in elucidating the impact of haplotype structures on phenotypic outcome and response.

Methods

scphaser assumes a diploid genome, for which there are two possible states of the DNA haplotype sequence. If a gene is mono-allelically expressed the genotype vector of such a cell is identical to the haplotype sequence. Cells in the variant-space, where each variant is a variable with the ASE as domain, with an imbalance in its allelic expression will be closer to the haplotype prototype vector towards which it is imbalanced. Determining which of the two underlying states a cell is closest to can then be viewed as a two-class clustering problem.

To solve this, we implemented an exhaustive search where every possible combination of the two possible states for each variant in a gene is evaluated, where the combination is chosen that minimize the variation of the resulting cell distribution. We also include PAM-clustering as an alternative option (R-package "cluster"). We also include an option to minimize the variation using discrete transcribed genotypes, instead of the continuous ASE, and a simple transcribed-genotype caller if allele read counts are input. The package also includes a weighing option, based on the read counts, as to account for sampling error. Thus, scphaser provides eight ways to conduct phasing as there are three binary options: clustering: {exhaustive, PAM}, input: {genotype, read allele counts} and weigh: {true, false}. Usage instructions are detailed in the vignette, as part of the R package.

References

Borel, C., *et al.* Biased allelic expression in human primary fibroblast single cells. *Am. J. Hum. Genet.* 2015;96(1):70-80.

Browning, S.R. and Browning, B.L. Haplotype phasing: existing methods and new developments. *Nat. Rev. Genet.* 2011;12(10):703-714.

Kumasaka, N., Knights, A.J. and Gaffney, D.J. Fine-mapping cellular QTLs with RASQUAL and ATAC-seq. *Nat. Genet.* 2016;48(2):206-213.

Marinov, G.K., *et al.* From single-cell to cell-pool transcriptomes: stochasticity in gene expression and RNA splicing. *Genome Res.* 2014;24(3):496-510.

Reinius, B. and Sandberg, R. Random monoallelic expression of autosomal genes: stochastic transcription and allele-level regulation. *Nat. Rev. Genet.* 2015;16(11):653-664.

van de Geijn, B., *et al.* WASP: allele-specific software for robust molecular quantitative trait locus discovery. *Nat. Methods* 2015;12(11):1061-1063.

Metagenomic proxy assemblies of single cell genomes

Andreas Bremges[1,2,*], Jessica Jarett[2], Tanja Woyke[2], Alexander Sczyrba[1,2]

[1] Center for Biotechnology and Faculty of Technology, Bielefeld University, 33615 Bielefeld, Germany
[2] U.S. Department of Energy Joint Genome Institute, Walnut Creek, CA 94598, USA

Background

Over 99% of the microbial species observed in nature cannot be grown in pure culture, making it impossible to study them using classical genomic methods. Metagenomics and single cell genomics are two complimentary approaches to study the microbial dark matter.

Metagenomics can obtain genome sequences from uncultivated microbes through direct sequencing of environmental DNA. Each genome's metagenomic coverage is constant and depends only on its abundance. A complementary approach to sequencing DNA of a whole microbial community is single cell genomics. Prior to sequencing of a single cell, its DNA needs to be amplified. This usually is done by multiple displacement amplification (MDA), introducing a tremendous coverage bias. Poorly amplified regions result in extremely low sequencing coverage or physical sequencing gaps. These parts of the genome cannot be reconstructed in the subsequent assembly step, and therefore genomic information is lost.

Results

Frequently, single amplified genomes (SAGs) and shotgun metagenomes are generated from the same environmental sample. We developed a fast, *k*-mer based recruitment method to sensitively identify metagenomic "proxy" reads representing the single cell of interest, using the raw single cell sequencing reads as recruitment seeds. By assembling metagenomic proxy reads instead of the single cell reads, we circumvent most challenges of single cell assembly, such as the aforementioned coverage bias and chimeric MDA products. In a final step, the original single cell reads are used for quality assessment of the proxy assembly.

On real and simulated data we show that, with sufficient metagenomic coverage, assembling metagenomic proxy reads instead of single cell reads significantly improves assembly contiguity while maintaining the original accuracy. By applying our method iteratively, we span physical sequencing gaps and are able to recover genomic regions that otherwise would have been lost. However, careful contamination screening is needed.

Conclusions

We developed kgrep, a new tool that naturally exploits the complementary nature of single cells and metagenomes to improve *de novo* assembly of single cell genomes.

* abremges@cebitec.uni-bielefeld.de

Bayesian latent variable models for single-cell trajectory learning

Kieran Campbell & Christopher Yau
University of Oxford

May 5, 2016

Background

The transcriptomes of single cells undergoing diverse biological processes - such as differentiation or apoptosis - display remarkable heterogeneity that is averaged over in bulk sequencing. Single-cell sequencing itself offers only a snapshot of these processes by capturing cells of variable and unknown progression through them. Consequently, one outstanding problem in single-cell genomics is to find an ordering of cells (known as their pseudotime) that best reflects their progression, for which several computational methods have been proposed.

To date, the vast majority of such methods emphasise transcriptome-wide 'data-driven' approaches that assume no prior knowledge of gene dynamics along the trajectory during inference. The suitability of the inferred trajectory is typically assessed by post-hoc examination of a set of marker genes to ensure the inferred behaviour aligns with prior assumptions. Furthermore, most current methods are algorithmic and rely on heuristics as opposed to probabilistic models, which in the context of bifurcations requires the pseudotimes to be first inferred prior to the identification of any bifurcation events.

Results

Here we introduce a general probabilistic framework for single-cell trajectory learning based on Bayesian non-linear factor analysis and apply it to two outstanding problems in single-cell analysis. Firstly, we demonstrate how such a framework may be used to integrate prior knowledge of gene behaviour in trajectory inference. By assuming a parametric form of gene expression evolution across pseudotime we can place informative priors on parameters that govern gene behaviour within a Bayesian statistical framework. Consequently, we remove the need for subjective post-inference checks and simultaneously solve related problems such as trajectory orientation and setting implicit length scales. We demonstrate how using such methods only a small panel of marker genes are required to achieve comparable results to transcriptome wide 'data-driven' alternatives. We further demonstrate how such a method can be used to recover trajectories corresponding to known pathways in the presence of heavily confounding effects.

The second application of our framework is to modelling bifurcations in single-cell data. By considering a Bayesian mixture of factor analysers we simultaneously infer both the pseudotimes and branching behaviour of the cells, which is unique compared to existing methods. We derive a Gibbs sampler that allows for fast inference across hundreds of cells while accounting for the zero inflation that is pertinent to single-cell RNA-seq data. Notably, by using a Bayesian framework we can integrate prior knowledge of branch-specific gene behaviour allowing for robust inference on challenging datasets.

Conclusions

We introduce a flexible Bayesian framework that solves several outstanding issues in single-cell trajectory learning. This framework uniquely provides a principled method for integrating prior knowledge of gene behaviour along single-cell trajectories and allows for such trajectories to be learned from a

small panel of marker genes. We also introduce the first statistical method for bifurcation inference that simultaneously infers both the pseudotimes of the cells as well as the bifurcation events, providing robust trajectories as well as full uncertainty estimates. We apply our methods to a range of both synthetic and real data, and more generally discuss the challenges of single-cell latent variable modelling including the connection of principal component analysis to both pseudotime inference and dropout rate. We conclude by motivating why such methods can be applied to a wide range of 'omics' data including modelling cancer progression and patient treatment outcomes.

Background

The number of sequenced genomes is growing exponentially, profoundly shifting the bottleneck from data generation to genome interpretation. Traits are often used to characterize and distinguish bacteria, and are likely a driving factor in microbial community composition, yet little is known about the traits of most microbes. We present Traitar, the microbial trait analyzer, a fully automated software package for deriving phenotypes from the genome sequence. Traitar accurately predicts 67 traits related to growth, oxygen requirement, morphology, carbon source utilization, antibiotic susceptibility, amino acid degradation, proteolysis, carboxylic acid use and enzymatic activity.

Results

Traitar uses L1-regularized L2-loss support vector machines for phenotype assignments, trained on protein family annotations of a large number of characterized bacterial species, as well as on their ancestral protein family gains and losses. We demonstrate that Traitar can reliably phenotype bacteria even based on incomplete single-cell genomes and simulated draft genomes. We furthermore showcase its application by characterizing two novel Clostridiales species based on genomes recovered from the metagenomes of commercial biogas reactors, verifying and complementing a manual metabolic reconstruction.

Conclusions

Traitar enables microbiologists to quickly characterize the rapidly increasing number of bacterial genomes. It could lead to models of microbial interactions in a natural environment and inference of the conditions required to grow microbes in pure culture. Our phenotype prediction framework offers a path to understanding the variation in microbiomes. Traitar is available at https://github.com/hzi-bifo/traitar.

Title: Curation, c*haracterization and quantification of a PacBio transcriptome*

Tardaguila Manuel[2] , de la Fuente Lorena[1], del Risco Hector[2], Martí Cristina[1], Pereira Cecile[2], Moreno Victoria[3], Rodríguez Susana[4], Conesa Ana[1,2]

1. Centro de Investigación Príncipe Felipe, Genomics of Gene Expression, Valencia, Spain
2. Institute for Food and Agricultural Sciences, Department of Microbiology and Cell Science, University of Florida, Gainesville, USA
3. Centro de Investigación Príncipe Felipe, Gene Expression and RNA Metabolism, Valencia, Spain
4. Centro de Investigación Príncipe Felipe, Neuronal and Tissue Regeneration, Valencia, Spain

Background:

Alternative splicing, a widespread means of creating functional diversity in higher eukaryotes, entails substantial challenges for its bioinformatic analysis. Paramount among these is the elaboration of the transcriptome to analyse, specially given the high similarity rate between isoforms and the incompleteness and/or variation in the annotation of the 5' and 3' ends of the mRNA. Here we have applied both PacBio (long reads) and Illumina (short reads) sequencing in a murine model of neural stem cell differentiation. PacBio sequencing detects whole transcripts (mean length resolved is 3000 bp) and is ideal to elaborate precise transcriptomes and perform isoform discovery. The trade off is the high error rate (around 5%) and the loss of quantification power. Complementarily, Illumina allows for quantification of expression and for the correction of error-prone long reads.

Results:

Classification of our PacBio transcriptome based on the splice pattern of isoforms reveals 60% of transcripts match annotated references in Refseq and ENSEMBL, 30% show novel splice junctions and 5% map to regions thought to be deprived of coding potential (genic introns and intergenic regions). Analysis of splicing features such as non canonical splicing rate, retrotranscription artifacts or Splice Junction coverage among others revealed that PacBio transcriptome needed further curation. We have developed a classifier to deal with this curation and results show that curated transcripts show better splicing features. Further characterization of the novel isoforms involved the evaluation of their peptide coverage using large databases of mass spectrometry profiles. Lastly, important expression associations can be made from this data: we found that most of multi-isoform genes expressed at least one additional annotated isoform at greater levels, in the majority of the cases it being the so called Principal Isoform, while a reduced subset of genes only expressed the novel isoform.

Conclusions:

Our results prove that the output of the PacBio Isoseq pipeline requires careful curation in order to eliminate isoforms showing abnormal features of splicing. After this curation has been done, the percentage of novel isoforms remains as high as 30% indicating the suitability of PacBio to perform the discovery of novel isoforms that are robust. Besides as we and others have found, the use of a filtered transcriptome instead of a global reference, diminishes the amount of quantification artifacts. Altogether these results shed light into the complex dynamics of alternative splicing and points to the necessity of using restricted transcriptomes to adequately analyze gene expression at the isoform level.

Resource-efficient Assembly of Large Genomes with Bloom Filter ABySS

Ben Vandervalk, Hamid Mohamadi, Justin Chu, Shaun D Jackman, Golnaz Jahesh, Lauren Coombe, Rene L Warren, Inanc Birol

Michael Smith Genome Sciences Centre

April 29, 2016

Background

Since the introduction of the de Bruijn graph assembly approach by Pevzner et al. in 2001, de Bruijn graph assemblers have become the dominant method for de novo assembly of large genomes. Nonetheless, assembling large genomes remains a challenging task. For instance, the estimated memory requirements for a human genome assembly with the ALLPATHS-LG assembler is 512GB of RAM. While distributed de Bruijn graph assemblers such as ABySS, Ray, and PASHA eliminate the requirement for a single large-memory machine by distributing the de Bruijn graph across multiple cluster nodes, these assemblers still require a computing cluster with a large amount of aggregate memory and a high-speed network fabric. While assemblers typically represent the de Bruijn graph as a hash table of k-mers, the Minia assembler (Chikhi et al., 2012) introduced a more compact probabilistic representation using a Bloom filter, which reduces the memory requirement by orders of magnitude and renders large genome assemblies feasible on a single commodity machine.

Results

Here we present two fundamental improvements to the ABySS assembler that reduce the memory and running time for large genome assemblies. First, as in Minia, we have reduced memory requirements by an order of magnitude through the use of a Bloom filter de Bruijn graph. While Minia

is a standalone unitig assembler, our new Bloom filter assembler is integrated with the existing ABySS pipeline, including downstream stages for contig building, mate pair scaffolding, and long read scaffolding. Second, we have reduced assembly time through the use of a specialized hash function called "ntHash". In our application, ntHash achieves runtimes that are orders of magnitude faster than standard hash functions through the use of a constant-time sliding window calculation, where the hash value of each k-mer is computed from the hash value of the k-mer that preceeds it. On a single 32-core machine with 120GB RAM, the new Bloom filter version of ABySS is able to assemble a modern 76X human dataset (SRA:ERR309932) and scaffold with MPET data (SRA:ERR262997) with an NG50 of 1.7 Mbp, wallclock time of 46 hours, and a peak memory usage of 102GB RAM.

Conclusions

While many implementations of de Bruijn graph assemblers are available, de novo assemblies of large genomes such as Homo sapiens still require heavy computational resources. Here we have demonstrated improvements to ABySS with respect to both memory usage and running time that significantly reduce the cost of assembling large genomes.

HiLive – Real-Time Mapping of Illumina Reads while Sequencing

Martin S. Lindner[1,#], Benjamin Strauch[1], Jakob Schulze[1], Piotr W. Dabrowski[1,2], Andreas Nitsche[2], Bernhard Y. Renard[1,*]

[1] - *Research Group Bioinformatics (NG 4), Robert Koch Institute, Berlin, Germany*
[2] - *Centre for Biological Threats and Special Pathogens, Robert Koch Institute, Berlin, Germany*
[#] - *current affiliation: Karius Inc., Menlo Park, CA, United States of America.*

Background

Next Generation Sequencing (NGS) is increasingly used in time critical setups, such as in clinical diagnostics or precision medicine. Today, the computational analysis of the massive amounts of data produced by modern devices is still a bottleneck on the way to the final interpretation of the experiment. Mapping reads to reference sequences is an essential step in many analysis pipelines. While read mapping algorithms have always been optimized for speed, they follow a sequential paradigm and only start after finishing of the sequencing run and conversion of files. The time while the sequencer is running is typically not used for data analysis.

We developed HiLive, the first general purpose read mapper that performs read mapping while the sequencer is still sequencing. HiLive makes use of the intermediate results generated by Illumina machines to perform read mapping and thereby drastically reduces crucial overall sample analysis time, e.g. in precision medicine.

Results

We present HiLive as a novel real time read mapper that is able to perform read mapping on the temporary, unfinished read data generated by Illumina sequencers. Such a strategy is facing mainly two problems: (i) Parallelism: > 1 billion reads are generated by the sequencer in parallel and need to be processed simultaneously to overcome the sequential paradigm of traditional read mappers. (ii) Incomplete information: Calculating the optimal alignment is not possible when the read is not completely sequenced. Therefore, many candidate alignments need to be stored for each read in the intermediate cycles. To address these problems, HiLive implements a k-mer based alignment strategy: the mapper continuously reads the intermediate BCL files created in each cycle of the instrument and extends initial k-mer matches by the increasingly produced data from the sequencer. We use exact and heuristic quality criteria to determine false alignments as early as possible without discarding true alignments. The overall memory footprint and required disk space is kept low by a slim implementation and data streaming.

We applied HiLive on real human transcriptome data to show that live mapping is technically possible and no compromise has to be made in comparison to traditional mappers. In our experiment, we mapped the 1.7 billion NGS reads generated in one Illumina HiSeq 1500 run to the human transcriptome. On a workstation size computer (32 cores), HiLive finished read mapping 9 min 53 s after the end of the sequencing run. Conversion of the BCL files to fastq files took already 48 min, and subsequent mapping with BWA took 12 h 31 min. Comparison to BLAST alignments shows that HiLive is on par with current read mappers, such as Bowtie 2, BWA, and Yara with slight advantages in sensitivity. These findings on the real data could be reproduced in an experiment based on simulated data.

Conclusions

We could show that live mapping of Illumina reads is technically and practically possible. Our tool HiLive allows a massive reduction in total sample analysis time by starting read mapping while the sequencer is still running. Although HiLive implements a completely different alignment strategy, the quality is comparable to other state of the art mappers.

HiLive is freely available from https://sourceforge.net/projects/hilive/ .

GRIDSS: sensitive and specific genomic rearrangement detection using positional de Bruijn graph assembly

Authors: Daniel L Cameron, Anthony T Papenfuss

Background

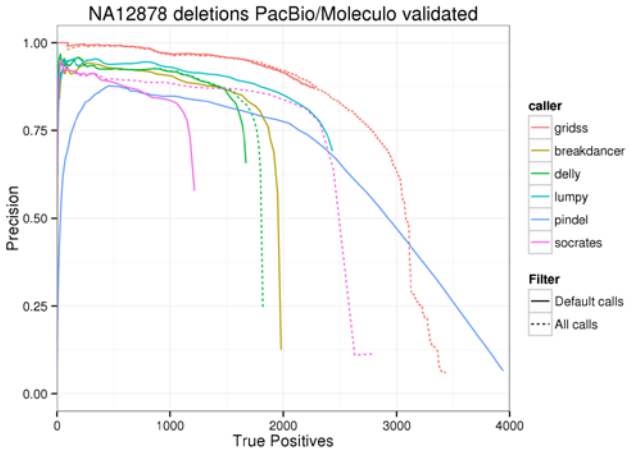
Many methods exist to identify structural variants (SVs) using high-throughput sequencing data with most methods using one or more of four approaches: read depth analysis (RD), discordantly-aligned read pair clustering (DP), split reads identification (SR), and assembly. RD approaches (e.g. CNVnator, Control-FREEC) are limited in their event size detection and cannot resolve breakpoint partners. DP approaches (e.g. BreakDancer, GASVPro) can be used to infer the presence of SVs but cannot in general identify exact breakpoint locations since the breakpoints occur in the unsequenced part of the fragments whereas SR approaches (eg CREST, Socrates) can obtain single nucleotide resolution by identifying breakpoint-spanning reads. Assembly-based methods perform either *de novo* assembly (e.g. cortex_var), targeted assembly based on previously identified candidates (e.g. SVMerge, TIGRA), or perform windowed assembly to detect small events (e.g. DISCOVAR, SOAPindel). These approaches are not mutually exclusive with some software incorporating two (e.g. DELLY) or three (e.g. LUMPY) of these approaches.

Here we describe GRIDSS, the Genome Rearrangement IDentification Software Suite, composed of an assembler, and a variant caller which combines assembly, split read and read pair evidence to identify structural variants. Our novel assembly approach performs genome-wide breakend assembly (that is, independent assembly of each side of each breakpoint) by using a genome-wide positional de Bruijn graph. Soft-clipped reads, split read, discordant read pairs, and read pairs with only one read mapped are assembled into the positional de Bruijn graph with the mapping locations of each read encoded as positional constraints within the graph itself. Post-assembly, we use the same realignment approach used to identify split reads from soft-clipped reads to identify the breakpoint supported by each breakend contig. Once identified, we used a probabilistic model to call variants from supporting assembly contigs, split reads, and discordant read pairs.

Results

To benchmark GRIDSS, we compared against BreakDancer, DELLY, LUMPY, Pindel, and Socrates on both simulated data and well-characterised cell lines. We simulated deletions, insertions, inversion, tandem duplication, and genomic fusions on 2x100bp sequencing at

varying levels of coverage. Above 8x coverage, GRIDSS sensitivity exceeds that of the other callers for events larger than 100bp, with Pindel showing highest sensitivity for small events. To compare performance on realistic data, we evaluated the callers on the Illumina Platinum Genomics 50x WGS NA12878 data (See Figure).



GRIDSS is able to almost halve the false discovery rate compared to other callers, with the highest scoring GRIDSS calls having a FDR close to zero. GRIDSS's execution time of 236 minutes is comparable to the 52, 82, 211, 489, and 2,184 minutes of SOCRATES, BreakDancer, LUMPY, DELLY, and Pindel respectively.

We have applied GRIDSS in multiple cancer contexts. Firstly, we have used GRIDSS to identify patient-specific somatic breakpoints. Secondly, we have used the single nucleotide precision of GRIDSS to identify complex compound rearrangements misclassified as simple events by a DP-based caller in 64 variants (5%) in tumour neochromosomes. Thirdly, we are using the multi-sample capability of GRIDSS to reconstruct somatic phylogenetic trees in both mouse xenograft and patient tumours.

Conclusion

GRIDSS achieves high sensitivity and specificity on simulated, cell line and patient tumour data. On NA12878 cell line data, GRIDSS halves the false discovery rate compared to other recent methods.

Our novel incorporation of assembly, split read and read pair evidence in the variant calling process is made possible by our approach of independently assembling each breakend. By using a genome-wide positional de Bruijn graph, we are able to perform untargeted assembly an order of magnitude faster than existing approaches. GRIDSS can perform combined variant discovery on multiple related samples and population data. GRIDSS is freely available at https://github.com/PapenfussLab/gridss.

An assembly approach utilizing next and third generation sequencing data for powerful structural variant detection

Xian Fan[1,2], Zechen Chong[2], Luay Nakhleh[1], Human Genome Structural Variation Consortium, Ken Chen[1,2*]

[1]Department of Computer Science, Rice University, Houston, Texas (USA)
[2]Department of Bioinformatics and Computational Biology, The University of Texas M.D. Anderson Cancer Center, Houston, Texas (USA)

Background

Detection of structural variations (SV), including both large and small (<50 bp) ones, is important in understanding human genetic diseases. Conventional approaches that utilize next generation sequencing (NGS) technologies (such as Illumina) have limited detection power due to short read length. Third generation sequencing (TGS) technologies, such as the Pacific BioScience (Pacbio) single molecule sequencing technology, facilitate SV identification by generating much longer reads. However, the long reads produced by TGS often have high sequencing error rates (~15%), which leads to 1) imprecise alignment to the reference, and 2) challenges in detecting small SVs.

It is therefore reasonable to develop computational approaches that combine the advantages of the NGS and TGS data in order to further improve the detection of SVs. However, while very few algorithms have been developed for this task thus far, none jointly utilizes both types of reads for discovery on the sequence level, nor targets novel insertions and small INDELs.

Result

We developed a hybrid assembly-based approach that utilizes both Illumina and Pacbio reads to discover SVs. The approach starts with an Illumina BAM file and Pacbio raw subreads. It extracts Illumina reads that cannot be well aligned to the reference, and aligns these reads to all Pacbio reads, aiming to extract Pacbio reads that span SVs. This process requires an aligner with high sensitivity in spite of high error rate in Pacbio reads and short length in Illumina reads. We utilized a customized version of BLASR for this purpose that achieved >90% success rate (percentage of the Illumina reads that have at least one high quality alignment to Pacbio reads). The pairwise alignments between Illumina and Pacbio reads form a bipartite graph, in which nodes represent the reads (Illumina and Pacbio reads are the two partite sets), and edges correspond to matches by alignment. We cluster the graph into connected components using a near linear graph-theoretic union-find algorithm. Each connected component contains a set of reads (including both Illumina and Pacbio) that have shared homology and likely originate from the same SV. We apply Celera Assembler to assemble the Pacbio reads in each connected component and produce contigs representing reconstructed alternative alleles. We align the contigs to the reference and identify putative SV breakpoints. Finally,

Illumina reads in the corresponding connected component are aligned to the assembled contigs to confirm the existence of the breakpoints. This method allows us to detect SVs of a wide range of sizes (11bp to > 10kbp), particularly INDELs in Short Tandem Repeats (STR) and large novel insertions.

To evaluate this approach, we ran it on the Pacbio and Illumina data generated from a haploid hydatidiform mole (CHM1) genome. An SV call set (A) was previously generated from the Pacbio data by a reference-alignment guided local assembly approach by Chaisson et al. We also generated SV call sets using Delly (B) and Lumpy (C) from Illumina data only. Our algorithm utilized 0.7% Illumina and 9% Pacbio reads and identified 3,268 large deletions (>50bp), 5,651 large insertions (>50bp), 13,223 small deletions (<=50bp), and 14,715 small insertions (<=50bp). 72% of large deletions identified by our method were also identified by at least one other method, which indicates a high specificity of our method. Additionally we detected 826 unique calls, which overlap well with known SVs in database of genomic variants (DGV, 87%) and STRs (65%), indicating a high sensitivity and specificity of our approach. Our method also identified 14 large (>500 bp) novel insertions (relative to build37) missed by Chaisson et al. but validated by build38. To evaluate small INDELs, we compared with Pindel and GATK. 70% deletions and 76% insertions in our call set were identified by at least one other method. The 3,933 novel deletions and 3,347 novel insertions we identified overlapped well with dbSNP (87% for deletion and 86% for insertion) and STR annotations (89% for deletion and 67% for insertion), indicating a high sensitivity and specificity of our approach.

We further applied our method to the three trios (YRI, PUR and CHS) in the 1000 Genomes Project (or Human Genome Structural Variation Consortium). Both Illumina and Pacbio reads were available for these trios. On average, we called ~26,000 SVs per sample and 20% of our calls are novel with respect to 4 other methods that analyze either only Illumina data (e.g., Delly, Pindel, Manta) or only Pacbio data.

Conclusion

We developed a novel method for SV detection through joint utilization of both NGS and TGS coverage at raw read level. Results obtained from analyzing a single haploid and 3 trio human samples indicate that our method can utilize the advantages of two platforms (accuracy of the Illumina reads and the length of the Pacbio reads), and generate high accuracy SVs with novel calls. In particular, our method can detect SVs of a wide size range, from 11bp to >10kbp, and is particularly effective at detecting large novel insertions not present on the reference, and small INDELs, which are challenging to other methods.

CAMSA: A Tool For Comparative Analysis And Merging Of Scaffold Assemblies*

Sergey Aganezov and Max A. Alekseyev

The George Washington University, Washington, DC

Background

Despite the recent progress in genome sequencing and assembly, many of the currently available assembled genomes come in a draft form. Such draft genomes consist of a large number of genomic fragments (*scaffolds*), whose positions and orientations along the chromosomes are unknown. The *scaffold assembly* problem asks for reconstruction of chromosomes from a set of scaffolds by identifying pairs of scaffolds extremities (*assembly points*) to be glued together. While there exists a number of methods for solving the scaffold assembly (using various computational and wet-lab techniques), they often can produce only partial error-prone assemblies.

Depending on the utilized information and the underlying techniques, different scaffold assembly methods may produce results that differ from each other. Moreover, some scaffold assemblers can produce only non-oriented assemblies, where the relative orientation of (some) scaffolds in assembly points is yet to be determined. It therefore becomes important to compare and merge scaffold assemblies produced by different methods, thus combining their advantages and highlighting potential conflicts for further investigation. These tasks may be labor intensive if performed manually.

We present CAMSA, a tool for comparative analysis and merging of scaffold assemblies. CAMSA takes as an input two or more assemblies of the same set of scaffolds and generates a comprehensive comparative report for them. The report not only contains multiple numerical metrics for the input assemblies, but also provides an interactive framework for their visual comparison and analysis. CAMSA is available for download from `https://cblab.org/camsa/`.

Methods

CAMSA interprets the input assemblies as sets of assembly points, and further analyzes and classifies individual assembly points by a numbers of characteristics (e.g., uniqueness, orientation, conflictedness, etc). Results of this analysis are then reported at the levels of whole assemblies and individual assembly points.

For the purpose of comparative analysis and visualization of the input scaffold assemblies, CAMSA utilizes the *multiple breakpoint graph* (MBG) data structure traditionally used for analysis of gene orders across multiple species [4]. The MBG in CAMSA is formed by directed *scaffold edges* and undirected *assembly edges* of different colors representing the different input assemblies (Fig. 1). While conventional MBG is constructed for sequences of *oriented* genes (where orientation is defined by the strand), in CAMSA we extend it to support sequences of non-oriented scaffolds.

In addition to generating a comprehensive comparison report, CAMSA also produces a *merged assembly* that is most consistent with all input assemblies. CAMSA can take into account the level of confidence of each assembly point in each input assembly, which can be specified as the *confidence weight* on the scale from 0 to 1 (with 1 being the default value). These confidence weights contribute to the weights of assembly (multi-)edges in the MBG, which are then used to construct the *merged assembly* as the maximal matching on assembly edges (shown as bold colored edges in Fig. 1). We further use the constructed merged assembly to identify orientation for some non-oriented assembly points that is most consistent across the input assemblies (e.g., in Fig. 1 for the blue non-oriented assembly point $(\overrightarrow{F}, \overleftarrow{G})$ it suggests orientation $(\overrightarrow{F}, \overrightarrow{G})$), as well as to resolve issues of varying resolution across different assemblies, i.e., when a scaffold is missing in one assembly but is present in another (e.g., in Fig. 1 the scaffold D is missing in the red assembly, but is present in the blue assembly as well as in the merged assembly).

Results

The results of scaffold assembly analysis in CAMSA are presented in the form of an interactive report for the set of input assemblies, and an interactive visualization of the input and merged assemblies. Extensive interactive filtering options

*The work is supported by the National Science Foundation under the grant No. IIS-1462107.

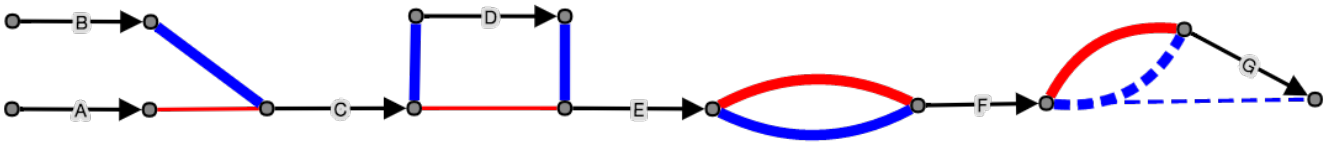


Figure 1: The MBG of "red" and "blue" assemblies of the same set of scaffolds $\{A, B, C, D, E, F, G\}$, where directed black edges correspond to scaffolds, red and blue edges correspond to assembly points, dashed edges represent alternative orientations for non-oriented assembly points, and bold edges indicate assembly points that participate in the merged assembly.

allow researchers to identify and work with groups of assembly points that are of most interest.

First section of the report produced by CAMSA focuses on comparison of each assembly to the others and presents several characteristics such as:

(i) number of *unique* assembly points (i.e., present only in one assembly);
(ii) percentage of *non-oriented* assembly points;
(iii) number of assembly points *shared* with other assemblies, with specification of particular subset of such assemblies (e.g., in Fig. 1 the assembly point $(\overrightarrow{E}, \overrightarrow{F})$ is shared by red and blue assemblies);
(iv) number of *conflicting* assembly points, i.e., scaffolds' extremities participating in different assembly points in other assemblies (e.g, in Fig. 1 the red assembly point $(\overrightarrow{A}, \overrightarrow{C})$ conflicts with the blue assembly point $(\overrightarrow{B}, \overrightarrow{C})$);
(v) proportion of assembly points that participate in the merged assembly.

Second section of the report addresses individual assembly points in the context of all input assemblies. For each assembly point P, CAMSA reports several characteristics such as:

(i) a set of *source assemblies* that contain P;
(ii) a flag specifying if P is oriented;
(iii) a set of non-source assemblies *conflicting* with P;
(iv) a subset of source assemblies that are uncertain about P (e.g., suggest alternative assembly points conflicting with P);
(v) a flag specifying if P is present in the merged assembly.

The interactive visualization of the input and merged assemblies is represented in the form of their MBG. This representation is dynamic with respect to the graph layout as well as the filtration of graph components.

Conclusions

CAMSA addresses the current deficiency of automated comparison and merging of multiple assemblies of the same scaffolds. Due to existence of various methods and techniques for scaffold assembly, identifying similarities and dissimilarities across different assemblies is beneficial both for developers of scaffold assembly algorithms and researchers improving genome assembly of specific organisms.

We remark that an alpha version of CAMSA is currently utilized in the study of Anopheles mosquito genomes, where multiple research laboratories (including ours) work on improving the existing assemblies for a number of mosquito species [5]. This project utilizes several scaffolding techniques [3, 1, 2], ranging from PacBio-based to homology-based assembly methods. CAMSA provides an automated framework for interactive comparison, analysis, and integration of constantly improving scaffold assemblies, thus helping the researchers to refine the resulting genome assemblies.

References

[1] Sergey Aganezov, Nadia Sydtnikova, AGC Consortium, and Max A. Alekseyev. Scaffold assembly based on genome rearrangement analysis. *Computational Biology and Chemistry*, 57:46–53, 2015.

[2] Yoann Anselmetti, Vincent Berry, Cedric Chauve, Annie Chateau, Eric Tannier, and Sèverine Bérard. Ancestral gene synteny reconstruction improves extant species scaffolding. *BMC Genomics*, 16(10):1–13, 2015.

[3] Lauren Assour and Scott Emrich. Multi-genome Synteny for Assembly Improvement. *Proceedings of 7th International Conference on Bioinformatics and Computational Biology*, pages 193–199, 2015.

[4] Pavel Avdeyev, Shuai Jiang, Sergey Aganezov, Fei Hu, and Max A. Alekseyev. Reconstruction of ancestral genomes in presence of gene gain and loss. *Journal of Computational Biology*, 23(3):1–15, 2016.

[5] D. E. Neafsey, R. M. Waterhouse, M. R. Abai, S. S. Aganezov, M. A. Alekseyev, et al. Highly evolvable malaria vectors: the genomes of 16 Anopheles mosquitoes. *Science*, 347(6217):1258522, 2015.

Genotyping somatic insertions and deletions

Louis J. Dijkstra[1,2,3], Johannes Köster[1], Tobias Marschall[4,5,*], Alexander Schönhuth[1,*]

[1] Life Sciences Group, Centrum Wiskunde & Informatica, Amsterdam, The Netherlands
[2] Computational Science Lab, Universiteit van Amsterdam, The Netherlands
[3] Department of High Performance Computing, ITMO University, St. Peterburg, Russia
[4] Center for Bioinformatics, Saarland University, Saarbrücken, Germany
[5] Max Planck Institute for Informatics, Saarbrücken, Germany
[*] Joint last authorship
alexander.schoenhuth@cwi.nl

May 8, 2016

Keywords. Cancer Genomes; Insertions and Deletions; Next-Generation Sequencing Data; Precision Oncology; Somatic Variants; Data Uncertainty

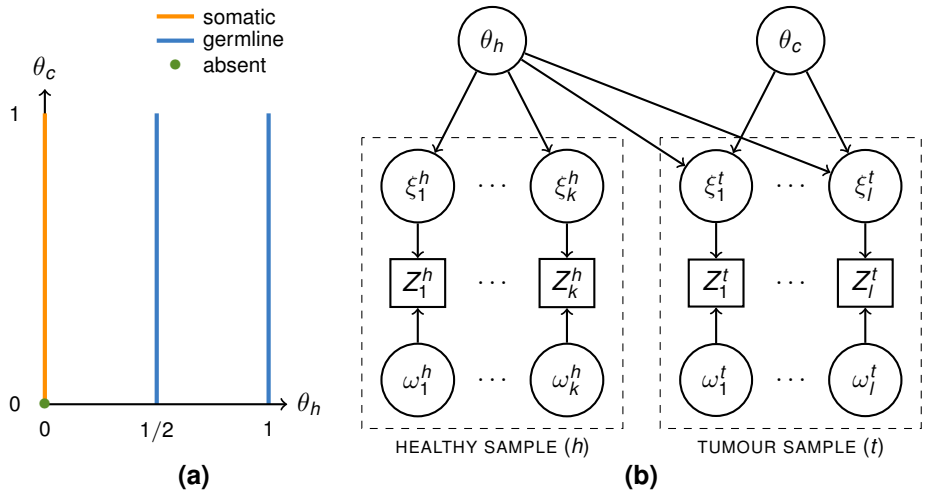
Background. Cancer is a genetic disorder in the first place; somatic mutations in the genome of an originally healthy cell allow for the maintenance of a potentially rapidly proliferating hetergeneous mix of cancer clones. This explains why in the recent past several thousands of cancer/control genome pairs have been sequenced, concerted by global consortia [7]; to date, the cancer genomes sequenced already amount to petabytes of data. The promises this massive pile of data holds for applications in precision oncology are enormous and have the potential to lead to drastically improved diagnosis and selection of therapy protocols.

While calling somatic single nucleotide variants (SNVs) can be done at both high recall and precision (e.g. [1, Mutect]), calling somatic indels has remained difficult and complex. Confounding factors, such as alignment and fragment length uncertainty, can pose substantial challenges already for germline indels. This becomes particularly disturbing for indels of length 30-150 bp (the *NGS twilight zone of indels*). In earlier work, we have shown how to resolve these issues and safely call and genotype substantial amounts of such indels in the frame of population-scale projects [4, 3, 5].

When calling and genotyping somatic indels, where genotyping refers to estimating the allele frequency of variants, cancer heterogeneity and data impurity make another confounding layer of issues. Heterogeneity and impurity of samples imply that estimating the allele frequency of somatic variants requires to appropriately quantify the inherent uncertainties. Only if this complex mix of disturbing factors has been appropriately disentangled, calling somatic indels at sufficient recall and precision is possible. Since, prior to our approach, there have been no methods to call *somatic twilight zone indels*, somatic variant databases are still virtually devoid of this type of genetic variation. Beyond this, genotyping somatic indels has also remained a substantial computational challenge in general.

Results. Here, we present a method, PROSIC (Postprocessing somatic indel calls), based on a Bayesian latent variable model (see Fig. 1) that aids in genotyping somatic indel calls while accounting for the above mentioned confounding factors of impurity, the unknown clonal structure, and alignment and typing uncertainty. Our method requires a list of potential somatic indel calls in VCF format, together with a cancer

and a matched normal BAM file as input. The output then is an annotated VCF where indel calls have been genotyped (by a VAF estimate) and been equipped with a Bayesian type a posteriori probability that the indel is somatic, as derived from the model.



(a) Genotype space. Genotypes need to be estimated for both cancer (θ_c) and control sample (θ_h). While $\theta_h \in \{0, \frac{1}{2}, 1\}$, representing absence, hetero- or homozygosity of the variant, $\theta_c \in [0, 1]$, reflecting that VAF's of somatic variants can cover the whole range due to cancer heterogeneity and impurity. **(b)**: The latent variable model, where $i \in \{1, ..., k\}, j \in \{1, ..., l\}$ index the alignments of the healthy and the cancer sample, respectively. Latent variables representing uncertainties (ω, ξ) and allele frequencies (θ_h, θ_c) are represented by circles; note that θ_h has an influence also on the cancer sample, which addresses impurity. Rectangles represent variables (Z_i^j) that can be immediately observed, such as alignment length and gaps.

We have evaluated our model on simulated data and on the datasets provided by the DREAM challenge (see `https://www.synapse.org/#!Synapse:syn312572`). We demonstrate that we can raise both recall and precision substantially, often achieving quite drastic improvements (more than 30% in recall and 30-40% in precision, reaching precision rates of 85-95%) in comparison to standard, best-practice somatic indel calling workflows provided by gold standard indel discovery methods such as Platypus [6], Pindel [8] and the HaplotypeCaller [2]. We also demonstrate that our tool compares very favorably with best practice pipelines on cancer/control cell line data. Finally, we point out ways how to substantially increase recall in the *somatic indel twilight zone* of 30-150 bp at precison rates of at least 80% which, to the best of our knowledge, is novel. The German Cancer Research Center (DKFZ) has submitted an official proposal that our tool will be integrated into the ICGC somatic indel calling pipelines to postprocess and genotype indel calls arising from the latest TCGA project (`https://tcga-data.nci.nih.gov/tcga/tcgaAbout.jsp`) on more than 2800 matched cancer/control genome pairs.

Conclusions. We present a statistical, latent variable model which allows to estimate allele frequencies of indels in matched cancer/control samples, and to derive Bayesian a posteriori probabilities for the indel calls to be somatic. In this, we take all disturbing data uncertainties, such as sample impurity, cancer heterogeneity, alignment and typing uncertainties into account, which also allows us to make good calls in relatively difficult-to-access regions of the human genome. When applying our model to indel callsets generated by gold standard indel discovery tools, we achieve substantial improvements over current best-practice workflows both in terms of recall and precision. In summary, we are providing a tool that allows to leverage ordinary, well-approved indel callers into high quality somatic indel callers. See `https://github.com/louisdijkstra/somatic-indel-calling` for software.

References

[1] K. Cibulskis, M.S. Lawrence, S.L. Carter, A. Sivachenko, D. Jaffe, C. Sougnez, S. Gabriel, M. Meyerson, E.S. Lander, and G. Getz. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nature Biotechnology*, 2013.

[2] Mark A DePristo, Eric Banks, Ryan Poplin, Kiran V Garimella, Jared R Maguire, Christopher Hartl, Anthony A Philippakis, Guillermo del Angel, Manuel A Rivas, Matt Hanna, Aaron McKenna, Tim J Fennell, Andrew M Kernytsky, Andrey Y Sivachenko, Kristian Cibulskis, Stacey B Gabriel, David Altshuler, and Mark J Daly. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, 43(5):491–498, May 2011.

[3] The Genome of the Netherlands Consortium. Whole-genome sequence variation, population structure and demographic history of the Dutch population. *Nature Genetics*, 2014.

[4] Hehir-Kwa et al. A high-quality reference panel reveals the complexity and distribution of structural genome changes in a human population. Technical report, bioRxiv:036897, 2016.

[5] Tobias Marschall, Iman Hajirasouliha, and Alexander Schönhuth. MATE-CLEVER: Mendelian-inheritance-aware discovery and genotyping of midsize and long indels. *Bioinformatics*, 29(24):3143–3150, 2013.

[6] A. Rimmer, H. Phan, I. Mathieson, Z. Iqbal, S.R.F. Twigg, WGS500 Consortium, A.O.M. Wilkie, G. McVean, and G. Lunter. Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nature Genetics*, 2014.

[7] The International Cancer Genome Consortium. International network of cancer genome projects. *Nature*, 464(7291):993–998, 2010.

[8] Kai Ye, Marcel H. Schulz, Quan Long, Rolf Apweiler, and Zemin Ning. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics*, 25(21):2865–2871, Nov 2009.

Title

Evaluation of strategies for somatic mutation discovery in tumor specimens without matched germline: effect of tumor content, sequencing depth, and copy number alterations

Authors

Rebecca F. Halperin[1], John D. Carpten[2], Jessica Aldrich[1], Winnie S. Liang[1], Jonathan Keats[3], Megan Russell[1], Daniel Enriquez[1], Ana Claasen[1], Irene Cherni[3], Seungchan Kim[3], David W. Craig[1],

[1]Neurogenomics Division, Translational Genomics Research Institute, Phoenix, AZ
[2]University of Southern California, Department of Translational Genomics, Los Angeles, CA
[1]Integrated Cancer Division, Translational Genomics Research Institute, Phoenix, AZ

Introduction

Large-scale multiplexed identification of somatic alterations in cancer has become feasible with next generation sequencing (NGS). By definition, somatic alterations are those that are found in the tumor and not the germline sequence, so the standard approach to somatic variant detection involves comparing the tumor sequence to the germline sequence of the same individual. However, in some situations, such as with archival samples, blood or other constitutional tissue samples are not available to obtain germline sequence. In order to identify somatic variants in such tumor samples, the tumor is typically compared to a reference sample, and then the variants that are found public germline variant databases are filtered out. However, all individuals will have some private germline variants not found in any database. Differences in allele frequencies between somatic and germline variants in impure tumors can also help to differentiate somatic and germline variants. Here we will examine the extent to which leveraging allele frequencies can help to overcome false positives due to private germline variants in tumor only calling.

Results

We developed a Bayesian framework to integrate the population frequency and allele frequency information. At each position, we determine the prior probability of a germline or somatic based on 1000 Genomes or COSMIC frequencies, respectively. We also estimate copy number, minor allele copy number, and clonal sample fraction in order to calculate expected allele frequencies of somatic and germline variants at each position. As expected, the higher the clonal sample fraction, the closer the expected allele frequencies are for somatic and germline variants. We also find that there also other combinations of tumor content and copy number state where the expected allele frequencies of somatic and germline variants are very similar.

Applying this framework to simulated data, we estimate coverage required for different tumor content and copy number states. For example, to detect about 90% of the somatic variants in a diploid region of a 50% tumor sample, we would only need 200X mean target coverage, but we would need 800X mean target coverage to achieve the same sensitivity in a 75% tumor sample, or 1600X for an 85% tumor sample. We then apply the framework to a set of nine cancer samples. We find that the observed

sensitivity correlates well with the expected sensitivity based on the coverage, the clonal sample fractions, and the copy number alterations. In silico dilutions and downsampling experiments also confirm the expected relationships between coverage, tumor content, and sensitivity.

We find that the Bayesian tumor only caller is able to greatly reduce false positives due to private germline variants, with greater than 95% of true private germline variants correctly classified as germline. The calling precision is also significantly improved with Bayesian approach, which has an average positive predictive value of greater than 70% compared to 35% with database filtering alone. Overall the accuracy of the Bayesian tumor only caller is greater than 99.9%

Conclusions

Our Bayesian tumor only calling approach can eliminate most false positives due to private germline variants. However, the sensitivity of the approach is dependent tumor content, coverage, and copy number alterations. The data presented here can be used to design tumor only sequencing experiments with appropriate coverage based on the sample characteristics.

A Bayesian Network Algorithm for Somatic Mutation and Germline Variant Identification from Tumor Molecular Profiling of Cancer Patients by High-Throughput Sequencing

Francisco M. De La Vega,[1,2] Sean Irvine,[3] David Ware[3], Kurt Gaastra[3], Yosr Bouhlai[1], Daniel Mendoza[1], Anna Vilborg[1], Yannick Pouliot[1], Federico Goodsaid[1], Austin So[1], and Len Trigg.[3]
[1]TOMA Biosciences, Foster City, CA 94404, USA, [2]Stanford University School of Medicine, Stanford, CA, USA, and [3]Real Time Genomics, Hamilton, New Zealand.

Background

Cancer tumor profiling by targeted resequencing of actionable cancer genes is rapidly becoming the standard approach for selecting targeted therapies in refractory cancer patients. In this scenario, DNA from a tumor FFPE sample is sequenced deeply by targeted next-generation sequencing (NGS) to uncover actionable somatic mutations in relevant cancer genes. Currently, clinical labs preforming such tests under the CLIA regulation, largely utilize analysis pipelines based in academic tools developed as part of the TCGA or ICGC projects, where tumor and germline specimens from cancer patients are sequenced in parallel to facilitate the identification of cancer somatic mutations *vs* germline variants. A major challenge that arises in the clinical scenario is the need to analyzing tumor-derived data in the absence of normal/germline tissue data, as the current standard of care only requires pathologists to obtain a biopsy of the tumor tissue[1]. This makes very difficult to distinguish between somatic and germline variants, leaving clinicians to resort to crude heuristic filtering procedures with unknown performance. Furthermore, recent benchmarking of somatic calling methods have shown poor performance and significant inconsistencies in the major published algorithms, even when provided with both tumor and normal tissue data[2].

Results

Here we present Bayesian network variant caller to identify both SNV and indel somatic mutations and germline variants from targeted resequencing data from tumor tissue samples. Our approach models the distribution of reads harboring germline and somatic mutations in cancer cells, estimates the contamination from normal tissue in tumor specimens, scores putative somatic mutation, and imputes germline variants present in the genome of cancer cells and contaminating normal cells, without matching normal tissue data. Our "tumor-only" caller can also utilize site- and allele-specific prior information to calculate the scores of somatic mutations, from sources such as databases of *bona fide* somatic mutations (e.g. COSMIC), catalogs of germline variation in populations (e.g. 1000 Genomes Project), and data from a panel of normal samples analyzed with the same assay platform to reduce systematic technology artifacts. This method has been developed in Java on top of the libraries of a previously developed variant caller.[3]

We validated our method by analyzing data obtained with the TOMA OS-Seq targeted enrichment assay for 130 cancer genes and then sequencing with the Illumina platform. Firstly, we obtained data from a gold standard sample for which a ground truth is available, the cell-line NA12878, upon which we simulated about 1,800 somatic mutations at variant allele fractions (VAF) ranging from 0.1 to 0.4, using the `bamsurgeon` software [4]. Secondly, we analyzed data from experiments where varying proportions of a reference sample (e.g. NA12878) is mixed with a constant amount of one of its parents, to simulate the behavior of tumor somatic mutations. Finally, we also analyzed data from cancer patient case triads, where normal, tumor and plasma cell free-DNA have been sequenced and we are able to compare the the results from the tumor-only caller vs the paired tumor/normal analysis also implemented in the software.

The ability to compare our results to a ground truth dataset permits us to evaluate our performance via Receiver Operator Characteristic (ROC) curves, where we can measure performance with the area under the curve, or true positive rates at a fixed FDR. Our initial evaluation of the caller showed that we can improve the AUC by providing priors for a database of somatic mutations, but the major benefit comes from utilizing a panel of normal samples. We can recover over 99% of true positives at a FDR of 1.6% when simulating mutations at a VAF of 0.4. As we reduce the VAF the separation in the improvements obtained by either of those methods decrease, as expected. As we evaluate the performance of our caller, it is important

to compare to other commonly used algorithms in cancer tumor profiling. We thus compared our results to the output from `FreeBayes`. We found that we can achieve >90% True Positive Rate (TPR) at 1.5% FDR while `FreeBayes` achieves only 15% TPR. At a 2% FDR, we achieve >99% TPR, while `FreeBayes` only achieves less than 80% TPR. While this is a work in progress and are un the process of evaluating additional datasets through our method and adjusting priors, we observe that our caller performs significantly better than other methods, and highlights the challenges of somatic mutation identification at low VAF.

Conclusions

We show that a Bayesian network approach is a very powerful method to infer somatic mutation calls from NGS data of mixed samples, such as tumor specimens, with the ability to decompose the mixture returning both somatic and germline variants calls, and leverage prior information in a natural and principled fashion. The Bayesian network approach allows not only to call somatic mutations, but to impute the germline genome to a considerable accuracy from the tumor sample. This is important information, as inherited susceptibility variants exist in cancer patients and this information should be used to both inform therapy and provide family counseling. Our method and ensuing software implementation provides a robust solution for a very common use case in clinical applications of NGS, where material form tumor biopsies from patients are analyzed to identify actionable somatic aberrations in the lack of normal sample. While we can strive to change the standard of practice by requiring a sample of the normal tissue to be sequenced in parallel to the tumor sample as done in research protocols for a paired tumor/normal analysis, these changes take many years[1]. In addition, even if these changes occur, this use case is still important to leverage the large scale biobanks of FFPE blocks that medical centers have accumulated for years together with clinical information and that are being sequenced to correlate molecular profiles, therapies, and outcomes retrospectively.

References

1. Topol, E. J. From Dissecting Cadavers to Dissecting Genomes. *Sci Transl Med* **5,** 202ed15–202ed15 (2013).
2. Alioto, T. S. *et al.* A comprehensive assessment of somatic mutation detection in cancer using whole-genome sequencing. *Nature Communications* **6,** 1–13 (2015).
3. Cleary, J. G. *et al.* Joint Variant and De NovoMutation Identification on Pedigrees from High-Throughput Sequencing Data. *Journal of Computational Biology* **21,** 405–419 (2014).
4. Boutros, P. C. *et al.* correspondence. *Nat. Genet.* **46,** 318–319 (2014).

Evolution of Structural Variation in Cancer Revealed by Read Clouds

Noah Spies, Ziming Weng, Alex Bishara, Justin M Zook,
Robert B West, Marc Salit, Arend Sidow

Background Structural variants, particularly distant translocations, are difficult to identify despite their fundamental importance in cancer and other diseases. Because any two genomic loci can be connected through a genomic rearrangement or translocation, the search space for structural variation is proportional to the square of the genome size, resulting in a massive multiple-testing problem for mammalian genomes. Even though current short-read technologies have very low rates of chimeric molecules and mismapping to the genome, these types of experimental and computational errors compound to result in high rates of false positives when searching genome-wide for structural variation. Furthermore, standard sequencing reads derive from short genomic fragments typically only several hundred base pairs in length, and thus cannot map uniquely to translocation breakpoints occurring in even moderately long repeat sequences.

Results The 10X Genomics platform generates barcoded short-reads from large genomic DNA fragments, which can then be clustered in silico to generate read clouds identifying the original large DNA fragments. We size-selected large (50–100kb) genomic DNA fragments from 7 spatially distinct tumor samples from a single sarcoma, as well as matched normal tissue, then applied the 10X platform to generate read clouds.

We have implemented new methods to identify structural variants from these read cloud data. We use the read cloud barcodes to identify candidate events where the similarity in barcode patterns between two loci is higher than expected given the distance between the loci. We then perform breakpoint refinement using the patterns of dropoff in observed long fragment density at the structural variant breakpoints.

Using this new method, we find structural variants that differ between sectors of the sarcoma, although most somatic structural variants (and

single-nucleotide variants) are shared across all samples in the tumor. Multiple, independent, ancestral chromothripsis events occurred in our sarcoma case, totaling hundreds of individual breakpoints shared between sectors.

To better understand these bursts of genome rearrangement, we have implemented a novel approach using patterns of read clouds to automatically reconstruct the order and orientation of complex structural variants involving many breakpoints. Furthermore, using the read cloud barcodes, we are able to identify all reads supporting a structural variant and assemble the full sequence of many of these complex structural variants (although this is still dependent on the local sequence complexity). This approach reveals that many of the complex structural variants involve the rearrangement of many short (several kb) genomic segments derived from distant locations on the same chromosome, forming new chromosomes. In the process of creating these neochromosomes, large intervening genomic segments are lost, resulting in a loss of heterozygosity.

Conclusions By harnessing the barcoded sequencing platform, we are able to phase and assemble complex genomic rearrangements, illuminating larger patterns of genome evolution in cancer. Because the read clouds derive from long DNA fragments, physical coverage of each breakpoint is substantially higher than for standard short-read data, resulting in a much higher signal-to-background. This approach is also able to identify structural variant breakpoints occurring in repetitive genomic regions, and can actually assemble the nucleotide sequences of these events. Finally, our results demonstrate that even very large (in this case, over 20 cm in length) tumors need not show substantial subclonal diversity, and that rather a series of extreme genomic rearrangements occurred early in tumor development.

VarMatch: A fast, parallel, and memory-efficient method for the variant matching problem

Chen Sun and Paul Medvedev

The Pennsylvania State University, USA

1 Introduction

Small variant ($\leq$ 30bp) calling is widely used in medical and genetic research to identify how genome mutations are related to phenotypes of interest. Variant matching is the problem of comparing different sets of variant calls, to determine the variants that are in common between the sets or unique to each set. Variant matching can be done to (1) compare the performance of different tools with respect to each other or with respect to a ground truth. (2) extract high-confidence variants for an individual by taking the intersection of calls from multiple callers, and (3) find variants that are shared or unique across different individuals.

A set of small variants is typically represented as a collection of VCF entries, where each entry contains a position of the reference genome and the alternate diploid allele (e.g. sequence) in the donor. The most straightforward variant matching algorithm is to directly match identical VCF entries. However, it can fail to match two different VCF entries that nevertheless result in the same diploid donor genome. Normalization and decomposition [1–3] have been used to alleviate these problems, however, there are still alternate representations for the same variant that are not matched [4]. An alternate approach is to formulate and solve an appropriate optimization problem that finds, roughly speaking, the largest number of matches [4]. This method can detect equivalent variants unmatched by heuristic algorithms, but still suffers from large memory usage.

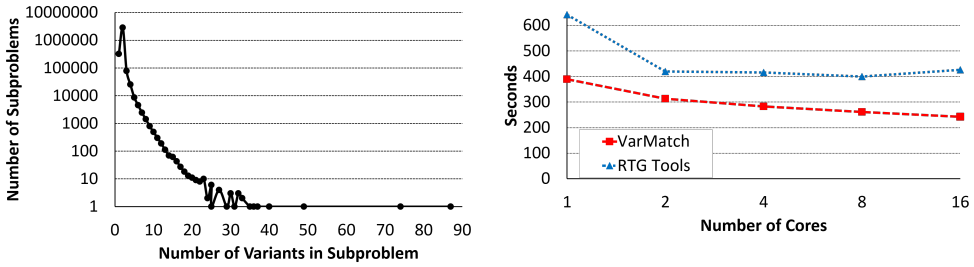
An additional limitation is that these approaches can only support maximizing the number of total matched VCF entries. However, this is sensitive to whether a tool represents complex variants as a single entry or as multiple, decomposed, entries. A more representation-invariable optimization criteria would be to maximize the number of matched nucleotides. In other cases, such as comparing multiple callers to a ground truth set, it is desirable to instead maximize the total number of matched entries from the ground truth set only.

2 Summary

To address these problems, we introduce a new algorithm VarMatch. VarMatch is an exact algorithm for variant matching that is guaranteed to find matching variants under a wide variety of optimization criteria. VarMatch employs a provably optimal divide and conquer strategy to partition the set of variants into disjoint subproblems. Because each subproblem is typically very small, we can use an exact dynamic programming algorithm similar to [4] (for the maximum number of matches optimization criteria) or even brute force (for other criteria) to solve each subproblem. While our algorithm has exponential running time in the worst case, we demonstrate it performs very fast in practice and uses an order of magnitude less memory than [4]. This can be crucial for applications in medical settings, where the software may be run on embedded processors or portable devices. VarMatch is also a parallel algorithm that scales over multiple processors and/or threads. Additionally, our divide and conquer strategy makes it easy to support any optimization criteria for doing matches, since even a brute force implementation is practical for the small subproblems. We have implemented several scoring functions in VarMatch: (1) maximize the total number of matched entries, (2) maximize the number of matched entries from one of the call sets, and (3) maximize the total number of matched bases. VarMatch is implemented as a user-friendly software package that will be available on GitHub if accepted.

3 Results

We consider the variant matching problem, roughly defined as follows: given a pair of variant sets $\langle \mathcal{V}, \mathcal{W} \rangle$, find subsets $V \in \mathcal{W}$ and $W \in \mathcal{W}$ such that applying V and W results in the same diploid sequence, and $f(V, W)$ is maximized. The function f can be almost any computable function, with the most natural



(a) (b)

Fig. 1: Effectiveness of problem partitioning (a) and parallelization (b) of VarMatch

one $f(V, W) = |V| + |W|$. If we consider a reference genome interval without any variants, we can split the input variants into those to the left and to the right of the interval. Our main theoretical result states that, given a sufficiently long and non-repetitive interval, the solution to the variant matching problem on $\langle \mathcal{V}, \mathcal{W} \rangle$ is equivalent to the union of the solutions to the problem on $\langle \mathcal{V}_{\text{left}}, \mathcal{W}_{\text{left}} \rangle$ and $\langle \mathcal{V}_{\text{right}}, \mathcal{W}_{\text{right}} \rangle$. This theorem is significant for two reasons. First, it leads to an exact, parallel, fast and low-memory divide-and-conquer algorithm that partitions the large problem into smaller subproblems which can be solved with a brute-force algorithm. Second, it allows the use of any reasonable optimization criteria, which other algorithms do not allow.

Table 1 illustrates evaluation of VarMatch on two published real data sets [2] with single thread, comparing the accuracy, memory usage(RAM) and running time of VarMatch to the normalization approach (based on Vt [1] followed by direct matching) and to RTG Tools [4]. For dataset CHM1 (Table 1a), we take variant call sets on the same sequencing data of the CHM1hTERT cell line. Variants were called separately by FreeBayes and HaplotypeCaller of GATK. For dataset NA12878 (Table 1b), we take variant call sets by Platypus and UnifiedGenotyper of GATK on NA12878 cell line. Both RTG Tools and VarMatch match more VCF entries than Vt at the cost of more resources, but VarMatch uses less running time and an order of magnitude less memory than RTG Tools.

Method	Matched Entries		RAM (Gb)	Time (s)
	FB	HC		
Vt	2,778,372	2,778,372	0.004	216
RTG Tools	2,843,004	2,911,802	48	642
VarMatch	2,843,004	2,911,802	4.7	389

(a) Dataset CHM1. Variants are called by Freebayes(FB) and HaplotypeCaller of GATK(HC).

Method	Matched Entries		RAM (Gb)	Time (s)
	PT	UG		
Vt	4,072,823	4,072,823	0.004	258
RTG Tools	4,228,302	4,414,044	34	836
VarMatch	4,228,302	4,414,044	5.5	704

(b) Dataset NA12878. Variants are called by Platypus(PT) and UnifiedGenotyper of GATK(UG).

Table 1: Comparison of VarMatch to two other variant matching methods on different datasets.

Figure 1a shows the effectiveness of our partitioning approach on dataset CHM1, VarMatch partitions 6,438,208 initial small variants into 3,272,206 subproblems, 99.9% of which have less than 9 variants in them. Figure 1b shows that on dataset CHM1 VarMatch scales with multiple threads, while with more threads I/O becomes the bottleneck($\sim$ 200 seconds).

References

1. Tan, A., Abecasis, G. R., and Kang, H. M. *Bioinformatics* , btv112 (2015).
2. Li, H. *Bioinformatics* **30**(20), 2841–2851 (2014).
3. Zook, J. M., Chapman, B., Wang, J., Mittelman, D., Hofmann, O., Hide, W., and Salit, M. *Nature biotechnology* **32**, 246–251 (2014).
4. Cleary, J. G., Braithwaite, R., Gaastra, K., Hilbush, B. S., Inglis, S., Irvine, S. A., Jackson, A., Littin, R., Rathod, M., Ware, D., et al. *bioRxiv* , 023754 (2015).

Low memory, fast, specific, sensitive, multi-reference sequence classification using Bloom filter maps

Justin Chu, Sarah Yeo, Ben Vandervalk, Golnaz Jahesh, Hamid Mohamadi, Chen Yang, Shaun Jackman, Rene Warren, Inanc Birol
Canada's Michael Smith Genome Sciences Centre, British Columbia Cancer Agency, Vancouver, BC V5Z 4S6, Canada. Contact: cjustin@bcgsc.ca

Background

Sequence classification is traditionally performed by alignments of sequencing reads onto a reference sequence set. Although alignment methodologies have the potential to map the location of these reads precisely, this information is not a prerequisite for classification and thus perform more computation than is needed. Hash table based methods provide fast access for classification but require a large amount of memory. To address these shortcomings, we previously proposed an efficient classification method, BioBloom Tools (Chu *et al.* 2014), that uses a low memory, probabilistic set membership query data structure called a Bloom filter.

Using a Bloom filter, elements of a sequence, such as k-mers, are queried to determine whether they are or not members of those decomposed from the reference set. The memory and time benefits of the data structure have spurned the development of classification tools such as FACs (Stranneheim *et. al.* 2010) and BioBloom Tools. However, querying for the set of origin between multiple reference sets requires the construction and usage of multiple Bloom filters leading to O(n) time complexity when querying, where n is the numbers of reference sets/filters. Here we present a new Bloom filter based data structure called a Bloom Map that can act as an associative array, storing and querying the identifier to the set of origin of a specific element in O(1) time .

Results

Conceptually, a Bloom map is a simply a Bloom filter with buckets that are larger than a single bit. In our implementation we first construct a normal Bloom filter by hashing our elements and set the corresponding buckets in the bit vector to 1. We then interleave rank information into the bit vector. Then, we fill an ID array the size of the population count of the filled bloom filter. Finally, we hash the elements again, but this time setting elements in the ID array with the corresponding IDs of the reference sets according to their rank in the bloom filter. This saves memory by effectively reducing the space of each empty bucket to a single bit. To query we check the bit vector and then use the rank information to look up the ID array for the identity of the queried element in O(1) time.

We compared our tool against a metagenomic classification tool called Kraken (Wood & Salzburg 2014), and its spaced seed counterpart Seed-Kraken (Brinda, *et. al.* 2015) on the NCBI bacterial database. Though our tool is designed for general purpose classification, we correctly classified the genus of 97% reads compared to to Kraken's 92% and Seed-Kraken 95%, while utilizing less memory (61 GB RSS + 0GB pre-cache) than both Kraken (73GB RSS + 66GB pre-cache) and Seed-Kraken (71GB RSS + 64GB pre-cache) on a read simulated dataset. The simulated dataset was constructed using dwgsim and on the genomes used in the bacterial database mimicking ~1mil 2x150bp Illumina reads. To investigate specificity we introduced 50282 random sequence reads into our simulated read set; Seed-Kraken incorrectly assigned a single random read to a genus, but both Kraken and our method managed to not

assign any random sequences to a genus. On 8 cores our tool took <2 minutes to run on our simulated 1mil bp dataset.

Hash collisions are dealt with by using logic such as a majority hit rule in addition to assigning heavily colliding reference IDs a mutual collision ID. Other features of our implementation is the utilization of a recursive rolling hash called ntHash for speed, as well as using complementary spaced seeds patterns instead of the traditional use of multiple hash functions to improve both sensitivity and specificity. Unlike Kraken our k-mer/seed sizes do not affect the memory of our method, which gives BBT the potential to reach a higher specificity by using longer seed k-mer/seed sizes.

Conclusions

Sequence classification to a set of known reference sequences has many applications in contamination screening, pathogen detection, metagenomics, and preprocessing for targeted assembly from shotgun sequence data. Here we present an efficient low memory alternative to hash tables for general purpose, multi-reference sequence classification with broad applications, including taxonomic characterization of bio-organisms from metagenomics samples.

RNA-Bloom: *de novo* RNA-seq assembly with Bloom filters

Ka Ming Nip[1,2], Justin Chu[1,2], Inanç Birol[1,3]
[1]Canada's Michael Smith Genome Sciences Centre, BC Cancer Agency, Vancouver, BC, Canada
[2]Bioinformatics Graduate Program, the University of British Columbia, Vancouver, BC, Canada
[3]Department of Medical Genetics, the University of British Columbia, Vancouver, BC, Canada

RNA-seq is primarily used in measuring gene expression, quantification of transcript abundance, and building reference transcriptomes. Without bias from a reference sequence, *de novo* RNA-seq assembly is particularly useful for building new reference transcriptomes, detecting fusion genes, and discovering novel transcripts. A number of approaches for de novo RNA-seq assembly were developed over the past six years, including Trans-ABySS, Trinity, Oases, IDBA-tran, and SOAPdenovo-Trans. Using 12 CPUs, it takes approximately a day to assemble a human RNA-seq sample and require up to 100GB of memory. While the high memory usage may be alleviated by distributed computing, access to a high-performance computing environment is a strict requirement for RNA-seq assembly.

Here, we present a novel de novo RNA-seq assembler, "RNA-Bloom," that utilizes Bloom filter-based data structures for compact storage of k-mer counts and the de Bruijn graph of two k-mer sizes in memory. Compared to existing approaches, RNA-Bloom can assemble a human RNA-seq sample with comparable accuracy using merely 10GB of memory, which is readily available on modern desktop computers. The de Bruijn graph of two k-mer sizes allows RNA-Bloom to effectively assemble both lowly and highly expressed transcripts. In addition, RNA-Bloom can assemble and quantify transcript isoforms without alignment of sequence reads, thus resulting in a quicker run-time than existing alignment-based protocols.