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.

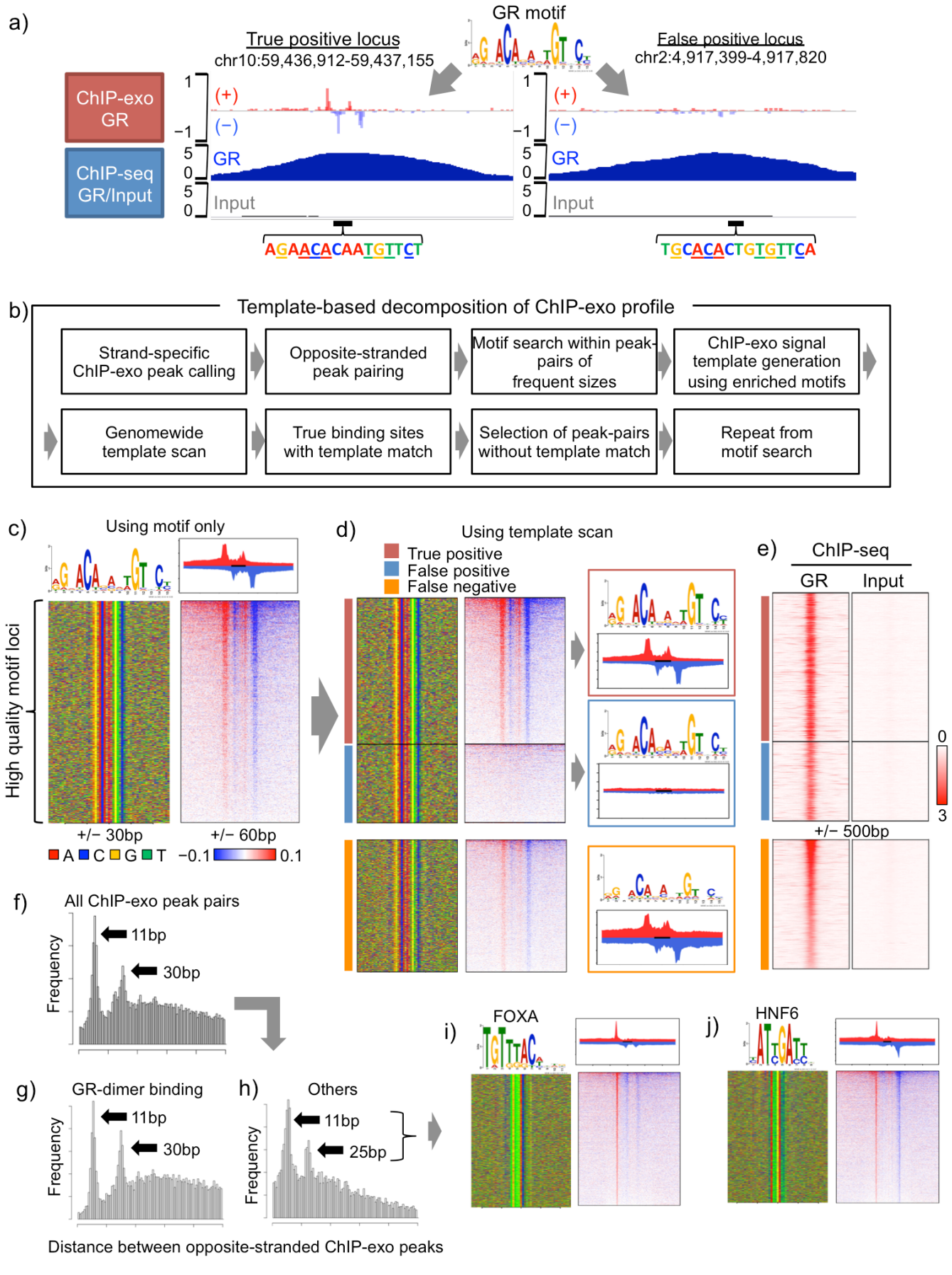
Then we selected ChIP-exo peak-pairs that do not match with the GRE ChIP-exo template (**Figure1f-h**) for the next round analysis. From the second motif search, we obtained HNF6 and FOXA motifs, which were not significantly enriched at the first round. Subsequent template scans revealed thousands of 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.

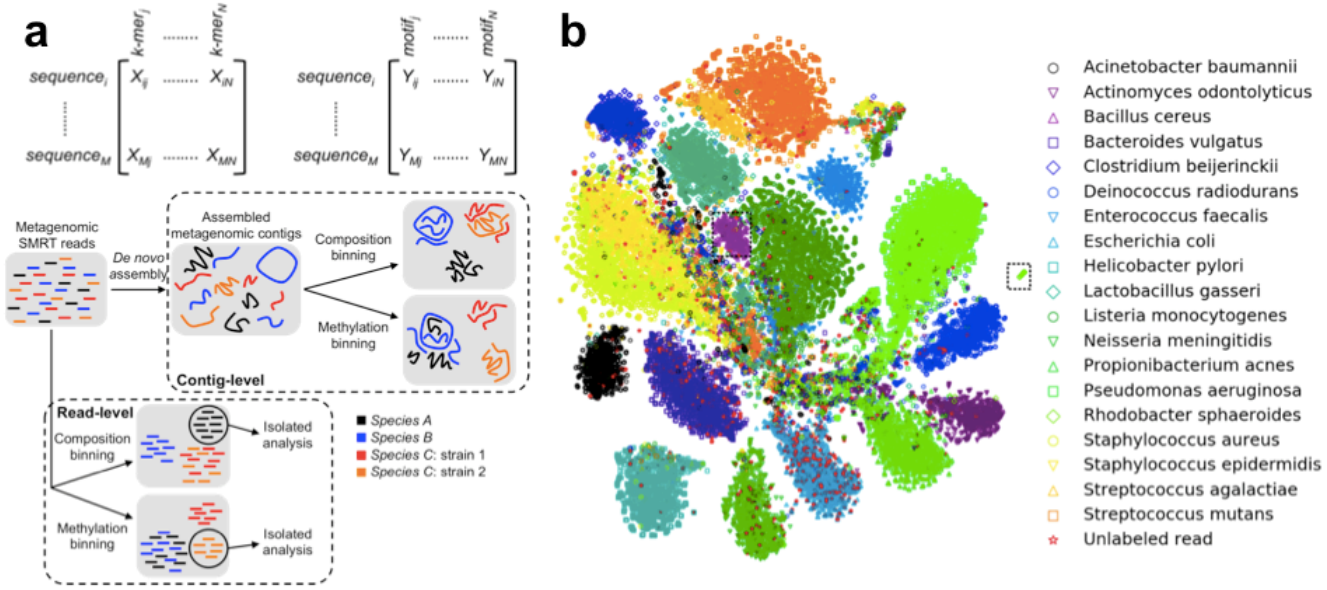


Figure 1: (a) Schematic of metagenomic SMRT binning approaches, where composition-based binning helps separate divergent species, while methylation-based binning helps untangle highly similar genomes. The two binning approaches can also be combined to leverage the strengths of each. (b) Composition-based 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.

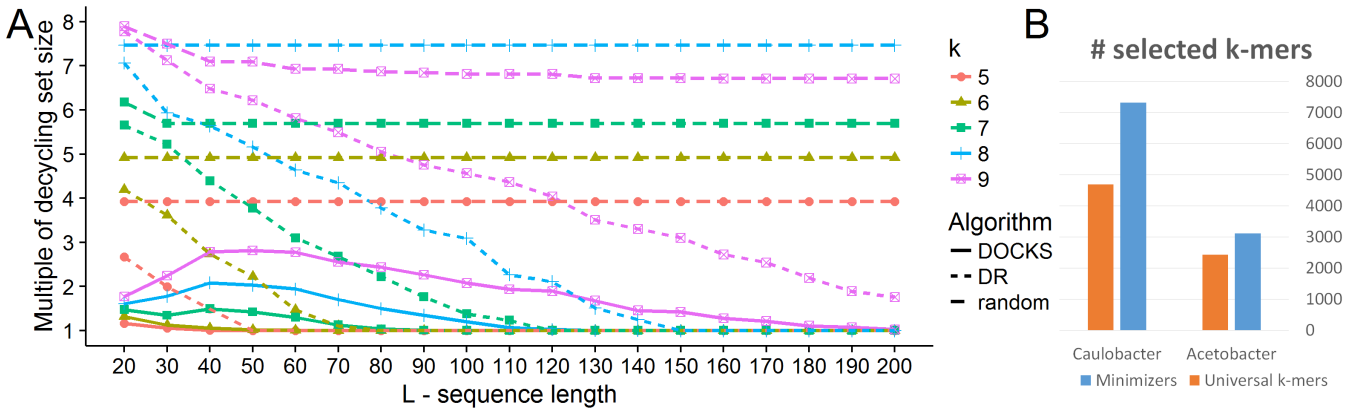


Fig. 1. Performance of DOCKS. A) For different combinations of k and L we ran DOCKS and two random procedures over the DNA alphabet. The results are shown in comparison to the size of the decycling set. When the ratio is 1, all the sequences avoiding the decycling set were of length shorter than L. DR: 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

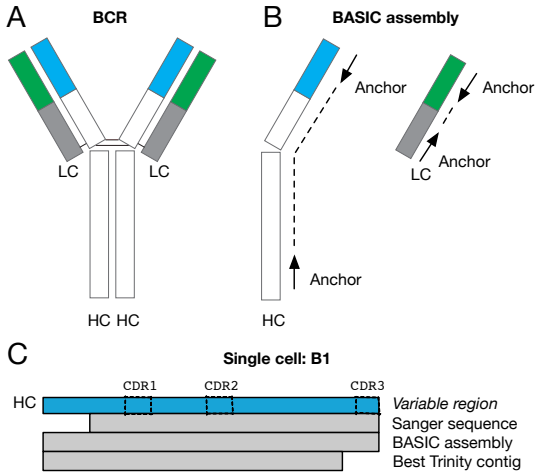


Fig. 1. A) The BCR is a large 'Y' shaped protein complex composed of two identical heavy chains (HC) and two identical light chains (LC). The variable regions are colored blue and green in the paired chains. The complementarity determining regions (CDRs) are those parts of the variable regions that participate 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.