High Throughput Sequencing
Algorithms and Applications
A special track of the ISMB/ECCB 2019 meeting
Basel, Switzerland, July 22-23, 2019

ISMB-ECCB 2019 HiTSeq Track Proceedings

Basel, Switzerland
July 22-23, 2019
http://www.hitseq.org

Organizers:

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

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

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

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

Gang Fang
Mount Sinai School of Medicine, New York, NY, USA
E-mail: fanggang@gmail.com
Kjong Lehmann
ETH-Zürich, Zürich, Switzerland
E-mail: kjong.lehmann@inf.ethz.ch

Layla Oesper
Carleton College, Northfield, MN, United States E-mail: loesper@carleton.edu

Gunnar Rätsch
ETH-Zürich, Zürich, Switzerland
E-mail: raetsch@inf.ethz.ch

Beta-binomial modeling of CRISPR pooled screen data identi es target genes with greater sensitivity and fewer false negatives

Hyun-Hwan Jeong[1], Seon Young Kim1, Maxime W.C. Rousseaux[2], Huda Y. Zoghbi[3] and Zhandong Liu[1]

[1] Baylor College of Medicine
[2] University of Ottawa
[3] Howard Hughes Medical Institute

1 Background

The simplicity and cost-e ectiveness of CRISPR (Clustered Regularly Interspaced Short Palindromic Re-peats) technology have made high-throughput pooled screening approaches accessible to virtually any lab. Analyzing the large sequencing data derived from these studies, however, still demands considerable bioinfor-matics expertise. Various methods have been developed to lessen this requirement, but there are still three tasks for accurate CRISPR screen analysis that involve bioinformatic know-how if not prowess: designing a proper statistical hypothesis test for robust target identi cation, developing an accurate mapping algorithm to quantify sgRNA levels, and minimizing the parameters necessary that need to be ne-tuned.

2 Results

We have developed a new algorithm, called CRISPRBetaBinomial or CB^2 (https://CRAN.R-project. org/package=CB2). In CB^2 , we adapted a beta-binomial model [Baggerly et al., 2003] with a modi ed Students t-test to measure di erences in single-guide RNA (sgRNA) levels, followed by Fishers combined probability test [Fisher, 1925] to estimate the gene-level signi cance. We compared CB^2 with eight state-of-the-art methods (HiTSelect, MAGeCK, PBNPA, PinAPL-Py, RIGER, RSA, ScreenBEAM, and sgRSEA) on benchmark datasets [Evers et al., 2016, Sanson et al., 2018] evaluating gene essentiality using di erent technologies: CRISPRn (CRISPR nuclease gene knockout via Cas9) and CRISPRi (CRISPRinterference, a CRISPR/Cas9 system with a catalytically inactive Cas9 fused to the transcriptional repressor KRAB which results in gene repression). Based on the beta-binomial distribution, which is better suited to sgRNA data, CB^2 outperformed all other methods at every FDR cut-o level, and all other methods lost their detection powers at more rigorous FDRs (Figure 1). In other words, all methods demonstrated a small type-I error due to the strong lethality phenotype of the CRISPR assay, but CB^2 demonstrated a signi cantly lower type-II error than the other methods. Across all paradigms tested with di erent FDR cut-o s, CB^2 performed the best, with a much larger F1-score and recall. CB^2 also accommodates staggered sgRNA sequences, and it provides more accurate alignment then other alignment methods without parameter tuning using an adaptive hash-mapping algorithm. In conjunction with CRISPRcloud framework (http://crispr.nrihub. org), CB^2 will bring CRISPR screen analysis within reach for a wider community of researchers.

3 Conclusion

The advent of CRISPR/Cas9 systems heralded a new era of large-scale screening approaches. Over the past three years, there has been tremendous growth in the number of pooled genetic screens. The number of datasets for CRISPR/Cas9 screens in Gene Expression Omnibus have more than tripled each year (39 datasets in 2015, 121 datasets in 2016, and 408 datasets in 2017). Much of this has been due to the widespread availability of large-scale genome-wide perturbation libraries via the non-pro t repository Addgene (https: //www.addgene.org/) and resource sharing between labs. However, the computational burden of CRISPR pooled screen data analysis has not been so amenable to a cheap and a widely accessible solution. In this study, we provide a novel algorithm { CB^2 { that provides a powerful and robust analysis of the data and does not require heavy computation for both hypothesis test and sgRNA abundance quanti cation. Moreover, to the best of our knowledge, CB^2 outperforms other CRISPR/Cas9-screen analysis programs and will accelerate discovering novel biological ndings from CRISPR pooled screens with the cooperation of CRISPRcloud.

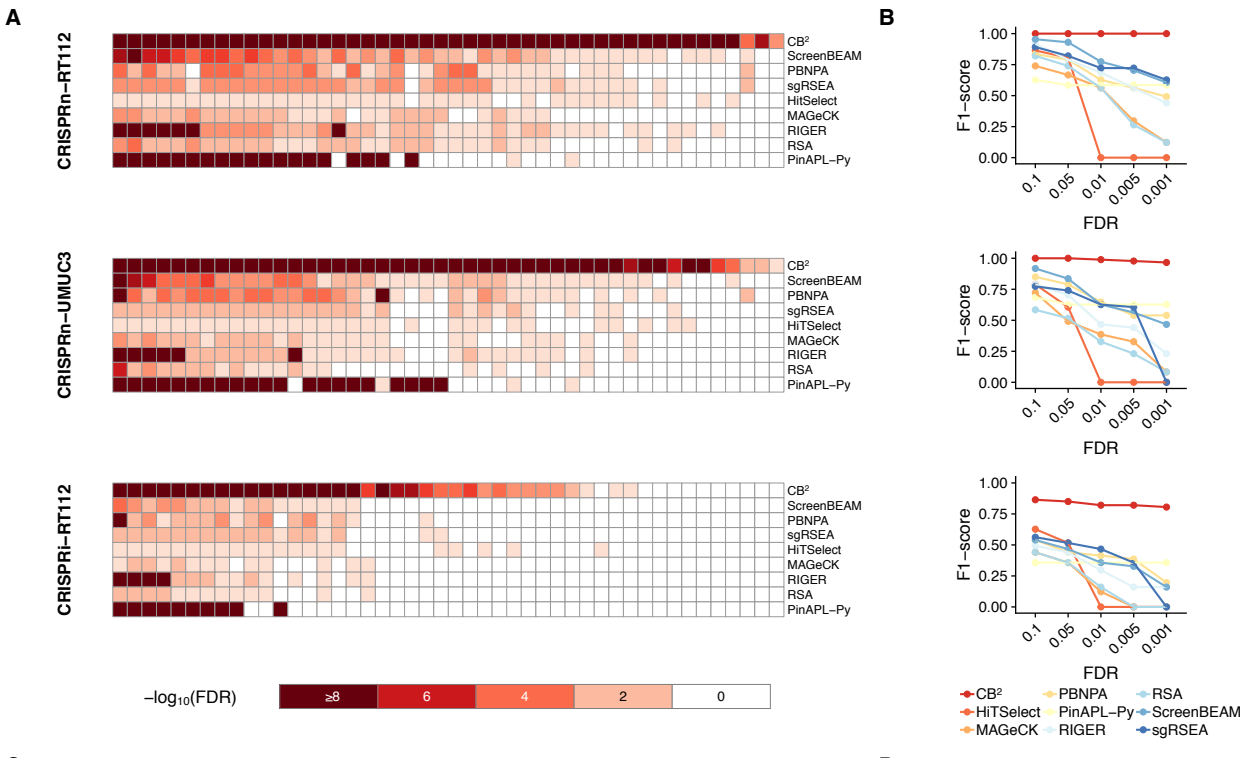


Figure 1: CB^2 offers robust target identification with high precision and recall. Benchmark results using data from [Evers et al., 2016] (A) Heatmaps illustrate FDRs of gene statistics from each of nine leading high-complexity pooled screen analysis tools. (B) F1-score measurements at different FDR cut-offs across all methods. At commonly used FDR cut-offs, CB2 can identify most of the essential genes with high rates of precision and recall.

References

K. A. Baggerly, L. Deng, J. S. Morris, and C. M. Aldaz. Differential expression in sage: accounting for normal between-library variation. *Bioinformatics*, 19(12):1477–1483, 2003.

B. Evers, K. Jastrzebski, J. P. Heijmans, W. Grernrum, R. L. Beijersbergen, and R. Bernards. Crispr knockout screening outperforms shrna and crispri in identifying essential genes. *Nature biotechnology*, 34 (6):631, 2016.

R. A. Fisher. Theory of statistical estimation. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 22, pages 700–725. Cambridge University Press, 1925.

K. R. Sanson, R. E. Hanna, M. Hegde, K. F. Donovan, C. Strand, M. E. Sullender, E. W. Vaimberg, A. Goodale, D. E. Root, F. Piccioni, et al. Optimized libraries for crispr-cas9 genetic screens with multiple modalities. *Nature communications*, 9(1):5416, 2018.

Subpopulation detection and their comparative analysis across single cell experiments with PopCorn

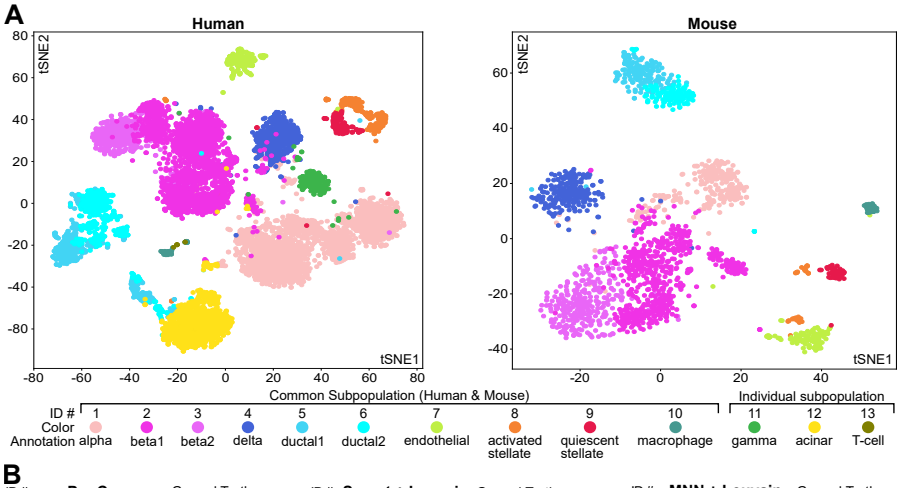
Yijie Wang, Jan Hoinka and Teresa Przytycka
National Center of Biotechnology Information, National Library of Medicine, NIH

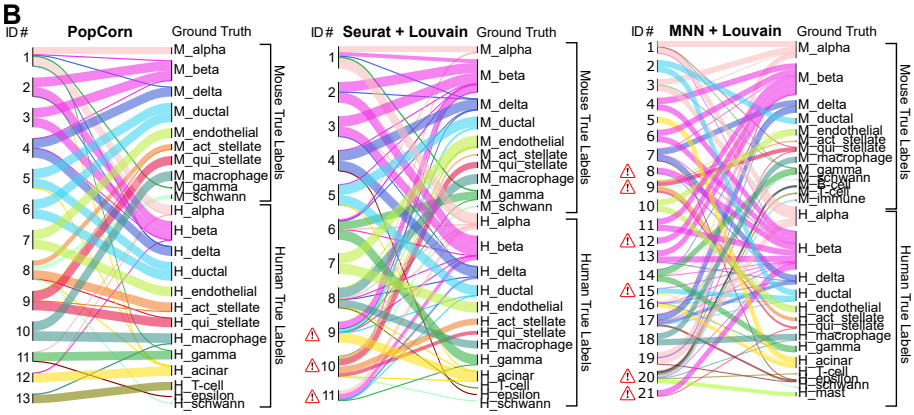
One of the key applications of scRNA-seq technology concerns the identifica-tion of subpopulations of cells present in a sample, and comparing such subpopula-tions across multiple samples/experiments. This conceptually natural task, is com-plicated by technical and biological noise which can obscure the true biological similarities and differences between the samples. To address this need, we intro-duce a computational method, PopCorn (single cell sub-Populations Comparison). Leveraging the information from all input data sets, PopCorn performs these two tasks simultaneously by optimizing a joint objective function.

First, any pair of cells from two different experiments/samples are connected by an edge that measures their similarity. Next, if two cells are in the same exper-iment, a more complex relation called subpopulation co-membership propensity graph, is estimated. Informally, subpopulations should be defined in such a way that cells in the same subpopulation should all be similar to each other while si-multaneously distinguishing themselves from other cells. To strike the balance, PopCorn uses the idea of Googles personalized PageRank. For each individual, this method measures personal preferences (ranking) towards specific pages on the World Wide Web which are recorded in a personalised PageRank vector. Here, we substitute the network of web pages by a cell-to-cell expression similarity graph and for each cell we estimate its preference (a “vote”) of which other cells should be included in the same subpopulation with itself. This voting takes into account the above mentioned criterion that the expression pattern of cells within a subpop-ulation should be consistent but distinct from the expression of the cells outside the subpopulation. The resulting preferences of all cells are then combined to define the edge weights of the subpopulation co-membership propensity graph. Thus, this graph takes into account not only the relation of each pair of cells to themselves but also their relation to the rest of the cells in the same experiment.

Finally, both types of edges are combined into one graph which is then par-titioned into connected subgraphs that define both: the subpopulations and their mapping across the experiments.

We tested the performance of PopCorn in three distinct settings. First, we demonstrated its potential in identifying and aligning subpopulations from single cell data from human and mouse pancreatic single cell data (?). Next, we applied PopCorn to the task of aligning biological replicates of mouse kidney single cell data (?). PopCorn achieved the best performance over the previously published tools. Finally, we applied it to comparing populations of cells from cancer and healthy brain tissues reveling the relation of neoplastic cells to neural cells and astrocytes.





C **Comparison between all competing methods**

	Rccsp	Rucct	Ruict	Acc_H	Acc_M	Acc_Avg
PopCorn	**10/10**	**8/10**	**2/7**	**93.4%**	**92.2%**	**92.8%**
Seurat+Louvain	8/10	7/10	0/7	79.2%	85.8%	82.5%
MNN+Louvain	19/21	7/10	1/7	82.9%	74.9%	78.9%

Figure 1: Comparison of performance of PopCorn, Seurat + Louvain, and MNN + Louvain methods on human and mouse pancreatic cell data (A) Two t-SNE plots for human and mouse scRNA-seq data sets, respectively. Colors indicate different cell annotations which are determined from literature provided labels. Cells of identical color denote subpopulations identified by PopCorn. (B) Sankey diagrams of the resulting mapping between identified subpopulations (left) to ground truth labels (right) for PopCorn, Seurat, and MNN. The width of the flow bar is proportional to the purity score (see STAR methods). Incorrectly identified and mapped subpopulations are annotated by exclamation marks adjacent to the ID number. (C) Comparison of PopCorn, Seurat, and MNN on several metrics (see STAR methods for a detailed definition).

Detection and assembly of novel sequence insertions using Linked-Reads

Dmitry Meleshko[1,2], Patrick Marks[3], Stephen Williams[3], and Iman Hajirasouliha[2,4,*]

[*]Corresponding author

Availability: Software is freely available at https://github.com/1dayac/novel_insertions
Contact: imh2003@med.cornell.edu

As a result of efforts in advancing DNA sequencing technologies and related algorithm developments, the field of personal genomics has been revolutionized in the past decade. Leveraging next-generation sequencing technologies, whole genome sequencing (WGS) has shown unprecedented promise in characterizing variants among human genomes. However, current methods are still unable to assemble a large fraction of structural variants due to limitations of short-reads in resolving repetitive regions of the genome effectively (Chaisson *et al.* (2015b); Huddleston and Eichler (2016)).

Long-read sequencing technologies have recently become commercially available. These techniques promise the ability to call structural variants and improve de novo assembly (Chaisson *et al.* (2015a); Sedlazeck *et al.* (2017)). While these technologies offer much longer reads than traditional short-read technologies, their base-pair error rates are higher than Illumina short reads (10-15% vs. 0.3% error) (Koren *et al.* (2012)). Additionally, long-read technologies have still much higher costs and often require high amount of DNA. This makes long-reads impractical for large-scale screenings of whole genome samples.

Low-cost and high-accurate Linked-Read technologies have emerged recently to improve the ability of standard short-read sequencing in determining whole genomes. In Linked-Read sequencing, DNA molecules are sheared into long fragments, and barcoded short reads from these long fragments are produced in such a way that reads from a long fragment share the same barcode. Most recently, Linked-Reads proved to be useful in multiple applications including but not limited to genome assembly (Weisenfeld *et al.* (2017)), genome phasing (Zheng *et al.* (2016)), or large-scale somatic SV detection (Spies *et al.* (2017)). The contribution of Linked-Reads to SV detection is, however, still limited to large structural variations. In particular, virtually none of the available SV detection algorithms attempt to characterize mid-size novel insertions (as small as only 300 bp and up to a few thousand bp in size).

In the past, several approaches attempted novel sequence insertion detection using standard short-read whole genome sequencing data. All algorithms are based on the idea of assembling reads that are not aligned on the reference genome and connecting these assembled sequences with potential insertion breakpoints on the reference genome. All these approaches are, however, limited to conventional paired-end sequencing data and it turns out to be problematic to correctly locate insertions in the repetitive regions of the genome, because the size of anchors is limited by the small fragment size. Our main objective here is to develop a novel technique that can leverage barcodes and long fragment information encoded in Linked-Read sequencing to achieve much longer anchors.

We introduce an integrated mapping-based and assembly-based method, which is significantly more accurate than existing short-read methods for novel insertion discovery. While our method is less efficient in run-time that existing short-read methods, it is indeed more efficient compared to the recent Linked-Read algorithms that use whole-genome *de novo* assembly such as (Wong *et al.* (2018)) because it uses only a very small fraction of informative Linked-Reads. Our Linked-Read method is able to characterize one of the most challenging classes of SVs with a reasonable additional cost to standard short-read sequencing.

Our method is based on a novel idea that the barcode information encoded in Linked-Reads can be used to reconstruct **long anchors** that can be unambiguously placed on the reference genome. This allows finding exact break-point positions on the reference even in repeat regions. Our approach is based on the local assembly of multiple barcodes originated from the same genomic loci. An overview of our technique is also shown in Figure 1.

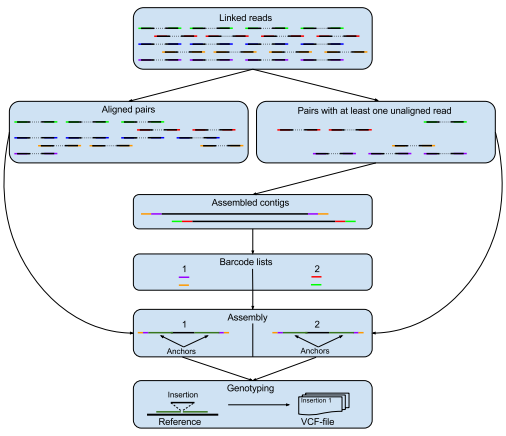


Figure 1: An overview of the steps in the Novel-X method is shown.

The input for Novel-X is a reference genome and a BAM-file. We refer to the set of the reads from the input BAM-file as *original reads* and to the BAM-file itself as *original BAM*. A pre-processing step in Novel-X is the extraction of paired-end reads from the original BAM that cannot be aligned to the reference genomes or have poor alignments. Intuitively, novel insertion sequences should consist of reads that do not align anywhere on the reference. We choose paired-end reads in which at least one end is not aligned to the reference genome, or has the mapping quality below 10, or has more than 20% of soft-clipped bases. For simplicity, we collectively call this set of unaligned reads as $\mathcal{U}$. Reads from $\mathcal{U}$ correspond to novel sequence insertions and anchor sequences.

We first use the Velvet *de novo* assembler (Zerbino (2010)) to assemble $\mathcal{U}$. Ideally, the resulting assembly contigs would belong to sequences of novel insertions. If needed, we can perform a contamination removal procedure similar to what was previously done in NovelSeq (Hajirasouliha *et al.* (2010)). i.e. perform

a BLAST search against nt/nr database and filter out all contigs that align to non-human references. We call the remaining contigs *orphan contigs*.

For each orphan contig c, we first align the reads from $\mathcal{U}$ to c and filter read alignments with low-quality scores or with a large fraction (>20%) of soft-clipped or hard-clipped sequences. Let $R(c)$ be the set of filtered barcoded-reads aligned to c. We denote $B(c)$ as the set of all barcodes in $R(c)$. We extract and store every read from the set of original reads whose barcode is in $B(c)$. The information about barcodes of remaining reads is, however, extracted and aggregated separately for each orphan contig.

In order to reconstruct anchors and automatically connect them to novel sequences, for each barcode list we search the original BAM for reads that have a barcode from the barcode list and extract them. Then, we reassemble each set of extracted reads separately. While we understand that different long fragments from different places can share identical barcodes, only regions of our interest would receive enough sequence coverage and can be assembled into contiguous sequences.

The last step of the pipeline is the detection of the positions on the reference genome where the novel insertions took place. We use Minimap2 for aligning the resulting assemblies to the reference genome. Alignments that are adjacent with respect to the reference genome are analyzed for insertion signatures. We keep only insertions longer than 300 bp with at least one anchor exceeding 300 bp to prevent false calls. All found insertion are stored in a vcf (Variant Call Format) file.

In order to evaluate the utility of Novel-X on real data, we performed experiments with a CHM1 dataset. The dataset is available at
https://support.10xgenomics.com/de-novo-assembly/datasets/2.0.0/chm.

In summary, for the CHM1 dataset, Novel-X identified 314 insertions longer than 300 bp. To compare results for different novel insertion callers we compared WGS novel sequence insertion callers with the calls obtained with SMRT-SV algorithm (see Table 1). While Pamir finds more insertions than Novel-X and PopIns, it tends to call shorter insertions. Novel-X finds more insertion of size greater than 500 bp compared to other callers. These findings are consistent with our theory because longer insertion sequences recruit more barcodes than short ones and their assembly will more likely produce long anchors. Another encouraging validation of our method is that about 80% of Novel-X calls overlap with SMRT-SV calls while the amount of agreement with Pamir and PopIns is below 45%. Indeed Pamir and PopIns are more likely to produce false positive calls.

Novel insertions detection results for insertions longer than 300 bp				
Length (bp)	SMRT-SV	Novel-X	Pamir (Kavak *et al.* (2017))	PopIns (Kehr *et al.* (2016))
300-499	1919	97 (78, **80%**)	324 (**121**, 36%)	156 (25, 16%)
500-999	1144	139 (**115**, **83%**)	76 (51, 67%)	151 (33, 22%)
1000-1999	598	56 (**49**, **88%**)	5 (4, 80%)	89 (13, 15%)
≥2000	608	22 (**12**, 55%)	2 (2, **100%**)	21 (6, 29%)
Total(≥300)	4269	314 (**254**, **81%**)	407 (178, 44%)	417 (77, 18%)

Table 1: Length breakdown and comparison between SMRT-SV, Pamir and PopIns, and Novel-X for CHM1. The numbers in brackets indicate the count of overlaps with SMRT-SV calls and the percentage of the overlapping calls.

Finally, we ran our method on the well-known CEPH/HapMap NA12878 diploid genome and compared it with a recently de novo assembly based method Wong *et al.* (2018) and SMRT-SV calls (see Table 2). For the NA12878 sample, Novel-X found 219 novel sequence insertions with mean length 778 bp and a total length of 170 kbp.

Novel insertions detection results for insertions longer than 300 bp			
Length (bp)	SMRT-SV	Novel-X	NUI
300-499	2661	73 (**50**, 68%)	9 (8, **89%**)
500-999	1462	112 (**83**, 74%)	17 (16, **94%**)
1000-1999	1072	28 (**14**, 50%)	2 (2, **100%**)
≥2000	1228	6 (**2**, 33%)	3 (**2**, **67%**)
Total(≥300)	6423	219 (**149**, 68%)	31 (28, **90%**)

Table 2: Length breakdown and comparison between the PacBio based tool, SMRT-SV, Linked-Read method NUI-pipeline and our Linked-Read method Novel-X on NA12878. The numbers in brackets indicate the count of overlaps with SMRT-SV calls.

Funding DM is supported by the Tri-Institutional Training Program in Computational Biology and Medicine (via NIH training grant 1T32GM083937). This work was also supported by start-up funds (Weill Cornell Medicine) and a US National Science Foundation (NSF) grant under award number IIS-1840275 to IH.

Conflict of Interest IH and DM have none to declare. PM and SW are employees of 10x Genomics.

References

Chaisson, M. J. P. *et al.* (2015a). Genetic variation and the de novo assembly of human genomes. *Nat Rev Genet*.

Chaisson, M. J. P. *et al.* (2015b). Resolving the complexity of the human genome using single-molecule sequencing. *Nature*, **517**, 608–611.

Hajirasouliha, I. *et al.* (2010). Detection and characterization of novel sequence insertions using paired-end next-generation sequencing. *Bioinformatics*, **26**(10), 1277–1283.

Huddleston, J. and Eichler, E. E. (2016). An incomplete understanding of human genetic variation. *Genetics*, **202**(4), 1251–1254.

Kavak, P. *et al.* (2017). Discovery and genotyping of novel sequence insertions in many sequenced individuals. *Bioinformatics*, **33**(14), i161–i169.

Kehr, B. *et al.* (2016). Popins: population-scale detection of novel sequence insertions. *Bioinformatics*, **32**(7), 961–967.

Koren, S. *et al.* (2012). Hybrid error correction and de novo assembly of single-molecule sequencing reads. *Nat Biotechnol*, **30**(7), 693–700.

Sedlazeck, F. J. *et al.* (2017). Accurate detection of complex structural variations using single molecule sequencing. *bioRxiv*.

Spies, N. *et al.* (2017). Genome-wide reconstruction of complex structural variants using read clouds. *Nature Methods*, **14**, 915 EP –.

Weisenfeld, N. I. *et al.* (2017). Direct determination of diploid genome sequences. *Genome research*, **27**(5), 757–767.

Wong, K. H. *et al.* (2018). De novo human genome assemblies reveal spectrum of alternative haplotypes in diverse populations. *Nature communications*, **9**(1), 3040.

Zerbino, D. R. (2010). Using the velvet de novo assembler for short-read sequencing technologies. *Curr Protoc Bioinformatics*, **Chapter 11**, Unit 11.5.

Zheng, G. X. Y. *et al.* (2016). Haplotyping germline and cancer genomes with high-throughput linked-read sequencing. *Nature Biotechnology*, **34**, 303 EP –.

Bspliced: a Bayesian hierarchical model for differential splicing accounting for sample-to-sample variability and mapping uncertainty

Simone Tiberi and Mark D Robinson

Institute of Molecular Life Sciences and SIB, University of Zurich

Alternative splicing plays a fundamental role in the biodiversity of proteins as it allows a single gene to generate several transcripts and, hence, to code for multiple proteins. However, variations in splicing patterns can be involved in diseases. When investigating differential splicing (DS) between conditions, typically healthy vs disease, scientists are increasingly focusing on differential transcript usage (DTU), i.e. in changes in the proportion of transcripts.

A big challenge in DTU analyses is that, unlike gene level studies, the counts at the transcript level, which are of primary interest, are not observed because most reads map to multiple transcripts. Tools such as *salmon* or *kallisto* allow, via expectation maximization (EM) algorithms, to estimate the expected number of fragments originating from each transcript. Most DTU methods (e.g., *DRIMSeq*, *BayesDRIMSeq* and *SUPPA2*) follow a *plug-in* approach and take the estimated counts as input by treating them as real transcript counts, thus neglecting the uncertainty in the estimates. In order to overcome this issue, some methods, such as *cjBitSeq* and *casper*, consider what transcripts each read is compatible with (also called equivalence class); nevertheless, none of these tools allows for sample-specific proportions (i.e., they assume all samples to share the same transcript relative abundance).

To overcome the limitations of current methods for DTU, we present Bspliced, an R package* to perform DTU based on RNA-seq data. Bspliced uses a Bayesian hierarchical model, with a Dirichlet-multinomial structure, to explicitly model the variability between samples. Our tool inputs the equivalence class of each read, by treating the allocations of reads to the transcripts as latent variables. When a read is compatible with more than one gene, the gene allocation is also treated as a latent variable. The parameters of the model are inferred via Markov chain Monte Carlo (MCMC) techniques where, via a data augmentation procedure, we alternately sample the Dirichlet-multinomial parameters and the latent variables.

To ensure that the MCMC posterior chains have converged, we assess the stationarity of the full log-posterior density via Heidelberg and Welch's convergence diagnostic. Despite the computational complexity of full MCMC algorithms, the core of our method is coded in C++, which makes Bspliced highly efficient and feasible to run on a laptop, even for complex model organisms.

In order to test for DTU, at both transcript and gene level, we approximate the posterior densities of the parameters by a multivariate normal distribution and apply a multivariate Wald test. Our method tests for DTU at both transcript and gene level, allowing scientists to investigate what specific transcripts are differentially used in selected genes. Furthermore, our tool is not limited to two group comparisons and also allows to test for DTU when samples belong to more than two groups.

We will show how, both in simulation studies and experimental data analyses, the proposed methodology outperforms existing methods (e.g., see Figure 1).

*We will submit Bspliced to Bioconductor in April 2019.

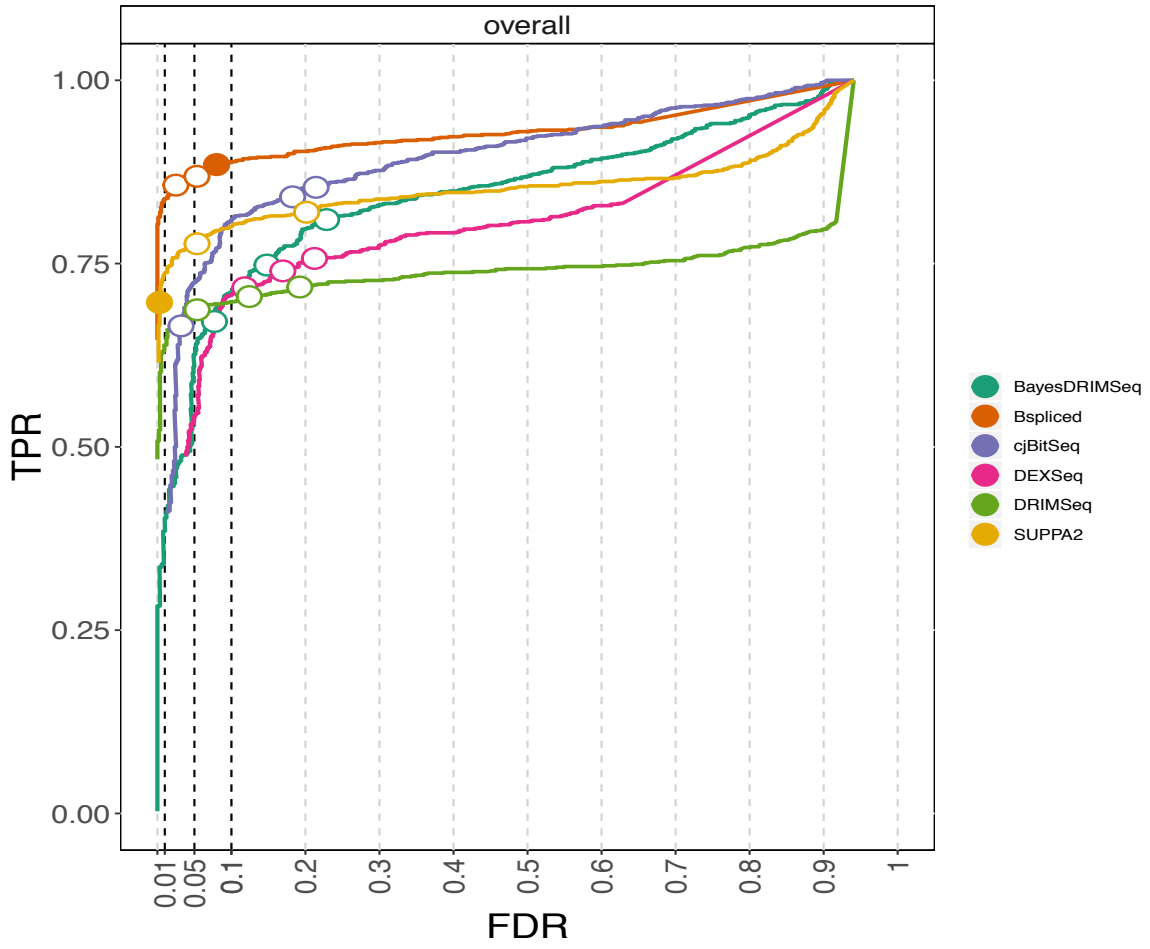


Figure 1: True positive rate (TPR) vs. false discovery rate (FDR) computed for several methods for DTU in a 6 vs 6 RNA-seq simulation study from a human genome. For any given FDR threshold, Bspliced provides a significantly higher TPR than any other method considered. We obtained similar results in all simulation and experimental data analyses we performed.

Accurate determination of node and arc multiplicities in de Bruijn graphs using conditional random fields

Aranka Steyaert, Pieter Audenaert, Jan Fostier

Background De Bruijn graphs are used in many bioinformatics tools, e.g. genome assembly, read correction and variant detection, because they efficiently represent the overlap between sequences. The nodes correspond to k-mers of the sequences, while arcs represent $k+1$-mers such that the first (resp. last) k nucleotides coincide with those of the node that the arc points from (resp. points to) [1]. In the read-based de Bruijn graph that results from a sequencing experiment, each node (resp. arc) has a *coverage* corresponding to the number of times its k-mer (resp. $k+1$-mer) is present in the read-set. This read-based graph is a noisy representation of the (unknown) genomic sequence underlying the reads. We call the number of times a k-mer (resp. $k+1$-mer) is present in the genomic sequence the *multiplicity* of the corresponding node (resp. arc). The multiplicity of a node/arc is reflected in its coverage, however, coverage variability and coverage biases complicate the identification of the true multiplicities. To accurately infer the underlying genome de Bruijn graph from a read-based graph, we want to label all nodes and arcs with a multiplicity: for true nodes/arcs this multiplicity corresponds to their repeat copy-number, while erroneous nodes/arcs should have multiplicity zero.

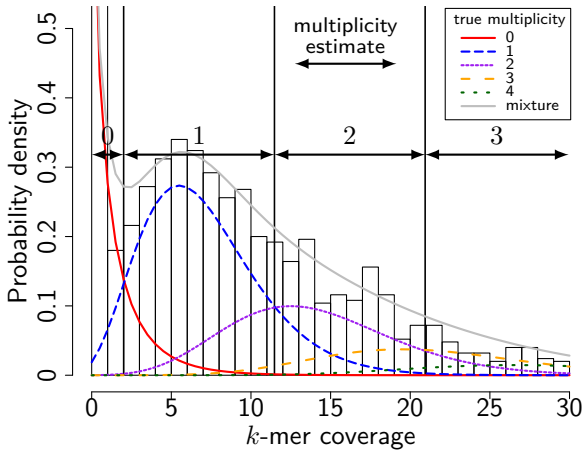


Figure 1: Mixture of negative binomial distributions fitted to a k-mer histogram. The distinct components and derived multiplicity intervals are shown.

Current methodology to infer node and arc multiplicities often uses a k-mer histogram approach. A k-mer histogram is obtained by counting at each k-mer coverage present in the data, the number of k-mers that occurs with that coverage. A mixture of distributions is fitted to this histogram, such that each component corresponds to a multiplicity (see Figure 1). To infer multiplicities, hard intervals of coverage are selected based on the mixture model. However, because the components that constitute the mixture model overlap, inference based on a k-mer histogram alone is error-prone. Nodes and arcs that have a coverage near the interval boundaries can be assigned an erroneous multiplicity. This, in turn, leads to erroneous conclusions in the applications that use a de Bruijn graph, such as the deletion of low-coverage true k-mers that results in a more fragmented assembly [2].

Erroneous multiplicity assignments can be identified by using a known property of genome de Bruijn graphs: at each node in the graph, the multiplicity of that node has to equal the sum of the multiplicities of the incoming arcs and the sum of the multiplicities of the outgoing arcs [3]. This property, which we call 'conservation of flow of multiplicity', also holds in the presence of sequencing errors when the spurious nodes/arcs are assigned multiplicity zero.

We present a statistical model that incorporates coverage of individual nodes/arcs as well as the conservation of flow property, to label the nodes/arcs of a de Bruijn graph with multiplicities. We believe that this methodology can be a useful addition to the many bioinformatics tools that make use of de Bruijn graphs.

Results We build a conditional random field (CRF) model that incorporates both local evidence (coverage) and evidence present in a neighbourhood surrounding a node or arc (conservation of flow), in a high-dimensional statistical model. This model is used to assign the most likely multiplicity to nodes/arcs in a de Bruijn graph. The model contains the following variables: the unknown multiplicities, $\mathbf{Y} = \{Y_1, \ldots, Y_N\}$, which we need to infer, and the coverages of the arcs, $\mathbf{X} = \{X_1, \ldots, X_N\}$, which are observed. We model the conditional distribution $P(\mathbf{Y}|\mathbf{X})$ as a product of factors φ such that $P(\mathbf{Y}|\mathbf{X}) = \frac{1}{Z(\mathbf{X})} \prod_{i=1}^{M} \varphi_i(\mathbf{D_i})$, $(\mathbf{D_i} \subseteq \mathbf{X} \cup \mathbf{Y}, \mathbf{D_i} \nsubseteq \mathbf{X})$ [4]. We use two types of factors. The first type, $\varphi(Y, X)$, relates the coverage of a

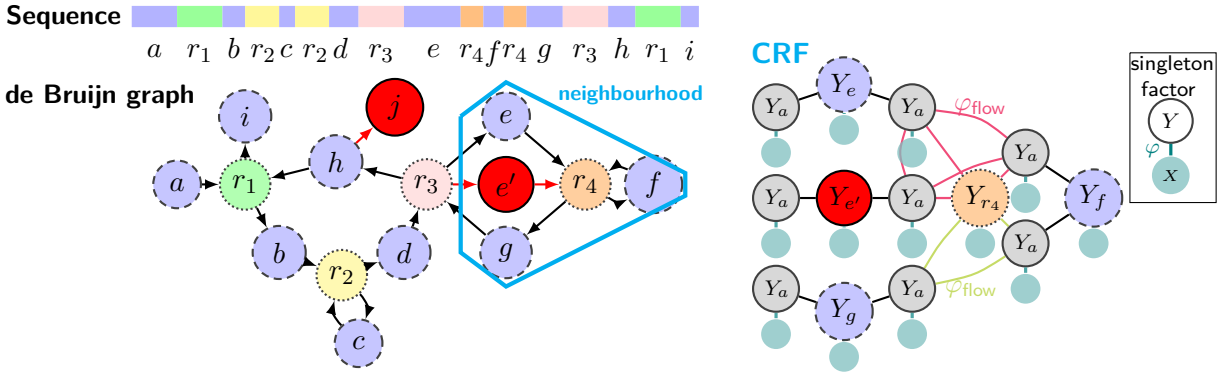


Figure 2: Small genome sequence with repeats with corresponding de Bruijn graph. The red nodes in the de Bruijn graph correspond to sequencing errors. For a neighbourhood of size 1 around r_4 the CRF is shown. Nodes Y_n correspond to nodes n in the de Bruijn graph and nodes Y_a correspond to the arcs. Each CRF node Y is connected to a CRF node X through a φ factor that represents the coverage information. All connections between Y-nodes arise from conservation of flow factors, their corresponding cliques are shown for the φ_{flow}-factors of node Y_{r_4}.

node/arc to its expected multiplicity based on a mixture model fitted to a k-mer histogram. A second type of factor $\varphi_{\text{flow}}(Y_n, \{Y_a\}_{a\in\text{in}(n)})$ / $\varphi_{\text{flow}}(Y_n, \{Y_a\}_{a\in\text{out}(n)})$ implies a conservation of flow. Its arguments are the multiplicity variable for a node n and multiplicity variables for all its incoming/outgoing arcs a. This factor assigns a high value when the conservation of flow property holds and a low value otherwise. The conditional random field model provides us with a graph representation of the probability distribution. Herein, nodes represent variables and arcs connect variables such that a clique is formed whenever variables co-occur in a factor. This graph representation allows us to rely on graph algorithms to perform inference more efficiently. As exact inference methods are computationally intensive, we estimate the multiplicity labels for each node and arc in the de Bruijn graph based on a CRF built for a selected neighbourhood around that node/arc. A neighbourhood of size s around node n then contains all nodes reachable by a path of length $\leq s$ from n and all their incoming and outgoing arcs (see Figure 2 for an example).

		P. aeruginosa		H. sapiens chr. 21	
coverage		accuracy	iters	accuracy	iters
15×	cut-off	87.77	25	62.20	25
	CRF	96.16	8	67.36	18
30×	cut-off	73.25	25	69.47	25
	CRF	98.54	5	76.49	11
75×	cut-off	98.02	15		
	CRF	99.19	4		

Table 1: Accuracy and number of EM-iterations needed to label nodes in a de Bruijn graph with a multiplicity using a k-mer histogram cut-off or a CRF model. Datasets were downsampled to different levels of coverage. H. sapiens chr. 21 has coverage below $50\times$, hence, no results are shown for $75\times$ coverage.

We use the CRF in an EM-setting and alternatingly determine model parameter estimates and multiplicity labels. Table 1 shows results for a P. aeruginosa dataset (ERR330008) and a H. sapiens chr. 21 dataset (Illumina data library). For both datasets a well-characterised reference sequence is available (ref. ID ERR330008, resp. HG19) [5], which we used to determine the true multiplicities in the de Bruijn graph. This allows us to determine the accuracy as the proportion correctly assigned labels to the nodes in a de Bruijn graph. We notice a higher accuracy as well as a need for fewer EM-iterations when using the CRF model.

Conclusions By using a CRF to incorporate contextual information in addition to the k-mer histogram for single nodes, we consistently obtain better accuracy when labelling nodes in a de Bruijn graph with their multiplicity. With more efficient inexact inference techniques we will be able to use larger neighbourhoods, thus increasing the accuracy even further. This framework can be valuable in bioinformatics tools that rely on a de Bruijn graph.

References

[1] P. A. Pevzner et al., "An Eulerian path approach to DNA fragment assembly," *PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES OF THE UNITED STATES OF AMERICA*, vol. 98 (17), pp. 9748–9753, aug 2001.

[2] M. Heydari et al., "Evaluation of the impact of Illumina error correction tools on de novo genome assembly," *BMC BIOINFORMATICS*, vol. 18, aug 2017.

[3] P. A. Pevzner and H. Tang, "Fragment assembly with double-barreled data," *Bioinformatics*, vol. 17 (suppl 1), pp. S225–S233, 2001.

[4] D. Koller and N. Friedman, *Probabilistic Graphical Models: Principles and Techniques*, Massachusetts Institute of Technology, 2009.

[5] P. Greenfield et al., "Blue: correcting sequencing errors using consensus and context," *BIOINFORMATICS*, vol. 30 (19), pp. 2723–2732, oct 2014.

Characterizing chromatin landscape from aggregate and single-cell genomic assays using flexible duration modeling.

Mariano I. Gabitto[1, †], Anders Rasmussen[1], Orly Wapinski[2,3,4], Kathryn Allaway[2,3,4], Nicholas Carriero[1,5], Gordon J. Fishell [2,3,4], Richard Bonneau[1,6,7, †]

1. Center for Computational Biology, Flatiron Institute, Simons Foundation, New York, NY, 10010, USA
2. New York University, Neuroscience Institute and the Department of Neuroscience and Physiology, Smilow Research Center, NY, NY 10016, USA
3. Department of Neurobiology, Harvard Medical School, Boston, MA 02115
4. Stanley Center at the Broad, Cambridge, MA 02142.
5. Scientific Computing Core, Flatiron Institute, Simons Foundation, New York, NY, 10010, USA
6. New York University, Center for Data Science, NY, NY, 10010
7. New York University, Department of Biology, NY, NY, 10012, USA.
[†.] Co-corresponding authors. Correspondence to: mgabitto@flatironinstitute.org and rb133@nyu.edu

Abstract:

ATAC-seq has become a leading technology for probing the chromatin landscape of single and aggregated cells. Distilling functional regions from ATAC-seq and other similar genomic technologies presents diverse analysis challenges, due to the relative sparseness of the data produced and the interaction of complex noise with multiple chromatin structure scales. Methods commonly used to analyze chromatin accessibility datasets are adapted from algorithms designed to process different experimental technologies, disregarding the statistical and biological differences intrinsic to the ATAC-seq technology. Here, we present a Bayesian statistical approach that uses Hidden Semi-Markov models to better model the duration of functional and accessible regions, termed ChromA. We demonstrate the method on multiple genomic technologies, with a focus on ATAC-seq data. ChromA annotates the cellular epigenetic landscape by integrating information from replicates, producing a consensus de-noised annotation of chromatin accessibility. ChromA can analyze single cell ATAC-seq data, improving cell type identification and correcting many biases generated by the sparse sampling inherent in single cell technologies. We validate ChromA on multiple technologies and biological systems,

including mouse and human immune cells and find it effective at recovering accessible chromatin, establishing ChromA as a top performing general platform for mapping the chromatin landscape in different cellular populations from diverse experimental designs.

We will also discuss new work, not present in the early preprint provided (via link) below to extend this model to CRISPR screens, single cell cut&run, DNA methylation and other genomic technologies aimed at chromatin state and function.

An early draft of this work (currently under review) is available at:

https://www.biorxiv.org/content/10.1101/567669v1

Software availability

A Python implementation of ChromA is available for download on GitHub: http://github.com/marianogabitto/ChromA . The website will be updated periodically with new versions.

PipelineOlympics: Benchmarking of processing workflows for bisulfite sequencing data

Toth Reka[1], Yassen Assenov, Karl Nordström, Angelika Merkel, Edahi Gonzalez-Avalos, Matthias Bieg, Stephen Kraemer, Murat Iskar, Helene Kretzmer, Lelia Wagner, Lilian Leiter, Giuseppe Petroccino, Anand Mayakonda, Kersten Breuer, Gideon Zipprich, Lena Weiser, Philip Kensche, Renata Jurkowska, Christian Lawerenz, Ivo Buchhalter, Steve Hoffmann, Simon Heath, Marc Zapatka, Joern Walter, Matthias Schlesner, Christoph Bock, Christoph Plass and Pavlo Lutsik

[1] Division of Epigenomics, German Cancer Research Center (DKFZ)

Background

DNA methylomes are widely used as epigenomic blueprints in basic life science research, but also hold promise as clinically relevant biomarkers [1]. Whole genome bisulfite sequencing (WGBS) is a state-of-the-art method for the genome-scale assessment of DNA methylation levels, with a growing number of alternative protocols, such as tagmentation based (T-) WGBS and PBAT, widely used for bulk, low-input and, more recently, single-cell analysis. However, many experimental factors might affect the final methylation calls, and therefore basic data processing – read trimming, alignment and site-wise estimation of DNA methylation levels – is crucial for downstream analysis, such as identification of differentially methylated regions. Despite the high relevance, there is little consensus as to which of the numerous software tools and workflows guarantee optimal data processing results, as benchmarking studies are limited and scarce. The main reason is the lack of adequate real-world reference datasets, allowing an unambiguous evaluation of pipeline performance. Furthermore, the complexity of pipelines and heterogeneous data influx make it hard to implement the benchmarking results as a universal data processing solution for all types of applications. *PipelineOlympics* is a collaborative effort of 15 leading European and American labs to comprehensively benchmark bisulfite sequencing software, aiming to provide ultimate data processing guidelines for each of the popular wet-lab protocols.

Results

PipelineOlympics is schematically depicted in **Figure 1**. To obtain a unique reference dataset, we applied several WGBS protocols to four specimens well characterized in a previous benchmarking study of targeted DNA methylation assays, that resulted in highly accurate DNA methylation measurements at 47 selected genomic regions [2]. Overall, up to 1 TB of primary sequencing data was generated. In the currently ongoing pilot phase of the project, we utilized this gold-standard data set to compare the performance of 10 different workflows, developed and routinely used by the participants, on two protocols (standard WGBS with Illumina X-ten and, T-WGBS on HiSeq 4000). Tested workflows covered a broad selection of read trimming tools (*Trimmomatic*, *cutadapt*, *skewer*), aligners (*bwa-meth*, *bwa-mem*, *GSNAP*, *bowtie*, *GEM*, *segemehl*, *bsmap*), post-alignment filtering tools (*Picard*, *deduplicator*, *sambamba*), and methylation callers (*MethylDackel*, *methylCtools*, *BisSNP*, *meth-caller*, *bscall*, *BAT*, *methratio*). As a rule, each workflow was executed by its respective developers to ensure optimized parameter settings.

We evaluated the workflows twofold: (i) by relating the obtained calls to the gold standard methylation measurements of selected loci, and (ii) through cross-workflow comparison to

draw genome-wide inference. To assess the workflow accuracy on the gold standard set, we developed a set of novel metrics and visualization approaches. Our major performance metric was *cumulative absolute distance from the gold-standard corridor.* Moreover, in order to establish the final rank of the pipelines, their errors were weighted by the mappability of the regions and samples.

The majority of pipelines showed high accuracy, with an on-par performance of the top-scorers, and several less accurate outliers. As expected, wet-lab protocol and sequencing depth had a major impact upon the accuracy, with standard bulk protocol (WGBS) resulting in significantly better calls then the low-input protocol (T-WGBS). The correlations between pipelines were above 0.95 for X-ten WGBS, but reached levels as low as 0.88 in the case of T-WGBS. Nonetheless, higher sequencing depth did not always translate into higher accuracy, neither for standard WGBS, nor for T-WGBS. Both methods resulted in higher error rates in case of the tumor samples. The main reason for this can be the lower stability of the tumor genome; copy number alterations and somatic mutations might disrupt the methylation calling. For T-WGBS the average absolute distances per pipelines can be as high as 0.1 in tumors and only up to 0.06 in normal.

Using the pilot evaluation results as a primer, in the second phase of the project a systematic mapping of the combinatorial workflow space is currently being carried out. Briefly, the already tested and novel pipelines wrapped using the CWL specification [3] are executed in a uniform compute cloud environment that would also allow to asses computational performance. In addition, we are working on infrastructural solutions (R package, containers) to transform the present study into a persistent and extensible benchmarking resource.

Conclusions

The pilot phase of the *PipelineOlympics* project allowed us to evaluate bisulfite processing workflows based on real biological data along with highly accurate gold standard reference measurements. To our knowledge, this is the first WGBS benchmarking study of such kind. It is expected to give a significantly better feedback of the workflow performance as compared to benchmarks using simulated data, or those with inadequate reference data (e.g. methylation arrays). Our ongoing, systematic benchmarking is going to close remaining unaddressed gaps and pave the way to optimal processing pipelines for currently used and future WGBS protocols. Finally, the design of our study will allow the constant expansion of the set of included pipelines and will aid scientists in navigating through the ample data analysis options.

References:

1. Smith Z. and Meissner A., *Nat Rev Genet*, 2013
2. BLUEPRINT Consortium, *Nat. Biotech.*, 2016.
3. Common Workflow Language, v1.0. Specification, *CWL working group*. https://w3id.org/cwl/v1.0/

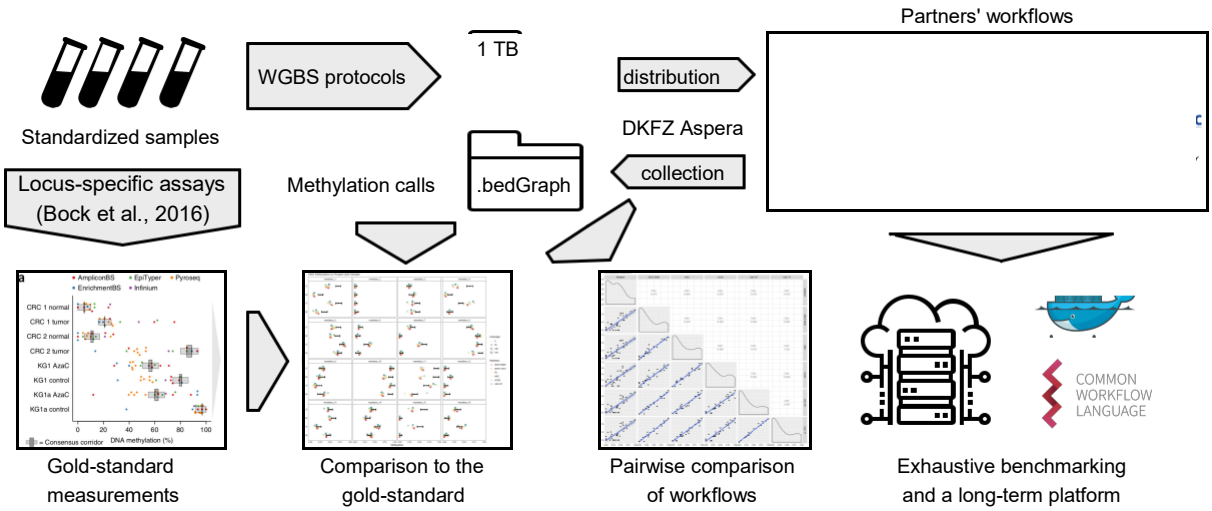


Figure 1. Scheme of *PipelineOlympics* activities.

Free vectorized icons downloaded from flaticon.com.

Characterization of large-scale structural variants using Linked-Reads

Fatih Karaoglanoglu , Camir Ricketts , Ezgi Ebren , Marzieh Eslami Rasekh ,
Iman Hajirasouliha * and Can Alkan *

Alterations of DNA content and organization larger than 50 bp, commonly referred to as genomic structural variations (SVs) [2], are among the major drivers of evolution [6], and diseases of genomic origin [9]. Despite decades of research they remain difficult to accurately characterize contributing to our lack of full understanding of the etiology of complex diseases.

Recently a Linked-Read sequencing method called the 10x Genomics system (10xG) was introduced as an alternative method to generate highly accurate Illumina short reads data with additional long-range information [7]. The ability of extracting long range information from accurate and inexpensive but short read sequencing data makes Linked-Read sequencing attractive for various applications. It has been used for genome scaffolding [10], haplotype-aware assembly [7], metagenomics [3], single cell transcriptome profiling [8] and regulatory network clustering [1], haplotype phasing [7], and genome structural variation discovery [4, 5]. Despite the advances in SV discovery using various technologies, detecting complex SV such as balanced rearrangements (i.e., inversions and translocations), and segmental duplications (SDs) remains challenging due to mapping ambiguity. Currently no Linked-Read based method exists to *anchor* a new SD (i.e. find their insertion locations).

Here we present **novel algorithms** to discover deletions, inversions, translocations, and large (>40 Kbp) direct and inverted interspersed SDs using Linked-Read sequencing data. We redesign and extend upon VALOR [4] and use split molecule and read pair signatures (Figure 1) to detect SDs and estimate the insertion sites of the new SD paralogs, and further include read depth signature to filter potential false positives caused by incorrect mappings. We implemented our new algorithms as the VALOR_2 software package. Briefly, VALOR_2 differs from the former version of VALOR through: 1) it can characterize segmental duplications, translocations, and deletions, 2) it incorporates read depth information to improve predictions and reduce false calls, and 3) it provides full support to alignment files (i.e., BAM) generated from 10xG Linked-Read data sets.

Using simulated data sets we show that VALOR_2 achieves high precision and recall (94% and 82%, respectively) for segmental duplications, 98% and 76% for large inversions, and 93% and 71% for translocations. We also applied VALOR_2 to the genomes of NA12878, and a Yoruban trio in addition to a haploid genome (CHM1) sequenced with the 10xG platform [7]. Of the several tools we tested, VALOR_2 had the largest number of validated inversions in the NA12878 genome while predicting the second lowest number of total inversions (only LUMPY, which only called 7 inversions, has fewer). This result further highlights the superior precision and recall of VALOR_2. Additionally VALOR_2 was very useful in identifying large scale duplications by exploiting Linked-Read information in the NA12878 sequencing data. We predicted multiple direct segmental duplications and inverted duplications with chromosomes 1 and 16 containing both classes of duplications.

Funding This work was supported by a grant by TÜBİTAK (215E172) and an EMBO Installation Grant (IG-2521) to C.A. This work was also supported by start-up funds (Weill Cornell Medicine) and a National Science Foundation (NSF) grant under award number IIS-1840275 to I.H. C.R. received support from the Tri-Institutional Training Program in Computational Biology and Medicine (via NIH training grant 1T32GM083937).
Availability: VALOR_2 source code is available at https://github.com/BilkentCompGen/valor, and a Docker image is available at https://hub.docker.com/r/alkanlab/valor

References

[1] Sara Aibar, *et al.* SCENIC: single-cell regulatory network inference and clustering. *Nature methods*, 14:1083–1086, November 2017.

[2] Can Alkan, *et al.* Genome structural variation discovery and genotyping. *Nat Rev Genet*, 12(5):363–376, May 2011.

[3] David C. Danko, *et al.* Minerva: an alignment and reference free approach to deconvolve linked-reads for metagenomics. *Genome Res.*, 2018.

[4] Marzieh Eslami Rasekh, *et al.* Discovery of large genomic inversions using long range information. *BMC Genomics*, 18:65, January 2017.

[5] Patrick Marks, *et al.* Resolving the full spectrum of human genome variation using linked-reads. *BioRxiv*, p. 230946, 2017.

[6] Tomas Marques-Bonet, *et al.* A burst of segmental duplications in the genome of the African great ape ancestor. *Nature*, 457(7231):877–881, Feb 2009.

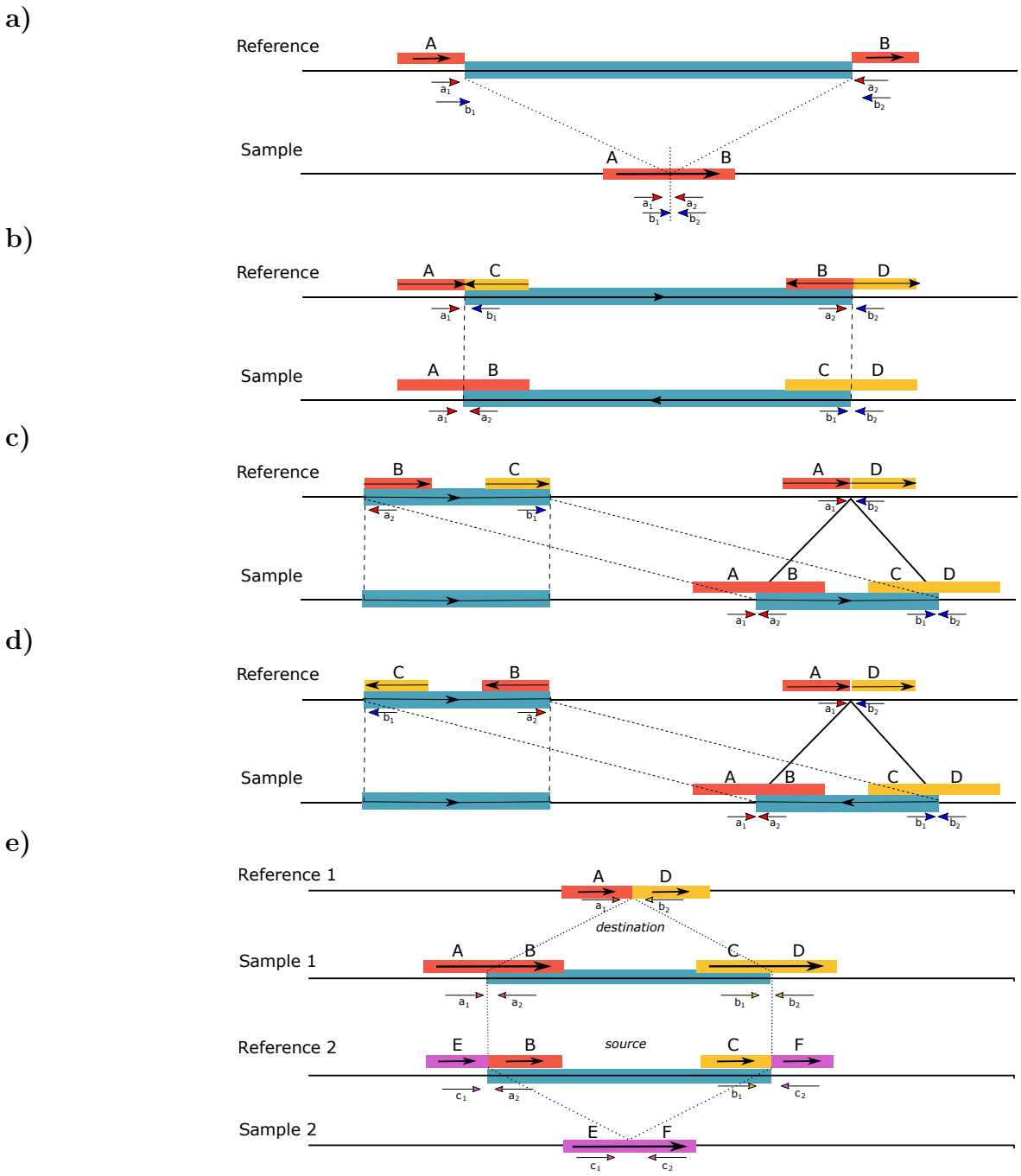


Figure 1: Split molecule and read pair sequence signatures used in VALOR$_2$. a) Deletion, b) inversion, c) interspersed duplication in direct orientation, d) inverted duplication, e) translocation. In each case, the large molecules that span the SV breakpoints are split into two mapped regions. Note that, it is not possible to determine the mapped strand of the split molecules shown here. In (e), section including B and C is moved to between A and D. We do not show inverted translocations here for simplicity. From the perspective of the reference genome (i.e., mapping), A,B,C,D,E,F are defined as *submolecules*, A/B, C/D, and E/F pairs are *candidate splits*, and A/B-C/D quadruple is a *split molecule pair*.

[7] Yulia Mostovoy, *et al.* A hybrid approach for de novo human genome sequence assembly and phasing. *Nature methods*, 13:587–590, July 2016.

[8] Daniel A Skelly, *et al.* Single-cell transcriptional profiling reveals cellular diversity and intercommunication in the mouse heart. *Cell reports*, 22:600–610, January 2018.

[9] Paweł Stankiewicz and James R. Lupski. Structural variation in the human genome and its role in disease. *Annu Rev Med*, 61:437–455, 2010.

[10] Sarah Yeo, *et al.* ARCS: scaffolding genome drafts with linked reads. *Bioinformatics*, 34:725–731, March 2018.

Bayesian deconvolution of somatic clones and pooled individuals with expressed variants in single-cell RNA-seq data

Yuanhua Huang1, Davis McCarthy[1], Raghd Rostom[2], Sarah Teichmann[2] and Oliver Stegle[1]

[1] EMBL-European Bioinformatics Institute

[2] Wellcome Sanger Institute

Extended abstract

Decoding the clonal substructures of somatic tissues sheds light on cell growth, development and differentiation in health, ageing and disease. DNA-sequencing, either using bulk or using single-cell assays, has enabled the reconstruction of clonal trees from frequency and co-occurrence patterns of somatic variants. However, approaches to systematically characterize phenotypic and functional variations between individual clones are not established. Here we present cardelino (https://github.com/PMBio/cardelino), a Bayesian method for inferring the clonal tree configuration and the identity of individual cells from single-cell RNA-seq (scRNA-seq) data with a Gibbs sampler (see full preprint [1]). Briefly, cardelino models the expressed variant alleles in single cells as a clustering method, with clusters corresponding to somatic clones with (unknown) mutation states (Fig. 1A). Critically, cardelino can integrate a guide clonal tree configuration derived from external data, e.g., bulk or single-cell DNA sequencing data, as scRNA-seq data alone is usually very sparse and additional data helps.

Initially, we assess the accuracy of cardelino using simulated data that mimics typical real data. By default, we used a guide clone configuration with 10% errors compared to the simulation truth. Alongside assessing the performance of cardelino, we compare the results with two alternative approaches: SCG [2] without using the guide clone configuration and Demuxlet [3] instead assuming the guide configuration fully correct. In the default setting, cardelino achieves high overall performance (Precision- Recall AUC=0.947; Fig. 1B), remarkably outperforming SCG and Demuxlet.

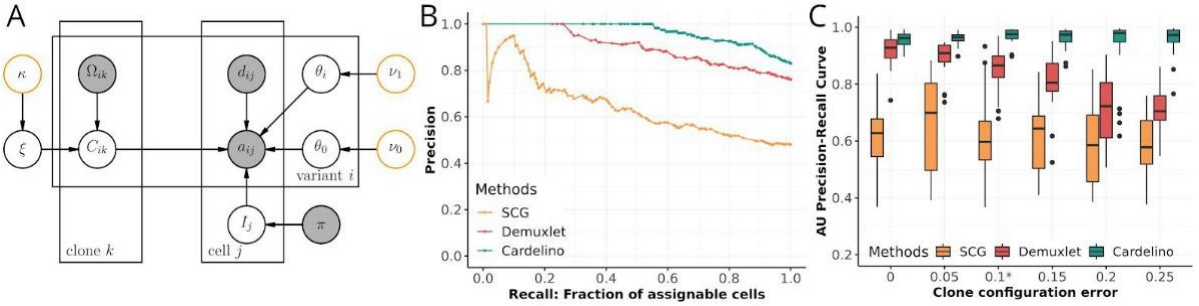


Figure 1 | Cardelino model for somatic clone inference with scRNA-seq. (a) Graphical representation of the *cardelino* model. The clonal tree configuration matrix C is an unknown variable and follows a Bernoulli prior distribution encoded by an guide tree configuration Ω and an error rate ξ. The clone configuration C and cell identity I together encode the genotype c of each variant i in each cell j. The alternative allelic read count a out of total depths d follows a binomial distribution with parameter $\theta_{i,Ij}$ if $c_{i,Ij} = 1$, otherwise θ_0. Shaded nodes represent observed variables; unshaded nodes represent unknown variables; yellow circled nodes represent fixed hyper parameters. **(b)** The precision-recall curves for three methods on a simulated data set with 10% error in the guide clone configuration. The simulation setting follows typical real observations: 4 clones, 10 variants per branch, 25% of variants with read coverage, 200 cells, 20 repeat experiments **(c)** The area under the precision-recall curve when varying the error rate in the guide clone configuration.

We further explore the effect of a variety of key dataset characteristics on cell assignment (Fig 1 in [1]), especially the error rate in the guide clone configuration. Fig. 1C here shows that Demuxlet suffers when there is a high error rate in the guide clone configuration, while cardelino is robust to such errors and keeps excellent performance (AUPRC>0.97) even with 25% error rate in the guide clone configuration, thanks to its ability to identify and correct these errors.

After validating our model using simulations, we apply cardelino to matched scRNA-seq and bulk exome sequencing data from 32 human dermal fibroblast lines, identifying hundreds of differentially expressed genes between cells from different somatic clones (Fig 3 in [1]). These genes are frequently enriched for cell cycle and proliferation pathways, indicating a key role for cell division genes in non-neutral somatic evolution (Fig 4 in [1]).

A similar problem to the clone reconstruction is donor deconvolution, i.e., demultiplexing cells from pooled scRNA-seq experiments by using common genetic (similar to somatic) variants. Existing demultiplexing strategies, e.g., Demuxlet [3], rely on access to complete genotype data from the set of pooled samples, which greatly limits the applicability of such methods, in particular when genetic variation is not the primary object of study. To address this, we modified cardelino model and introduced a variational inference method (named Vireo, see full preprint [4]), efficiently and accurately demultiplexing data from pooled experimental designs. Our model can be applied to dataset with partial or without any genotype information of the pooled samples. Using simulations and results on real data (Fig 2; more in Fig 2-3 in [4]), we demonstrate the robustness of our model and illustrate the utility of multi-sample experimental designs for common expression analyses.

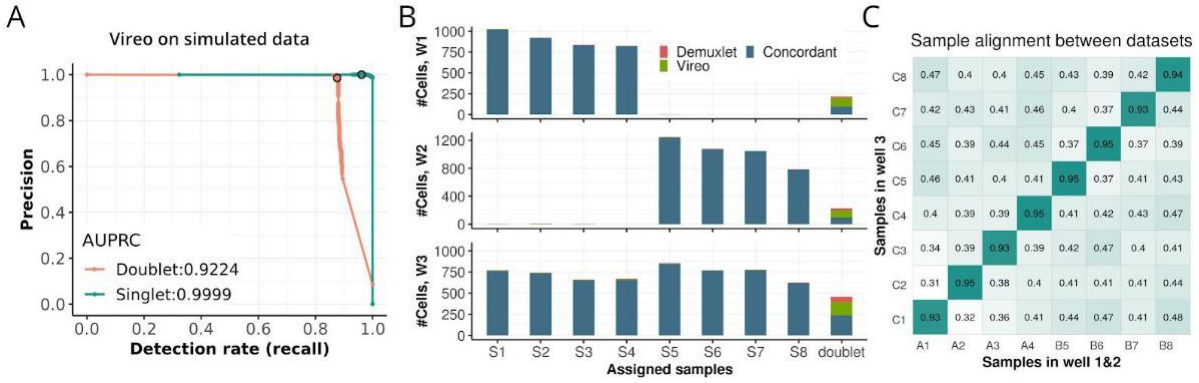


Figure 2 | Performance of Vireo in demultiplexing without genotype reference data. (A) Precision-recall curve for the assignment of singlet-cells and doublet detection in a simulation of 8 input samples. The 90% thresholds are highlighted with black circles. **(B)** Concordance of singlet assignment and doublet detection between Vireo without genotype data and Demuxlet applied with complete genotype reference on three experimental batches (i.e., wells). Bars denote the number of cells assigned to each cell, either considering cells that were consistently assigned by Vireo and Demuxlet (blue), or assigned exclusively by Vireo (green) or Demuxlet (red). **(C)** Alignment of samples between batches 1&2 and batch 3 when , when applying Vireo separately. Values in the heatmap denote the fraction of concordant genotype states between pairs of samples from separate Vireo runs.

Reference

[1] McCarthy, et al. (2018). Cardelino: Integrating whole exomes and single-cell transcriptomes to reveal phenotypic impact of somatic variants. bioRxiv, 413047

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

[3] Kang, et al. (2018) Multiplexed droplet single-cell RNA sequencing using natural genetic variation. Nat Biotech, 36(1):89

[4] Huang, et al. (2019). Vireo: Bayesian demultiplexing of pooled single-cell RNA-seq data without genotype reference. bioRxiv, 598748

Genotyping structural variations using long reads data

Lolita Lecompte, Pierre Peterlongo, Dominique Lavenier and Claire Lemaitre

INRIA, France

Background

Structural variations (SV) are characterized as genomic segments of a least 50 base pairs (bp) long, that are rearranged in the genome. There are several types of SV such as deletions, insertions, duplications, inversions, translocations. This kind of polymorphism have been shown involved in many biological processes, particularly diseases or evolution [1]. Databases referencing such variants grow as new variants are discovered, at this time dbVar, the reference database of human genomic SVs [2], contains 35,428,724 variant calls, illustrating that many SVs have already been discovered and characterized in the human population. In this context, it becomes very interesting and informative to evaluate for a given newly sequenced individual if its genome holds already known SVs. This is commonly known as the SV genotyping problem.

Such genotyping methods already exist for short reads data: for instance, SVtyper [3], SV^2 [4]. Though short reads are often used to discover and genotype SVs, this is well known that their short size make them ill-adapted for predicting large SVs or SVs located in repeated regions. Third generation sequencing technology, such as Paci c Biosciences (PB) and Oxford Nanopore Technologies (ONT), can produce long reads data compared to Next Generation Sequencing technologies. Despite their higher error rate, long reads are crucial in the study of SVs. Indeed, the size range of this data can reach a few kilobases to megabases, thus long reads can extend over rearranged SV sequences as well as over the repeated sequences often present at SV's breakpoint regions.

Following long reads technology's development, many SV discovery tools have emerged, such as Sni es [5]. To our knowledge there is currently no tool that can perform genotyping from a set of known SVs with long reads data. Thus, there is a need to develop accurate and e cient methods to genotype SVs with long reads data, especially in the context of clinical diagnoses.

Results

Method We propose a novel method that aims at assigning a genotype for a set of already known SVs in a given individual sample sequenced with long reads data. In other words, the method assesses if each SV is present in the given individual, and if so, how many variant alleles it holds, ie. whether the individual is heterozygous or homozygous for the particular variant. The method is described and implemented here for only one type of SV, the deletions, but the principle can be easily generalized to other types of SVs. We also provide an implementation of this method in the tool named Biskoul.

The principle of the method is based on: 1) Generating reference sequences that represent the two possible alleles of each SV. The reference allele (allele 0) is therefore the sequence of the deletion with adjacent sequences at each side, and the alternative allele (allele 1) consists in the joining of the two previous adjacent sequences. 2) Then, sequenced long reads are aligned on all previously generated references, using Minimap2 [6], speci cally designed for long erroneous reads. 3) An important step of our method consists in selecting informative alignments, in order to remove i) uninformative alignments, that is those not discriminating between the two possible alleles, and ii) spurious false positive alignments, that are mainly due to repeated sequences. 4) Finally, for each SV, the allele frequency is measured based on the number of supporting alignments, in order to estimate genotype.

Evaluation on simulated data Biskoul was assessed on PB simulated long reads for the human chromosome 1, with 1,000 real characterized deletions found in dbVar [2], ranging from 50 to 10,000 bp, equally distributed among the three di erent genotypes (0/0, 0/1, 1/1).

Biskoul achieved 95.8 % precision, it correctly assigned genotypes to 942 over 987 predicted deletions. Most erroneous genotypes concern deletions of small size (less than 100 bp), as expected these are harder to genotype than longest deletions. As a matter of fact, the precision is of 85.4 % for deletions smaller than 100 bp versus 97.9 % for deletions greater than 500 bp. The remaining false positive deletions of size 100 bp, were manually investigated, and most of them occur in regions with a high density of mobile elements.

Comparison with SV discovery approaches Then we assessed if these simulated deletions could be easily detected and genotyped by a long read SV discovery tool. We applied here the best to date such tool, Sniffles [5] to the chromosome 1 simulated read dataset. As expected, none of the 333 simulated deletions with 0/0 genotypes were assigned a genotype in the Sni es output call set, since a discovery tool naturally only reports present variants. Surprisingly, among the 667 deletions simulated with either a 0/1 or 1/1 genotype, only 406 were discovered by Sni es, which gives a recall of only 60.9 %. Interestingly, Sni es also mis-predicts the genotype of the discovered deletions, assigning most of the 1/1 discovered deletions (n = 254, 81 %) as heterozygous. This highlights the fact that Sni es, a SV discovery tool, is much less precise for the genotyping task than a dedicated genotyping tool.

Application to real human data Biskoul was also applied on real ONT data [7] for the whole human genome of NA12878 individual. As the set of deletions to genotype, we used the merged SV call set provided by the Genome in a Bottle (GiAB) consortium for NA12878 (Mt. Sinai School of Medicine dataset), where only SVs predicted by all methods were kept. This set of known variants contains 1,685 deletions.

Biskoul assigned a genotype to 1,684 deletions, of which 1,514 (90 %) were genotyped exactly as in GiAB. Biskoul took 1h46m on this dataset, including 1h42 for the alignment with Minimap2 parallelized on 40 cpu and with a maximum RAM memory of 6.5 Gbytes.

Conclusions

In this work, we provide a novel SV genotyping approach for long reads data, that is fast and accurate on both simulated and real datasets. This work also demonstrated that this is crucial to develop dedicated SV geno-typing methods, since SV discovery methods are not as e cient and precise to genotype variants once SVs have been discovered. The approach is implemented for the moment only for deletion variants in the Biskoul software. However, this proof of principle on deletion variants is a rst step before generalizing the approach for all types of SVs. Insertion variants are simply the counterpart of deletions, and inversions and transloca-tions are SVs even more balanced than insertions/deletions regarding the number of breakpoints (with exactly two breakpoints per allele). Therefore, for all these types of SVs, the method will be easily generalized. Our method lls a gap and now enables SV genotyping using long reads for clinical diagnosis or population geno-typing. Biskoul is available at https://data-access.cesgo.org/index.php/s/6EhOdOBsVNRr72n, under GNU A ero GPL licence.

References

[1] James R Lupski. Structural variation mutagenesis of the human genome: Impact on disease and evolution. Environmental and molecular mutagenesis, 56(5):419{436, 2015.

[2] Lon Phan, Je rey Hsu, Le Quang Minh Tri, Michaela Willi, Tamer Mansour, Yan Kai, John Garner, John Lopez, and Ben Busby. dbvar structural variant cluster set for data analysis and variant comparison. F1000Research, 5, 2016.

[3] Colby Chiang, Ryan M Layer, Gregory G Faust, Michael R Lindberg, David B Rose, Erik P Garrison, Gabor T Marth, Aaron R Quinlan, and Ira M Hall. Speedseq: ultra-fast personal genome analysis and interpretation. Nature methods, 12(10):966, 2015.

[4] Danny Antaki, William M Brandler, and Jonathan Sebat. Sv2: accurate structural variation genotyping and de novo mutation detection from whole genomes. Bioinformatics, 34(10):1774{1777, 2017.

[5] Fritz J Sedlazeck, Philipp Rescheneder, Moritz Smolka, Han Fang, Maria Nattestad, Arndt von Haeseler, and Michael C Schatz. Accurate detection of complex structural variations using single-molecule sequencing. Nat Methods, 15(6):461{468, 2018.

[6] Heng Li. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics, 34(18):3094{3100, 2018.

[7] Miten Jain, Sergey Koren, Karen H Miga, Josh Quick, Arthur C Rand, Thomas A Sasani, John R Tyson, Andrew D Beggs, Alexander T Dilthey, Ian T Fiddes, et al. Nanopore sequencing and assembly of a human genome with ultra-long reads. Nature biotechnology, 36(4):338, 2018.

GraphAligner: Rapid and Versatile Sequence-to-Graph Alignment

Mikko Rautiainen and Tobias Marschall
Max Planck Institute for Informatics

Background

Graph-based methods have been of growing interest in genomic analysis. Graphs provide a natural way of expressing variation or uncertainty in a genome [1, 2]. They can be used for diverse applications such as genome assembly, error correction and structural variation genotyping. With the growing usage of graphs, methods for handling graphs efficiently are becoming more important. In particular, sequence alignment is one of the most fundamental operations in genome analysis and used in most applications, including error correction [3], genome assembly [4], and graph-based haplotype phasing [5].

Results

In this work we present our tool GraphAligner for aligning long reads to genome graphs. GraphAligner uses a seed-and-extend approach and combines novel strategies for banded sequence-to-graph alignment with our previous algorithmic advances on bit-parallel sequence-to-graph alignment [6].

The seeding is based on finding exact matches fully contained inside a node, which we found to be highly effective for long reads. We interface with the MUMmer4 [7] API for finding exact matches. Our read mapper GraphAligner is based on the bit -parallel sequence-to-graph method from our previous work[6]. In order to scale it to large genomes, we have designed and implemented a novel banded alignment approach. In contrast to sequence-to-sequence alignment, where efficient banded alignment schemes are easy to implement [8], the situation is more complex for graphs, where the graph topology can render simple banding approaches infeasible. To address this challenge, our algorithm dynamically determines which cells of the DP table are to be examined based on the alignment scores.

Table 1 shows the runtime and memory use of aligning several datasets and graphs. Bacterial, gene and chromosomal scale graphs can easily be handled with a regular laptop, and human whole genome graphs with a modest computing server. We compared our alignment approach against the vg toolkit[9]. We built a variation graph of the whole human genome by taking the reference GRCh38 and variant calls from the Human Genome Structural Variation Consortium [10] and using vg to build a graph with the alternate alleles. We extracted the subgraph corresponding to chromosome 22. Then we selected PacBio reads from the individual HG00733 by aligning them to the reference with minimap2[11] and taking the reads which aligned to chromosome 22 and randomly downsampled them to 10x coverage. We aligned the selected reads to the chr22 graph using both vg and GraphAligner. Finally we filtered out alignments with an identity less than 60% as spurious mappings. Table 3 shows the results. We see that GraphAligner is almost 50 times faster when indexing the graph and almost 9 times faster during subsequent read mapping, while successfully aligning almost three times more sequence.

To show how better alignment can improve downstream analyses, we built a hybrid error correction pipeline for long reads. This uses the same method as the long read corrector LoRDEC[3]: short reads are used to build a de Bruijn graph (DBG), long reads are aligned to the graph and the aligned path is extracted as the corrected read. We ran LoRDEC using the settings suggested in their paper[3]. For our method, we self- corrected the Illumina reads using Lighter [12], build the de Bruijn graph from them using Bcalm[13], and aligned the long reads to this graph using GraphAligner. To evaluate the error rate, we aligned the corrected reads to the reference using BWA [14] and measured the number of mismatches. We compared our pipeline to LoRDEC on E. Coli. We also ran our error correction pipeline for fruit fly, the major histocompatibility complex (MHC) region of HG002 and the whole genome of HG00733. Table 2 shows the results. Although we use the same idea as LoRDEC, the improved alignment method means that we can align more reads faster and more accurately: While LoRDEC uses 4001 CPU-seconds to achieve an error rate of 0.057%, our pipeline finishes in 415 CPU-seconds and delivers an error rate of 0.0064% (E.coli). Due to the rapid speed, GraphAligner now enables error correction on whole human genomes.

Conclusion

As sequence alignment is one of the most fundamental operations in genome analysis, better alignment methods will produce many downstream benefits. GraphAligner is a tool for rapidly aligning long reads to genome graphs about one order of magnitude faster than existing methods. GraphAligner is open source on Github (https://github.com/maickrau/GraphAligner) and available on bioconda [15].

Tables

Graph	Graph size (bp)	Sequence (bp)	Cov	Aligned	CPU-time	Peak memory
E. Coli DBG	5,390,452	98,213,822	21x	83,537,974 (85%)	188 sec	1.0 Gb
Fruit fly DBG	174,176,070	29,306,290,844	167x	18,845,672,801 (64%)	173 h	30 Gb
HG00733 DBG	2,978,668,577	305,778,524,405	101x	198,877,101,999 (65%)	1508 h	142 Gb
HG002 MHC	5,778,089	112,313,805	19x	109,270,932 (97%)	353 sec	0.4 Gb
Human variation graph	3,097,741,781	305,778,524,405	101x	207,641,933,617 (68%)	663 h	71 Gb
Human chr22 variation graph	50,914,444	5,401,948,911	106x	3,141,016,398 (58%)	4.5 h	3.4 Gb
		540,244,732	10x	316,580,602 (59%)	26.2 min	3.3 Gb

Table 1: Performance of GraphAligner on different graphs and datasets.

Dataset	Correction	Sequence (bp)	Error rate	CPU-time	Peak memory
E. Coli	None	98,213,822	13.9%	-	-
	LoRDEC	80,323,902	0.057%	4001 sec	2.0 Gb
	GraphAligner	83,537,974	0.0064%	415 sec	2.0 Gb
Fruit fly	None	29,306,290,844	9.9%	-	-
	GraphAligner	18,845,672,801	1.6%	176 h	30 Gb
HG00733 whole genome	None	305,778,524,405	11.8%	-	-
	GraphAligner	198,877,101,999	1.6%	1534 h	142 Gb
HG002 MHC	None	112,313,805	14.6%	-	-
	GraphAligner	109,270,932	0.2%	439 sec	1.4 Gb

Table 2: Error correction. The CPU-time and peak memory for GraphAligner measures the whole pipeline, including short read self-correction, graph construction, indexing and alignment.

		Indexing		Alignment	
Aligner	Aligned (bp)	CPU-time	Peak memory	CPU-time	Peak memory
vg 1.13.0 map	109,124,508 (20%)	33.2 min	9.3 Gb	229.1 min	4.0 Gb
GraphAligner	314,511,746 (58%)	0.7 min	2.8 Gb	25.5 min	3.3 Gb

Table 3: Comparison of GraphAligner and vg version 1.13.0 on the chromosome 22 variation graph. Aligned column counts only alignments with an identity at least 60%.

References

[1] The Computational Pan-Genomics Consortium. Computational pan-genomics: status, promises and challenges. Briefings in bioinformatics, 19(1):118–135, 2016.

[2] Benedict Paten, Adam Novak, Jordan Eizenga, Erik Garrison: Genome graphs and the evolution of genome inference. Genome research 27(5):665–676, 2017.

[3] Leena Salmela and Eric Rivals. Lordec: accurate and efficient long read error correction. Bioinformatics, 30(24):3506–3514, 2014.

[4] Dmitry Antipov et al. hybridSPAdes: an algorithm for hybrid assembly of short and long reads. Bioinformatics 32.7:1009-1015, 2015.

[5] Shilpa Garg et al. A graph-based approach to diploid genome assembly. Bioinformatics 34.13: i105-i114, 2018.

[6] Mikko Rautiainen, Veli Mäkinen, Tobias Marschall, Bit-parallel sequence-to-graph alignment, Bioinformatics, btz162, https://doi.org/10.1093/bioinformatics/btz162

[7] Guillaume Marçais, Arthur L Delcher, Adam M Phillippy, Rachel Coston, Steven L Salzberg, and Aleksey Zimin. Mummer4: A fast and versatile genome alignment system. PLoS computational biology, 14(1):e1005944, 2018.

[8] Kun-Mao Chao, William R Pearson, and Webb Miller. Aligning two sequences within a specified diagonal band. Bioinformatics, 8(5):481–487, 1992.

[9] Erik Garrison, Jouni Sirén, Adam M Novak, Glenn Hickey, Jordan M Eizenga, Eric T Dawson, William Jones, Shilpa Garg, Charles Markello, Michael F Lin, et al. Variation graph toolkit improves read mapping by representing genetic variation in the reference. Nature biotechnology, 2018.

[10] Mark Chaisson et al. Multi-platform discovery of haplotype-resolved structural variation in human genomes. BioRxiv, page 193144, 2018.

[11] Heng Li. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 34.18: 3094-3100, 2018..

[12] Li Song, Liliana Florea, and Ben Langmead. Lighter: fast and memory-efficient sequencing error correction without counting. Genome biology, 15(11):509, 2014.

[13] Rayan Chikhi, Antoine Limasset, and Paul Medvedev. Compacting de bruijn graphs from sequencing data quickly and in low memory. Bioinformatics, 32(12):i201–i208, 2016.

[14] Heng Li and Richard Durbin. Fast and accurate short read alignment with burrows–wheeler transform. Bioinformatics, 25(14):1754–1760, 2009.

[15] Björn Grüning et al. Bioconda: sustainable and comprehensive software distribution for the life sciences. Nature Methods 15: 475-476, 2018.

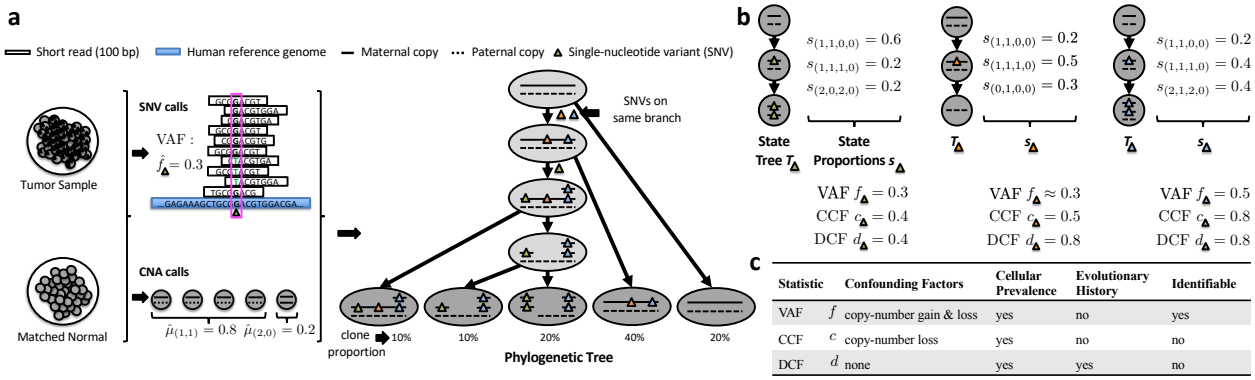
Descendant Cell Fraction: Copy-aware Inference of Clonal Composition and Evolution in Cancer

Mohammed El-Kebir[1,*], Simone Zaccaria[2], and Benjamin J. Raphael[2,*]

[1]Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana, IL 61801
[2]Department of Computer Science, Princeton University, Princeton, NJ 08540
[*]Correspondence: melkebir@illinois.edu, braphael@princeton.edu

A tumor results from an evolutionary process where somatic mutations accumulate in a population of cells. This process gives rise to a tumor that is a mixture of distinct clones, distinguished by somatic mutations including single-nucleotide variants (SNVs), copy-number aberrations (CNAs), and other changes. The standard approach to identify such clones is to cluster SNVs that have similar *cancer cell fractions* (CCFs), defined as the proportion of tumor cells harboring the mutation. The key assumption of this approach is that SNVs with similar CCFs have occurred on the same phylogenetic branch. There are, however, two key deficiencies: (1) the CCF cannot be unambiguously inferred from DNA sequencing data; (2) the CCF does not account for loss of mutations, which is common in tumors with CNAs (Fig. 1). Thus, the standard approach might lead to incorrect reconstructions of tumor clonal architectures, which in turn might lead to incorrect conclusions in downstream analyses.

Here, we define a novel quantity, the *descendant cell fraction* (DCF) that addresses these deficiencies in a rigorous manner, providing a summary statistic for both the prevalence *and* the evolutionary history of an SNV. That is, SNVs with the same DCF are likely to have occurred on the same branch of the phylogenetic tree describing the evolution of the tumor. We introduce DeCiFer, an algorithm to simultaneously infer evolutionary histories of individual SNVs and clusters SNVs by their corresponding DCFs under the principle of parsimony. Underpinning DeCiFer is an elegant embedding of the high-dimensional space of evolutionary histories of SNVs onto the low-dimensional DCF space. On simulated data, we show that DeCiFer more accurately clusters SNVs than existing methods and infers evolutionary histories with high recall. On a metastatic prostate cancer dataset, we show that DeCiFer's use of the DCF to cluster SNVs results in more parsimonious evolutionary and migration histories of these metastatic cancers. Thus, DeCiFer enables more accurate quantification of intra-tumor heterogeneity and improves inference of tumor evolution.



Statistic		Confounding Factors	Cellular Prevalence	Evolutionary History	Identifiable
VAF	f	copy-number gain & loss	yes	no	yes
CCF	c	copy-number loss	yes	no	no
DCF	d	none	yes	yes	no

Fig. 1: a, Calling of SNVs and CNAs in bulk DNA samples yield variant allele frequencies $\hat{f}$ (VAFs) for SNVs and copy-number proportions $\mu_{(x,y)}$ for CNAs. **b,c,** The prevalence and evolutionary history of an SNV is described by unobserved state proportions $\mathbf{s}$ and a state tree T, respectively. State proportions are summarized by different statistics. The VAF, or the fraction of chromosomal copies that harbor the SNV, is confounded by copy-number gain and loss (green and orange). The cancer cell fraction (CCF), or the proportion of cells that harbor an SNV, is confounded by copy-number loss (orange and blue). The descendant cell fraction (DCF) is the proportion of extant cells that descend from the cell that introduced the SNV. The DCF summarizes both a state tree T and state proportions $\mathbf{s}$, and is not confounded by mutation loss: the orange and blue SNV have identical DCF $d = 0.8$, reflecting their common ancestry.

Fast and accurate bisulfite alignment and methylation calling for mammalian genomes

Jonas Fischer[1,2] and Marcel H. Schulz[1,2,3]

[1]Max Planck Institute for Informatics, Saarbrücken, Germany
[2]MMCI Cluster of Excellence, Saarbrücken, Germany
[3]Institute for Cardiovascular Regeneration, Goethe University, Fankfurt, and German Centre for Cardiovascular Research (DZHK), Partner site RheinMain, Frankfurt, Germany

April 11, 2019

Background

Whole Genome Bisulfite Sequencing (WGBS) is considered the gold standard for genome wide, high resolution DNA methylation measurements. With the ongoing advances in Next Generation Sequencing techniques, for example single-cell methylomes, an evergrowing amount of WGBS sequencing data is produced for different organisms, tissues, and cell types. However, while sequencing throughput has increased, alignment and methylation calling algorithms have not been adapted to the increasing demands, which causes a serious bottleneck in current applications. Furthermore, widely used tools do not resolve the mapping ambiguity introduced by WGBS, which comes at the cost of accuracy of the called methylation rates.

Results

Here, we present a novel approach called **FAME** (**F**ast and **A**ccurate **ME**thylation calling), which combines bisulfite read alignment and methylation calling into one task. We designed a specialized index structure based on sequence k-mers that can store mammalian sized genomes efficiently, while allowing fast lookups of candidate matching positions for bisulfite converted reads. To allow for constant time traversal of the index structure, gapped k-mers are hashed using a fast rolling hash function called *ntHash* [1], specifically tailored for genomic sequences. We further designed fast filters to prune redundant or repetitive information in the index, which drastically reduces the search space for read alignment.

Using this index, we can find candidate matching regions for a read by looking up all gapped k-mers of a queried read efficiently using rolling hash functions. Smart filtering of the candidate regions based on q-grams allows us to reduce time spent on false positive candidates. To carry out exact indel-based alignment, we extended the Shift-And based pattern matching automata to allow for asymmetric C/T mapping to resolve ambiguity introduced by bisulfite conversion. Once a unique best alignment is found, methylation levels are directly estimated in the data structure, thus avoiding excessive I/O for writing large Bam files as well as additional postprocessing time for methylation calling.

We compared FAME against the state of the art with synthetic and real data sets. The synthetic data consisted of 25 million reads sampled from the human chromosome 22 mimicking the WGBS protocol, such that ground truth methylation rates were known. The real data included 437 million PE WGBS reads and EPIC beadchip microarray data for LNCaP cells taken from [2], where the EPIC array measurements were taken as ground truth. Difference to ground truth values was computed using root mean square error (RMSE) ovre all existing CpGs. The quality of the methylation rates of FAME are on par with the most accurate competitor, while FAME provides an order of magnitude faster processing time as can be seen in Fig. 1.

Conclusion

We suggested a new method, FAME, for calling methylation rates of WGBS data that is based on a novel index structure specifically tailored for methylation data, and a bitvector matching that resolves the WGBS mapping ambiguity. We proved on both synthetic and real data that the quality of methylation rates called by FAME

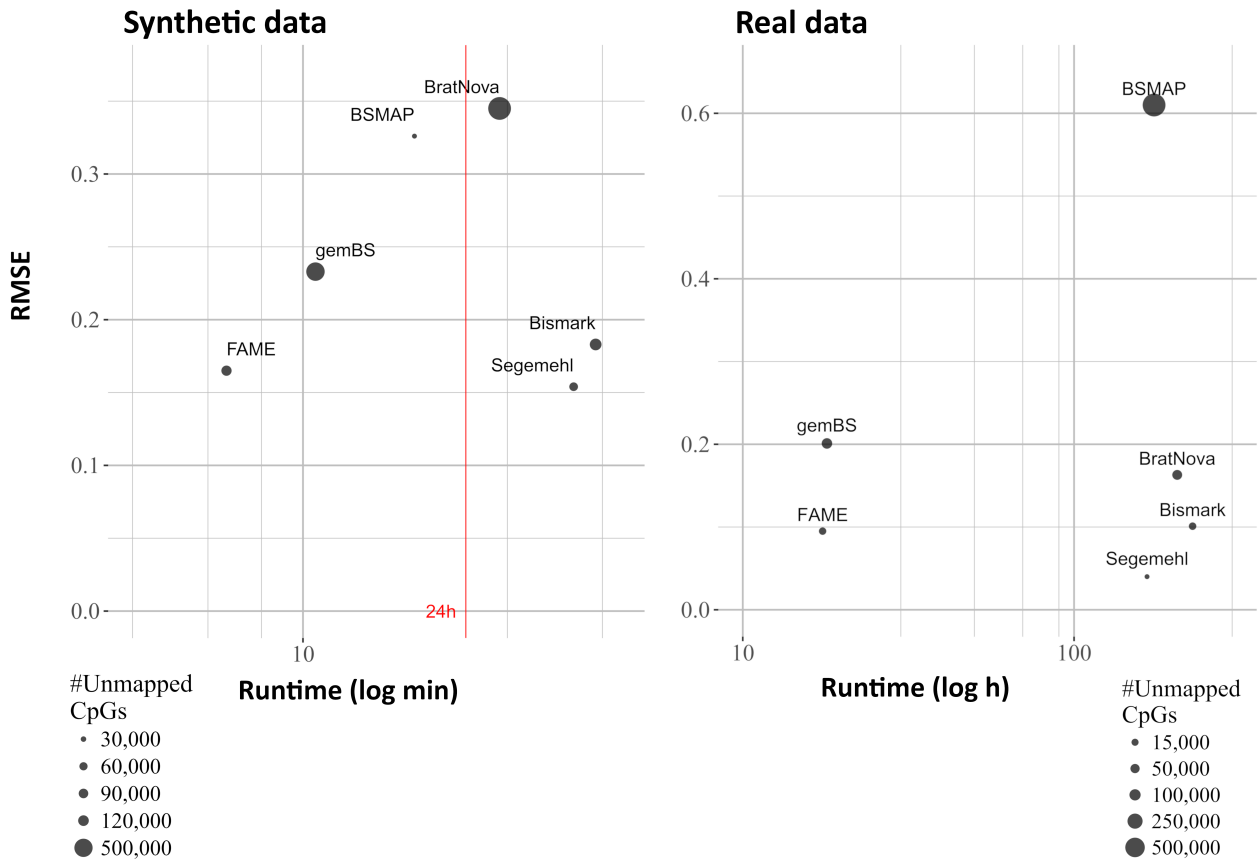


Figure 1: *Method comparison.* Analysis of the runtime and accuracy of state of the art bisulfite aligners. Results for synthetic (25 million PE reads, left) and real data (437 million PE reads, right) comparing runtime (x-axis) against accuracy of the predicted methylation rates as RMSE (y-axis). The size of each point indicates the number of unmapped CpGs.

are on par with the most accurate state of the art aligner, but FAME processes data an order of magnitude faster. Furthermore, due to the novel index structure, FAME does not require extensive I/O or realignment and thus naturally suits the task of aligning extensive WGBS single cell data. Hence, FAME paves the way for methylation calling of large-scale datasets and is ideal for cloud computing. FAME is open source and free to use and can be downloaded from https://github.com/FischerJo/FAME.

References

[1] H. Mohamadi, J. Chu, B. P. Vandervalk, and I. Birol. ntHash: recursive nucleotide hashing. *Bioinformatics*, 32(22):3492–3494, Nov 2016.

[2] R. Pidsley, E. Zotenko, T. J. Peters, M. G. Lawrence, G. P. Risbridger, P. Molloy, S. Van Djik, B. Muhlhausler, C. Stirzaker, and S. J. Clark. Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling. *Genome Biol.*, 17(1):208, Oct 2016.

ImmunoPepper: Generating Neoepitopes from RNA-Seq data

Matthias Hüser, Jiayu Chen, Andre Kahles and Gunnar Rätsch
ETH Zurich

Motivation:
RNA-Sequencing has enabled the high-throughput assessment of a cell's transcriptome. However, in many situations this technique is used as a stand-in, specifically to measure the pool of mRNA as a proxy for the cell's proteome. In addition to quantitative features, such as gene expression, it is mainly qualitative features such as splicing forms, resulting in distinct protein isoforms, that are of interest. The problem of predicting the set of expressed transcript forms from shotgun sequencing data is inherently hard to solve and has been widely addressed in the past [1,2]. For some applications, however, it is not necessary to generate full protein isoforms and one is only interested in the local variability of the whole proteome. Especially in the context of personalized cancer therapy, assessing the proteome's variability and predicting immunogenicity of peptide fragments sampled from the proteome are a central concern.

Method:
We present the software tool *ImmunoPepper*, that generates the set of all plausible peptides from a splicing graph, derived from a given RNA-Seq sample. This splicing graph contains both annotated as well as novel splicing variation. The set of peptides is generated through combinatorial traversal of all exon pairs, using all reading frames implied through the propagation of all annotated translation start sites of the gene along all paths in the graph. In addition, the generated peptide set can be personalized with germline and somatic variants and takes un-annotated introns into account. The comprehensive set of peptides can then be used for further downstream analyses such as domain annotation or computational immunology. To facilitate analysis with standardized tools for MHC binding prediction, we provide output for unique k-mer sets of all generated peptides, where typical k-mer lengths reach from 8 to 22. For each peptide additional quantitative metadata is provided that can be used for filtering and increasing the specificity of the predicted neoepitopes. This metadata includes RNA-Seq expression support, but also the support for any of the given genome variants.

Results:
The core algorithm of our software has already been successfully applied to a large cancer patient cohort [3], predicting the set of splicing derived neoepitopes predominantly observed in cancer samples. The *ImmunoPepper* software built around this core is a much more generalized implementation that includes a substantial amount of new features, such as a more complete traversal of complex graphs, the inclusion of single exon genes and the generation of peptides across multiple splice junctions. These new contributions lead to an increase of sensitivity of more than 50%. At the same time, a better tracking of metadata allows for fine-grained filtering and an increase of specificity for the prediction of sample-specific peptides.
We demonstrate the versatility of *ImmunoPepper* with applications to a set of 63 cancer samples from the The Cancer Genome Atlas cohort contrasted with samples from the Genotype Tissue Expression project; and the analysis of 5 mouse tumor samples in comparison to more than 300 background samples taken from mouse ENCODE and a reference study [4]. In both cases, we can demonstrate the existence of sample-specific (tumor-specific) splicing-derived peptides that can give rise to (tumor-specific) neoepitopes.

Availability:
The software *ImmunoPepper* is implemented in Python3 and is available open source on the GitHub platform.

Literature:

1. Garber M, Grabherr MG. Computational methods for transcriptome annotation and quantification using RNA-seq. Nat Methods. 2011;8: 469–477.
2. Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, et al. A survey of best practices for RNA-seq data analysis. Genome Biol. 2016;17: 13.
3. Kahles A, Lehmann K-V, Toussaint NC, Hüser M, Stark SG, Sachsenberg T, et al. Comprehensive Analysis of Alternative Splicing Across Tumors from 8,705 Patients. Cancer Cell. Elsevier; 2018;34: 211–224.e6.
4. Li B, Qing T, Zhu J, Wen Z, Yu Y, Fukumura R, et al. A Comprehensive Mouse Transcriptomic BodyMap across 17 Tissues by RNA-seq. Sci Rep. nature.com; 2017;7: 4200.

Haplotype Threading:
Accurate Polyploid Phasing from Long Reads

Sven Schrinner[1], Rebecca Serra Mari[2], Jana Ebler[2], Gunnar W. Klau[1] and Tobias Marschall[2]

[1] Heinrich Heine University Düsseldorf
[2] Center for Bioinformatics, Saarland University, Saarbrücken

Background

Polyploid genomes have more than two homologous sets of chromosomes. Polyploidy is common to many plant species, including important food crops like potato, wheat and maize. Resolving these genomes at the haplotype level is crucial for understanding the evolutionary history of polyploid species [7] and for designing advanced breeding strategies [5]. While phasing diploid genomes using long reads has become a routine step, polyploid phasing still presents considerable challenges [2]. Higher ploidy increases the complexity of the underlying computational problem: In the diploid case, assembling one haplotype over all heterozygous variants directly determines the complementary second haplotype. For genomes of higher ploidy, this is not possible, especially since in certain regions, two or more haplotypes can be identical. The Minimum Error Correction (MEC) model [3], which is the most common and successful formalization for diploid haplotype phasing, does not address such locally identical haplotypes. Approaches for polyploid phasing based on MEC hence struggle in such regions and, beyond that, face the challenge that dynamic programming techniques for solving diploid MEC [4] become infeasible in practice.

Results

Method. Here, we present WHATSHAP POLYPHASE, a novel two-stage approach that overcomes these challenges and produces accurate haplotypes for polyploid genomes using data from single-molecule sequencing technologies. See Fig. 1 for an overview.

In the **first phase**, we use cluster editing [8] to find clusters of reads that are likely to originate from the same haplotype. Therefore we compute a position-dependent similarity score for each pair of reads and construct a graph using the reads as nodes and the scores as edge weights. The size of the graph makes it infeasible to solve cluster editing to optimality, so we propose a fast iterative heuristic that produces accurate clusters. In this first phase, we intentionally do not make assumptions on the ploidy. If multiple haplotypes are locally identical, their reads from this region might form a single cluster.

In the **second phase**, we perform the actual haplotype assembly by **threading** a bundle of haplotypes through the clusters obtained in the first phase (see Fig. 1). More formally, threading one haplotype through the set of clusters consists in picking one cluster for each variant position. In contrast to the MEC model, this approach allows us to take the coverage into account: Our model allows to thread multiple haplotypes through a single cluster of reads from multiple haplotypes due to local similarity. To determine a threading for all haplotypes, the model takes the following factors into account: (i) The read coverage of the cluster should be explained by the number of haplotypes that are threaded through this cluster, and deviations from this are penalized; (ii) On each position the consensus of the visited clusters has to match the input genotype, if possible; (iii) Switching between clusters is penalized, to encourage haplotypes to stay in the same cluster as long as possible. We propose a novel dynamic programming approach able to rapidly find

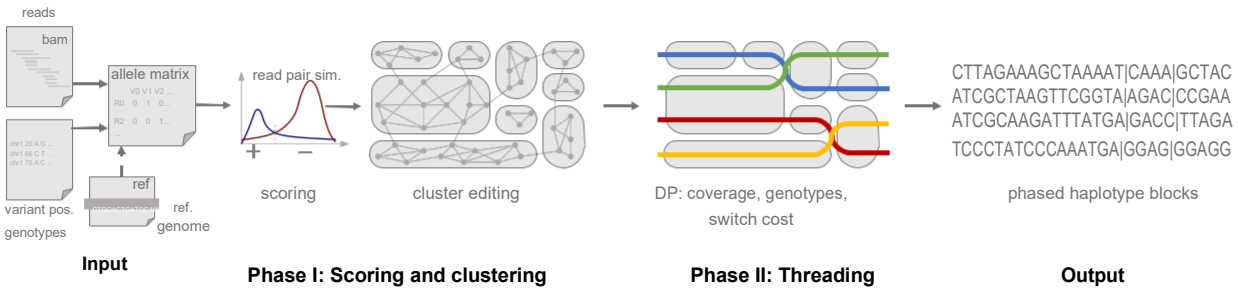


Figure 1: WHATSHAP POLYPHASE. The input allele matrix results from a given BAM and VCF file and an optional realignment step. Phase I: Statistical scoring of each read pair results in a weighted graph. Cluster editing produces read clusters. Phase II threads k haplotypes through the clusters (here k = 4) balancing coverage violations and switch costs while respecting the genotype information. This results in blocks of k phased haplotypes.

rately reconstruct haplotypes of polyploid organisms while properly handling genomic regions of similar or identical haplotypes.

Experimental Results. For evaluation, we generated a tetraploid version of human chromosome 22 by combining sequencing data of two individuals (NA19240 and HG00514), for which ground truth haplotype information is available [1]. We simulated reads at different coverages and additionally produced equivalent datasets by merging real sequencing reads of these two individuals.

We compare WHATSHAP POLYPHASE to H-POPG [6], a state-of-the-art tool for polyploid phasing, and evaluate the performances of both tools based on the switch error rate (SE), block-wise Hamming rate (HR) and N50 of the phased blocks, see Table 1. WHATSHAP POLYPHASE achieves lower switch error rates than H-POPG for both real and simulated reads, as well as lower block-wise Hamming rates. The latter is still quite high in both cases, which is caused by switch errors in the middle of large blocks. Generally, WHATSHAP POLYPHASE avoids uncertain phase connections and opts to split blocks instead. This results in smaller but better quality blocks compared to H-POPG.

	SE (%)		HR (%)		N50 (bp)	
coverage	WH-PP	H-POPG	WH-PP	H-POPG	WH-PP	H-POPG
20	1.45	2.08	16.63	26.06	45 907	927 570
40	0.88	1.27	20.03	23.98	247 502	1 029 048

(a) real read data

	SE (%)		HR (%)		N50 (bp)	
coverage	WH-PP	H-POPG	WH-PP	H-POPG	WH-PP	H-POPG
20	1.74	2.56	13.45	26.93	24 558	852 018
40	0.72	1.17	20.01	23.77	330 736	927 570
80	0.49	0.81	20.57	23.62	720 984	1 216 882
120	0.46	0.71	21.14	22.06	798 580	1 134 439

(b) simulated read data

Table 1: Results on real (a) and simulated (b) datasets. Performances are based on the switch error rate (SE), block-wise Hamming rate (HR) and N50 for the block size. Total length of the chromosome is 50.8Mb

Conclusions

We present a new two-stage approach for polyploid phasing. The first phase clusters reads based on their similarity using a position-dependent statistical scoring scheme. The second phase threads haplotypes through the clusters and takes coverage and genotype information into account. Our model departs from the popular MEC paradigm, which has been successful for phasing diploid genomes, and results in the first algorithm designed to specifically handle locally identical haplotypes. Current challenges lie in eliminating switch errors causing high Hamming rates and in scaling the algorithm to ploidies above six.

Our method WHATSHAP POLYPHASE delivers haplotype reconstructions with 30% lower error rates compared to the state-of-the-art tool H-POPG on an artificial tetraploid benchmark genome. Our algorithm is implemented as part of the widely used open source tool WHATSHAP and is hence ready to be included in production settings. We are presently exploring the ability of WHATSHAP POLYPHASE to phase polyploid plant genomes and to create maps of identical haplotype regions.

References

[1] M. J. P. Chaisson et al. Multi-platform discovery of haplotype-resolved structural variation in human genomes. BioRxiv, page 193144, 2018.

[2] G. W. Klau and T. Marschall. A guided tour to computational haplotyping. In Unveiling Dynamics and Complexity, Lecture Notes in Computer Science, pages 50–63. Springer, June 2017.

[3] R. Lippert et al. Algorithmic strategies for the single nucleotide polymorphism haplotype assembly problem. Brief. Bioinform., 3(1):23–31, March 2002.

[4] M. Patterson et al. WhatsHap: Weighted haplotype assembly for Future-Generation sequencing reads. J. Comput. Biol., 22(6):498–509, June 2015.

[5] R. G. F. Visser et al. Possibilities and challenges of the potato genome sequence. Potato Res., 57(3-4):327–330, December 2014.

[6] M. Xie et al. H-PoP and H-PoPG: heuristic partitioning algorithms for single individual haplotyping of polyploids. Bioinformatics, 32(24):3735–3744, December 2016.

[7] J. Yang et al. Haplotype-resolved sweet potato genome traces back its hexaploidization history. Nat Plants, August 2017.

[8] C. T. Zahn Jr. Approximating symmetric relations by equivalence relations. J. Soc. Indust. Appl. Math., 12(4), December 1964.