High Throughput Sequencing
Algorithms and Applications

A special track of the ISMB 2024 meeting

Montreal, QC, Canada, July 13-14, 2024

ISMB 2024 HiTSeq Track Proceedings

Montreal, QC, Canada
July 13-14, 2024
https://www.hitseq.org

Organizers:

Can Alkan, Ph.D.
Bilkent University, Bilkent, Ankara, Turkey
E-mail: calkan@cs.bilkent.edu.tr

Christina Boucher, Ph.D.
University of Florida, Gainesville, FL, USA
E-mail: cboucher@cise.ufl.edu

Broňa Brejová, Ph.D.
Comenius University in Bratislava, Slovakia
E-mail: brejova@dcs.fmph.uniba.sk

Ana Conesa, Ph.D.
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, Ph.D.
Dr. Dirk Evers Consulting, Heidelberg, Germany
E-mail: dirk.evers@gmail.com

Kjong Lehmann, Ph.D.
Centre of Medical Technology, Aachen, Germany
E-mail: kjong.lehmann@inf.ethz.ch

Ana Isabel Castillo Orozco
McGill University, Montreal, Canada
E-mail: ana.castillo.2091@gmail.com

Kristoffer Sahlin, Ph.D.
Stockholm University, Stockholm, Sweden
E-mail: ksahlin@math.su.se

Neil Peterman (Foundation Medicine), Eliana Polisecki (Foundation Medicine), Nicole Lambert (Foundation Medicine) and Alexander Robertson (Foundation Medicine). *A method for selecting a set of differentially methylated regions for tumor response monitoring by liquid biopsy in multiple tumor types.*

Abstract. Patients undergoing systemic anti-cancer therapy suffer from side effects and financial toxicity. Tumor response monitoring (TRM) is critical to monitor the tumor's progression to determine if the patient is truly benefiting from the treatment. Liquid biopsy-based monitoring of tumor response via circulating tumor DNA (ctDNA) promises many improvements over existing imaging technologies. Tracking ctDNA levels based on sequence mutations is difficult when ctDNA levels are low or there are few mutations. This is not a problem when tracking methylation based on differentially methylated regions (DMRs), because tumors have widespread aberrant methylation. While existing DMR algorithms find differential methylation between healthy and cancerous tissue, TRM requires detecting ctDNA at low levels relative to cell-free DNA (cfDNA) from healthy origin. We developed a computational method for selecting DMRs from cfDNA whole genome bisulfite sequencing (WGBS) data based on the principles of maximizing signal-to-noise ratio, handling uncertainty from a limited number of training samples. We applied this method to WGBS data from a set of patients with lung, breast, colorectal, or prostate cancer as well as healthy controls and show concordance with sequence-mutation estimates of ctDNA. While there is overlap in the DMRs between tumor types, sensitivity performance is maximized when using a DMR set specific to the patient's tumor type. This method tells which regions of the epigenome are most potentially informative, given the diversity of methylation patterns both in the healthy population's cfDNA and in ctDNA in patients with the same tumor type.

Keywords: cfDNA, liquid biopsy, cancer, methylation, epigenetics, ctDNA, NGS, sequencing

Dongmei Li (University of Rochester Medical Center), Pinxin Liu (University of Rochester), Shijian Deng (University of Texas at Dallas) and Zidian Xie (University of Rochester Medical Center). *scMAE: A Deep Learning-Based Method for Differential Analysis in Single-Cell RNA Sequencing.*

Abstract. Differential gene expression (DGE) analysis stands as a crucial step in the single-cell RNA sequencing (scRNA-seq) data analysis pipeline, offering insights into novel cell types and gene signatures contributing to cellular heterogeneity. While numerous statistical methods exist for DGE analysis in scRNA-seq data, none explicitly consider gene-gene interactions within this context. In response, we introduce scMAE, a novel approach leveraging Mask AutoEncoder (MAE), to enhance DGE analysis in scRNA-seq data by incorporating gene-gene interactions. scMAE is a deep-learning algorithm rooted in transformer architecture, commonly employed in Natural Language Processing (NLP) and Computer Vision (CV). Through simulation studies, scMAE demonstrates superior performance compared to many existing statistical methods for DGE analysis in scRNA-seq data.

Keywords: Deep learning, Differential analysis, Single-cell RNA sequencing

Xiuhui Yang (Mcgill University), Koren Mann (Mcgill University), Hao Wu (Shandong University) and Jun Ding (Mcgill University). *scCross: A Deep Generative Model for Unifying Single-cell Multi-omics with Seamless Integration, Cross-modal Generation, and In-silico Exploration.*

Abstract. Single-cell multi-omics illuminate intricate cellular states, yielding transformative insights into cellular dynamics and disease. Yet, while the potential of this technology is vast, the integration of its multifaceted data presents challenges. Some modalities have not reached the robustness or clarity of established scRNA-seq. Coupled with data scarcity for newer modalities and integration intricacies, these challenges limit our ability to maximize single-cell omics benefits. We introduce scCross: a tool adeptly engineered using variational autoencoder, generative adversarial network principles, and the Mutual Nearest Neighbors (MNN) technique for modality alignment. This synergy ensures seamless integration of varied single-cell multi-omics data. Beyond its foundational prowess in multi-omics data integration, scCross excels in single-cell cross-modal data generation, multi-omics data simulation, and profound in-silico cellular perturbations. Armed with these capabilities, scCross is set to transform the field of single-cell research, establishing itself in the nuanced integration, generation, and simulation of complex multi-omics data.

Keywords: single cell, muti-omics, cross-modal generation, in-silico perturbations, multimodal integration, generative adversarial network, autoencoder

Jia Li (Vanderbilt University Medical Center), Yu Shyr (Vanderbilt University Medical Center) and Qi Liu (Vanderbilt University Medical Center). *An Adaptive K-Nearest Neighbor Graph Optimized for Single-cell and Spatial Clustering.*

Abstract. Unsupervised clustering is crucial for characterizing cellular heterogeneity in single-cell and spatial transcriptomics analysis. While conventional clustering methods have difficulty in identifying rare cell types, approaches specifically tailored for detecting rare cell types gain their ability at the cost of poorer performance for grouping abundant ones. We introduce aKNNO, a method to identify abundant and rare cell types simultaneously based on an adaptive k-nearest neighbor graph with optimization. Unlike traditional kNN graphs, which require a predetermined and fixed k value for all cells, aKNNO selects k for each cell adaptively based on its local distance distribution. This adaptive approach enables accurate capture of the inherent cellular structure. Through extensive evaluation across 38 simulated scenarios and 20 single-cell and spatial transcriptomics datasets spanning various species, tissues, and technologies, aKNNO consistently demonstrates its power in accurately identifying both abundant and rare cell types. Remarkably, aKNNO outperforms conventional and even specifically tailored methods by uncovering both known and novel rare cell types without compromising clustering performance for abundant ones. Most notably, when utilizing transcriptome data alone, aKNNO delineates stereotyped fine-grained anatomical structures more precisely than integrative approaches combining expression with spatial locations and/or histology images, including GraphST, SpaGCN, BayesSpace, stLearn, and DR-SC.

Keywords: adaptive k-nearest neighbor graph, single-cell and spatial transcriptomics, clustering, rare cells

Stephen Hwang (XDBio Program, Johns Hopkins School of Medicine), Nathaniel K. Brown (Department of Computer Science, Johns Hopkins University), Omar Y. Ahmed (Department of Computer Science, Johns Hopkins University), Katharine Jenike (Department of Computer Science, Johns Hopkins University), Sam Kovaka (Department of Computer Science, Johns Hopkins University), Michael C. Schatz (Department of Computer Science, Johns Hopkins University) and Ben Langmead (Department of Computer Science, Johns Hopkins University). *Compressed Indexing for Pangenome Substring Queries.*

Abstract. There is a growing number of collections of genomes, including pangenomes and taxonomic databases. However, approaches to study sequence homology and genome evolution from these growing collections are limited by computationally intensive reference-graph construction, large index sizes, or fixed-length substring queries. Here, we present Maximal Exact Match Ordered (MEMO), a tool to index a pangenome and allow user-specified region and arbitrary length-k queries for per-position, per-genome k-mer presence/absence (membership) and per-position count of the number of genomes with each k-mer (conservation). While allowing arbitrary length-k queries, MEMO has smaller indexes and faster queries than existing fixed k-mer length indexes. MEMO indexes 89 human autosomal haplotypes in 2.04 GB, over 8x smaller than existing approaches. MEMO queries 31-mer conservation across the human leukocyte antigen (HLA) locus, a highly variable 3.75 Mbp region on chr6, in 13.89 seconds–over 2.5x faster than existing approaches. In summary, MEMO's small index size, lack of k-mer length dependence, and efficient queries make it a flexible and efficient tool for querying and visualizing substring conservation within pangenomes.

Keywords: Pangenomics, Comparative genomics, Compressed indexing

Chen-Hsiang Yeang (Academia Sinica). *Assessing transcriptomic heterogeneity of single-cell RNASeq data by bulk-level gene expression data.*

Abstract. Single-cell and bulk-level RNA sequencing data provide complementary merits and shortcomings. We propose a modeling framework to integrate bulk-level and single-cell RNASeq data to infer their transcriptomic heterogeneity. Contrary to the standard approaches of factorizing the bulk-level data with one algorithm using single-cell RNASeq data as references, we employ multiple deconvolution algorithms to factorize the bulk-level data, constructed the probabilistic graphical models of cell-level gene expressions from the decomposition outcomes, and compared the log-likelihood scores of these models in single-cell data. We term this framework backward deconvolution as inference operates from coarse-grained bulk-level data to fine-grained single-cell data. We selected six deconvolution algorithms and validated backward deconvolution in four datasets. In the in-silico mixtures of mouse sc-RNASeq data, the log-likelihood scores of the deconvolution algorithms were strongly anticorrelated with their errors of mixture coefficients and cell type specific gene expression signatures. In the true bulk-level mouse data, the sample mixture coefficients were unknown but the log-likelihood scores were strongly correlated with accuracy rates of inferred cell types. In datasets of breast cancer and low-grade gliomas, we compared the log-likelihood scores of three simple hypotheses about the expression patterns of the cell types underlying the sample subtypes. The model that tumors of each subtype were dominated by one cell type persistently outperformed an alternative model that each cell type had elevated expression in one gene group and tumors were mixtures of those cell types.

Keywords: single-cell RNASeq data, deconvolution, probabilistic graphical models, heterogeneity

Byungwook Lee (KOBIC). *Bio-Express: Bioinformatics system for massive genomic data analysis.*

Abstract. The rapidly increasing amounts of data available from the new high-throughput methods have made data processing without automated pipelines infeasible. Integration of data and analytic resources into workflow systems provides a solution to the problem, simplifying the task of data analysis. To address the challenge, we developed a cloud-based workflow management system called Bio-Express to provide fast and cost-effective analysis of massive genomic data. We implemented complex workflows making optimal use of high-performance compute clusters. Bio-Express allows users to create multi-step analyses using drag-and-drop functionality and modify parameters of pipeline tools. Users can also import the Galaxy pipelines into Bio-Express. Bio-Express is a hybrid system that enables users to utilize both traditional analysis tools and MapReduce-based big data analysis programs in a single pipeline simultaneously. Thus, the execution of analytics algorithms can be parallelized, which can speed up the whole process. We also developed a high-speed data transmission solution, GBox, to transmit a large amount of data at a fast rate. GBox has a file transfer speed up to 10 times faster than standard FTP and HTTP. The computer hardware for Bio-Express consists of 800 CPU cores and 800TB of storage, enabling 500 jobs to run simultaneously. Bio-Express is a scalable, cost-effective, and publicly available web service for large-scale genomic data analysis. Bio-Express supports the reliable and highly scalable execution of sequencing analysis workflows in a fully automated manner. The Bio-Express cloud server is freely available for use from https://www.bioexpress.re.kr/.

Keywords: Could computing, Analysis pipeline, Bio-Express

James Bonfield (Wellcome Sanger Institute), Tony Burdett (European Bioinformatics Institute, European Molecular Biology Laboratory), Peter Clapham (Wellcome Sanger Institute), Josh Cudby (University of Cambridge), Robert Davies (Wellcome Sanger Institute), Richard Durbin (University of Cambridge), David Holland (Wellcome Sanger Institute), Aditya Jain (University of Cambridge), James McCafferty (Wellcome Sanger Institute), Yanisa Sunthornyotin (European Bioinformatics Institute, European Molecular Biology Laboratory), Andrew Whitwham (Wellcome Sanger Institute), Orson Ye (University of Cambridge), David Yuan (European Bioinformatics Institute, European Molecular Biology Laboratory) and Sergii Strelchuk (University of Cambridge). *Quantum Computing for Genomic Analysis.*

Abstract. Many essential tasks in genomic analysis are extremely difficult for classical computers due to problems inherently hard to be solved efficiently with classical (empirical) algorithms. Quantum computing offers novel possibilities with algorithmic techniques capable of achieving provable speedups over existing classical exact algorithms in large scale genomic analyses. This research utilizes PhiX174, SARS-CoV-2, and human genome data to explore quantum algorithms and data encoding techniques to pave the way for the analysis with better time and space efficiency.

We take a two-pronged approach:

Algorithm Development: We will design novel quantum algorithms for MSA subproblems and heuristic methods (QAOA) for de novo assembly.

Data Encoding and State Preparation: We develop efficient quantum circuits to encode genomic data and reduce the computational overhead with a variety of techniques, including tensor network. It facilitates data encoding into quantum states for Machine Learning applications.

Starting with the PhiX174 genome, we will test our quantum algorithms and prove their speedup compared to classical methods. This allows us to scale the approach to larger and more complex genomes like SARS-CoV-2 and the human genome. We'll develop efficient encoding strategies and optimize quantum circuits to minimize resource needs for the current hardware. We are using advanced tensor network contraction methods for more efficient simulation of the circuits.

This project paves the way for utilizing quantum computing to unlock the vast potential of genomics in healthcare. By overcoming the computational blockage in the classical approach, we aim to achieve a deeper understanding of human health and pathogens.

Keywords: tensor network, genomic analysis, quantum algorithms, data encoding, state preparation

Kevin Berg (University of Regensburg), Lygeri Sakkelaridi (University of Regensburg), Teresa Rummel (University of Regensburg) and Florian Erhard (University of Regensburg). *Identifying host cell factors and pathway modulators leveraging cellular heterogeneity.*

Abstract. The cellular response to challenges like a virus infection or drug treatment is influenced by a wide variety of host factors and the naturally occurring heterogeneity among genetically identical cells can drastically affect the resulting cellular state. Understanding the determinants of such outcomes is pivotal for comprehending those processes. Traditional methods for studying these phenomena often rely on complete gene knock-outs or siRNA-mediated knock-downs resulting in non-physiological conditions or may cause cellular stress and activation of pattern recognition pathways, respectively.

Here, we propose Heterogeneity sequencing (Heterogeneity-seq) as an alternative approach to circumvent these disadvantages by exploiting the inter-cellular heterogeneity at the time of perturbation to identify cellular states and genes that are either beneficial or unfavorable for the resulting cellular response. To determine the transcriptional state of an individual cell before and after perturbation, we use scSLAM-seq, a method that combines metabolic labeling with scRNA-seq to distinguish unlabeled, “old” RNA from newly synthesized, labeled RNA. This temporal resolution allows us to detect potential pathway modulators by correlating gene expression before perturbation with the resulting response level afterwards. To validate our approach on a well-studied pathway, we analyzed a glucocorticoid treated time-series dataset and found both long known and recently identified factors that modulate the strength of the glucocorticoid response. Our findings strongly suggest that Heterogeneity-seq is capable of identifying promising pathway modulators in scSLAM-seq experiments by taking advantage of intercellular heterogeneity and without additional experimental work load.

Keywords: scRNA-seq, SLAM-seq, scSLAM-seq, intercellular heterogeneity

Tanya Karagiannis (Tufts Medical Center), Ye Chen (Tufts Medical Center), Sarah Bald (Boston University), Albert Tai (Tufts University), Sofiya Milman (Albert Einstein College of Medicine), Stacy Anderson (Boston University School of Medicine), Thomas Perls (Boston University School of Medicine), Daniel Segre (Boston University), Paola Sebastiani (Tufts Medical Center) and Meghan Short (Tufts Medical Center).
Parallel systems comparison of metagenomics data of age and longevity .

Abstract. Shotgun metagenomics sequencing of gut microbial samples plays a crucial role in understanding health and disease, notably age-related changes. Various well-validated methods exist for processing sequence data into taxonomic profiles, including marker-gene-based (e.g., MetaPhlAn) and k-mer based (e.g. Kraken, Bracken) approaches. Despite the availability of many tools, often studies rely on a single taxonomic profiler. As previous studies document, substantial differences between classification approaches can impact taxa identified and thus the reported taxonomic profiles. Moreover, variations in database selection, preprocessing, and quality control procedures can result in misclassification and false positive outcomes. To highlight and examine these challenges, we present a case study in which we applied two popular taxonomic profiling methods on stool metagenomics samples from two cohorts of aging. We applied identical pre-processing, quality control steps, and ran MetaPhlAn4 and Kraken2 in parallel. Species numbers differed between MetaPhlAn4 (36,822 species) and Kraken2 (17,572 taxa) databases. Kraken2 identified more species (1591-2332) compared to MetaPhlAn4 (787-898) in both cohorts, which matches findings from previous benchmarking studies. Notably, there was limited overlap in the species identified by the two profilers within each cohort (114-245). Classification differences influenced downstream analyses such as alpha diversity, with MetaPhlAn4 producing results concordant with previous longevity studies, and Kraken2 exhibiting higher sensitivity to the study population. This analysis suggests that each method captures unique aspects of the data, emphasizing the value of employing multiple profilers for comprehensive analysis and the need for approaches to facilitate meaningful integration of results generated from different profilers and study populations.

Keywords: shotgun metagenomics sequencing, taxonomic profiling and analysis, aging

Chisato Ishiwata (Dept. Environ. Info. Stud., Keio Univ.; Inst. Adv. Biosci., Keio Univ.), Phillip Yamamoto (Syst. Biol. Prog. Grad. Sch. Media & Governance, Keio Univ.; Inst. Adv. Biosci., Keio Univ.), Hiroyuki Nakamura (Spiber Inc.), Akio Tanikawa (Lab. Biodiv. Sci., Fac. Agric., Univ. Tokyo.) and Nobuaki Kono (Grad. Sch. Media & Governance, Keio Univ.; Dept. Environ. Info. Stud., Keio Univ.; Inst. Adv. Biosci., Keio Univ.). *Determination of Mitochondrial genome, search for spider silk candidate gene of Heptathela kimurai and Large-scale phylogenetic analysis of spiders.*

Abstract. Spiders exhibit significant potential for biomaterial applications due to their ability to produce up to seven different types of silk, each with unique mechanical properties. Previous research has shown that ancestral spiders did not construct aerial webs and produced a limited range of silk types. Considering these insights, unraveling the evolutionary lineage of spider silk constitutes a fundamental theme in arachnology. However, the scarcity of molecular information on ancestral spiders has impeded research into the evolutionary lineage of the Araneae.

Therefore, in this study, we performed a whole genome sequencing of the ancestral spider Heptathela kimurai from the family Heptathelidae. We successfully sequenced the entire mitochondrial genome, identified candidate silk gene sequences, and compiled molecular information. The silk gene sequence of H. kimurai was estimated through a multi-omics approach, considering both nucleotide sequence and proteome analysis.

We then created a phylogenetic tree with other spider species and clarified the evolutionary position of the ancestral spider and the major spider lineage. The candidate silk gene sequences of H. kimurai suggest a high likelihood of phylogenetic affinity with genes found in scorpions, which are positioned upstream within the arachnid order.

In the future, we will study the points where the types of silk produced by spiders diverged into seven types, their evolutionary history and ecological significance, and the web-building behavior of spiders, using the revealed large spider lineage and the evolutionary system of spider silk genes.

Keywords: Spider, Massively parallel sequencer, Long-read sequencer, Multiomics approach

Nicolas Gustavo Gaitan Gomez (Universidad de los Andes) and Jorge Duitama (Universidad de los Andes).
A graph clustering algorithm for detection and genotyping of structural variants from long reads.

Abstract. Structural variants (SVs) are genomic polymorphisms defined by their length (>50 bp). The usual types of SVs are deletions, insertions, translocations, inversions, and copy number variants. SV detection and genotyping is fundamental given the role of SVs in phenomena such as phenotypic variation and evolutionary events. Thus, methods to identify SVs using long-read sequencing data have been recently developed.

We present an accurate and efficient algorithm to predict germline SVs from long-read sequencing data. The algorithm starts collecting evidence (signatures) of SVs from read alignments. Then, signatures are clustered based on a Euclidean graph with coordinates calculated from lengths and genomic positions. Clustering is performed by the DBSCAN algorithm, which provides the advantage of delimiting clusters with high resolution. Clusters are transformed into SVs and a Bayesian model allows to precisely genotype SVs based on their supporting evidence. This algorithm is integrated into the single sample variants detector of the Next Generation Sequencing Experience Platform, which facilitates the integration with other functionalities for genomics analysis. We performed multiple benchmark experiments, including simulation and real data, representing different genome profiles, sequencing technologies (PacBio HiFi, ONT), and read depths.

The results show that our approach outperformed state-of-the-art tools on germline SV calling and genotyping, especially at low depths, and in error-prone repetitive regions. We believe this work significantly contributes to the development of bioinformatic strategies to maximize the use of long-read sequencing technologies.

Keywords: structural variants, bioinformatics, genotyping, graph algorithms, genomics

Maryam Ghareghani (Freie Universität Berlin, Max Planck institute for molecular genetics), Lion Ward Al Raei (Freie Universität Berlin, Max Planck institute for molecular genetics), Hossein Moeinzadeh (Max Planck institute for molecular genetics), Nico Alavi (Freie Universität Berlin, Max Planck institute for molecular genetics), Jakob Hertzberg (Max Planck institute for molecular genetics) and Martin Vingron (Freie Universität Berlin, Max Planck institute for molecular genetics). *Genotyping tandem repeats: Introducing TandemTwister, a rapid and universal tool for long-read sequencing technologies.*

Abstract. Tandem repeats are segments of the genome sequence consisting of consecutively repeated units with polymorphic copy numbers and mutations in repeat units. They have been used in DNA fingerprinting for their hypervariability in the population. They have also been associated with complex traits and various genetic disorders, including neurodegenerative disease and developmental disorders. Having a fast and scalable tool for accurately genotyping these variants is indeed essential for understanding their functional impact, phenotype, and disease association.

We developed TandemTwister, a novel algorithm written in C++, as a highly scalable and parallelized tool for the genotyping of copy numbers in tandem repeats. We evaluated the performance of TandemTwister on different sequencing technologies in the Ashkenazi trio on a set of ~900k tandem repeat region annotations. For each Hifi sample, TandemTwister ran in ~5 minutes on 32 CPU cores, with 99,5% recall and 93.1% Mendelian consistency and 99.1% accuracy considering an off-by-one error.

In addition, we genotyped tandem repeat copy numbers of the HGSVC2 population cohort including 70 assembled haplotypes of 35 samples from different populations, and analysed population clustering and inheritance patterns of tandem repeats and haplotype blocks in the three trio sets.

Furthermore, we validated the ability of TamdemTwister to detect pathogenic repeat expansions in a cohort of 31 samples with neurodegenerative and developmental disorders.

TandemTwister is an accurate scalable tool for genotyping tandem repeats. It is the fastest genotyping tool to date and suited for all long-read sequencing technologies as well as assembled genomes.

Keywords: Tandem-repeats, genotyping, long-read sequencing

Michael Hall (University of Melbourne), Ryan Wick (University of Melbourne), Louise Judd (University of Melbourne), An Nguyen (University of Melbourne), Eike Steinig (University of Melbourne), Ouli Xie (University of Melbourne), Mark Davies (University of Melbourne), Torsten Seemann (University of Melbourne), Timothy Stinear (University of Melbourne) and Lachlan Coin (University of Melbourne).
Benchmarking reveals superiority of deep learning variant callers on bacterial nanopore sequence data.

Abstract. Variant calling is fundamental in bacterial genomics, underpinning the identification of disease transmission clusters, the construction of phylogenetic trees, and antimicrobial resistance prediction. This study presents a comprehensive benchmarking of SNP and indel variant calling accuracy across 14 diverse bacterial species using Oxford Nanopore Technologies (ONT) and Illumina sequencing. We generate gold standard reference genomes and project variations from closely related strains onto them, creating biologically realistic distributions of SNPs and indels.

Our results demonstrate that ONT variant calls from deep learning-based tools delivered higher SNP and indel accuracy than traditional methods and Illumina, with Clair3 providing the most accurate results overall. We investigate the causes of missed and false calls, highlighting the limitations inherent in short reads and discover that ONT's traditional limitations with homopolymer-induced indel errors are absent with high-accuracy basecalling models and deep learning-based variant calls. Furthermore, our findings on the impact of read depth on variant calling offer valuable insights for sequencing projects with limited resources, showing that 10x depth is sufficient to achieve variant calls that match or exceed Illumina.

In conclusion, our research highlights the superior accuracy of deep learning tools in SNP and indel detection with ONT sequencing, challenging the primacy of short-read sequencing. The reduction of systematic errors and the ability to attain high accuracy at lower read depths enhance the viability of ONT for widespread use in clinical and public health bacterial genomics.

Keywords: variant calling, bioinformatics, benchmark, nanopore, bacterial genomics

Sophie-Marie Wind (Institute of Medical Informatics - University of Muenster), Julian Varghese (Institute of Medical Informatics - University of Muenster) and Carolin Walter (Institute of Medical Informatics - University of Muenster). *Benchmarking of 4C-seq algorithm combinations on published an simulated data.*

Abstract. The spatial organization of chromatin plays a crucial role in gene regulation and is associated with cancer development and progression. Since chromatin modifications are reversible, a deeper understanding of these might be used for the development of new cancer therapies.

Circular chromosome conformation capture sequencing (4C-seq) is a cutting-edge next-generation sequencing technique that can uncover the three-dimensional structure of selected genomic loci at high resolution, enabling the identification of chromatin loops between genes and regulatory elements. The bioinformatic analysis of these data is highly complex and prone to bias. Benchmarking of available 4C-seq algorithms has revealed that none of them performed adequately for all use cases and that different algorithms should be used for an optimized analysis.

To uncover how potential synergies between different algorithms can be optimally exploited, we present a comprehensive benchmarking study examining different combination strategies of existing 4C-seq algorithms. The benchmarking is based on simulated datasets and already published datasets with validated 4C-seq interactions. The algorithms are combined in various constellations using different approaches such as union, intersection and majority vote. The performance of these combination strategies is then thoroughly evaluated and compared using key metrics such as precision, recall and F1 score.

A union approach with fourSig and r3C-seq showed a generally high recall for all datasets considered, while a majority vote algorithm based on r3Cseq, peakC and fourSig generally indicated higher precision and a decent F1 score.

Combined algorithms have benefits over individual algorithms and therefore seem promising for the optimization of 4C-seq analysis.

Keywords: next generation sequencing, 4C-seq, chromatin conformation capture, benchmarking, performance evaluation

Pavel Sumazin (Baylor College of Medicine) and Mohammad Panah (Baylor College of Medicine).
Effective methods for bulk RNA-seq deconvolution using scnRNA-seq transcriptomes .

Abstract. . RNA profiling technologies at single-cell resolutions, including single-cell and single-nuclei RNA sequencing (scRNA-Seq and snRNA-Seq, scnRNA-Seq for short), can help characterize the composition of tissues and reveal cells that influence key healthy and disease functions. However, the use of these technologies is challenging because of their relatively high costs and exacting sample collection requirements. Computational deconvolution methods that infer the composition of bulk-profiled samples using scnRNA-Seq-characterized cell types can broaden the applications of scnRNA-Seq, but their effectiveness remains controversial. We produced the first systematic evaluation of deconvolution methods on datasets with either known or scnRNA-Seq-estimated compositions. Our analyses revealed biases that are common to scnRNA-Seq 10X Genomics assays and illustrated the importance of accurate and properly controlled data preprocessing and method selection and optimization. Moreover, our results suggested that concurrent RNA-Seq and scnRNA-Seq profiles can help improve the accuracy of both scnRNA-Seq preprocessing and the deconvolution methods that employ them. Indeed, our proposed method, Single-cell RNA Quantity Informed Deconvolution (SQUID), combined RNA-Seq transformation and a dampened weighted least squares deconvolution approach to consistently outperform other methods in predicting the composition of cell mixtures and tissue samples. Moreover, our analysis suggested that only SQUID could identify outcomes-predictive cancer cell subtypes in pediatric acute myeloid leukemia and neuroblastoma datasets.

Keywords: snRNA-seq, scRNA-seq, RNA-seq, deconvolution

Jonathan Mandl (Bar-Ilan University), Marcus Bluestone (Massachusetts Institute of Technology), Scott Longwell (Stanford University), Polly Fordyce (Stanford University) and Yaron Orenstein (Bar-Ilan University). *PrimerDesigner: Designing efficient primers for protein synthesis with no cross-hybridization risk.*

Abstract. Efficient protein synthesis of large protein libraries is key to making high-throughput biochemistry scalable. A recent breakthrough in protein synthesis utilized microarrays containing more than 170,000 short oligonucleotides to dramatically reduce costs by harvesting and then hybridizing oligonucleotides to long coding sequences. Effective hybridization of short oligonucleotides to long sequences requires efficient primer design with no cross-hybridization (Figure 1). However, existing primer-design methods for complete coverage of a coding sequence with maximum efficiency and no cross-hybridization use suboptimal heuristics, are restricted by license fees, and do not optimize globally over multiple proteins.

Here, we developed PrimerDesigner to design the most efficient primers for large protein libraries with complete coding-sequence coverage without cross-hybridization. We first proved that primer design with no-cross hybridization is NP-hard, even for a single protein. Consequently, we defined the primer-design problem as an Integer Linear Programming (Figure 2) and implemented an efficient dynamic-programming-based algorithm to search for cross-hybridization risks within a protein and across proteins. Additionally, we relaxed the problem of designing primers for a single protein or multiple variants of the same protein by assuming the cross-hybridization risk in a single protein is negligible. PrimerDesigner produced an optimal solution in only a few minutes for a single protein, and in a few hours for multiple proteins (Figures 3,4). We expect PrimerDesigner to greatly improve protein synthesis efficiency without cross-hybridization risks. Such synthesis will advance protein research through high-throughput screening and enable the design of proteins with improved functions for therapeutics and other applications.

Keywords: Primer design, NP-hard, Integer linear programming

Rongting Huang (The University of Hong Kong), Xianjie Huang (HKU) and Yuanhua Huang (University of Hong Kong). *Comprehensive benchmarking of copy number variations analysis methods for single-cell and spatial transcriptomic data.*

Abstract. Somatic copy number variations (CNVs) play a critical role in the genetic landscape of cancer diseases and are increasingly analyzed by single-cell RNA sequencing (scRNA-seq) data for dissecting cellular heterogeneity at both genetic and transcriptomic levels. A few computational methods have been proposed to detect CNVs from scRNA-seq data, some even achieving haplotype specificity. However, there is an urgent demand of a systematic benchmarking of these CNV detection methods across different tumor complexity, sequencing setting and source of ground truth. Moreover, there is no assessment of their potential applicability and performance in spatial transcriptomics (ST), which is a revolutionized tool for understanding the tumor-immune co-evolution within a microenvironment. Therefore, we conducted a comprehensive benchmarking study over five prominent methods (inferCNV, CopyKAT, CaSpER, Numbat, and XClone) for CNV detection in scRNA-seq data with a perspective towards their extension into ST. To support examining diverse settings, we developed a simulator scCNASimulator, a unique tool support reads synthesis for allelic specific CNV simulation. Also, we compiled a collection of real-world scRNA-seq and ST data sets covering multiple cancer types with gold-standard or high-quality CNV annotations. These scenarios include varying levels of sequencing depth, cell population heterogeneity, with/without allele-specific CNV types (copy gain, copy loss, Loss of heterozygosity), and potential whole genome duplications. Our benchmarking system with such large amount of data hence provides realistic assessments of each method's accuracy, sensitivity, and specificity, offering a detailed guideline for users to select tools for their sample characteristics, data properties, and scientific demands.

Keywords: Somatic copy number variations, benchmarking, scRNA-seq, spatial transcriptomics

Limeng Pu (Computational Biology, St. Jude Children's Research Hospital), Karol Szlachta (Center for Applied Bioinformatics, St. Jude Children's Research Hospital), Virginia Valentine (Cytogenetics, St. Jude Children's Research Hospital), Xiaolong Chen (Pediatric Translational Medicine Institute, Shanghai Children's Medical Center), Jian Wang (Computational Biology, St. Jude Children's Research Hospital), Dennis Kennetz (DSRIAB Engineering, Roche), Daniel Putnam (Computational Biology, St. Jude Children's Research Hospital), Sivaraman Natarajan (Computational Biology, St. Jude Children's Research Hospital), Li Dong (Computational Biology, St. Jude Children's Research Hospital), Thomas Look (Department of Pediatric Oncology, Dana-Farber Cancer Institute, Harvard Medical School), Marcin Wlodarski (Hematology, St. Jude Children's Research Hospital), Lu Wang (Pathology, St. Jude Children's Research Hospital), Steven Burden (Cytogenetics, St. Jude Children's Research Hospital), John Easton (Computational Biology, St. Jude Children's Research Hospital), Xiang Chen (Computational Biology, St. Jude Children's Research Hospital) and Jinghui Zhang (Computational Biology, St. Jude Children's Research Hospital). *Seq2Karyotype (S2K): A Method for Deconvoluting Heterogeneity of Copy Number Alterations Using Single-Sample Whole-Genome Sequencing Data.*

Abstract. The Seq2Karyotype (S2K) algorithm is developed for in-silico karyotyping using single-sample whole-genome sequencing (WGS) data, addressing limitations in current methods for detecting copy number alterations (CNAs) in cancer diagnostics. Traditional approaches using cytogenetic imaging offer detailed chromosomal band resolution, beneficial for assessing tumor heterogeneity but lack precision in locus-specific mapping. S2K improves upon this by modeling coverage and allelic imbalance in high-quality heterozygous SNPs to establish reference diploid regions, segmenting deviations, and fitting empirical data to models representing various CNAs to estimate clonality.

To validate S2K, analyses were performed on benchmark cell lines COLO829 (melanoma) and HCT1395 (breast cancer), with comparisons to single-cell WGS and additional validation through methods like FISH and spectral karyotyping. Results showed S2K could effectively replicate known populations and detect new CNAs, albeit with some differences in clonality estimates compared to single-cell analyses. Further application of S2K to 17 neuroblastoma cell lines, 24 pediatric AML samples, and two MDS cases demonstrated its utility in identifying both known cytogenetic events and previously undetected phenomena like copy-neutral loss-of-heterozygosity and mosaic uniparental disomy (UPD). Notably, S2K detected significant intra-tumor heterogeneity, indicating its potential impact on the design and interpretation of cancer research and treatment strategies.

Overall, S2K offers a robust tool for analyzing CNAs and tumor heterogeneity using WGS data, providing insights that are critical for advancing cancer diagnostics and personalized medicine.

Keywords: whole genome sequence, next-generation sequencing, copy number alterations, karyotyping, tumor heterogeneity, clonality

Taha Shahroodi (TU Delft, ETH Zurich), Gagandeep Singh (AMD Research), Mahdi Zahedi (TU Delft), Haiyu Mao (ETH Zurich), Joel Lindegger (ETH Zurich), Can Firtina (ETH Zurich), Stephan Wong (TU Delft), Said Hamdioui (TU Delft) and Onur Mutlu (ETH Zurich). *Swordfish: A Framework for Evaluating Deep Neural Network-based Basecalling using Computation-In-Memory with Non-Ideal Memristors.*

Abstract. Basecalling, an essential step in many genome analysis studies, relies on large Deep Neural Networks (DNNs) to achieve high accuracy. Unfortunately, these DNNs are computationally slow and inefficient, leading to considerable delays and resource constraints in the sequence analysis process. A Computation-In-Memory (CIM) architecture using memristors can significantly accelerate the performance of DNNs. However, inherent device non-idealities and architectural limitations of such designs can greatly degrade the basecalling accuracy, which is critical for accurate genome analysis. To facilitate the adoption of memristor-based CIM designs for basecalling, it is important to (1) conduct a comprehensive analysis of potential CIM architectures and (2) develop effective strategies for mitigating the possible adverse effects of inherent device non-idealities and architectural limitations.

This paper proposes Swordfish, a novel hardware/software co-design framework that can effectively address the two aforementioned issues. Swordfish incorporates seven circuit and device restrictions or non-idealities from characterized real memristor-based chips. Swordfish leverages various hardware/software co-design solutions to mitigate the basecalling accuracy loss due to such non-idealities. To demonstrate the effectiveness of Swordfish, we take Bonito, the state-of-the-art (i.e., accurate and fast), open-source basecaller, as a case study. Our experimental results using Swordfish show that a CIM architecture can realistically accelerate Bonito for a wide range of real datasets by an average of 25.7×, with an accuracy loss of 6.01\%.

Keywords: basecalling, computation in memory (CIM), processing in memory (PIM), memristors, non-ideality

Timofey Prodanov (Institute for Medical Biometry and Bioinformatics, Heinrich Heine University, 40225 Düsseldorf, Germany) and Tobias Marschall (Institute for Medical Biometry and Bioinformatics, Heinrich Heine University, 40225 Düsseldorf, Germany). *Targeted genotyping of complex polymorphic genes using short and long reads.*

Abstract. The human genome contains numerous highly polymorphic loci, rich in tandem repeats and structural variants. There, read alignments are often ambiguous and unreliable, resulting in hundreds of disease-associated genes being inaccessible for accurate variant calling. In such regions, structural variant callers show limited sensitivity, k-mer based tools cannot exploit full linkage information of a sequencing read, and gene-specific methods cannot be easily extended to process more loci. Improved ability to genotype highly polymorphic genes can increase diagnostic power and uncover novel disease associations.

We present a targeted tool Locityper, capable of genotyping complex polymorphic loci using both short- and long-read whole genome sequencing, including error-prone ONT data. For each target, Locityper recruits WGS reads and aligns them to possible locus haplotypes (e.g. extracted from a pangenome). By optimizing read alignment, insert size, and read depth profiles across haplotypes, Locityper efficiently estimates the likelihood of each haplotype pair. This is achieved by solving integer linear programming problems or by employing stochastic optimization.

Across 256 challenging medically relevant loci and 40 HPRC Illumina datasets, 95% Locityper haplotypes were accurate (QV, Phred-scaled divergence, ≥33), compared to 27% accurate haplotypes, reconstructed from the phased NYGC call set. In leave-one-out (LOO) evaluation, Locityper produced 60% accurate haplotypes, a fraction that will increase with larger reference panels as >91% haplotypes were very close (ΔQV≤5) to best available haplotypes. Overall, 82% 1KGP trio haplotypes were concordant. Finally, across 36 HLA genes LOO Locityper correctly predicted protein product in 94% cases, outperforming the specialized HLA-genotyper T1K at 78%.

Keywords: genotyping, challenging genes, pangenome

Hugo Magalhães (Institute for Medical Biometry and Bioinformatics, Medical Faculty, and Center for Digital Medicine, HHU, Düsseldorf), Timofey Prodanov (Institute for Medical Biometry and Bioinformatics, Medical Faculty, and Center for Digital Medicine, HHU, Düsseldorf), Jonas Weber (Institute of Medical Microbiology and Hospital Hygiene, HHU, Düsseldorf), Gunnar Klau (Algorithmic Bioinformatics, HHU, Düsseldorf) and Tobias Marschall (Institute for Medical Biometry and Bioinformatics, Medical Faculty, and Center for Digital Medicine, HHU, Düsseldorf). *Sequence-to-graph alignment based copy number calling using a flow network formulation.*

Abstract. Variation of copy number (CN) between individuals has been associated with phenotypic differences. Consequently, CN calling is an important step for disease association and identification, as well as in genome assembly. Traditionally, sequencing reads are mapped to a linear reference genome, after which CN is estimated based on observed read depth. This approach, however, leads to inconsistent CN assignments and is hampered by sequences not represented in a linear reference. To address this issue, we propose a method for CN calling with respect to a graph genome using a flow network formulation.

The tool processes read alignments to any bidirected genome graph, and calculates CN probabilities for every node according to the Negative Binomial distribution and total base pair coverage across the node. Integer linear programming is then employed to find a maximum likelihood flow through the graph, resulting in CN predictions for each node. This way, the method achieves consistent CN assignments across the graph.

The proposed method is capable of processing a wide variety of input graphs and read mappings from different sequencing technologies. We processed reads aligned to a Verkko assembly graph for HG02492 (HGSVC) using high coverage mixed HiFi and ONT-UL reads in under 2 hours using one thread and <2Gb peak memory. For 18% nodes, the method produced different CN values than those expected from read depth alone, showcasing how the graph topology informs CN assignment. Further applications include CN assignment as part of diploid/polyploid (pan)genome assembly workflows.

Keywords: graph, flow network, copy number, pangenomes

Samantha Yuen (Maisonneuve-Rosemont Hospital Research Center, Department of Medicine, University of Montreal, Montreal, QC, Canada), Nicolas Paradis-Isler (Maisonneuve-Rosemont Hospital Research Center, Department of Medicine, University of Montreal, Montreal, QC, Canada) and Yoshiaki Tanaka (Maisonneuve-Rosemont Hospital Research Center, Department of Medicine, University of Montreal, Montreal, QC, Canada). *Uncovering the Role of Retrotransposons in Glioblastoma Multiforme Subtype Transitions.*

Abstract. Glioblastoma multiforme (GBM) is a deadly adult brain tumor with a one-year survival rate upon diagnosis. It is classified into three major subtypes: proneural, mesenchymal, and classical, which display different genetic abnormalities, clinical outcomes, and therapeutic responses. A tumor comprises multiple subtypes which transition to others upon GBM treatment. However, it remains unclear how transitioning to poor-outcome subtypes results in therapeutic failure.

We seek to reveal molecular characteristics of GBM subtype transition in the context of both genes and transposable elements (TEs). The latter insert copies of themselves throughout the genome and are abnormally expressed in cancer stem cells. To better understand their roles in different subtypes of GBM, we characterized them using publicly available GBM transcriptome datasets in a high-throughput fashion. We then infer key genes contributing to the transition between GBM subtypes by developing a probabilistic and explainable hidden Markov model trained on sequences of ordered single cell gene expression profiles.

Our analysis revealed three isolated clusters of proneural GBM. Upon further investigation, one proneural subgroup was enriched in human-specific TEs such as SINE-VNTR-Alu. Gene ontology of this subgroup showed enrichment in cancer-related terms such as stress response, proliferation, and hypoxia. We demonstrate that including TE information with human transcriptome profiles allows us to capture potential features not present in other mammalian models. Intratumor heterogeneity underlies the complexity of GBM research, therefore, uncovering molecular features in subtype transitions helps to understand and avoid therapies that transform tumors into more aggressive subtypes.

Keywords: scRNA-seq, machine learning, oncology, personalized medicine, mobile genetic elements

Yanis Asloudj (LaBRI), Fleur Mougin (Bordeaux Population Health) and Patricia Thebault (LaBRI). *Embrace the bias: extrinsic variability in scRNA-seq clustering analyseis is informative.*

Abstract. Clustering analyses are fundamental in single-cell data science. Hundreds of methods have been developed to conduct this analysis, but they all generate different results. Benchmarks and reviews make this issue obvious, and they show that no single clustering method outperforms all the others. To generate clustering results robust to the method used, scRNA-seq ensemble clustering algorithms are developed.

Existing ensemble algorithms tackle this issue by minimizing the differences between multiple clustering solutions.

In our work, we investigate a novel alternative approach. We name ”extrinsic variability” the differences across clustering solutions that are due to methodological choices. Unlike the state of the art, we hypothesize that the extrinsic variability is not to be minimized, but rather leveraged to prevent over-clustering.

To verify our hypothesis, we have developed scEVE, an algorithm that embraces this approach. We apply it on a human glioblastoma scRNA-seq dataset, and we compare its performance to three state-of-the-art ensemble algorithms, on three diverse scRNA-seq datasets.

We present scEVE, and we showcase its functionalities on the public glioblastoma dataset. Incidentally, we reveal the existence of a sub-cluster of cancer cells, that we characterize biologically. Then, we show that scEVE is at worst a middle performer, and at best a top performer, with regards to the existing ensemble

methods.

Overall, our work shows that the extrinsic variability is informative, and we release scEVE, a novel scRNA-seq ensemble clustering algorithm, that also addresses two main challenges in scRNA-seq clustering, by generating a multi-resolution clustering with explicit consensus values.

Keywords: scRNA-seq, ensemble clustering, graph theory, extrinsic variability

Jack Fiore (Systems Genomic Section, Laboratory of Parasitic Diseases, NIAID/NIH, Bethesda, MD, United States), Wei Wang (Systems Genomic Section, Laboratory of Parasitic Diseases, NIAID/NIH, Bethesda, MD, United States), Stephanie Banakis (Systems Genomic Section, Laboratory of Parasitic Diseases, NIAID/NIH, Bethesda, MD, United States), Ted Ross (Center for Vaccines and Immunology, University of Georgia, Athens, GA, United States), Katherine Johnson (Systems Genomic Section, Laboratory of Parasitic Diseases, NIAID/NIH, Bethesda, MD, United States) and Elodie Ghedin (Systems Genomic Section, Laboratory of Parasitic Diseases, NIAID/NIH, Bethesda, MD, United States). *Characterizing age-specific spatial and temporal dynamics of influenza infection in a ferret model.*

Abstract. Influenza virus is a major public health concern that continues to cause significant morbidity and mortality each year. There remain significant gaps in predicting disease severity in high-risk populations distinguished by age, weight, and co-morbidities who represent much of the population's disease burden. Understanding how pathology develops in these vulnerable populations during influenza infection remains crucial to reduce adverse outcomes and fatalities. While influenza is widespread in the human population, studies with human subjects are often hindered by several factors including unclear histories of vaccination and infection, high dropout rates, and limited access to tissue biopsies. Previous work established that different aged ferrets mimic the age-specific infection dynamics observed in human influenza infections. To better understand the age-related factors associated with influenza virus disease severity and susceptibility, we sequenced the miRNA, mRNA, and genomic viral RNA (vRNA) from samples collected longitudinally from young, adult, and aged naïve female ferrets infected with a 2009 pandemic H1N1 influenza A strain (A/California/07/09). We defined age-specific temporal responses to influenza using transcriptomic data collected from the whole blood, upper lung, and lower lung of ferrets to identify predictors of pathology early in infection. Future and ongoing work will focus on using the miRNA and mRNA data from the upper and lower respiratory tissues to define how differences in local microenvironments shape the intra-host evolution of influenza virus.

Keywords: Influenza virus, RNA-Seq, Host-pathogen interactions

Nathalie Bonin (University of Maryland), Adena Collins (University of Maryland) and Mihai Pop (University of Maryland). *A Novel Plasmid Reconstruction and Identification Pipeline for the NCBI Pathogen Detection Database.*

Abstract. Food-borne illnesses cause 128K hospitalizations and $15.5B in economic impact annually. Because of this, the Food and Drug Administration curates genomes of the most common pathogens in the NCBI Pathogen Detection database (NCBI-PD), currently totaling 1.8M isolates. These data are the basis of over 1,220 published research studies and applications to-date on detection and treatment of food-borne illness.

A recent pilot study on a Salmonella outbreak revealed the critical role of plasmids (mobile DNA isolated from the main chromosome) in epidemiological tracking. Despite this, plasmids remain under-used in pathogen identification and tracing because NCBI-PD genomes are assembled through a method which ineffectively captures plasmid sequences. Moreover, prevailing plasmid assembly tools have yet to be benchmarked for large, high-throughput datasets or are unable to distinguish between chromosomal and plasmid sequences.

Our project aims to re-analyze a randomly-chosen subset of raw sequences from the NCBI-PD genomes to include pathogen-associated plasmids. First, we are benchmarking existing plasmid reconstruction methods. Next we are incorporating scalable and reproducible methods for sequence assembly, sorting, and clustering into current plasmid reconstruction methods and databases. Our workflow will annotate and validate the plasmids we assembled before re-linking them to their NCBI-PD bacterial isolates resulting in an improved and robust data set. Upon reaching significant improvements with this preliminary subset, we plan to expand our analysis to the entire database.

Keywords: Assembly, Plasmid, Pathogen, High-throughput

Gagandeep Singh (AMD), Mohammed Alser (ETH Zurich), Kristof Denolf (AMD), Can Firtina (ETH Zurich), Alireza Khodamoradi (AMD), Meryem Banu Cavlak (ETH Zurich), Henk Corporaal (Eindhoven University of Technology) and Onur Mutlu (ETH Zurich). *RUBICON: A Framework for Designing Efficient Deep Learning-Based Genomic Basecallers.*

Abstract. Nanopore sequencing generates noisy electrical signals that need to be converted into a standard string of DNA nucleotide bases using a computational step called basecalling. The performance of basecalling has critical implications for all later steps in genome analysis. Therefore, there is a need to reduce the computation and memory cost of basecalling while maintaining accuracy. We present RUBICON, a framework to develop efficient hardware-optimized basecallers. We demonstrate the effectiveness of RUBICON by developing RUBICALL, the first hardware-optimized mixed-precision basecaller that performs efficient basecalling, outperforming the state-of-the-art basecallers. We believe RUBICON offers a promising path to develop future hardware-optimized basecallers.

Keywords: genomics sequencing, basecalling, hardware acceleration, machine learning, deep neural network

Xiongbin Kang (Bielefeld University), Jialu Xu (College of Biology, Hunan University), Xiao Luo (College of Biology, Hunan University) and Alexander Schoenhuth (Bielefeld University). *Hybrid-hybrid correction of errors in long reads with HERO.*

Abstract. Although generally superior, hybrid approaches for correcting errors in third-generation sequencing (TGS) reads, using next-generation sequencing (NGS) reads, mistake haplotype-specific variants for errors in polyploid and mixed samples. We suggest HERO, as the first “hybrid-hybrid” approach, to make use of both de Bruijn graphs and overlap graphs for optimal catering to the particular strengths of NGS and TGS reads. Extensive benchmarking experiments demonstrate that HERO improves indel and mismatch error rates by on average 65% (27 - 95%) and 20% (4 - 61%). Using HERO prior to genome assembly significantly improves the assemblies in the majority of the relevant categories.

Keywords: Sequencing Error Correction, Third-Generation Sequencing, Hybrid Error Correction, Overlap Graphs, De Bruijn Graphs, Genome Assembly

Huan Huang (University of Surrey) and Tom Thorne (University of Surrey). *Non-parametric distribution clustering to find out gene group behaviours in the treatments targeting glioblastoma.*

Abstract. Glioblastoma multiforme (GBM), a highly aggressive brain tumour, poses significant challenges due to its low survival rates. Leveraging spatial transcriptomic data obtained from GBM and surrounding tissue in mice via 10X Genomics' Visium platform, we aim to unravel gene-specific profiles associated with treatment efficacy.

Many existing data analysis pipelines involve normalizing, dimensionality reduction, and clustering individual "spots" within spatial transcriptomic data using machine learning approaches. However, these methods fall short of capturing gene-specific information. To address this limitation, we propose a novel non-parametric clustering algorithm based on a Bayesian hierarchical model that models the distribution of gene expression within clusters. By employing the Dirichlet Process, we automatically cluster spots according to their marginal probabilities derived from negative binomial models of transcriptomic data.

Our approach applies variational inference, enabling scalability for large datasets, and greatly reducing the computational resources required compared to Markov Chain Monte Carlo (MCMC) methods. Through the resulting clusters, we seek to identify genes directly linked to sub-population specific behaviours within the tumour microenvironment. Our findings hold promise for improved GBM therapies, higher survival rates, and enhanced patient quality of life.

Keywords: Glioblastoma multiforme, non-parametric distribution-based clustering, Bayesian hierarchical model, machine learning, Dirichlet Process, variational inference, spatial transcriptomics

Nayoung Park (Konkuk University), Minji Gu (Konkuk University) and Jaebum Kim (Konkuk University).
ROQ: a new measure for more accurately filtering mapped reads.

Abstract. Mapping reads to a reference genome is a critical step in utilizing next-generation sequencing data. Typically, mapped reads are filtered based on their mapping quality (MAPQ) score, which indicates alignment reliability. However, our simulation-based study, where simulated short reads from a human genome assembly were mapped back to the same genome, demonstrated that high MAPQ scores do not always correspond to accurate mapping of reads to their original genomic regions. The correlation between the MAPQ scores and the overlap proportion of the actual and mapped locations was only 0.6237. This highlights the need for a new measure for more accurately filtering mapped reads. To this end, we introduced a new measure, read overlapping quality (ROQ), designed to quantitatively measure the proportion of overlapping of the genomic region where a read originated and the mapped region of the read, and developed a machine-learning model based on XGBoost to predict the ROQ score given various information of mapped reads. When evaluated using simulation data, our model achieved an MSE of 0.0006 and an R2 of 0.9279 with a correlation of 0.9633 between the ROQ scores and the overlap proportion of the actual and mapped locations. These results suggest that ROQ can serve as a superior measure to MAPQ for filtering mapped reads. The use of ROQ will contribute to a better understanding of genomes by dramatically reducing false-positive interpretation of reads.

Keywords: High-throughput sequencing, Read alignment, Mapping quality, Alignment filtering

Yoshitaka Sakamoto (Division of Genome Analysis Platform Development, National. Cancer Center Research Institute), Masahiro Sugawa (Division of Genome Analysis Platform Development, National Cancer Center Research Institute), Ai Okada (Division of Genome Analysis Platform Development, National Cancer Center Research Institute), Yotaro Ochi (Department of Pathology and Tumor Biology, Graduate School of Medicine, Kyoto University), Yosuke Tanaka (Division of Cellular Signaling, National Cancer Center Research Institute), Yasunori Kogure (Division of Molecular Oncology, National Cancer Center Research Institute), Kenichi Chiba (Division of Genome Analysis Platform Development, National Cancer Center Research Institute), Wataru Nakamura (Division of Genome Analysis Platform Development, National Cancer Center Research Institute), Junji Koya (Division of Hematology, Department of Medicine, Keio University School of Medicine), Hiroyuki Mano (Division of Cellular Signaling, National Cancer Center Research Institute), Seishi Ogawa (Department of Pathology and Tumor Biology, Graduate School of Medicine, Kyoto University), Keisuke Kataoka (Division of Molecular Oncology, National Cancer Center Research Institute) and Yuichi Shiraishi (Division of Genome Analysis Platform Development, National. Cancer Center Research Institute). *Personalized reference genome-based approach can detect structural variants accurately in cancer genomes .*

Abstract. Accurate detection of structural variants (SVs) in cancer genomes is important to understand the mechanisms of tumorigenesis and cancer progression. However, analyzing individually diverse regions in human genomes, such as telomeres and centromeres, is challenging by current human reference genome-based approach. Recently, the combination of long-read sequencing technologies, such as PacBio HiFi (HiFi) and Oxford Nanopore Technologies Ultra-long read sequencing (ONT UL), along with trio or Hi-C data has enabled nearly comprehensive analysis of the human genome spanning from telomere to telomere.

We developed a novel approach based on patient-derived personalized reference genomes to detect somatic SVs accurately. First, we constructed a personalized diploid reference genome by de novo genome assembly with HiFi, ONT UL and Hi-C data from normal samples using Hifiasm. Then, haplotype-aware alignment of tumor and normal data to the personalized reference genome was conducted using haplotype-specific k-mers. Finally, somatic SV calling was performed by Nanomonsv. We also characterized the SVs by converting the breakpoint coordinates to the current human reference genome-based coordinates and annotating the breakpoints by RepeatMasker.

We applied our method to two cancer cell lines (H2009 and HCC1954) and their respective matched-control cell lines. Personalized reference genome-based SV calling results covered more than 80% of the GRCh38-based SV calling results. Furthermore, our approach identified approximately 100 SVs in repeat regions, such as centromeres, which were not detected by the GRCh38-based approach.

In summary, this approach can provide a comprehensive view of the cancer genome structures and provide new insights into cancer genome studies.

Keywords: structural variant, cancer genome, long-read sequencing

Wim Cuypers (University of Antwerp), Julia Gauglitz (University of Antwerp), Halil Ceylan (University of Antwerp), Eline Turcksin (University of Antwerp), Nicky de Vrij (University of Antwerp - Institute of Tropical Medicine Antwerp), Tessa de Block (Institute of Tropical Medicine Antwerp), Koen Vercauteren (Institute of Tropical Medicine Antwerp), Kevin K. Ariën (Institute of Tropical Medicine Antwerp), Philippe Selhorst (Institute of Tropical Medicine Antwerp), Wout Bittremieux (University of Antwerp) and Kris Laukens (University of Antwerp). *SquiDBase: a centralized community resource of microbial paired squiggle-sequence data.*

Abstract. Fast, flexible, and unbiased systems are crucial for advancing microbial monitoring and management. Nanopore sequencing has shown significant potential in this area. Most researchers currently analyze base-called data. However, storing and sharing raw nanopore data, also known as 'squiggles', offers inherent benefits. Maintaining a repository of squiggle data enables updated base-calling as new algorithms are developed, optimization of species-specific base-calling algorithms, and presents a unique resource for benchmarking data processing and analysis approaches.

We have designed SquiDBase, a structured datalake that contains squiggle vectors and their associated metadata. Our data preprocessing pipeline, powered by NextFlow, eliminates human genetic information from the uploaded samples. We will showcase the current status of the database, including the user interface. The latter enables the upload of raw POD5 files as well as structured metadata to promote reusability, and the ability to download data and query within the database. The initial scope of SquiDBase focuses on microbiological data from a variety of sources, including pathogens detected in human blood, such as Plasmodium, but also cultured viral isolates (SARS-CoV-2 variants, Dengue virus, Monkeypox virus, etc.) which we sequenced using the R10.4.1 chemistry. These represent important datasets for community benchmarking. We invite the research community to contribute data to this publicly available resource. Collaborations are welcome to help expand this community tool.

Keywords: Nanopore sequencing, Raw sequencing data, Database, Pathogens

Jens Zentgraf (Saarland University) and Sven Rahmann (Saarland University). *Xengsort2: Ultrafast accurate xenograft sorting.*

Abstract. With an increasing number of patient-derived xenograft (PDX) models being created and subsequently sequenced to study tumor heterogeneity, there is a similarly increasing need for methods to separate reads originating from the graft (human) tumor and reads originating from the host species' (mouse) surrounding tissue.

We present xengsort2, a strongly improved method and tool in comparison to our previous xengsort.

We reduced construction time and memory for the index and introduced a new classification method.

We parallelize the index by splitting the hash table into multiple sub-tables.

Each k-mer is assigned to one specific subtable using a single hash function.

After this, we insert the k-mer into the subtable using Cuckoo hashing.

This allows us to use one thread per subtable without communication between subtables.

Furthermore, we created a producer-consumer method to provide k-mers fast enough to the subtables.

The producer-consumer method can read multiple files in parallel and hand the reads over to a number of independent threads, which split the reads into k-mers.

We developed a new classification strategy, not only based on the number of k-mers from host and graft, but also on the number of bases in each read that are covered by such k-mers.

We show that the Wall time of the index step is reduced by a factor of 5. In addition, the CPU time to classify reads is reduced by a factor of 10 in comparison to alignment based methods

Keywords: k-mers, Cuckoo hashing, xenograft sorting

Alessandro Brandulas Cammarata (University of Lausanne). *Robust inference of gene expression state across cells and cell types.*

Abstract. Single-cell RNA sequencing (scRNA-seq) is revolutionizing our understanding of cellular heterogeneity but is susceptible to technical and biological noise, which can hide true gene expression profiles. We developed a new method that estimates scRNA-seq noise and signal by leveraging reads mapped to selected intergenic regions. We apply an imputation technique to predict the expression counts found in the selected intergenic regions in each cell. We then use expression levels in those regions to assess the background noise found for each cell in a given experiment. Gene expression in individual cells is then determined via testing if each gene is significantly different from the background, classifying genes as expressed (present) or unexpressed (absent). Such method is more stable than the usually used 1 CPM threshold of presence usually used since it adapts to each specific cell noise. The p-values of expression are then aggregated first at the cell type level and subsequently across cell types to provide a comprehensive gene expression profile for entire organs. Validation of our approach against gold-standard datasets in mouse liver, human lung, and drosophila testis demonstrates that our tool achieves comparable accuracy (true positive rate and false positive rate) to traditional bulk RNA sequencing for detecting gene expression state methods, without sacrificing the level of detail of single-cell measurements. These results suggest that our method can effectively reduce noise interference in scRNA-seq, providing a robust platform for gene expression analysis from individual cells and cell types, across diverse biological systems.

Keywords: single-cell, Background noise, Expression profiling, Cell type aggregation

Yunhee Jeong (German Cancer Research (DKFZ)), Clarissa Gerhäuser (German Cancer Research (DKFZ)), Guido Sauter (Universitätsklinikum Hamburg Eppendorf, Institut für Pathologie), Thorsten Schlomm (Charité – Universitätsmedizin Berlin), Karl Rohr (Heidelberg University) and Pavlo Lutsik (Catholic University (KU) Leuven). *MethylBERT: Broadly applicable read-level tumour DNA methylation pattern analysis and tumour purity estimation method using a language model.*

Abstract. DNA methylation (DNAm) is a key epigenetic mark that shows profound alterations in cancer. Read-level methylomes enable more in-depth DNAm analysis due to the broad genomic coverage and preservation of rare cell-type signals, compared to array-based data such as EPIC/450K array. Here, we propose MethylBERT, a novel Transformer-based read-level methylation pattern analysis model. MethylBERT identifies tumour-derived sequence reads by classifying reads based on their methylation patterns and genomic sequence. Based on the classification probability, the method estimates tumour purity within bulk samples.

MethylBERT outperforms existing deconvolution methods in the evaluation using simulated bulks and demonstrates high read classification accuracy regardless of methylation pattern complexity and read length. Moreover, despite a very low circulating tumour DNA (ctDNA) content (<1%), MethylBERT can distinguish colorectal and pancreatic cancer patients at stage Ⅱ-Ⅳ from healthy donors using their blood plasma samples. This highlights its applicability for non-invasive early cancer diagnosis. We also show that MethylBERT can be used for tumour metastasis analysis without tumour reference sequencing data, which is often not available. MethylBERT trained only with normal prostate epithelium accurately estimates tumour-derived read proportions in lymph node samples collected from hormone-sensitive metastatic prostate cancer cases.

Overall, MethylBERT represents a significant advancement in read-level methylome analysis. In particular, due to the classification performed on individual reads, MethylBERT is robust to read depth differences and applicable to all types of bisulfite sequencing (BS-seq) data including whole genome BS-seq, scBS-seq, and targeted BS-seq. In addition, the long-read simulation result shows its potential for nanopore sequencing analyses.

Keywords: Bisulfite sequencing, DNA methylation, Cancer epigenomics, circulating tumour DNA, Language model, Machine learning

Sudhanva Shyam Kamath (Indian Institute of Science, Bangalore), Mehak Bindra (Indian Institute of Science, Bangalore), Debnath Pal (Indian Institute of Science, Bangalore) and Chirag Jain (Indian Institute of Science, Bangalore). *Telomere-to-telomere assembly by preserving contained reads.*

Abstract. Automated telomere-to-telomere (T2T) de novo assembly of diploid and polyploid genomes remains a formidable task. A string graph is a commonly used assembly graph representation in the overlap-based algorithms. The string graph formulation employs graph simplification heuristics, which drastically reduce the count of vertices and edges. One of these heuristics involves removing the reads contained in longer reads. However, this procedure is not guaranteed to be safe. In practice, it occasionally introduces gaps in the assembly by removing all reads covering one or more genome intervals. The factors contributing to such gaps remain poorly understood. In this work, we mathematically derived the frequency of observing a gap near a germline and a somatic heterozygous variant locus. Our analysis shows that (i) an assembly gap due to contained read deletion is an order of magnitude more frequent in Oxford Nanopore reads than PacBio HiFi reads due to differences in their read-length distributions, and (ii) this frequency decreases with an increase in the sequencing depth. Drawing cues from these observations, we addressed the weakness of the string graph formulation by developing the RAFT assembly algorithm. RAFT fragments reads and produces a more uniform read-length distribution. The algorithm retains spanned repeats in the reads during the fragmentation. We empirically demonstrate that RAFT significantly reduces the number of gaps using simulated datasets. Using real Oxford Nanopore and PacBio HiFi datasets of the HG002 human genome, we achieved a twofold increase in the contig NG50 and the number of haplotype-resolved T2T contigs compared to Hifiasm.

Keywords: Genome assembly, Overlap graphs, Phasing, Long reads, Graph connectivity

Atsushi Takeda (Waseda University), Daisuke Nonaka (The University of Tokyo), Yuta Imazu (Waseda), Tsukasa Fukunaga (Waseda) and Michiaki Hamada (Waseda University). *REPrise: de novo interspersed repeat detection using inexact seeding.*

Abstract. Motivation:

Interspersed repeats occupy a large part of many eukaryotic genomes, and thus their accurate annotation is essential for various genome analyses. Database-free de novo repeat detection approaches are powerful for annotating genomes that lack well-curated repeat databases. However, existing tools do not yet have sufficient repeat detection performance.

Results:

In this study, we developed REPrise, a de novo interspersed repeat detection software program based on a seed-and-extension method. Although the algorithm of REPrise is similar to that of RepeatScout, which is currently the de facto standard tool, we incorporated three unique techniques into REPrise: inexact seeding, affine gap scoring and loose masking. Analyses of rice and simulation genome datasets showed that REPrise outperformed RepeatScout in terms of sensitivity, especially when the repeat sequences contained many mutations. Furthermore, when applied to the complete human genome dataset T2T-CHM13, REPrise demonstrated the potential to detect novel repeat sequence families.

Keywords: repeat detection, sequence alignment, transposable elements, algorithm

Julien Faure-Levesque (Université de Sherbrooke), Pierre-Étienne Jacques (Université de Sherbrooke), Sébastien Rodrigue (Université de Sherbrooke), Dominick Matteau (Université de Sherbrooke), Jérémy Gagnon (Université de Sherbrooke) and Audrey Poirier (Université de Sherbrooke). *GENESYS: A Synthetic DNA Design and Assembly Tool.*

Abstract. Background: Shifting from the classical genomic approach, which focuses on dissecting and analyzing individual components of a living system, one could explore an alternative method: reconstructing and synthesizing the genomes of living organisms to understand them in their entirety. This method, inspired by synthetic genomics, combines engineering with molecular biology to engineer DNA and enhance our understanding of life's fundamental principles. Experimental genome reconstruction and editing are challenging and resource-intensive, requiring significant time and investment. Additionally, the absence of advanced bioinformatic tools that can predict optimal DNA assembly solutions further complicates successful genome assembly.

Results: To address these challenges, we started developing the Genome Engineering Synthesis Suite (GENESYS) as a strategic initiative to facilitate the creation of synthetic genomes. GENESYS includes a full suite of genome editing tools: six modules —Mutation, Deletion, Insertion, Recoding, Reorganization, and Refactoring— and a genome construction program, the Constructor. Constructor identifies amplifiable DNA segments from the original genome along with optimal primers, as well as synthetic DNA fragments, and finds the best possible assembly solution using graph theory algorithms. The development of GENESYS has facilitated the design of many synthetic versions of the M13 bacteriophage and the synthetic reconstruction process of the Mesoplasma florum genome.

Conclusions :

GENESYS enhances our ability to engineer DNA and deepens our understanding of life's fundamental principles. Once fully developed, GENESYS will streamline the design and construction of synthetic genomes, enabling the development of organisms for applications like drug synthesis, biofuel production, bioremediation, or carbon fixation.

Keywords: Synthetic Genomics, Genome Reconstruction and Editing, Bioinformatics tool

Johanna Schmitz (Center for Bioinformatics, Saarland Informatics Campus, Saarland University, 66123 Saarbrücken, Germany), Nihit Aggarwal (Department of Genetics, Saarland University , 66123 Saarbrücken, Germany), Lukas Laufer (Department of Genetics, Saarland University , 66123 Saarbrücken, Germany), Jörn Walter (Department of Genetics, Saarland University , 66123 Saarbrücken, Germany), Abdulrahman Salhab (Department of Genetics, Saarland University , 66123 Saarbrücken, Germany) and Sven Rahmann (Center for Bioinformatics, Saarland Informatics Campus, Saarland University, 66123 Saarbrücken, Germany). *EpiSegMix: a flexible distribution hidden Markov model with duration modeling for chromatin state discovery.*

Abstract. Motivation

Automated chromatin segmentation based on ChIP-seq (chromatin immunoprecipitation followed by sequencing) data reveals insights into the epigenetic regulation of chromatin accessibility. Existing segmentation methods are constrained by simplifying modeling assumptions, which may have a negative impact on the segmentation quality.

Results

We introduce EpiSegMix, a novel segmentation method based on a hidden Markov model with flexible read count distribution types and state duration modeling, allowing for a more flexible modeling of both histone signals and segment lengths. In a comparison with existing tools, ChromHMM, Segway, and EpiCSeg, we show that EpiSegMix is more predictive of cell biology, such as gene expression. Its flexible framework enables it to fit an accurate probabilistic model, which has the potential to increase the biological interpretability of chromatin states.

Availability and implementation

Source code: https://gitlab.com/rahmannlab/episegmix.

Published paper

https://doi.org/10.1093/bioinformatics/btae178

Keywords: hidden Markov model, chromatin segmentation, epigenetics, histone modifications, probabilistic modeling

Carolin Walter (Institute of Medical Informatics), Sarah Sandmann (Institute of Medical Informatics) and Julian Varghese (Institute of Medical Informatics). *4CassowRy: 4C-seq-applied scale-space transformation with R/Shiny.*

Abstract. Circular Chromosome conformation capture with sequencing (4C-seq) is a specialized next-generation sequencing technique that allows to explore spatial contacts for genomic loci of interest with high resolution. Since 4C-seq data is complex and prone to technical biases, filtering steps as well as smoothing and window approaches are frequently employed to decrease noise levels and assist in the interpretation of the actual signal. Depending on the 4C-seq interactions' strength and structure, arbitrary parameter choices for these steps may help or hinder the overall analysis, and multi-scale solutions may therefore offer a more comprehensive understanding of the interaction landscape at a chosen locus.

We present 4CassowRy, a 4C-seq-applied scale-space transformation with R/Shiny that allows to conduct and adapt a multi-scale near-cis visualization for 4C-seq data based on Witkin's scale-space filtering. Gaussian smoothing operators with increasing strengths are used to create signal curves that can be analyzed for inflection points and allow to differentiate between more stable 4C-seq features, or “peaks”, and regions that are dominated by background noise. Two-dimensional tesselation maps subsequently enable the user to combine information of multiple smoothing processes, and to assess a multi-scale feature representation for 4C-seq sample data.

Using R/Shiny's flexibility, 4CassowRy includes filter options and adaptive smoothing ranges for interactive near-cis exploration and visualization. Fragment-based import of 4C-seq data is supported, and preprocessing functions as well as export routines are included. For datasets with multiple samples and conditions, a DESeq2-based approach offers information regarding differential interactions. Consequently, 4CassowRy can help to facilitate 4C-seq near-cis analyses.

Keywords: Chromosome conformation capture, 4C-seq, visualization, differential interaction analysis

Clara Inserte (Institute of Medical Informatics, University of Münster, Münster, 48149, Germany), Kornelius Kerl (Department of Pediatric Hematology and Oncology, University Children's Hospital Münster, 48149 Münster, Germany), Julian Varghese (Institute of Medical Informatics, University of Münster, Münster, 48149, Germany) and Sarah Sandmann (Institute of Medical Informatics, University of Münster, Münster, 48149, Germany). *Comparison of integration strategies in fixed RNA profiling for comprehensive tumor analysis.*

Abstract. In cancer research, exploring the similarities and differences between distinct tumor entities is essential to advance our understanding of the disease. For this goal, single-cell analysis has been increasingly used. A recent advancement in this field is fixed RNA profiling, which enhances the scalability of sample processing by enabling extended storage and batch processing.

Bioinformatics analysis of this data requires integration of the samples to eliminate disrupting batch effects. However, the standard approach in single-cell analysis - considering each sample as an independent experiment - may lead to overcorrection when applied to fixed RNA data. Previous studies have evaluated different integration methods for single-cell transcriptomics data using healthy tissue as a benchmark, yet the applicability of these findings to tumor tissue and fixed RNA data remains uncertain.

Comparing three methods – Harmony, CCA and RPCA – and three strategies – considering each sample, batch or a combined annotation of batch and subtype as an independent experiments - on the basis of real fixed RNA data from 44 choroid plexus tumors (3 subtypes) and 33 posterior-fossa ependymomas (3 subtypes), we established a guideline for future integrations of fixed RNA data from different tumor types. Considering batches as independent experiments and using Harmony or RPCA methods yield the best results in terms of batch effect correction while preserving the biological differences between tumor subtypes.

Our comparisons provide valuable insights into the selection of integration strategies, paving the way for enhanced accuracy and reliability of fixed RNA data analysis across diverse tumor types.

Keywords: fixed RNA, integration, cancer, batch correction

Annu Annu (University of Montreal, Montreal, QC, Canada), Levi Adams (RWJMS Institute for Neurological Therapeutics, Rutgers-Robert Wood Johnson Medical School, Piscataway, NJ, USA), Yoon-Seong Kim (RWJMS Institute for Neurological Therapeutics, Rutgers-Robert Wood Johnson Medical School, Piscataway, NJ, USA) and Yoshiaki Tanaka (Maisonneuve-Rosemont Hospital Research Centre, University of Montreal, Montreal, QC, Canada). *Cross-Species Single-Cell Analyses Uncovers Species-Specific Molecular Dynamics During Aging.*

Abstract. Aging is characterized by the loss of biological and physiological functions over time. Brain aging is characterized by memory impairment, decline in cognitive functions and motor coordination. Along with these changes, brain aging is considered as a major risk factor in neurodegenerative diseases, such as Alzheimer's disease and Parkinson's disease. Therefore, a better understanding of molecular changes during aging is essential for aiding or preventing the age-related impairment of brain functions.

With technological advancements, single cell RNA seq(scRNA-seq) studies have been performed on various animal models and human postmortem biospecimens to address the age-related molecular changes. However, the brain is the most different tissues morphologically and functionally across species, and it remains unclear how the age-related molecular changes are conserved. To identify those similarity and differences, we performed an integrative scRNA-seq study on midbrain of mouse and human containing young and old age data. From individual studies, it was observed that oligodendrocytes are the most populated cells in the midbrain of both mouse and human. Through integrative analysis, we found that some clear differences can be seen in various cells of mouse and human data. Additionally, it was found that endothelial and microglial cells have most differently expressed genes compared to other cells. Taken together, our analysis shows the impact of aging on different brain cells in both human and mouse.

Keywords: Aging, scRNA-seq, human, mouse, species-specific, brain

Hufsah Ashraf (Institute for Medical Biometry and Bioinformatics, Medical Faculty and Center for Digital Medicine. HHU, Düsseldorf), Jana Ebler (Institute for Medical Biometry and Bioinformatics, Medical Faculty and Center for Digital Medicine. HHU, Düsseldorf) and Tobias Marschall (Institute for Medical Biometry and Bioinformatics, Medical Faculty and Center for Digital Medicine. HHU, Düsseldorf). *Allele detection using k-mer-based sequencing error profiles.*

Abstract. Correct detection of alleles carried by sequencing reads is vital for variant genotyping and haplotype phasing. In comparison to short-reads, long-reads span larger regions, and hence also access more repetitive regions of the genome. However, long-reads often come with high rates of systematic sequencing errors, which makes the alignments at variant sites unreliable. Thus, allele detection is not a trivial task, especially for single-nucleotide polymorphisms and indels. We propose a new method for allele detection, k-merald, built on the idea that technology-specific sequencing error profiles can provide insights to distinguish a variant allele from a sequencing error, hence improving the allele detection accuracy. k-merald first learns the error profiles by traversing aligned reads in the non-variant regions of the genome. It then employs a k-mer-based alternative to global sequence alignment, aligning strings of consecutive k-mers generated from the respective sequences. Instead of using a fixed cost value, mismatches are penalized using the learned error model, leading to k-mer mismatches representing common sequencing errors being allowed in the alignment at a low cost. We observed that k-merald improves allele detection accuracy leading to better genotyping performance as compared to edit-distance-based allele detection, with a decrease of 18% and 24% in error rate for high-coverage Oxford-Nanopore and PacBio-CLR sequencing reads for sample HG002, respectively. We additionally observed a prominent improvement in genotyping performance for low coverage sequencing data. For 3× coverage Oxford-Nanopore sequencing data, the genotyping error rate reduced from 34% to 31%, corresponding to a 9% decrease.

Keywords: long read sequencing, sequencing error profile, genotyping, k-mer

Mena Kamel (Sanofi Digital. Sanofi Digital has fully funded this project. Authors not connected in any way to 10x Genomics.), Amrut Sarangi (Sanofi Digital. Sanofi Digital has fully funded this project. Authors not connected in any way to 10x Genomics.), Pavel Senin (Sanofi Digital. Sanofi Digital has fully funded this project. Authors not connected in any way to 10x Genomics.), Sergio Villordo (Sanofi Digital. Sanofi Digital has fully funded this project. Authors not connected in any way to 10x Genomics.), Ana Solbas (Sanofi Digital. Sanofi Digital has fully funded this project. Authors not connected in any way to 10x Genomics.), Mathew Sunaal (Sanofi Digital. Sanofi Digital has fully funded this project. Authors not connected in any way to 10x Genomics.), Het Barot (Sanofi Digital. Sanofi Digital has fully funded this project. Authors not connected in any way to 10x Genomics.), Seqian Wang (Sanofi Digital. Sanofi Digital has fully funded this project. Authors not connected in any way to 10x Genomics.), Luis Cano (Sanofi, Precision Medicine & Computational Biology), Ziv Bar-Joseph (Sanofi Digital. Sanofi Digital has fully funded this project. Authors not connected in any way to 10x Genomics.) and Albert Pla Planas (Sanofi Digital. Sanofi Digital has fully funded this project. Authors not connected in any way to 10x Genomics.).

Analysis and Modeling for Visium High Definition (HD) Slides.

Abstract. Spatial transcriptomics (ST) is an evolving field of study that enables researchers to explore the spatial distribution of the genes within tissue. Despite their benefits, ST technologies have different limitations including transcriptome coverage, resolution, and cost. Recently, Visium HD was released, addressing the shortcomings of sequencing-based ST methods by enabling whole transcriptome reads at a (sub-)cellular resolution. In parallel, the field of digital pathology is also evolving with the rapid advancements in computer vision. Many models are already available to segment cells with reasonable generalizability across tissue types. However, inferring cell types without fine tuning the models for every new tissue type is challenging.

To address these issues, we developed the first pipeline combining sub-cellular transcript counts provided by Visium HD and cell segmentation models to infer cell types. The workflow consists of three steps: (1) Cell segmentation, (2) Bin-to-cell assignment, and (3) Cell type inference. Using the predicted cell types, spatial statistical tests can be run to identify cell interactions and distribution patterns around tissue landmarks.

To validate the proposed workflow, four publicly available FFPE Visium HD samples are analyzed. Preliminary results show significant agreement between the predicted cell labels and pathologist-drawn annotations, indicating the viability of the proposed workflow.

This is the first tissue-type agnostic workflow combining cell segmentation and Visium HD output to infer cell types across whole tissue sections. Further studies in this line may open the possibility of generating large scale datasets combining image and gene data without requiring any manual annotation.

Keywords: Spatial Transcriptomics, Cell Segmentation, NGS Analysis

Liudmyla Kondratova (University Of Florida) and Ana Conesa (Institute for Integrative Systems Biology,).

Post-quality control transcript expression levels estimation on long-reads derived transcriptomes.

Abstract. Long-read sequencing technologies, such as PacBio and Oxford Nanopore, enable the generation of transcript sequences spanning full-length transcripts, offering valuable insights into complex isoforms and transcript structures. However, long-read derived transcriptomes are susceptible to artifacts such as sequencing errors, chimeric reads, and incomplete coverage, which can significantly impact transcript quantification accuracy.

SQANTI3 is a comprehensive tool designed to assess qualitative features of long-read transcript models, consisting of three key components: transcript classification, artifact filtering, and rescue. During the artifact filtering step, SQANTI3 leverages transcript junction characteristics, transcript ends, and orthogonal data support to identify artifact-containing transcripts and remove or replace them with suitable reference sequences. Following these steps expression values must be recalculated accordingly.

Here, we introduce a novel method for adjusting long-read RNA-seq transcript expression levels post-quality control. Our approach redistributes reads from transcript models labeled as artifacts to new transcripts based on their splice junction chains and coverage matching. To validate our method, we apply it to samples containing various combinations of kidney and brain mice tissues and estimate tissue-specific expression levels based on their ratios. We show that the combination of a highly curated transcriptome definition, together with a splice-aware quantification strategy significantly improves quantification estimates. This method is aimed to be integrated into the SQANTI3 quality control pipeline, enhancing the precision of transcript expression level estimation.

Keywords: Long reads quantification, Quality control, SQANTI3

Rija Zaidi (University College London Cancer Institute) and Simone Zaccaria (University College London Cancer Institute). *Accurate and robust bootstrap inference of single-cell phylogenies by integrating sequencing read counts.*

Abstract. Recent single-cell DNA sequencing (scDNA-seq) technologies have enabled the parallel investigation of thousands of individual cells. This is required for accurately reconstructing tumour evolution, during which cancer cells acquire a multitude of different genetic alterations. Although the evolutionary analysis of scDNA-seq datasets is complex due to their unique combination of errors and missing data, several methods have been developed to infer single-cell tumour phylogenies by integrating estimates of the false positive and false negative error rates. This integration relies on the assumption that errors are uniformly distributed both within and across cells. However, this assumption does not always hold; error rates depend on sequencing coverage, which is not constant within or across cells in a sequencing experiment due to, e.g., copy-number alterations and the replication status of a cell, limiting the accuracy of existing methods.

To address this challenge, we developed a novel single-cell phylogenetic method that integrates raw sequencing read counts into a statistical framework to robustly correct the errors and missing data. Specifically, our method includes bootstrapping to robustly correct for high error frequency genomic positions and a fast probabilistic heuristic based on hypothesis testing to distinguish the remaining errors from truly observed genotypes. We demonstrate the improved accuracy and robustness of our method compared to existing approaches across several simulation settings. To demonstrate its impact, we applied our method to 42,009 breast cancer cells and 19,905 ovarian cancer cells, revealing more accurate phylogenies consistent with larger genetic alterations.

Keywords: single-cell DNA sequencing, tumour evolution, phylogeny, tumour heterogeneity

Christopher Saunders (Pacific Biosciences), James Holt (Pacific Biosciences), Daniel Baker (Pacific Biosciences), Juniper Lake (Pacific Biosciences), Jonathan Belyeu (Pacific Biosciences), Zev Kronenberg (Pacific Biosciences), William Rowell (Pacific Biosciences) and Michael Eberle (Pacific Biosciences).
Improving long-read structural variant discovery and genotyping with local haplotype modeling.

Abstract. We describe a new structural variant (SV) calling method for mapped high-quality long reads. This method emphasizes assembly of local SV haplotypes and their utilization in all downstream sample merging and genotyping steps, improving accuracy compared to more variant-focused approaches.

Assessing our method against the GIAB draft SV benchmark based on the T2T-HG002-Q100 diploid assembly shows substantial gains in accuracy (precision: 0.981, recall: 0.961) compared to state-of-the-art SV callers such as pbsv (precision: 0.938, recall: 0.916) and Sniffles2 (precision: 0.962, recall: 0.935) on HiFi WGS input. Furthermore, on the GIAB Challenging Medically Relevant Genes benchmark our method reduces the combined false positive and negative SV count to 4, compared to 19 and 15 with pbsv and Sniffles2, respectively.

Joint-genotyping accuracy was evaluated on 10 HiFi WGS samples comprising the 2nd and 3rd generations of the CEPH-1463 pedigree. The known inheritance pattern for this pedigree enables assessment of SV genotype accuracy. From high genotype-quality calls, our method yields 26199 concordant and 5687 discordant SV alleles (82.2% concordance). This is over 6000 more concordant alleles at substantially higher percent concordance compared to the closest method, Sniffles2, with 19922 concordant and 13486 discordant alleles (68.7% concordance).

In addition to improving accuracy, our assembly-focused approach allows all calls to be made with single-base precision, enabling breakpoint insertion and homology annotations for all variant types. Through the combination of high accuracy and detailed breakpoint modeling, we offer improved options for WGS sample analysis with high-quality long reads.

Keywords: Structural Variants, Long-read Sequencing, Variant Calling, Whole Genome Sequencing, Joint Genotyping

Zhezheng Song (The Pennsylvania State University), Tasfia Zahin (The Pennsylvania State University) and Mingfu Shao (The Pennsylvania State University). *Accurate detection of tandem repeat in error-prone long reads.*

Abstract. Tandem repeats in human genomes are well-established markers associated with various diseases. While short tandem repeats (STRs) have been extensively studied, the exploration of long tandem repeats (LTRs) has been limited due to challenges with short-read sequencing technologies, which cannot cover multiple units of a repeat. Conversely, long-read sequencing can cover entire repeat regions but suffers from high error rates, complicating accurate repeat unit reconstruction.

We propose a novel method for determining tandem repeat units from long, error-prone reads. Our approach involves three steps. Initially, we use SubseqHash, our newly developed locality-sensitive hashing (LSH) algorithm designed for error-prone long reads, to identify potential repeat regions. Following this, the repeat region undergoes self-alignment to generate nucleotide-level equivalence classes, which help in constructing kmer-level equivalence classes. Kmers within the same class are expected to derive from identical positions across different units. Finally, a de Bruijn graph (dBG) is constructed, by only gluing equivalent kmers. We then identify a cycle in the dBG whose minimized weight is maximized, representing the sequence of a unit.

The novelty of our approach lies in the use of LSH to accurately pinpoint regions with repeats and in defining equivalent kmers through alignment, effectively handling identical kmers that pose challenges for other methods. Tested on synthetic gene sequences, our method demonstrates high sensitivity and accuracy in detecting tandem repeats, thereby enhancing the study of LTRs and their implications in genetic diseases.

Keywords: Tandem repeats, Locality-sensitive hashing, Sequence alignment, De Bruijn graph

Jonathan Bard (State University of New York at Buffalo), Norma Nowak (State University of New York at Buffalo), Satrajit Sinha (State University of New York at Buffalo) and Michael Buck (State University of New York at Buffalo). *Maximizing accuracy of cellular deconvolution. (ACeD).*

Abstract. Bulk RNA-sequencing has been a mainstay for biomedical research since its inception. In cancer alone, the TCGA project has examined 33 cancer types with over 20,000 samples. Each sample has a wealth of patient information associated with it, from survival records to several data modalities including copy number, microbiome, methylation and transcriptomic profiling at the bulk tissue level. However, the challenge with bulk tissue profiling, like RNA-seq, is that the assay measures the average expression across all the cells in the sample, thus hiding cellular heterogeneity. Leveraging cellular deconvolution, these datasets can be used to infer cell type composition and molecular heterogeneity. However, accurate deconvolution is contingent upon using a high-quality single-cell reference dataset with proper cell-type cluster resolution. Therefore, there is a fundamental need for methodology to quantify single-cell dataset quality for deconvolution with optimization of cell-type cluster resolution. To address this challenge, we developed a novel computational strategy to identify the optimal cell-type clustering resolution that maximizes deconvolutional performance. Our R-based software package (ACeD) provides the research community with a valuable toolset to evaluate reference set quality and optimize data upstream of reference-based deconvolution algorithms, enhancing our analysis and understanding of the tumor microenvironment.

Keywords: genomics, single-cell, deconvolution, sequencing, bioinformatics, oncology, tumor microenviornment

Arvid Gollwitzer (ETH Zurich), Mohammed Alser (ETH Zurich), Joel Bergtholdt (ETH Zurich), Joel Lindegger (ETH Zürich), Maximilian-David Rumpf (ETH Zurich), Can Firtina (ETH Zurich), Serghei Mangul (University of California, Los Angeles) and Onur Mutlu (ETH Zurich & Carnegie Mellon University).
MetaTrinity: Enabling Fast Metagenomic Classification via Seed Counting and Edit Distance Approximation.

Abstract. Metagenomics, the study of genome sequences of diverse organisms cohabiting in a shared environment, has experienced significant advancements across medical and biological fields. Metagenomic analysis is crucial, for instance, in clinical applications such as infectious disease screening and the diagnosis and early detection of diseases such as cancer. A key task in metagenomics is to determine the species present in a sample and their relative abundances. Currently, the field is dominated by either alignment-based tools, which offer high accuracy but are computationally expensive, or alignment-free tools, which are fast but lack the needed accuracy for many applications. In response to this dichotomy, we introduce MetaTrinity, a tool based on heuristics, to achieve a fundamental improvement in accuracy-runtime tradeoff over existing methods. We benchmark MetaTrinity against two leading metagenomic classifiers, each representing different ends of the performance-accuracy spectrum. On one end, Kraken2, a tool optimized for performance, shows modest accuracy yet a rapid runtime. The other end of the spectrum is governed by Metalign, a tool optimized for accuracy. Our evaluations show that MetaTrinity achieves an accuracy comparable to Metalign while gaining a 4x speedup without any loss in accuracy. This directly equates to a fourfold improvement in runtime-accuracy tradeoff. Compared to Kraken2, MetaTrinity requires a 5x longer runtime yet delivers a 17x improvement in accuracy. This demonstrates a 3.4x enhancement in the accuracy-runtime tradeoff for MetaTrinity. This dual comparison positions MetaTrinity as a broadly applicable solution for metagenomic classification, combining advantages of both ends of the spectrum: speed and accuracy.

Keywords: Metagenomics, Microbiome, Taxonomic Classification, Clinical Genomics, Read Mapping, Containment Search

Arvid Gollwitzer (ETH Zurich), Maximilian-David Rumpf (ETH Zurich), Mohammed Alser (ETH Zurich), Joel Lindegger (ETH Zürich), Nour Almadhoun Alserr (ETH Zürich), Can Firtina (ETH Zurich), Serghei Mangul (University of California, Los Angeles) and Onur Mutlu (ETH Zurich & Carnegie Mellon University).
SequenceLab: A Comprehensive Benchmark of Computational Methods for Comparing Genomic Sequences.

Abstract. Computational complexity is a key limitation of genomic analyses. Thus, over the last 30 years, researchers have proposed numerous fast heuristic methods that provide computational relief. Comparing genomic sequences is one of the most fundamental computational steps in most genomic analyses. Due to its high computational complexity, optimized exact and heuristic algorithms are still being developed. We find that these methods are highly sensitive to the underlying data, its quality, and various hyperparameters. Despite their wide use, no in-depth analysis has been performed, potentially falsely discarding genetic sequences from further analysis and unnecessarily inflating computational costs. We provide the first analysis and benchmark of this heterogeneity. We deliver an actionable overview of the 11 most widely used state-of-the-art methods for comparing genomic sequences. We also inform readers about their advantages and downsides using thorough experimental evaluation and different real datasets from all major manufacturers (i.e., Illumina, ONT, and PacBio). SequenceLab is publicly available at https://github.com/CMU-SAFARI/SequenceLab.

Keywords: Genome Analysis, Heuristic Methods, Benchmarking, Genomic Sequence Comparison, Bioinformatics, Metagenomics

Varuni Sarwal (UCLA), Seungmo Lee (UCLA), Jianzhi Yang (USC), Sriram Sankararaman (UCLA), Mark Chaisson (USC), Eleazar Eskin (UCLA) and Serghei Mangul (USC). *VISTA: An integrated framework for structural variant discovery.*

Abstract. Structural variation (SV), refers to insertions, deletions, inversions, and duplications in human genomes. With advances in whole genome sequencing (WGS) technologies, a plethora of SV detection methods have been developed. However, dissecting SVs from WGS data remains a challenge, with the majority of SV detection methods prone to a high false-positive rate, and no existing method able to precisely detect a full range of SVs present in a sample. Here, we report an integrated structural variant calling framework, VISTA (Variant Identification and Structural Variant Analysis) that leverages the results of individual callers using a novel and robust filtering and merging algorithm. In contrast to existing consensus-based tools which ignore the length and coverage, VISTA overcomes this limitation by executing various combinations of top-performing callers based on variant length and genomic coverage to generate SV events with high accuracy. We evaluated the performance of VISTA on comprehensive gold-standard datasets across varying organisms and coverage. We benchmarked VISTA using the Genome-in-a-Bottle (GIAB) gold standard SV set, haplotype-resolved de novo assemblies from The Human Pangenome Reference Consortium (HPRC), along with an in-house PCR-validated mouse gold standard set. VISTA maintained the highest F1 score among top consensus-based tools measured using a comprehensive gold standard across both mouse and human genomes.In conclusion, VISTA represents a significant advancement in structural variant calling, offering a robust and accurate framework that outperforms existing consensus-based tools and sets a new standard for SV detection in genomic research.

Keywords: Bioinformatics, Computational Biology, Machine Learning, Structural Variation

Fatemeh Mohebbi (University of Southern California), Mohammad Vahed (University of Southern California), Cecilia Liu (University of Southern California), Jiaqi Fu (University of Southern California) and Serghei Mangul (University of Southern California). *Assessing the robustness and reproducibility of RNA-seq quantification tools.*

Abstract. One of the fundamental steps in RNA-Seq analysis is to estimate the abundance of genes and transcripts in biological samples. Thus far, numerous quantification tools have been developed to accurately estimate gene and transcript expression levels. Inconsistencies in gene and transcript quantification could have significant implications for the accuracy of diagnostic or therapeutic decisions. It is, however, difficult to achieve accurate and consistent gene expressions due to the presence of experimental variations. Currently, it is unknown which types of experimental variations RNA-Seq quantification tools can mitigate and maintain consistent results and which they cannot account for. Exiting efforts attempting to assess the consistency and reproducibility of quantification tools' results are limited and some of the widely used quantification tools such as Salmon and Kallisto have not undergone a thorough consistency assessment. In this study, we have developed a framework with a scoring metric scheme to evaluate the robustness and reproducibility of RNA-Seq quantification tools.

We studied ten popular quantification tools and compared the consistency of their gene and transcript expression estimates across both synthetic and real technical replicates. Incorporating both types of replicates allowed us to observe the effect of different experimental variations and the inconsistencies introduced by the tools themselves on the quantification results.

Our analysis revealed a notable disparity in the ability of the quantification tools to maintain consistent estimation of gene and transcript expressions across both technical and synthetic replicates. Importantly, expectation-maximization and mapping-based tools were more effective at maintaining consistency compared to pseudoalignment tools.

Keywords: RNA-seq, Gene expression, Reproducibility

Seong Woo Han (University of Pennsylvania), San Jewell (University of Pennsylvania), Andrei Thomas-Tikhonenko (University of Pennsylvania) and Yoseph Barash (University of Pennsylvania).
Contrasting and Combining Transcriptome Complexity Captured by Short and Long RNA Sequencing Reads.

Abstract. High-throughput short-read RNA sequencing has given researchers unprecedented detection and quantification capabilities of splicing variations across biological conditions and disease states. However, short-read technology is limited in its ability to identify which isoforms are responsible for the observed sequence fragments and how splicing variations across a gene are related. In contrast, more recent long-read sequencing technology offers improved detection of underlying full or partial isoforms but is limited by high error rates and throughput, hindering its ability to accurately detect and quantify all splicing variations in a given condition.

To better understand the underlying isoforms and splicing changes in a given biological condition, it's important to be able to combine the results of both short and long-read sequencing, together with the annotation of known isoforms. To address this need, we develop MAJIQ-L, a tool to visualize and quantify splicing variations from multiple data sources. MAJIQ-L combines transcriptome annotation, long reads based isoform detection tools output, and MAJIQ (Vaquero-Garcia et al. (2016, 2023)) based short-read RNA-Seq analysis of local splicing variations (LSVs). We analyze which splice junction is supported by which type of evidence (known isoforms, short-reads, long-reads), followed by the analysis of matched short and long-read human cell line datasets. Our software can be used to assess any future long reads technology or algorithm, and combine it with short reads data for improved transcriptome analysis.

Keywords: Short read RNA seq, Long read RNA seq, Transcriptome complexity, Unified visualization

Marjorie Roskes (Weill Cornell Medicine), Alexander Martinez Fundichely (Weill Cornell Medicine), Weiling Li (Weill Cornell Medicine), Sandra Cohen (Weill Cornell Medicine), Hao Xu (McGill University), Shahd Elnaggar (Barnard College), Anisha Tehim (Cornell University), Metin Balaban (Princeton University), Chen Khuan Wong (Memorial Sloan Kettering Cancer Center), Yu Chen (Memorial Sloan Kettering Cancer Center), Ben Raphael (Princeton University) and Ekta Khurana (Weill Cornell Medicine).
Evolution of genomic and epigenomic heterogeneity in prostate cancer from tissue and liquid biopsy.

Abstract. Castration Resistant Prostate Cancer (CRPC) is an aggressive disease that is highly plastic. Although histologically there are two subtypes of CRPC: adenocarcinoma and neuroendocrine, we have shown it has four distinct molecular subtypes exhibiting differential chromatin and transcriptomic profiles. These are CRPC-AR (androgen receptor dependent), CRPC-WNT (Wnt pathway dependent), CRPC-SCL (stem-cell like), and CRPC-NE (neuroendocrine). During treatment with AR signaling inhibitors, patient tumors can evolve to different subtypes. Clinical identification of these subtypes and mechanistic understanding of the genomic and epigenomic heterogeneity accompanying this evolution is a huge challenge. To address this, we have amassed a unique cohort of 60 CRPC patients with various subtypes from whom cell-free DNA (cfDNA) was collected at various clinically relevant time points and whole-genome sequencing (WGS) was performed. For 24 of these patients, time-matched tissue RNA-seq was performed. We estimated epigenetic/transcriptomic heterogeneity in tissue by deconvolution of bulk RNA-seq data. We performed nucleosomal profiling from cfDNA WGS to infer tumor chromatin accessibility and estimate each epigenetic subtype's fractional contribution. We can detect the different subtypes in cfDNA and find that CRPC-SCL patients exhibit more heterogeneity than other subtypes in both tissue and cfDNA, likely indicating the transitory state of this subtype. We calculated allele-specific, genome-wide copy number alterations in cfDNA, and can track the parallel evolution of genomic and epigenomic events, e.g. AR gains track with increasing CRPC-AR fraction over time. Our study shows that, beyond biomarker development, cfDNA WGS can be used for characterizing the epigenomic and genomic evolution of patient tumors.

Keywords: cfDNA, tissue RNA-seq, tumor evolution, epigenetic profiling, deconvolution, prostate cancer

Tolulope Adeyina (The University of Texas at El Paso), Jonathon Mohl (The university of Texas at El paso) and Philip Lavretsky (The University of Texas at El Paso). *What type of duck are you? Identifying ancestry based on limited number of SNPs .*

Abstract. Understanding mallard ancestry can help inform conservational efforts for various duck species. This study explores the genetic complexities of various duck breeds using sophisticated machine learning methods. It focuses on breed identification, a common practice in animal genetics and breeding, which is increasingly leveraging artificial intelligence and high-throughput genomic data. Utilizing five duck breeds, comprising 559 individuals and with a total 40,401 SNPs, a challenge lies in identifying the most informative SNPs for optimal breed prediction accuracy and precision. Initial exploration involved the use of PCA components in K-means clustering for breed categorization, which yielded a high silhouette score (0.822), indicating cohesive clusters. A combination of variable selection techniques was employed, including stacking logistic regression with L2 regularization on some filter-based methods of feature selection. The logistic regression selected 15,390 features. These were further reduced to 74 through a combination of three statistical tests: chi-squared test, ANOVA F-test, and mutual information score. The refined feature selection methods significantly improved the accuracy of the Random Forest classifier from 55% to 95.5%. This project provides valuable insights into duck genomics and paves the way for the development of an efficient and user-friendly ancestry analysis framework to analyze a larger high-throughput sequencing project.

Keywords: Mallard Ancestry, Duck Breeds, Machine Learning Methods, High-throughput Genomic Data, Ancestry Analysis Framework

Can Firtina (ETH Zurich), Maximilian Mordig (Max Planck Institute for Intelligent Systems, ETH Zurich,), Joël Lindegger (ETH Zurich), Harun Mustafa (ETH Zurich, University Hospital Zurich, Swiss Institute of Bioinformatics), Sayan Goswami (ETH Zurich), Stefano Mercogliano (ETH Zurich), Yan Zhu (University of Toronto, ETH Zurich), Andre Kahles (ETH Zurich, University Hospital Zurich, Swiss Institute of Bioinformatics) and Onur Mutlu (ETH Zurich). *Rawsamble: Overlapping and Assembling Raw Nanopore Signals using a Hash-based Seeding Mechanism.*

Abstract. Although raw nanopore signal mapping to a reference genome is widely studied to achieve highly accurate and fast mapping of raw signals, mapping to a reference genome is not possible when the corresponding reference genome of an organism is either unknown or does not exist. To circumvent such cases, all-vs-all overlapping is performed to construct de novo assembly from overlapping information. However, such an all-vs-all overlapping of raw nanopore signals remains unsolved due to its unique challenges such 1) generating multiple and accurate mapping pairs per read, 2) performing similarity search between a pair of noisy raw signals, and 3) performing space- and compute-efficient operations for portability and real-time analysis.

We introduce Rawsamble, the first mechanism that can quickly and accurately find overlaps between raw nanopore signals without translating them to bases. We find that Rawsamble can 1) find overlaps while meeting the real-time requirements with throughput on average around 200,000 bp/sec, 2) share a large portion of overlapping pairs with minimap2 (37.12% on average), and 3) lead to constructing long assemblies from these useful overlaps. Finding overlapping pairs from raw signals is critical for enabling new directions that have not been explored before for raw signal analysis, such as de novo assembly construction from overlaps that we explore in this work. We believe these overlaps can be useful for many other new directions coupled with real-time analysis.

Rawsamble is integrated in RawHash and available at https://github.com/CMU-SAFARI/RawHash.

Keywords: nanopore sequencing, all-vs-all overlapping, raw nanopore signal analysis, adaptive sampling, read mapping, sequence analysis

Dottie Yu (Department of Quantitative and Computational Biology, USC), Ram Ayyala (Department of Quantitative and Computational Biology, USC), Sara Sadek (California State University), Likhitha Chittampalli (USC School of Pharmacy), Mina Jung (Department of Quantitative and Computational Biology, USC), Junghyun Jung (Department of Clinical Pharmacy, University of Southern California), Hafsa Farooq (Georgia State University), Abdullah Nahid (USC School of Pharmacy), Grigore Boldirev (Department of Computer Science, Georgia State University,), Yiting Meng (USC School of Pharmacy), Sungmin Park (Department of Computer Science and Engineering, Dongguk University-Seoul), Austin Nguyen (Oregon Health & Science University, Biomedical Engineering,), Alex Zelikovsky (Department of Computer Science, Georgia State University), Nicholas Mancuso (Population and Public Health Sciences, Keck School of Medicine, USC), Jong Wha Joo (Department of Computer Science and Engineering, Dongguk University-Seoul), Reid Thompson (Oregon Health & Science University), Houda Alachkar (Department of Clinical Pharmacy, School of Pharmacy, University of Southern California) and Serghei Mangul (University of California, Los Angeles). *A rigorous benchmarking of alignment-based HLA typing algorithms for RNA-seq data.*

Abstract. Precise identification of human leukocyte antigen (HLA) alleles is essential for clinical and research purposes, yet remains challenging due to HLA loci polymorphism. Despite the accessibility of Next-Generation Sequencing (NGS) data, numerous computational tools predict HLA types from RNA sequencing (RNA-seq) data. However, comprehensive benchmarking using realistic gold standards is lacking. To address this, we rigorously compared 12 HLA callers across 682 RNA-seq samples from eight datasets, using molecularly defined gold standards for HLA-A, -B, -C, -DRB1, and -DQB1 loci. We assessed accuracy metrics, optimized parameters, and scrutinized discrepancies at allele and loci levels. Notably, we examined performance across European and African groups, finding higher accuracy for European samples. We also evaluated computational efficiency in terms of CPU runtime and RAM usage. Our study aims to guide clinicians and researchers in selecting appropriate HLA callers.

Keywords: hla, benchmarking, RNA-seq, tool development, Immunology, MHC

Can Firtina (ETH Zurich), Melina Soysal (ETH Zurich), Joël Lindegger (ETH Zurich) and Onur Mutlu (ETH Zurich). *RawHash2: Mapping Raw Nanopore Signals Using Hash-Based Seeding and Adaptive Quantization.*

Abstract. Summary: Raw nanopore signals can be analyzed while they are being generated, a process known as real-time analysis. Real-time analysis of raw signals is essential to utilize the unique features that nanopore sequencing provides, enabling the early stopping of the sequencing of a read or the entire sequencing run based on the analysis. The state-of-the-art mechanism, RawHash, offers the first hash-based efficient and accurate similarity identification between raw signals and a reference genome by quickly matching their hash values. In this work, we introduce RawHash2, which provides major improvements over RawHash, including a more sensitive quantization and chaining implementation, weighted mapping decisions, frequency filters to reduce ambiguous seed hits, minimizers for hash-based sketching, and support for the R10.4 flow cell version and various data formats such as POD5 and SLOW5. Compared to RawHash, RawHash2 provides better F1 accuracy (on average by 10.57% and up to 20.25%) and better throughput (on average by 4.0× and up to 9.9×) than RawHash.

Availability and Implementation: RawHash2 is available at https://github.com/CMU-SAFARI/RawHash. We also provide the scripts to fully reproduce our results on our GitHub page.

Keywords: nanopore sequencing, raw nanopore signal analysis, adaptive sampling, read mapping, sequence analysis

Simon Jeanneau (Université de Sherbrooke), Antoine Champie (Université de Sherbrooke), Amélie De Grandmaison (Université de Sherbrooke), Sébastien Rodrigue (Université de Sherbrooke) and Pierre-Étienne Jacques (Université de Sherbrooke). *Systematic identification of synthetic lethal relations in Escherichia coli using High-Througput Transposon Mutagenesis (HTTM).*

Abstract. Background: Despite the commonness of the model bacterium Escherichia coli in research, nearly a third of its genes remain of unknown function. To study a gene and the cellular processes in which it is involved, inactivating mutations are typically made in the organism. Notably, this low-throughput approach has produced the Keio collection, encompassing around 3800 E. coli K-12 mutant strains and revealing 299 essential genes. A more efficient way to identify essential genes is transposon mutagenesis where cells harboring disruptions in essential genes do not survive, while non-essential genes are filled with transposon-originated reads after high-throughput sequencing. We recently developed the High-Throughput Transposon Mutagenesis (HTTM) method, yielding unprecedented coverage therefore allowing us to confidently identify genetic interactions at large scale. Previously, comprehensive gene pair deletions in yeast have identified numerous negative gene interactions and shed light on gene functions.

Results: We applied the HTTM method in approximately 1200 strains of the Keio collection, querying ~5 million of potential genetic interactions. After rigorous data treatment, this allowed us to identify the majority of known synthetic lethal interactions and identify hundreds of new ones that are under investigation, showcasing its utility and transformative potential in genomic research.

Conclusions: Our pioneering work probing systematically for the first time all potential prokaryotic gene interactions promises to not only enhance the functional annotation of numerous genes but also unravel complex, previously misunderstood biological pathways. Additionally, it will uncover surprising genetic interactions with potential applications across various fields including synthetic biology and novel antibiotic targets.

Keywords: Gene Interactions, Transposon Mutagenesis, Escherichia coli

Yaqi Su (UC Berkeley), Zhejian Yu (Zhejiang University), Siqian Jin (Zhejiang University), Zhipeng Ai (Zhejiang University), Ruihong Yuan (Zhejiang University), Xinyi Chen (Zhejiang University), Ziwei Xue (Zhejiang University), Yixin Guo (Zhejiang University), Di Chen (Zhejiang University), Hongqing Liang (Zhejiang University), Zuozhu Liu (Zhejiang University) and Wanlu Liu (Zhejiang University).
Comprehensive Assessment of mRNA Isoform Detection Methods for Long-Read Sequencing Data.

Abstract. The advancement of Long-Read Sequencing (LRS) techniques has significantly increased the length of sequencing to several kilobases, thereby facilitating the identification of alternative splicing events and isoform expressions. Recently, numerous computational tools for isoform detection using long-read sequencing data have been developed. Nevertheless, there remains a deficiency in comparative studies that systemically evaluate the performance of these tools, which are implemented with different algorithms, under various simulations that encompass potential influencing factors. In this study, we conducted a benchmark analysis of thirteen methods implemented in nine tools capable of identifying isoform structures from long-read RNA-seq data. We evaluated their performances using simulated data, which represented diverse sequencing platforms generated by an in-house simulator, RNA sequins (sequencing spike-ins) data, as well as experimental data. Our findings demonstrate IsoQuant as a highly effective tool for isoform detection with LRS, with Bambu and StringTie2 also exhibiting strong performance. These results offer valuable guidance for future research on alternative splicing analysis and the ongoing improvement of tools for isoform detection using LRS data.

Keywords: Long-read RNA sequencing, Benchmark study of computational methods, Detection of mRNA isoforms, Alternative splicing

Yushan Hu (University of Victoria), Li Xing (University of Saskatchewan), Xiaojian Shao (Digital Technologies Research Centre, National Research Council Canada), Ziying Liu (Digital Technologies Research Centre, National Research Council Canada), Don Sin (UBC Centre for Heart Lung Innovation, St. Paul's Hospital) and Xuekui Zhang (University of Victoria). *Automated Cell Annotation Tool from an Integrated Cell Atlas of Bronchoalveolar Lavage in Healthy Control.*

Abstract. Analyzing single-cell data from bronchoalveolar lavage (BAL) samples, collected by a minimally invasive and safe procedures, offers an unprecedented opportunity to dissect the heterogeneity of immune cells in lung at a single-cell resolution. However, recent single-cell RNA sequencing (scRNA-seq) studies have been hindered by a limited number of healthy donors and inconsistencies in cell type annotations. Despite the existence of large-scale human lung cell atlases, no such atlas has been created for the BAL samples. Leveraging the human lung cell atlas (HLCA) core dataset, we developed an automatic cell type annotation model tailored for BAL scRNA-seq data. Intra-data cross-validation testing and independent validation demonstrated that our model outperforms state-of-the-art automatic annotation tools for the majority of cell types in BAL, with an F1 score exceeding 0.9. Applying our model to a large in-house cohort of BAL scRNA-seq data, comprising 283,915 cells from 6 patients with chronic obstructive pulmonary disease (COPD) and 24 non-COPD patients, revealed notable changes in cell type proportions between the two groups. Furthermore, we constructed the most extensive BAL single-cell atlas by applying our automatic annotation model to integrated datasets from both publicly available and large-scale in-house BAL scRNA-seq datasets. In summary, our study has developed an effective automated annotation tool for human BAL scRNA-seq data and provides a comprehensive single-cell transcriptomic atlas of BAL samples in both healthy and diseased states.

Keywords: Bronchoalveolar Lavage, Single-cell RNA-seq, Automatic Cell Annotation, Lung Disease, Cell Atlas

Nicolas Jacquin (IRIC, Université de Montréal) and Sébastien Lemieux (IRIC, Université de Montréal).
K-mer Walking: An Efficient Reference-Free Algorithm for Flanking Sequence Reconstruction.

Abstract. With the rapid expansion of transcriptomics as a field, it's easy to forget that the references we use to analyze our transcriptomics data are incomplete, and that they introduce bias by filtering out reads that don't align to their reference. But reads outside of what our references consider protein-coding can still be of great biological significance, and sometimes are even found to code for non-canonical proteins, as outlined by Laumont et al. (2016). This latent part of the transcriptome is underexplored, and more tools to further facilitate its analysis could greatly improve our understanding of proteomics in general.

When working with non-canonical protein-coding sequences, the only well supported way to retrieve flanking sequences would be to align those sequences to a reference, keeping hits outside of the annotated region, and use the reference sequences that flank those hits. However, this can not only be resource intensive and hard to perform on a systematic scale for multiple samples, it also means that if the real flanking sequences are sufficiently different from the ones in the reference, the results could be completely wrong.

In this work, we showcase a new tool we are currently developing that allows for fast retrieval of flanking sequences of protein-coding RNA within the raw data of multiple RNASeq experiments, without the need of ever aligning those reads to a reference.

Keywords: Transcriptomics, RNA-Seq, Reference-free, Immunopeptidomics, K-mer

Daniel Gladish (Ottawa Hospital Research Institute), Theodore Perkins (Ottawa Hospital Research Institute), Steven Tur (Wisconsin Blood Cancer Research Institute), Marjorie Brand (Wisconsin Blood Cancer Research Institute) and Satrajit Chatterjee (University of Ottawa). *Attempt at Improving Single CEll Variational ANeuploidy analysis for application in T-cell acute lymhphoblastic leukemia.*

Abstract. T-cell acute lymphoblastic leukemia (T-ALL) is a rare and aggressive hematological cancer that is highly heterogeneous. It is associated with poor prognosis in adults and those with relapsed or refractory disease. Copy number alterations (CNAs) are large gains or deletions, ranging from 1 kb to whole chromosome arms. CNAs are observed in nearly all cancers and can contribute to disease development and progression. CNA data is used in the clinical setting for prognostication, risk stratification, and informing treatment. Studies have shown that CNAs are prevalent in T-ALL and that they often span oncogenes and tumour suppressors known to be recurrently altered in the disease.

Over the past two decades, CNA calling algorithms have been developed that predict the locations of CNAs from microarray and bulk next-generation sequencing data. Since the Drop-Seq publication in 2015, many algorithms designed for the analysis of single-cell RNA sequencing data have been published. In my current work, I am using the algorithm Single CEll Variational ANeuploidy analysis (SCEVAN) on expression data obtained from diagnostic and relapsed samples taken from a T-ALL patient. Although the authors quote high accuracy rates in their benchmarking, accuracy has been unsatisfactory and poor cross-technique reproducibility has been observed. The Perkins Lab is currently experimenting with ways of augmenting performance. Accuracy was not improved with the application of stricter filtration cutoffs. Despite setbacks, our efforts have the potential to help obtain valuable data on problematic T-ALL subpopulations, which can help improve outcomes by identifying novel drug targets.

Keywords: CNAs, Cancer, T-ALL

Xiaofei Carl Zang (Pennsylvania State University), Xiang Li (Pennsylvania State University), Kyle Metcalfe (Element Biosciences), Tuval Ben-Yehezkel (Element Biosciences), Ryan Kelley (Element Biosciences) and Mingfu Shao (Pennsylvania State University). *Anchorage accurately assembles full-length anchor-flanked synthetic long reads.*

Abstract. Modern sequencing technologies allow for the addition of short-sequence tags, known as anchors, to both ends of a captured molecule. Anchors are useful in assembling the full-length sequence of a captured molecule by accurately determining its endpoints. One representative of such anchor-enabled technology is LoopSeq Solo, a synthetic long read (SLR) sequencing protocol. LoopSeq Solo also achieves ultra-high sequencing depth and high purity of short reads covering the entire captured molecule. Despite the availability of many assembly methods, constructing full-length contigs from these anchor-enabled, ultra-high coverage sequencing data remains challenging due to the complexity of the underlying assembly graphs and the lack of specific algorithms leveraging anchors. We present Anchorage, a novel assembler that performs anchor-guided assembly for ultra-high depth sequencing data. Anchorage starts with a kmer-based approach for precise estimation of molecule lengths. It then formulates the assembly problem as finding an optimal path that connects the two vertices in the underlying De Bruijn graph determined by anchors, where optimality is defined as maximizing the weight of the smallest edge while matching the estimated length. Anchorage uses a dynamic programming algorithm to efficiently find the optimal path. Our evaluations show that Anchorage outperforms existing methods, reliably assembling full-length contigs with anchor sequences at both ends. Anchorage fills the gap of assembling anchor-enabled data. We anticipate its broad use as anchor-enabled technologies become prevalent.

Keywords: Assembly, de Bruijn Graph, Synthetic Long Reads, Anchor-guided Assembly

Nika Mansouri Ghiasi (ETH Zurich), Mohammad Sadrosadati (ETH Zurich), Harun Mustafa (ETH Zurich), Arvid Gollwitzer (ETH Zurich), Can Firtina (ETH Zurich), Julien Eudine (ETH Zürich), Haiyu Mao (ETH Zurich), Joel Lindegger (ETH Zürich), Meryem Banu Cavlak (ETH Zurich), Mohammed Alser (ETH Zurich), Jisung Park (POSTECH) and Onur Mutlu (ETH Zurich & Stanford University). *MegIS: High-Performance and Low-Cost Metagenomic Analysis with In-Storage Processing.*

Abstract. Metagenomics has led to significant advancements in many fields. Since the species present in a metagenomic sample are not known in advance, metagenomic analysis commonly involves the key tasks of determining the species present in a sample and their relative abundances. These tasks require searching large metagenomic databases containing information on different species' genomes. Metagenomics suffers from significant data movement overhead from the storage system to the rest of the system. In-storage processing can be a fundamental solution for reducing data movement overhead. However, designing an in-storage processing system for metagenomics is challenging because none of the existing approaches can be directly implemented in storage effectively due to the hardware limitations of modern SSDs.

We propose MegIS, the first in-storage processing system designed to significantly reduce the data movement overhead of the end-to-end metagenomic analysis. MegIS is enabled by our lightweight design that effectively leverages and orchestrates processing inside and outside the storage system. Through our detailed analysis of the end-to-end metagenomic analysis pipeline and careful hardware/software co-design, we address in-storage processing challenges for metagenomics via specialized and efficient 1) task partitioning, 2) data/computation flow coordination, 3) storage technology-aware algorithmic optimizations, 4) lightweight in-storage accelerators, and 5) data mapping. MegIS outperforms the state-of-the-art performance- and accuracy-optimized software metagenomic tools by 2.7x-37.2x and 6.9x-100.2x, respectively, while matching the accuracy of the accuracy-optimized tool. MegIS achieves 1.5x-5.1x speedup compared to the state-of-the-art metagenomic hardware tool, while achieving significantly higher accuracy.

Keywords: Metagenomic Analysis, Storage Systems, Data Movement Overhead

Anna-Sophie Fiston-Lavier (University of Montpellier - Institut of Sciences and Evolution (ISEM)), Shadi Shahatit (University of Montpellier - Institut of Sciences and Evolution (ISEM)) and Jean Monlong (UC Santa Cruz Genomics Institute, University of California, Santa Cruz, Santa Cruz, California 95060, USA.).
Pangenomic approaches to provide a better understanding of the dynamics of transposable elements in human.

Abstract. Humans originated on the African continent and migrated out of Africa, requiring adaptive processes. Comparative analyses of can therefore be expected to reveal signatures of selective forces that indicate potential adaptive loci. While the patterns of transposable element (TE) variation in populations can be used to trace the evolutionary history of our species, their evolutionary dynamics during the out-of-Africa event has not been well studied. Using the promising recent human pangenome reference, built using 47 human individuals from eight populations, we aim (1) to characterise the frequency distribution of TEs along the population and possible factors influencing their evolution; (2) and to examine TE loci under positive selection between African and non-African populations.

Our initial results show a non-random distribution of TE across chromosomes, with an over-representation in intergenic and intronic regions, mostly detected as very rare TE insertions. Recombination rate and distance between TEs have little or no effect on their frequency spectrum. As TEs also show patterns of population stratification, we were able to identify TE loci containing putative candidate genes under positive selection. Candidate genes were associated with immune, metabolic and developmental pathways and functions. All of these elements suggest features of local human adaptation. Further analysis will be required to achieve a more robust functional understanding of human evolution. Future larger sampling with more samples per population will allow a better resolution of genetic diversity and study of the impact of TEs to such evolutionary processes.

Keywords: Transposable element, Pangenomics, Evolution, Human, Selection

Vinzenz May (Berlin Institute of Health @ Charite Berlin), Dieter Beule (Berlin Institute of Health @ Charite Berlin) and Manuel Holtgrewe (Berlin Institute of Health @ Charite Berlin). *SvirlPool. Long, lifted, pooled: Reliable structural variant detection on multiple samples.*

Abstract. Structural variants (SVs) play a critical role in rare disease diagnostics, cancer genomics, and understanding genome variation in populations. Current SV detection methods using long reads often face challenges in detecting very large or complex SVs, especially in difficult regions like repeats or low complexity areas.

Our approach introduces a novel method that combines alignment- and assembly-based techniques. We begin by identifying candidate regions in read-to-reference alignments and then assemble clusters of relevant parts of aligned reads. These assembled fragments are aligned to a chosen reference sequence for SV-calling. Importantly, our workflow allows for precise matching of shared SVs across multiple samples by pooling sequencing data.

In benchmark tests using long-read whole genome sequencing datasets, our method demonstrates competitive results compared to existing methods (>95% F1 score on the Tier-1 Genome in the Bottle SV calling benchmark). A significant advantage is the reliable detection of SVs across samples, essential for pedigree and matched tumor-normal analysis, where other tools cannot yet make full use of the statistical power of multiple samples.

In conclusion, our method offers precise SV calling with long reads, ideal for rare disease research, diagnostics, and cancer genomics. Additionally, it can leverage the continual improvements in reference genomes, promising broader applicability in genomics research.

Keywords: structural variants, structural variant detection, long reads, genomics, rare disease, cancer genomics

Ayah Shevchenko (National Institute of Standards and Technology), Sierra Miller (National Institute of Standards and Technology), Samantha Maragh (National Institute of Standards and Technology), Simona Patange (National Institute of Standards and Technology), Justin Zook (National Institute of Standards and Technology), Nate Olson (National Institute of Standards and Technology), Jamie Almeida (National Institute of Standards and Technology), Hua-Jun He (National Institute of Standards and Technology), Natalia Kolmakova (National Institute of Standards and Technology) and Patricia Kiesler (National Institute of Standards and Technology). *Performance of variant detection technologies in the first NIST Genome Editing Consortium Interlab Study.*

Abstract. Background

Moving genome editing technologies into medical practice requires robust quantitative assays, with associated controls and data tools. The U.S. National Institute of Standards and Technology (NIST) Genome Editing Consortium (GEC) convenes experts across academia, industry, and government to address precompetitive genome editing measurements and standards needed to increase confidence in evaluating and utilizing these technologies in research and commercial products.

Results

Presented here is the first NIST GEC Interlab Study, wherein NIST provided participants with 5 qualified DNA or cell mixture samples, containing variants ranging in size from SNVs to indels tens of kilobases long, at frequencies of 0.1 - >30%. Participants analyzed samples at provided genomic loci while blinded to variant sequence and frequency, then provided results to NIST. For all variants, successfully called and uncalled, raw data was manually inspected to determine variant presence.

Technologies used by participants included: bulk targeted short and long read NGS, single cell targeted and genome wide NGS, targeted and genome wide DNA imaging, and capillary electrophoresis fragment analysis. 15 workflows were assessed, with 43 TB of data received.

Conclusions

Workflow performance differed across participants, even within similar NGS approaches. However, certain small indels proved challenging across NGS workflows, including those similar in structure to gene editing outcomes. Additionally, while most variant allele frequency (VAF) calls were close to true values, overall VAF accuracy decreased as indel size increased. Finally, at least a third of all uncalled variants were found in raw data, suggesting room for improvement in variant calling.

Keywords: genome editing, precision medicine, genetic variation analysis, large variant detection

Nicholas Strieder (Leibniz Institute for Immunotherapy). *Evaluating batch removal performance by spiking-in cells in single cell transcriptomic analyses.*

Abstract. Due to the high costs of single cell sequencing, recently lab researchers tend to apply single cell omics experiments to more and more specific cells types or subsets of cells, isolated by FACS sorting techniques. Cells from a limited number of these FACS-isolations will be later analyzed together. However, the differences in transcriptional patterns between the subpopulations will be at a far lower level then differences in cell types from PBMC isolates most batch-removal tools for scRNA data were developed for. In this work we analyze the beneficial effect of spiking-in a different rather complementary cell type from the the same individual to further help discriminate technical from biological variability.

Keywords: scRNA, batch effect, spikeIn

Yasuhiro Kojima (National Cancer Center Japan), Haruka Hirose (Tokyo Medical and Dental University), Shuto Hayashi (Tokyo Medical and Dental University) and Teppei Shimamura (Tokyo Medical and Dental University). *Inferring cell state dynamics dependent on extrinsic factors using deep generative model.*

Abstract. Recent high accessibility to single cell transcriptome analysis yields the comparative observation across multiple experimental conditions and multi-modal observation integrating transcriptional molecular profiles with other molecular layers such as chromatin accessibility. Although these advanced observations enhanced the identification of condition specific populations and the correlation structures across molecular layers, existing computational methodologies were not suitable for dissecting the generation process of such populations or correlation structures, which is possibly regulated or intervened by the covariates. Here, we present a deep generative model of covariate-dependent cell state dynamics, ExDyn, which realized counterfactual estimation of cell state dynamics for varying covariates in a single cell state. Demonstrating the ability of ExDyn to estimate differential cell state dynamics in simulated dataset, we utilized it for revealing the complex relationships between gene expression dynamics and its covariates in several real datasets. In application to squamous carcinoma dataset, we showed that subpopulation of fibroblasts induced cell state dynamics toward invasive cancer cells. We demonstrated the ability of ExDyn to dissect the complex relationships between transcriptome dynamics and the various covariates.

Keywords: Deep generative model, Single cell transcriptome, Spatial transcriptome, RNA velocity

Elham Salehisiavashani (Ottawa Hospital Research Institute, Ottawa, Ontario, Canada), Nicholas D Cober (Ottawa Hospital Research Institute, Ottawa, Ontario, Canada), Elmira Safaei Qamsari (Ottawa Hospital Research Institute, Ottawa, Ontario, Canada), Yupu Deng (Ottawa Hospital Research Institute, Ottawa, Ontario, Canada), Anu Situ (Ottawa Hospital Research Institute, Ottawa, Ontario, Canada) and Duncan J. Stewart (Ottawa Hospital Research Institute, Ottawa, Ontario, Canada). *Exploring mechanisms of pulmonary venous occlusive disease using single-cell transcriptomics.*

Abstract. Pulmonary veno-occlusive disease (PVOD) is a rare subset of pulmonary arterial hypertension (PAH) that, in its hereditary form, is caused by biallelic mutations in Eif2ak4/GCN2. It is characterized by widespread obliterative changes in the pulmonary vascular bed involving not only arterioles but also venules and capillaries. Unlike PAH, PVOD lacks effective treatment and is universally fatal without lung transplantation. Mitomycin-C (MMC) chemotherapy has been shown to be associated with PVOD in patients and reproduces this disease when administered to rats.

In a SU5416/chronic hypoxia rat model of severe PAH, scRNA-seq revealed a disease-specific, de-differentiated endothelial cell (dDEC) population characterized by loss of endothelial identity and reduced tight junction protein expression. dDECs were primed for endothelial to mesenchymal transition (EndMT), as shown by RNA velocity and trajectory analysis. Male and female rats received MMC injections (2-2.5 mg/kg) on Days 0 and 7. At week 5, right ventricular systolic pressure (RVSP) and Fulton Index were elevated compared to controls. Histology showed widespread obliterative remodeling and capillary hemangiomatosis. Flow cytometry of dispersed lung cells from control and PVOD rats revealed a significant reduction in CD31 positive ECs, indicating EC identity loss. Cells were barcoded, and scRNA-seq will focus on endothelial cell changes compared to the PAH model.

Our study aims to uncover molecular and cellular mechanisms of PVOD. We anticipate that, as in PAH, single-cell transcriptomics will reveal a dDEC cluster primed for EndMT as well as novel EC populations involved in venous and capillary lesions.

Keywords: Pulmonary veno-occlusive disease, Eif2ak4/GCN2 pathway, Single-cell RNA-sequencing

Félix-Antoine Trifiro (Université de Sherbrooke), Manon Stepanoff (Université Sherbrooke) and Marie Brunet (Université de Sherbrooke). *Detection of small genomic structural alterations using data-driven personalized annotations.*

Abstract. Despite novel sequencing technologies, the diagnostic rate of inherited developmental disorders (IDD) hovers around 50%. This surprising number suggests that some genomic alterations are overlooked by current analytical methods. During sequencing analyses, reads that do not agree with the reference genome annotation are discarded. Here, we suggest that these discordant read pairs (DRP) can highlight genomic alterations otherwise undetected.

We developed an analytical pipeline to build personalized genomic architectures and highlight structural alterations. We applied it on a cohort of 100 children with IDD. Whole exome sequencing data is aligned to the reference genome (hg38) using the standard analytical pipeline (BWA aligner). DRPs are retrieved using the MELT algorithm, and pairs with a low quality or low complexity are filtered out. Genome alignment of DRPs is validated using the slower but more precise Smith-Waterman algorithm. Confident DRPs are then used to build a personalized genomic architecture supported by the sequencing data. Initial results from 40 patients highlight that 25% of patients have a significantly high proportion of discordant read pairs. In 39 patients we detected the presence of a TDG retrocopy, absent from reference genomes, but known to be present in certain ethnicities (positive control of our pipeline). Detailed analyses unveiled a retrocopy of NBPF12 in one patient. NBPF genes have expanded in the human lineage and copy number variations are associated with developmental disorders corresponding to the patient's phenotype.

Our pipeline highlighted genomic structural alterations undetected by current methodologies and relevant to the diagnosis of patients with IDD.

Keywords: RNA-seq, Retrotransposition, Genomic

Weiling Li (Weill Cornell Medicine), Marjorie Roskes (Weill Cornell Medicine), Alexander Martinez-Fundichely (Weill Cornell Medicine), Sandra Cohen (Weill Cornell Medicine) and Ekta Khurana (Weill Cornell Medicine). *Predicting tumor gene expression from cell-free DNA whole-genome sequencing.*

Abstract. When a cell dies, it releases cell free DNA (cfDNA) into the bloodstream. Nucleosome protected regions survive in plasma, while nucleosome-depleted regions (NDRs) are degraded. Genome-wide studies have shown that nucleosome-depleted regions are present at the transcription start sites of active genes. This results in lower depth of sequencing coverage near transcription start sites (TSSs) and diversity of cfDNA fragment size values observed from cfDNA whole-genome sequencing (WGS). Thus, we incorporated multiple cfDNA features in a machine learning model for gene expression prediction. For training, we used matched cfDNA WGS and RNA-Seq profiling at the same time point. We then tested our model on 9 castration-resistant prostate cancer (CRPC) patients with RNA-seq and cfDNA WGS at matched time point. The concordance was evaluated using Pearson correlation between predicted gene expression and gene expression observed from RNA-seq of tissue site/s. We also compared our results with the most recent gene expression prediction paper (EPIC-seq [Esfahani et al, 2022]). 8 out of 9 test samples in our model show higher correlations than EPIC-Seq (when grouped by 10), most of which range from 0.54 to 0.65 (random expected would be zero). Further application of our method is predicting cancer subtype by pathway enrichment. Based on our predicted gene expression scores for these CRPC patients and androgen receptor vs. neuroendocrine pathways from [Tang et al, 2022], we can predict CRPC subtype for each patient, which agree with the histology classification.

Keywords: cell-free DNA, liquid biopsy, gene expression, cancer