HiTSeq

High Throughput Sequencing
Algorithms and Applications
A special track of the ISMB/ECCB 2025 meeting
Liverpool, United Kingdom, July 20-24, 2025

ISMB-ECCB 2025 HiTSeq Track Proceedings

Liverpool, United Kingdom
July 23-24, 2025
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

TALKS

Kendell Clement (University of Utah). *MutSuite: A Toolkit for Simulating and Evaluating Mutations in Aligned Sequencing Reads.*

Abstract. Simulated sequencing reads containing known mutations are essential for developing, testing, and benchmarking mutation detection tools. Most existing simulation tools introduce mutations into synthetic reads and then realign them to a reference genome prior to downstream analysis. However, this realignment step can obscure the true position of insertions and deletions, introducing ambiguity and potential error in evaluation. In particular, the alignment process can shift the apparent location of insertions and deletions, complicating efforts to assess recall and precision of variant callers.

To address this limitation and support the development of more accurate and sensitive mutation detection algorithms, we developed MutSim, a tool that introduces substitutions, insertions, and deletions directly into aligned reads (e.g., in BAM files). By avoiding realignment, MutSim ensures that each simulated mutation remains at its exact specified position, enabling precise evaluation of variant caller performance.

MutSim is part of a larger toolkit we call MutSuite, which also includes MutRun, a companion tool that automates the execution of variant calling software on simulated datasets, and MutAgg, which aggregates and summarizes results across multiple variant callers for performance comparison. Together, these tools provide a robust and flexible framework for mutation simulation and benchmarking. MutSuite is open-source and freely available at: https://github.com/clementlab/mutsuite.

Keywords: Simulated sequencing reads, Sequencing read alignment, Variant caller evaluation

Tanmayee Narendra (University of Dundee), Giovanni Visonà (Max Planck Institute for Intelligent Systems), Crhistian de Jesus Cardona (University of Dundee), James Abbott (University of Dundee) and Gabriele Schweikert (University of Dundee). *Towards Personalized Epigenomics: Learning Shared Chromatin Landscapes and Joint De-Noising of Histone Modification Assays.*

Abstract. Epigenetic mechanisms enable cellular differentiation and the maintenance of distinct cell-types. They enable rapid responses to external signals through changes in gene regulation and their registration over longer time spans. Consequently, chromatin environments exhibit cell-type and individual specificity contributing to phenotypic diversity. Their genomic distributions are measured using ChIP-Seq and related methods. However, the chromatin landscape introduces significant biases into these measurements. Here, we introduce DecoDen to simultaneously learn shared chromatin landscapes while de-biasing individual measurement tracks. We demonstrate DecoDen's effectiveness on an integrative analysis of histone modification patterns across multiple tissues in personal epigenomes.

Keywords: Histone modifications, Chromatin landscapes, Integrative epigenomics

Wui Wang Lui (Johns Hopkins University) and Liliana Florea (Johns Hopkins University). *SpliSync: Genomic language model-driven splice site correction of long RNA reads.*

Abstract. We developed SpliSync, a deep learning method for accurate splice site correction in long read alignments. It combines a genomic language model, HyenaDNA, and a 1D U-net segmentation head, integrating genome sequence and alignment embeddings. SpliSync improves the detection of splice sites and introns and, when integrated with a short read transcript assembler, allows for improved transcript reconstruction, matching or outperforming reference methods like IsoQuant and FLAIR. The method shows promise for transcriptomic applications, especially in species with incomplete gene annotations or for discovering novel transcript variations.

Keywords: long RNA reads, deep learning, transcriptomics, splice sites, error correction, transcript reconstruction

Noam Teyssier (Arc Institute) and Alexander Dobin (Arc Institute). *BINSEQ: A Family of High-Performance Binary Formats for Nucleotide Sequences.*

Abstract. Modern genomics routinely generates billions of sequencing records per run, typically stored as gzip-compressed FASTQ files. This format's inherent limitations—single-threaded decompression and sequential parsing of irregularly sized records—create significant bottlenecks for bioinformatics applications that would benefit from parallel processing. We present BINSEQ, a family of simple binary formats designed for high-throughput parallel processing of sequencing data. The family includes BINSEQ, optimized for fixed-length reads with true random access capability through two-bit encoding, and VBINSEQ, supporting variable-length sequences with optional quality scores and block-based organization. Both formats natively handle paired-end reads, eliminating the need for synchronized files. Our comprehensive evaluation demonstrates that BINSEQ formats deliver substantial performance improvements across bioinformatics workflows while maintaining competitive storage efficiency. Both formats achieve up to 32x faster processing than compressed FASTQ and continue to scale with increasing thread counts where traditional formats quickly plateau due to I/O bottlenecks. These advantages extend to complex workflows like alignment, with BINSEQ formats showing 2-5x speedups at higher thread counts when tested with tools like minimap2 and STAR. Storage requirements remain comparable to or better than existing formats, with BINSEQ (610.35 MB) similar to gzip-compressed FASTA (647.29 MB) and VBINSEQ (509.89 MB) approaching CRAM (491.85 MB) efficiency. To facilitate adoption, we provide high-performance libraries, parallelization APIs, and conversion tools as free, open-source implementations. BINSEQ addresses fundamental inefficiencies in genomic data processing by considering modern parallel computing architectures.

Keywords: Data Structures, Parallel Processing, Nucleotide Sequences, Storage Formats, Binary, Genomics, Long-Read Sequencing, High-Throughput Algorithms

Ondřej Sladký (Charles University), Pavel Veselý (Charles University) and Karel Brinda (INRIA/IRISA Rennes).
FroM Superstring to Indexing: a space-efficient index for unconstrained k-mer sets using the Masked Burrows-Wheeler Transform (MBWT).

Abstract. The exponential growth of DNA sequencing data calls for efficient solutions for storing and querying large-scale k-mer sets. While recent indexing approaches use spectrum-preserving string sets (SPSS), full-text indexes, or hashing, they often impose structural constraints or demand extensive parameter tuning, limiting their usability across different datasets and data types. Here, we propose FMSI, a minimally parametrized, highly space-efficient membership index and compressed dictionary for arbitrary k-mer sets. FMSI combines approximated shortest superstrings with the Masked Burrows-Wheeler Transform (MBWT). Unlike traditional methods, FMSI operates without predefined assumptions on k-mer overlap patterns but exploits them when available. We demonstrate that FMSI offers superior memory efficiency for processing queries over established indexes such as SSHash, Spectral Burrows-Wheeler Transform (SBWT), and Conway-Bromage-Lyndon (CBL), while supporting fast membership and dictionary queries. Depending on the dataset, k, or sampling, FMSI offers 2–3x space savings for processing queries over all state-of-the-art indexes; only a space-optimized SBWT (without indexing reverse complement) matches its memory efficiency in some cases but is 2–3x slower. Overall, this work establishes superstring-based indexing as a highly general, flexible, and scalable approach for genomic data, with direct applications in pangenomics, metagenomics, and large-scale genomic databases.

Keywords: k-mer sets, shortest superstring, Burrows-Wheeler Transform, membership queries, compressed dictionary

Netanya Keil (University of Florida), Carolina Monzó (Institute for Integrative Systems Biology (I2SysBio), Spanish National Research Council (CSIC)), Lauren McIntyre (Institute for Integrative Systems Biology (I2SysBio), Spanish National Research Council (CSIC)) and Ana Conesa (Institute for Integrative Systems Biology (I2SysBio), Spanish National Research Council (CSIC)). *Quality assessment of long read data in multisample lrRNA-seq experiments using SQANTI-reads.*

Abstract. SQANTI-reads leverages SQANTI3, a tool for the analysis of the quality of transcript models, to develop a read-level quality control framework for replicated long-read RNA-seq experiments. The number and distribution of reads, as well as the number and distribution of unique junction chains (transcript splicing patterns), in SQANTI3 structural categories are informative of raw data quality. Multisample visualizations of QC metrics are presented by experimental design factors to identify outliers. We introduce new metrics for 1) the identification of potentially under-annotated genes and putative novel transcripts and for 2) quantifying variation in junction donors and acceptors. We applied SQANTI-reads to two different datasets, a Drosophila developmental experiment and a multiplatform dataset from the LRGASP project and demonstrate that the tool effectively reveals the impact of read coverage on data quality, and readily identifies strong and weak splicing sites. SQANTI-reads is open source and is available in versions ≥ 5.3.0 in the SQANTI3 GitHub repository.

Keywords: long-read RNASeq, splicing, quality control

Joao P. C. R. Mendonca (Rigshospitalet), Kristoffer Staal Rohrberg (Rigshospitalet, University of Copenhagen), Peter Holst (Hervolution Therapeutics) and Frederik Otzen Bagger (Rigshospitalet, Denmark Technical University). *Landscape of The Dark Genome's variants and their influence on cancer.*

Abstract. Human endogenous retroviruses (HERVs) are remnants of ancient viral infections that now make up ~8% of the human genome. Although typically silenced, HERVs can become reactivated in cancer and are emerging as biomarkers and immunotherapeutic targets. However, their clinical utility is limited by challenges in resolving individual loci due to high sequence similarity, incomplete genome annotations, and an overreliance on linear reference genomes. To address this, we constructed a variational pangenome using long-read sequencing data from Genome in a Bottle and the Platinum Pedigree projects. This approach enables accurate detection of single nucleotide variants (SNVs), insertions/deletions (indels), and structural variants (SVs) in a reference-free manner, revealing polymorphic HERV insertions absent from the human reference genome. By integrating data from the Copenhagen Prospective Personalized Oncology (CoPPO) biobank, we link these variants to HERV expression in cancer, distinguishing potentially pathogenic variants from benign ones. We combine pangenome-informed annotations with locus-specific expression quantification tools to resolve HERV transcription at individual loci and connect specific sequence variants to tumorigenesis and immune modulation. Our findings enhance the resolution of HERV mapping across individuals and cancer types, uncovering previously inaccessible variation in a historically overlooked portion of the genome. This work not only improves our understanding of HERV-driven disease mechanisms but also lays the groundwork for variant-informed biomarker discovery and therapeutic targeting in precision oncology.

Keywords: cancer, human endogenous retroviruses, HERV, dark genome, variational graph genome, graph genome, long-reads, long read, variants, structural variants, polymorphisms, polymorphic, indels, insertion, deletion, single nucleotide variants, SNV, clinical, functional variants, regulatory function

Roland Faure (Institut Pasteur, Paris), Jean-François Flot (Université libre de Bruxelles) and Dominique Lavenier (CNRS / IRISA). *The Alice assembler: dramatically accelerating genome assembly with MSR sketching.*

Abstract. The PacBio HiFi technology and the R10.4 Oxford Nanopore flowcells are transforming the genomic world by producing for the first time long and accurate sequencing reads. The low error rate of these reads opens new venues for computational optimizations. However, genome and particularly metagenome assembly using high-fidelity reads still faces challenges. Current assemblers (e.g., Flye, hifiasm, metaMDBG) struggle to efficiently resolve highly similar haplotypes (homologous chromosomes, bacterial strains, repeats) while maintaining computational speed, creating a gap between rapid and haplotype-resolved methods.

We investigated this issue using on several dataset including a human gut microbiome sequencing and a diploid, finding that hifiasm_meta and metaFlye required over a month of CPU time to produce an assembly, while metaMDBG, which collapses similar strains, assembles the same dataset in four days.

We present Alice, a new assembler which introduces a new sequence sketching method called MSR sketching to bridge this gap and produce efficiently haplotype-resolved assemblies, for both genomic and metagenomic datasets. On the aforementioned human gut dataset, Alice completed the assembly in just 7 CPU hours. Furthermore, the analysis of the assemblies revealed that Alice missed <1% of abundant 31-mers (≥20x coverage), compared to >15% missed by both metaMDBG and hifiasm_meta.

Overall, our results indicate that Alice accelerates assembly dramatically while providing high quality assemblies, offering a powerful new tool for the field.

Keywords: High-fidelity long reads, Sketching, Genome assembly, Metagenome assembly, Strain separation

Leonardo Morelli (Laboratory of Chromatin Biology & Epigenetics, CIBIO, University of Trento, Trento, Italy), Stefano Cretti (Laboratory of Chromatin Biology & Epigenetics, CIBIO, University of Trento, Trento, Italy), Davide Cittaro (Center for Omics Sciences, Hospital San Raffaele, Milan, Italy), Tiago P. Peixoto (Inverse Complexity Lab, IT:U Interdisciplinary Transformation University, Linz, Austria) and Alessio Zippo (Laboratory of Chromatin Biology & Epigenetics, CIBIO, University of Trento, Trento, Italy). *Identification of interactions defining 3D chromatin folding from micro to meso-scale.*

Abstract. Understanding the structural principles of chromatin organization is a central challenge in computational epigenomics, largely due to the sparse, noisy, and complex nature of Hi-C data. Existing methods tend to focus either on local features, such as topologically associating domains (TADs), or global structures, like compartments. This methodological split often leads to poor agreement between models, limiting our ability to obtain a unified view of genome architecture. We introduce HiCONA, a novel graph-based framework that directly infers global 3D chromatin folding from both Hi-C contact maps and super resolution microscopy data. Unlike existing approaches, HiCONA optimizes a nested hierarchical representation of chromatin architecture by minimizing the entropy of the partition, thereby capturing the most informative and functionally relevant interactions. HiCONA enables simultaneous identification of topologically associating domains (TADs) and subcompartments using a single unified model, and performs robustly across gold-standard datasets. In benchmarking experiments, HiCONA recovers key chromatin contacts under both wild-type and cohesin-deficient conditions, offering insight into the structural consequences of architectural protein depletion. Furthermore, HiCONA provides a shared representation that facilitates direct comparison between imaging and sequencing-based data, bridging a major methodological gap in chromatin biology. By capturing chromatin folding from micro to mesoscale, HiCONA opens new avenues for understanding genome organization and its functional implications. This integrative and interpretable framework marks a significant advance in uncovering the forces that shape nuclear architecture, with potential applications in development, disease, and synthetic genome design.

Keywords: 3D chromatin Organization, Hi-C data analysis, DNA tracing data analysis, Multimodal data integration, Bayesian inference, Nested hierarchical clustering

Lucrezia Patruno (University College London Cancer Institute, Cancer Research UK Lung Cancer Centre of Excellence), Sophia Chirrane (University College London Cancer Institute, Cancer Research UK Lung Cancer Centre of Excellence) and Simone Zaccaria (University College London Cancer Institute, Cancer Research UK Lung Cancer Centre of Excellence). *POPSICLE: a probabilistic method to capture uncertainty in single-cell copy-number calling.*

Abstract. During tumour evolution, cancer cells acquire somatic copy-number alterations (CNAs), that are frequent genomic alterations resulting in the amplification or deletion of large genomic regions. Recent single-cell technologies allow the accurate investigation of CNA rates and their underlying mechanism by performing whole-genome sequencing of thousands of individual cancer cells in parallel (scWGS-seq). While several methods have been developed to identify the most likely CNAs from scWGS-seq data, the high levels of variability in these data make the accurate inference of point estimates for CNAs (i.e., a single value for the most likely copy number) challenging. Moreover, given that variability increases with increasing copy numbers, this is especially true when considering high amplifications and highly aneuploid cells, which play a key role in cancer. However, to date existing methods are limited to the inference of point estimates for CNAs in single cells and do not capture their related uncertainty.

To address these limitations we introduce POPSICLE, a novel probabilistic approach that computes the probability of having different copy numbers for every genomic region in each single cell. Using simulations, we show that POPSICLE improves ploidy and CNA inference for up to 20% of the genome in 90% of cells. Using a dataset comprising more than 60,000 of breast and ovarian cancer cells, we show how POPSICLE leverages uncertainty to improve the identification of genes that are recurrently highly amplified and might play a key role in tumour progression.

Keywords: Single-cell sequencing, Whole genome sequencing, Copy Number Alterations, Bayesian inference

Ghislain Fievet (Université de Lorraine), Julien Broséus (Université de Lorraine), David Meyre (Université de Lorraine) and Sébastien Hergalant (INSERM). *adverSCarial: a toolkit for exposing classifier vulnerabilities in single-cell transcriptomics.*

Abstract. Adversarial attacks pose a significant risk to machine learning (ML) tools designed for classifying single-cell RNA-sequencing (scRNA-seq) data, with potential implications for biomedical research and future clinical applications. We present adverSCarial, a novel R package that evaluates the vulnerability of scRNA-seq classifiers to various adversarial perturbations, ranging from barely detectable, subtle changes in gene expression to large-scale modifications. We demonstrate how five representative classifiers spanning marker-based, hierarchical, support vector machine, random forest, and neural network algorithms, respond to these attacks on four hallmarks scRNA-seq datasets. Our findings reveal that all classifiers eventually fail under different amplitudes of perturbations, which depend on the ML algorithm they are based on and on the nature of the modifications.

Beyond security concerns, adversarial attacks help uncover the inner decision-making mechanisms of the classifiers. The various attack modes and customizable parameters proposed in adverSCarial are useful to identify which gene or set of genes is crucial for correct classification and to highlight the genes that can be substantially altered without detection. These functionalities are critical for the development of more robust and interpretable models, a step toward integrating scRNA-seq classifiers into routine research and clinical workflows. The R package is freely available on Bioconductor (10.18129/B9.bioc.adverSCarial) and helps evaluate scRNA-seq-based ML models vulnerabilities in a computationally-cheap and time-efficient framework.

Keywords: Single-cell RNA-sequencing, Transcriptomics, Adversarial attacks, Explainable artificial intelligence (XAI), Machine learning, Cell annotation, Classifiers

POSTERS

Hasindu Gamaarachchi (University of New South Wales), James Ferguson (Garvan Institute of Medical Research), Hiruna Samarakoon (University of New South Wales), Kisaru Liyanage (University of New South Wales) and Ira Deveson (Garvan Institute of Medical Research). *Squigulator: Simulation of nanopore sequencing signal data with tunable parameters.*

Abstract. In silico simulation of high-throughput sequencing data is a technique used widely in the genomics field. However, there is currently a lack of effective tools for creating simulated data from nanopore sequencing devices, which measure DNA or RNA molecules in the form of time-series current signal data. Here, we introduce Squigulator, a fast and simple tool for simulation of realistic nanopore signal data. Squigulator takes a reference genome, a transcriptome, or read sequences, and generates corresponding raw nanopore signal data. This is compatible with basecalling software from Oxford Nanopore Technologies (ONT) and other third-party tools, thereby providing a useful substrate for development, testing, debugging, validation, and optimization at every stage of a nanopore analysis workflow. The user may generate data with preset parameters emulating specific ONT protocols or noise-free "ideal" data, or they may deterministically modify a range of experimental variables and/or noise parameters to shape the data to their needs. We present a brief example of Squigulator's use, creating simulated data to model the degree to which different parameters impact the accuracy of ONT basecalling and downstream variant detection. This analysis reveals new insights into the nature of ONT data and basecalling algorithms. We provide Squigulator as an open-source tool for the nanopore community at https://github.com/hasindu2008/squigulator.

Keywords: nanopore, simulation, signal

Christoph Bloß (Helmholtz Institute Freiberg for Resource Technology), Nora Schönberger (Helmholtz Institute Freiberg for Resource Technology), Purvi Jain (Helmholtz Institute Freiberg for Resource Technology), Gerda Techert (Helmholtz Institute Freiberg for Resource Technology) and Franziska Lederer (Helmholtz Institute Freiberg for Resource Technology). *Binder Hunters: Leveraging Phage Surface Display and NGS to Discover Novel Metal-Binding Peptides.*

Abstract. Introduction: Phage display is inherently sensitive, error-prone and labour-intensive. Advances in high-throughput sequencing technologies are providing insights into the peptide library space and the dynamics of directed evolution, improving our understanding of specific and non-specific binding behaviour. However, the identification of the best binding peptides is constrained by the limitations of traditional phage display. Selection-related and propagation-related library biases, off-target peptides and mutations in the phage propagation system often lead to the dominance of parasitic sequences resulting in an unrecorded collapse of library diversity and subsequent loss of high-affinity binders.
Methods: Using Differential Expression Analysis (DEA), meaningful differences in the evolution of the peptide population throughout consecutive biopanning rounds and between a target and control should be observable.
Results: This method enables the observation of meaningful differences in the development of the peptide population in successive rounds of biopanning and between a target and a control.
Discussion: This statistical model is a useful tool for clustering, analysing enrichment studies, the degree of relationship within the biopanning series and conserved peptide regions indicating a binding motif.

Keywords: Directed Evolution, Phage Surface Display, Next-Generation Sequencing, Differential Expression Analysis, Enrichment Analysis, Motif Finding, Resource Technology, Metal Recovery, Computational Biology, Bioinformatics

Runxuan Zhang (The James Hutton Institute). *Building a High-Resolution Barley Pan-Transcriptome to Uncover Genotype-Dependent Transcriptional Complexity.*

Abstract. We developed a barley pan-transcriptome by analyzing short- and long-read RNA sequencing datasets from 20 inbred genotypes, representing the diversity of domesticated barley across multiple tissues. To mitigate single-reference bias in transcript quantification, we generated genotype-specific reference transcript datasets (RTDs) and integrated them into a linear pan-genome framework, resulting in a pan-RTD. Our approach applies stringent filtering criteria to eliminate misassembled transcripts, enhance transcript diversity, and improve the accuracy of transcript-level quantification. This framework enables the categorization of transcripts into core, shell, or cloud groups.

Focusing on core transcripts (expressed in all genotypes), we identified significant variation in transcript abundance across tissues and genotypes, influenced by RNA processing, gene copy number, structural rearrangements, and promoter motif conservation. This study provides a systematic approach to analyzing transcriptional diversity across tissues and genotypes in barley, offering valuable insights into gene expression dynamics.

Reference

Guo, W., Schreiber, M., Marosi, V.B. et al. A barley pan-transcriptome reveals layers of genotype-dependent transcriptional complexity. Nat Genet 57, 441–450 (2025). https://doi.org/10.1038/s41588-024-02069-y

Keywords: transcript assembly, genotype, pan-transcriptome, gene expression, plant

Mounchili Njifon Aristide (Centre Pasteur du Cameroun), Abdou Fatawou Modiyinji (Centre Pasteur du Cameroun) and Richard Njouom (Centre Pasteur du Cameroun). *Identification of NS5B Resistance-Associated Mutations in Hepatitis C Virus Circulating in Treatment Naïve Cameroonian Patients.*

Abstract. NS5B polymerase inhibitors are the cornerstone of the current HCV (hepatitis C virus) infection treatment regimen. Treatment with direct-acting antivirals (DAAs) is quite successful. But they might be less effective if resistance-related mutations are present, especially if they affect non-structural polymerase protein 5B (NS5B). The purpose of this research was to find potential naturally occurring DAA (Direct-antibody action) mutations in the HCV NS5B gene linked to DAA resistance in patients with chronic hepatitis C from Cameroon who had not received any therapy. Whole blood samples were collected from patients with chronic hepatitis C from which plasma was subsequently separated and store at -800C for molecular analysis. The NS5B gene fragments were amplified using the designated primers and nucleotide sequences were acquired via the Sanger sequencing apparatus. The Geno2pheno 0.92 software was used to analyze the resistance profile. Analysis of NS5B sequences revealed three genotypes 1, 2 and 4 with numerous mutations. The significant S282T mutation, which confers high sofosbuvir resistance, was found in one patient, while the resistance-associated NS5B C316N polymorphism was found in 16 sequences. The Q309R mutation was detected in 19 genotype 1 sequences, and the L320F mutation was found in one genotype 4f sequence. Our investigation revealed that HCV patients who had not previously received DAA therapy exhibited a variety of NS5B gene alterations. Consequently, future treatment failure may be more likely due to these alterations.

Keywords: NS5B protein, hepatitis C virus, resistance associated, mutations, Cameroon

Carolina Monzó (Institute for Integrative Systems Biology (I2SysBio) - Spanish National Research Council (CSIC)), Carlos Blanco (Institute for Integrative Systems Biology (I2SysBio) - Spanish National Research Council (CSIC)), Alejandro Paniagua (Institute for Integrative Systems Biology (I2SysBio) - Spanish National Research Council (CSIC)), José Manuel Morante-Redolat (Departamento de Biología Celular, Biología Funcional y Antropología Física, Universidad de Valencia), Isabel Fariñas (Departamento de Biología Celular, Biología Funcional y Antropología Física, Universidad de Valencia) and Ana Conesa (Institute for Integrative Systems Biology (I2SysBio) - Spanish National Research Council (CSIC)). *Investigating Transcript Divergency in the Mouse Aging Brain Using Long-Read Transcriptomics.*

Abstract. Aging is a gradual decline in the overall function of organisms, leading to increased vulnerability to mortality and pathological states, including cognitive decline and neurodegenerative diseases. Various aspects of transcription including incorrect RNA processing leading to alternative isoform generation and increased transcriptional speed, have been found in the aging brain. Notably, splicing noise follows a distinctive pattern and increases with age, particularly affecting genes implicated in neurodegeneration. Transcript divergency (TD) represents a rare but real pool of RNA molecules in the transcriptome that originate from transcriptional errors and mis-splicing, reducing the productive fraction of the transcriptome. However, as the TD can only be extracted by full-length single-molecule transcript sequencing, whether TD is associated with decline processes remains to be elucidated.

We propose to investigate the role of TD in the aging brain using long-read RNA sequencing (lrRNA-seq). To do so, we have generated replicated datasets of young and old mice brains, employing both PacBio and Nanopore long-read sequencing technologies, alongside Illumina short-reads.

Our preliminary analyses using PacBio lrRNA-seq data, revealed that the ratio between known and novel transcript models (including noisy RNA molecules, intron retention events, technical errors etc.) in old mice increased in comparison to young animals for the majority of genes. We also found higher intron retention rates in old mice brains. These preliminary findings underscore the potential of lrRNA-seq in capturing TD signals in the aging brain. However, additional research is needed to separate the TD signal from technical errors in long reads sequencing data.

Keywords: long-reads, transcriptomics, aging brain, transcript divergency

Julieta Viglino (Universitat de Barcelona, Barcelona, Spain), Alejandro Torvisco (Universitat de Barcelona, Barcelona, Spain), Martijn Nawijn (University of Groningen, Netherlands; University Medical Center Groningen, Netherlands), Maarten van den Berge (University of Groningen, Netherlands; University Medical Center Groningen, Netherlands), Alvar Agusti (Universitat de Barcelona, Spain; IDIBAPS, Barcelona, Spain; CIBERES, Spain; Hospital Clinic, Barcelona, Spain), Sandra Casas (Institut d'Investigacions Biomediques August Pi i Sunyer (IDIBAPS), Barcelona, Spain) and Rosa Faner (Universitat de Barcelona, Spain; IDIBAPS, Barcelona, Spain; CIBERES, Spain). *TSS depth scores enable molecular profiling in Chronic Lung Disease.*

Abstract. Chronic obstructive pulmonary disease (COPD) is a heterogeneous condition characterized by lung damage. Profiling circulating cell-free DNA (cfDNA) may offer a novel, non-invasive approach for better disease understanding and molecular subtyping. This study aims to assess the utility of cfDNA profiling in COPD.

Plasma derived cfDNA was obtained from 18 normal lung function controls and 29 COPD patients with different airflow limitation ranges. In all samples, Whole Genome Sequencing (WGS) performed with 10x coverage. Custom transcription start sites (TSS) depth scores were developed and measured. With these scores we mapped cells of origin and obtained differential depth genes (DDGs) between cases and controls.

The main results were that immune cells were major contributors to cfDNA. An increased contribution of Serous glandular cells and Alveolar Type I cells was observed in patients (Wilcoxon Test, $p<0.05$). We identified 391 DDGs ($p<0.05$, $|log2fc|>0.5$). Hierarchical clustering of DDGs revealed two clusters: C#1 was increased in controls and enriched in cellular processes pathways; C#2 was increased in patients and associated with oxidative damage and impaired cell death.

This study is the first to identify differences in cfDNA in COPD vs. controls. The cfDNA from COPD individuals was able to reflect well known disease processes such as oxidative damage and cell death, but validation of the findings in expanded datasets is needed.

Funding: Supported by The European Research Council (ERC) under the Horizon Europe research and innovation program (Grant Agreement No. 101044387).

Keywords: cfDNA, Fragmentomics, Whole Genome Sequencing, Liquid biopsies, Chronic disease, Disease Profiling

Zhihan Zhu (Max Planck Institute for Molecular Genetics), Helene Kretzmer (Hasso Plattner Institute for Digital Engineering) and Steve Hoffmann (Leibniz Institute on Aging - Fritz Lipmann Institute (FLI)). *metilene3: Identifying DMRs Across Multiple Conditions with Auto-Classification.*

Abstract. DNA methylation serves as a critical epigenetic mark across numerous species, and identifying differentially methylated regions (DMRs) is essential for understanding the regulatory mechanisms of the genome. Current DMR detection methods generally require the labeling of samples, meaning that each sample needs to be associated with a category such as “cancer” or “healthy control.” This requirement limits the discovery of new epigenetic patterns, especially when such labels cannot be confidently assigned, as in clinical settings where molecular diagnoses, like tumor subtypes, are unavailable. Additionally, there is a shortage of methods that allow for the quick and accurate comparison of multiple groups. To address this significant gap, we introduce metilene3, a tool for rapidly and precisely identifying DMRs across various groups. It functions in both supervised mode, utilizing user-provided labels, and unsupervised mode, autonomously clustering samples without labels. By segmenting the genome based on multiple pairwise methylation difference signals, metilene3 enables sample classification and DMR-based inference of epigenetic relationships. We demonstrate its utility on diverse human datasets, showing that metilene3 reveals new potential regulatory elements and sample stratifications. Thus, metilene3 substantially expands the horizons of epigenomics research.

Keywords: epigenomics, DNA methylation, genome regulation, cancer, clustering

Alexander J. Petri (Stockholm University), Mahmud Sami Aydin (Stockholm University) and Kristoffer Sahlin (Stockholm University). *De novo clustering of extensive long-read transcriptome datasets with isONclust3.*

Abstract. Long-read sequencing techniques can sequence transcripts from end to end, greatly improving our ability to study the transcription process and enabling more detailed analysis of diseases such as cancer. While several well-established tools exist for long-read transcriptome analysis, most are reference-based and, limited by the reference genome. This prevents analysis of organisms without high-quality reference genomes and samples or genes with high variability (e.g., cancer samples or some gene families) from being analyzed to their full potential. In such settings, analysis using a reference-free method is favorable. The computational problem of clustering long reads by region of common origin is well-established for reference-free transcriptome analysis pipelines. Such clustering enables large datasets to be split up roughly by gene family therefore, an independent analysis of each cluster. However, none of tools can efficiently process the large amount of reads that are now generated by long-read sequencing technologies.

We present isONclust3, an improved algorithm over isONclust, to cluster massive longread transcriptome datasets at the gene family level. Like isONclust, IsONclust3 represents each cluster with a set of minimizers. However, unlike other approaches, isONclust3 dynamically updates the cluster representation during clustering by adding high-confidence minimizers from new reads assigned to the cluster. We show that isONclust3 yields results with higher or comparable quality to state-of-the-art algorithms but is 10-100 times faster on large datasets. Also, using a 256Gb computing node, isONclust3 was the only toolthat could cluster 37 million PacBio reads, which is a typical throughput of the recent PacBio Revio sequencing machine.

Decision: (conflict)
Keywords: long reads, transcriptomics, sequence clustering, minimizers

Ali Hamraoui (GenomiqueENS, Institut de biologie de l'ENS (IBENS), 75005 Paris, France), Audrey Onfroy (Institut Mondor de Recherche Biomédicale, Inserm U955 - Team 9, Créteil, France), Catherine Senamaud-Beaufort (GenomiqueENS, Institut de biologie de l'ENS (IBENS), 75005 Paris, France), Fanny Coulpier (Institut Mondor de Recherche Biomédicale, Inserm U955 - Team 9, Créteil, France), Piotr Topilko (Institut Mondor de Recherche Biomédicale, Inserm U955 - Team 9, Créteil, France), Stephane Le Crom (GenomiqueENS, Institut de biologie de l'ENS (IBENS), 75005 Paris, France), Sophie Lemoine (GenomiqueENS, Institut de biologie de l'ENS (IBENS), 75005 Paris, France) and Morgane Thomas-Chollier (GenomiqueENS, Institut de biologie de l'ENS (IBENS), 75005 Paris, France). *A systematic benchmark of bioinformatics methods for single cell and spatial RNA-seq Nanopore long reads data.*

Abstract. Alternative splicing is a pivotal mechanism contributing to transcriptome complexity and proteome diversity, influencing key biological processes and linked to various diseases. Recent methodological advances in single-cell RNA sequencing (scRNA-seq) using long-read technologies, such as Oxford Nanopore Technologies (ONT) and Pacific Biosciences (PacBio), enable the resolution of isoform diversity at the single-cell level, overcoming the limitations of short-read sequencing, which often fails to capture full-length transcripts.

We benchmarked multiple bioinformatics methods for processing Nanopore scRNA-seq data, including hybrid methods (Sicelore, Snuupy, and scNapBar) and long-read-only methods (FLAMES, Sockeye, Sicelore2.1, scNanoGPS, and Bambu), alongside isoform-level methods such as Isosceles and IsoQuant. Our evaluation spans various library preparation protocols and sequencing depths, assessing performance in terms of runtime, memory usage, scalability, barcode and UMI correction accuracy, transcript quantification, and biological insight at both gene and isoform levels.

Hybrid methods demonstrated high precision in cell barcode assignment but were limited by elevated sequencing costs and analytical complexity. In contrast, long-read-only pipelines—particularly wf-single-cell and Bambu—provided an effective balance between computational efficiency and read assignment accuracy, showing high concordance with short-read gene expression data. Furthermore, these methods surpassed CellRanger in cell type identification and annotation. Notably, Sicelore2.1 exhibited the most accurate read assignment to annotated transcripts and, when combined with IsoQuant, delivered the most reliable transcript assemblies.

This comprehensive benchmarking provides crucial insights for selecting optimal bioinformatics methods tailored to specific research objectives in single-cell and spatial transcriptomics.

Keywords: Single cell RNAseq, Spatial RNAseq, long-read, Nanopore, transcriptomics

Davide Bressan (University of Trento), Daniel Fernandez-Perez (IRB Barcelona), Alessandro Romanel (University of Trento) and Fulvio Chiacchiera (University of Trento). *SpikeFlow: A Framework for Standardized ChIP-Rx Data Analysis with Multiple Spike-in Normalization Methods.*

Abstract. ChIP-seq with exogenous spike-in chromatin, also known as ChIP-Rx, has emerged as an essential tool for investigating histone modification changes across biological conditions. Despite its widespread use, researchers face challenges in applying appropriate normalization methods, which differ from standard ChIP-seq workflows. In this work, we introduce SpikeFlow, an integrated Snakemake pipeline that addresses current ChIP-Rx data analysis limitations. SpikeFlow streamlines data processing by automating spike-in normalization while offering multiple scaling options to handle diverse experimental designs. Our tool performs standard and normalized peak calling, followed by differential analysis for both histone modifications and transcription factor binding sites. Comparative analysis showed that spike-in normalization for peak calling identifies regions with more consistent changes in bindings, potentially reducing false positives. The workflow generates a detailed analysis report with several quality control and plots that facilitate results interpretation. We compared SpikeFlow against established tools like DiffBind and SpikChIP and demonstrated comparable performance across diverse biological models while providing enhanced usability and integration. By integrating functionalities that were previously distributed across multiple tools, SpikeFlow streamlines ChIP-Rx data analysis and advances standardized analytical practices in the field.

Keywords: Sequencing, Epigenetics, ChIP-Seq, NGS, Workflow, Pipeline, Snakemake, Spikein

Maëlys Delouis (Institut Pasteur, Dpt Genome and Genetics, Spatial Regulation of Genomes), Romain Koszul (Institut Pasteur, Dpt Genome and Genetics, Spatial Regulation of Genomes) and Axel Cournac (Institut Pasteur, Dpt Genome and Genetics, Spatial Regulation of Genomes). *Eukaryotic plasmids computational identification.*

Abstract. Natural plasmids are genomic features often observed in bacteria, playing a role in metabolism, pathogenicity and the evolutionary trajectory of species (Smillie et al., 2010). However, known examples of plasmids are rare in eukaryotes, as only two systems have been clearly documented: the 2-micron plasmid found in the yeast S. cerevisiae (McQuaid et al., 2019) and other related yeasts, and the Ddp5 plasmid in the amoeba D. discoideum (Metz et al., 1983). The scarcity of eukaryotic plasmids has several possible causes: nuclear membranes acting as protective barriers, specific defense mechanisms, or such DNA sequences being overlooked since they are not systematically sought in assembly projects.

Using a combination of genomics datasets (Hi-C, ChIP-seq, RNA-seq) in various conditions, we recently showed that the 2-micron plasmid displays heterogeneous contacts with its host genome, with hotspots overlapping long non-transcribed regions (Girard et al., 2025). Interestingly, Ddp5 also presents a similar contact pattern along the amoeba genome. Based on these contact signatures, we hypothesized that revisiting published assemblies considering Hi-C data might unveil previously discarded sequences not included in the scaffolded genome that may display contacts suggesting a plasmidic nature.

We developed a pipeline that re-assembles unmapped assembly reads, building contigs absent from the reference genome. We then quantify, using Hi-C datasets, the contacts made by those sequences with the corresponding published reference genome. Profiles are fed to a model trained on contact profile dataset to sort the contigs into categories including chromosomal regions, mitochondrial or extra-nuclear organelle genomes, and potentially overlooked nuclear plasmids.

Keywords: genomics, eukaryotes, natural plasmid, Hi-C, assembly

Xiongbin Kang (MD Anderson Cancer Center, University of Texas), Wenhai Zhang (College of Biology, Hunan University), Yichen Li (College of Biology, Hunan University), Xiao Luo (College of Biology, Hunan University) and Alexander Schoenhuth (Bielefeld University). *HyLight: Strain aware assembly of low coverage metagenomes.*

Abstract. Different strains of identical species can vary substantially in terms of their spectrum of biomedically relevant phenotypes. Reconstructing the genomes of microbial communities at the level of their strains poses significant challenges, because sequencing errors can obscure strain-specific variants. Next-generation sequencing (NGS) reads are too short to resolve complex genomic regions. Third-generation sequencing (TGS) reads, although longer, are prone to higher error rates or substantially more expensive. Limiting TGS coverage to reduce costs compromises the accuracy of the assemblies. This explains why prior approaches agree on losses in strain awareness, accuracy, tendentially excessive costs, or combinations thereof. We introduce HyLight, a metagenome assembly approach that addresses these challenges by implementing the complementary strengths of TGS and NGS data. HyLight employs strain-resolved overlap graphs (OG) to accurately reconstruct individual strains within microbial communities. Our experiments demonstrate that HyLight produces strain-aware and contiguous assemblies at minimal error content, while significantly reducing costs because utilizing low-coverage TGS data. HyLight achieves an average improvement of 19.05% in preserving strain identity and demonstrates near-complete strain awareness across diverse datasets. In summary, HyLight offers considerable advances in metagenome assembly, insofar as it delivers significantly enhanced strain awareness, contiguity, and accuracy without the typical compromises observed in existing approaches.

Keywords: genome assembly, metagenome assembly, hybrid assembly, democratizing assembly, strain-aware assembly, low coverage sequencing

Jakob Heinz (Harvard University), Matthew Meyerson (Dana Farber Cancer Institute) and Heng Li (Dana Farber Cancer Institute). *Foldback Read Artifacts in Oxford Nanopore Datasets.*

Abstract. Cancer genomes undergo significant and complex genomic rearrangements, which long-read technologies from Oxford Nanopore Technologies (ONT) or Pacific Biosciences (PB) can help elucidate at the DNA and RNA levels. For long-read structural variation (SV) calling, we typically look for read "breakpoints," where fragments of the same read map to different locations on a reference genome. When doing so, we discovered an elevated number of ONT reads supporting foldback or inverted duplication SVs throughout the genome. In ONT direct-cDNA samples of the HCC1395, and K562 cell lines, approximately 15-25% of all reads supported a foldback event, while in matched ONT direct-RNA samples, at most one read supported a foldback event. We suspect that the elevated rate of foldbacks is not an actual biological event but rather a technical artifact. We analyzed cDNA samples from mouse brain and liver samples, HG002 and K562 cell lines, and gDNA metagenomic samples to explore this technical artifact further. We found numerous reads had known adaptor sequences between the foldback alignments. Foldback artifacts were most prevalent in direct-cDNA libraries (15-25%), observed at lower rates in metagenomic gDNA libraries (0.5-3%) and standard cDNA libraries (~0.1%), and absent in direct-RNA libraries. Unidentified foldback artifacts can lead to specificity issues in SV calling and tangles in assembly graphs. Here, we propose a quality control tool, the Break-inator, to flag foldback and chimeric reads present in the alignment file.

Keywords: ONT, Long-read Sequencing, Technical Artifacts, direct-cDNA

Jingyu Hao (Hong Kong University of Science and Technology), Juntao Zhao (Hong Kong University of Science and Technology), Yuqian Bian (Hong Kong University of Science and Technology) and Weichuan Yu (Hong Kong University of Science and Technology). *A Ground Truth-Free Scoring Framework for Merging Structural Variants Detection Results.*

Abstract. Structural variants (SVs) are associated with various diseases such as neurological disorders and cancers. Long-read sequencing has enabled us to detect SVs more easily than the 2nd generation sequencing techniques. Consequently, quite a few SV detection tools have been developed. However, these tools provide divergent detection results from the same samples. It is difficult for us to evaluate or compare SV detection results by using traditional metrics (e.g., precision and recall) because the ground truth is often unavailable in real applications.

To reduce the false positives, researchers often use callset-merging strategies to prioritize high-confidence variants for downstream analyses. However, current merging methods (e.g., SURVIVOR and Jasmine) also yield inconsistent results. Similarly, the lack of ground truth makes it difficult to choose a suitable merging method.

To address this issue, we propose to evaluate merging methods in real applications by using a ground truth-free scoring framework[1]. Concretely, our method evaluates the input trustworthiness of each merging method and then calculates the likelihood of the merged callset, which is the score of each merging method. The method with the highest score corresponds to the optimal merging strategy for a given scenario, whose callset is considered to be the most reliable. Experiment results on HG002 dataset show that our final callset has achieved 98.24% precision, surpassing the average of 94.57% for SV detection tools. By quantifying performance across callset-merging methods, our framework enhances confidence in merged results for subsequent downstream analyses.

[1] Fang, et al. ACMTIST 11.6 (2020): 1-24.

Keywords: Genomics, Structure variants detection, Structure variants merging, Long-read sequencing

Hyeyeong Hwang (Research Institute, National Cancer Center, 232 Ilsan-ro, Goyang-si, Gyeonggi-do, 10408, Republic of Korea), Daejin Hyung (Research Institute, National Cancer Center, 232 Ilsan-ro, Goyang-si, Gyeonggi-do, 10408, Republic of Korea), Namhee Yu (Research Institute, National Cancer Center, 232 Ilsan-ro, Goyang-si, Gyeonggi-do, 10408, Republic of Korea), Sehwa Hong (Research Institute, National Cancer Center, 232 Ilsan-ro, Goyang-si, Gyeonggi-do, 10408, Republic of Korea), Soo Young Cho (Hanyang University, 55 Hanyangdeahak-ro, Sangnok-gu, Ansan, Gyeonggi-do, 15588, Republic of Korea), Sang Myung Woo (Research Institute, National Cancer Center, 232 Ilsan-ro, Goyang-si, Gyeonggi-do, 10408, Republic of Korea) and Charny Park (Research Institute, National Cancer Center, 232 Ilsan-ro, Goyang-si, Gyeonggi-do, 10408, Republic of Korea). *Long-read sequencing to unveil aberrant transcriptome of pancreatic cancer.*

Abstract. Pancreatic cancer remains one of the deadliest malignancies, with a 5-year survival rate of only 13%. While multi-omics studies have identified critical variants and regulatory mechanisms in driver genes and oncogenes, effective treatments remain elusive. To explore the transcriptomic landscape of pancreatic cancer, we performed long-read sequencing (LR-seq) using the PacBio Revio system to uncover complex structural variants and alternative splicing events not observed from short-read sequencing. In our results, gene expression exhibited a high correlation (R = 0.63) compared to matched short-read RNA-seq, whereas the transcriptome varied, showing declining correlation (R = 0.43). Notably, novel isoforms possessing novel junctions comprised 16.79% of the total 150,904 isoforms, and 48.54% of them were already observed in the LR-seq results of our normal samples, GTEx, and ENCODE4. These novels were predominantly linked to tumor-associated genes (CD74, B2M, and DAXX) implicated in pancreatic cancer. Novels' expression profile could classify molecular subtypes of pancreatic cancer like known genes' expression and highlighted the disruption of the tumor microenvironment. Meanwhile, novel alternative start sites (ASS) exhibited tumor-specific peaks on promoter regions of histone modifications, and detected differential ASS events. We performed gene set enrichment tests using genes acquired from splicing event and isoform-specific expression analysis. Novel isoforms were involved in antigen processing and presentation function. In conclusion, LR-seq reveals global transcriptome disruptions driven by novel isoforms, which contribute to the molecular stratification of the disease. Additionally, these impair the interaction between pancreatic cancer cells and their microenvironment, implicated in a cold tumor phenotype.

Keywords: Long-read sequencing, Pancreatic cancer, Novel junction, Alternative splicing

Elena Cibola (University of Padova), Nicolò Gnoato (University of Padova), Paolo Martini (University of Brescia), Enrica Calura (University of Padova) and Chiara Romualdi (University of Padova). *A benchmark of tools for inferring Copy Number Alterations from single-cell RNA sequencing data.*

Abstract. Copy number alterations (CNAs) are large-scale genomic events that play a key role in cancer progression and treatment resistance. Single-cell RNA sequencing (scRNA-seq) offers the opportunity to infer CNAs at single-cell resolution, capturing tumor heterogeneity. However, the accuracy of tools developed for this purpose remains poorly characterized.

This work benchmarks four widely used tools for CNA inference from scRNA-seq data: InferCNV, SCEVAN, Numbat, and Xclone. Their output was compared to CNA profiles derived from bulk whole-genome sequencing (WGS), used as the ground truth, in high-grade serous ovarian cancer (HGSOC) samples. Among the evaluated tools, SCEVAN showed the highest accuracy.

The CNA profiles derived from scRNA-seq data were used in downstream analysis to quantify CNA signature activities and predict response to platinum-based chemotherapy. However, these predictions did not match those derived from bulk WGS data, revealing possible limitations in their current applicability at the level of single cell technology.

To enhance this benchmark, additional analyses are required on datasets with paired scWGS and scRNA-seq data from the same cells to better assess the real copy number profile at the level of single cell.

Overall, this benchmark provides guidance for selecting the optimal tool for CNA inference when working with scRNA-seq data, being particularly relevant for future studies investigating chromosomal instability and tumoral heterogeneity at the single-cell level.

Keywords: scRNA-seq, copy number alterations, ovarian cancer

Adrian Weich (Dermatology, Uniklinikum Erlangen, Friedrich-Alexander Universität (FAU) Erlangen-Nürnberg), Julio Vera-González (Dermatology, Uniklinikum Erlangen, Friedrich-Alexander Universität (FAU) Erlangen-Nürnberg) and Christopher Lischer (Hematology and Oncology, Uniklinikum Erlangen, Friedrich-Alexander Universität (FAU) Erlangen-Nürnberg). *Simulation-guided local coverage estimation for long-read DNA sequencing.*

Abstract. Motivation

Long-read DNA sequencing has become increasingly popular for whole-genome and targeted sequencing applications due to its ability to resolve complex structural variants and its reduced coverage requirements for variant detection. However, experimental planning for long-read DNA sequencing often lacks reliable a priori estimates of target region coverage, leading to costly and time-consuming pilot studies and biological replicates to meet expectations.

Results

We introduce a Monte Carlo-based simulation framework for estimating expected coverage of specific target regions in long-read DNA sequencing experiments. The simulation models regional coverage as a function of whole-genome sequencing depth and read length distribution, enabling data-driven predictions of regional sequencing outcomes.

We demonstrate the accuracy and utility of our framework through two case studies: First, we show that simulated coverage closely matches empirical long-read sequencing data. Second, we apply the simulation to heterogeneous cell populations, illustrating its applicability to complex experimental designs.

Our results show that the simulation can reliably reproduce observed coverage patterns and serves as a practical tool for both planning and interpretation. It enables researchers to optimize sequencing strategies for detecting rare variants, insertions, and other regions of interest, as well as to better rationalize variability in previous sequencing runs.

Availability and Implementation

The simulation is implemented entirely in Python and will be made freely available on PyPI upon publication. All iterations of the simulation are fully parallelized. The framework is CPU-bound, making it broadly applicable across a wide range of computing environments.

Keywords: Long-read sequencing, Whole-genome sequencing, Coverage Estimation, Monte Carlo Simulation, Experimental Design, High-throughput sequencing, Computational genomics, Oxford Nanopore Technologies, PacBio

Jan Thomas Schleicher (Department of Internal Medicine I, University Hospital Tübingen), Doreen Klingler (Department of Computer Science, University of Tübingen) and Manfred Claassen (Department of Internal Medicine I, University Hospital Tübingen). *The importance of accurate quantification of spliced and unspliced transcripts for 5'-sequencing data.*

Abstract. Single-cell RNA sequencing (scRNA-seq) allows for the detailed analysis of dynamic cellular processes. In particular, this has been enabled by the estimation of RNA velocity, the derivative of gene expression, from separate count matrices for different splice states, providing information about a cell's immediate future even in snapshot data. Useful velocity estimates strongly depend on accurate counts for spliced and unspliced transcripts. However, despite considerable advances in scRNA-seq, velocyto, the standard tool for spliced and unspliced mRNA molecule quantification, has not been updated to account for peculiarities of new protocols, such as popular approaches based on 5' chemistry. In this work, we demonstrate that velocyto incorrectly assigns counts to transcripts on the opposite DNA strand, resulting in inaccurate quantification of spliced and unspliced molecules coinciding with severe underestimation of total counts per cell. By comparing velocyto to alevin-fry, a quantification method compatible with 5'-sequencing data, we show that this limitation can result in substantial deviations in inferred velocities and differing interpretations. Furthermore, we present tidesurf, a command line tool for the quantification of spliced and unspliced transcript molecules from scRNA-seq libraries. Tidesurf takes aligned reads as its input, similar to velocyto, but is aware of the differing read orientation between 3' and 5' protocols. We demonstrate its accuracy on various publicly available 10x Genomics Chromium datasets with either 3' or 5' chemistry, whereas velocyto's results are highly erroneous for the latter. Considering broader applicability, our results highlight tidesurf as a potential replacement for velocyto.

Keywords: RNA velocity, velocyto, scRNA-seq, single-cell RNA sequencing, 5'-sequencing, splice-state aware expression quantification

Mohammad Darbalaei (Bioinformatics and Computational Biophysics, University of Duisburg-Essen, Essen, Germany), Philip Dujardin (Department of Medical Oncology, West German Cancer Center, University Hospital Essen, 45147 Essen, Germany), Thomas Mühlenberg (Department of Medical Oncology, West German Cancer Center, University Hospital Essen, 45147 Essen, Germany), Julia Zummack (Department of Medical Oncology, West German Cancer Center, University Hospital Essen, 45147 Essen, Germany), Sebastian Bauer (Department of Medical Oncology, West German Cancer Center, University Hospital Essen, 45147 Essen, Germany), Barbara M. Grüner (Department of Medical Oncology, West German Cancer Center, University Hospital Essen, 45147 Essen, Germany) and Daniel Hoffmann (Bioinformatics and Computational Biophysics, University of Duisburg-Essen, Essen, Germany). *A computational Bayesian method for efficient estimation of treatment resistance of cancer cells using barcoded cell mixtures.*

Abstract. We present a new computational Bayesian framework for efficiently quantifying the individual treatment resistance of cancer cells with distinct genotypes and resistance mutations in heterogenous cell pools in vitro and in vivo based on DNA barcoding technology. This model then allows deconvolution to estimate treatment resistance of individual cancer cell lines.

For this, the count-compositional DNA barcode data of cancer cell mixtures under treatment is modeled with a Dirichlet-multinomial hierarchical model that accounts for sampling variability and compositional constraints. In parallel, absolute growth dynamics of the cell mixtures in vitro (confluency) and in vivo (tumor volume) is modeled with appropriate noise terms and multilevel architectures to account for multiple sources of variation. Inference of treatment effects on individual cell lines is based on the joint conditional probability distribution, combining relative growth of mixture components and absolute growth of total mixtures. Thus, this joint model integrates the necessary information for deconvolution of absolute growth of individual cell lines and quantitatively propagates uncertainty of the integrated measurements.

Individual cell line experiments were used for validation and alignment with clinical resistance patterns confirmed biological findings.

To the best of our knowledge, this is the first statistical framework that directly quantifies treatment resistance using DNA barcoded cell mixtures. This method enables robust estimation of clone-specific resistance metrics in complex, mixed-cell populations and serves as a computational pipeline for efficient, quantitative assessment of treatment response. The approach allows for mutation-specific therapeutic decision-making while significantly reducing in vivo experiments.

Keywords: Cellular barcoding, Bayesian model, therapeutic decision

Jiayi Yao (Rigshospitalet), Yuliu Guo (Rigshospitalet), Miyako Kodama (Rigshospitalet), Drew Kaley Ann Thompson (Rigshospitalet), Emilie Sofie Engdal (Rigshospitalet), Alban Laus Obel Slabowska (Rigshospitalet) and Frederik Otzen Bagger (Rigshospitalet). *Benchmarking structural variant callers for short-read whole-genome sequence data.*

Abstract. Structural variations (SVs) are major contributors to human genetic diversity, and have demonstrated significant relevance in clinical diagnostics. Consequently, benchmarking SV calling tools is essential to identify optimal methods that ensure accurate and reliable genomic interpretation.

In this study, we evaluated 15 tools covering diverse SV calling approaches, including manta, Dysgu, Delly, Lumpy, TIDDIT, PopDel, GridSS, CNVpytor, Octopus, TARDIS, Smoove, SvABA, Wham, GATK gCNV, and Dragen. Two datasets were used for benchmarking: 1) HG002 from Genome in a Bottle Consortium 2) four samples from the Platinum Pedigree dataset, representing two generations of an extended family. Each tool was run with default parameters, and performance was assessed using Truvari.

In the GIAB dataset, several tools demonstrated high precision, reaching up to 0.9. Dragen, Dysgu and Manta led in recall performance, ranging from 0.4 to 0.6. Overall, Dragen achieved the highest F1 score, followed by Dysgu and Manta. For the Platinum Pedigree samples, both precision and recall declined, with precision ranging from 0.4 to 0.9 and recall from 0.1 to 0.25. Specifically, we observed tool-specific biases toward SV types, highlighting the need for tool integration. Notable, combining SVs called by Dysgu, Octopus, Manta and GRIDSS improved recall from 0.492 (Dysgu alone) to 0.564, while maintaining the F1 score despite increased false positives. Ongoing work will focus on optimizing parameters to further enhance the performance of combinations of SV callers.

Keywords: structural variant caller, benchmarking, short-read, whole-genome sequencing

Erkan Narmanli (Institut Curie) and Joshua Waterfall (Institut Curie). *Alignment- free classification for bulk RNASeq samples* .

Abstract. Differential gene expression analysis is a widespread technique for characterizing and classifying RNA-seq samples, particularly in biomedicine and phylogenetics. Standard “alignment-first” methods create, upstream of any comparative study, common frames of reference for comparison between samples, by estimating gene or mutation expression based on reference genomes or transcriptomes. However, this first processing step results in a significant loss of raw sequencing information. Here, we present an alignment-free classification pipeline, based directly on read expression across a reference read cohort. This method is faster, more accurate and allows better interpretation of classification results. Benchmarking on 28,000 samples from TCGA and GTEx shows results comparable to or even better than those obtained using alignment-first methods, using a hundred randomly selected reads, i.e. less than 0.002% of reads in a fastq file. With these results, we advocate the use of reference read banks: metameromes; coupled with an alignment-last strategy to analyse significant reads in classification.

Keywords: kmer, alignment-free, RNA-seq, classification, reads, TCGA, GTEx

Nadine Bestard Cuche (University of Edinburgh), Kellie Horan (University of Edinburgh), Vanessza Fentor (University of Edinburgh), Meryam Beniazza (University of Edinburgh), Andrea Corsinotti (University of Edinburgh) and Anna Williams (University of Edinburgh). *Multimodal Spatial Transcriptomics with VisiumHD: Assessing the Effects of Xenium Preprocessing on Data Integrity.*

Abstract. Spatial transcriptomics is increasingly applied to investigate localized biological processes in complex tissues. In this study, we analyse mouse brain sections from a focal subpial cortical grey matter demyelination model of multiple sclerosis using 10x Genomics VisiumHD. Our dataset includes fresh-frozen VisiumHD samples and "VisiumHD-postXenium" samples, which were processed after Xenium in situ transcriptomics. This design allows us to assess how Xenium processing impacts VisiumHD data quality and downstream analysis. The matched Xenium dataset will be presented in a companion poster. To support cell type annotation and cross-modality comparison, we also generated a matched single-nucleus RNA-seq (snRNA-seq) data from biologically matched tissue samples.

We observe that VisiumHD-postXenium samples consistently have reduced transcript and read counts compared to fresh VisiumHD, with a decrease in mitochondrial (mt) transcript percentage. Given the short half-life of mt RNAs, this may reflect degradation during the Xenium workflow and highlights the importance of considering RNA stability when designing multimodal protocols.

We are currently assessing whether differential expression patterns between control and demyelinated lesion regions are reproducible across both spatial platforms. For example, preliminary inspection using Xenium Explorer suggests lesion-associated microglia contain neuronal transcripts, potentially reflecting phagocytosis.

Together, this work highlights key technical considerations for spatial transcriptomic integration and contributes to best practices for multi-platform analysis.

Keywords: spatial transcriptomics, single-cell RNAseq, single-nuclei RNAseq, Xenium, Visium HD

Theo Brunet (Aix-Marseille University, CNRS, IBDM UMR7288, Marseille, France), Rikesh Jain (Aix-Marseille University, CNRS, LCB UMR7283, Marseille, France), Tam Mignot (Aix-Marseille University, CNRS, LCB UMR7283, Marseille, France) and Bianca Habermann (Aix-Marseille University, CNRS, IBDM UMR7288, Marseille, France). *Modeling of the metabolism of bacterial predation.*

Abstract. Predation significantly shapes ecosystems by mediating nutrient transfer and balancing predator-prey interactions. Unlike well-studied macroscopic predation, bacterial predation remains less understood despite its ecological importance. Bacteria employ diverse predatory strategies, such as secreting lytic compounds or direct invasion (10.1111/1462-2920.13171), while prey species evolve defensive responses.

This study investigates how prey bacteria metabolically respond to predation, focusing on the predator-prey model of Myxococcus xanthus and Escherichia coli. M. xanthus is a socially complex predator capable of consuming various bacteria and fungi. Through co-evolution experiments with collaborators, an enhanced predatory M. xanthus strain and a resistant E. coli strain (E7) were developed.

To identify the resistance mechanisms in E7, transcriptomic data from wild-type (WT) and E7 E. coli, both alone and interacting with M. xanthus, were analyzed. These data were integrated into the E. coli metabolic model iML1515 (10.1038/nbt.3956) via the E-flux method (10.1371/journal.pcbi.1000489), using flux balance analysis to simulate metabolic responses (10.1038/nbt.1614).

Preliminary metabolic flux analysis indicates increased reactive oxygen species (ROS), particularly superoxide, in E7. Experimental results confirm ROS as critical in E. coli's defense. Flux sampling revealed that E7's metabolic state when alone resembles WT E. coli under predation, suggesting E7 proactively adopts a defensive metabolic profile rather than reacting post-threat.

This study demonstrates how integrating omics data into metabolic models enables comparison across metabolic states, providing insights into bacterial predation defense mechanisms.

Keywords: Metabolic Modeling, Bacterial Predation, Transcriptomics

Adriano Fonzino (1) University of Bari Aldo Moro, Bari, Italy), Bruno Fosso (1) University of Bari Aldo Moro, Bari, Italy), Grazia Visci (1) University of Bari Aldo Moro, Bari, Italy), Carmela Gissi (1) University of Bari Aldo Moro, Bari, Italy 2) Consiglio Nazionale delle Ricerche, Bari, Italy), Graziano Pesole (1) University of Bari Aldo Moro, Bari, Italy 2) Consiglio Nazionale delle Ricerche, Bari, Italy) and Ernesto Picardi (1) University of Bari Aldo Moro, Bari, Italy 2) Consiglio Nazionale delle Ricerche, Bari, Italy). *NanoListener and NanoSpeech: ab initio detection of multiple epitranscriptomic modifications in ONT dRNA reads.*

Abstract. Epitranscriptomics modifications are emerging as important factors in fine-tuning gene expression and regulation, and have been linked to a variety of human diseases. Oxford Nanopore direct RNA sequencing (dRNA) offers the unique opportunity to chart the epitranscriptome at single-molecule resolution. Nonetheless, the accurate identification of RNA modifications remains a challenging task. Indeed, most of the available tools require time-consuming and computationally intensive pre-processing steps consisting of raw signals basecalling by modification-unaware software, re-squiggling, and modification calling by additional models. In this context, we developed a computational pipeline to train modification-aware basecallers, enabling the direct identification of modified nucleotides in native RNA reads using only raw current measurements. Our pipeline is based on two brand-new programs, NanoListener and NanoSpeech. NanoListener can build robust training datasets from various organisms and synthetic constructs. NanoSpeech, instead, implements a transformer model based on the speech-to-text translation paradigm and uses an expanded dictionary to detect ab initio canonical and modified ribonucleotides. By training NanoSpeech on datasets comprising both modified and unmodified in vitro transcribed constructs (IVT), we were able to basecall ab initio up to nine different modifications. Our pipeline can also deal with reads generated by the latest RNA004 sequencing kit.

Keywords: DEEP LEARNING, BIOINFORMATICS, SIGNAL PROCESSING, EPITRANSCRIPTOMICS, LONG READS, NATIVE RNA SEQUENCING

Sharon Natasha Cox (University of Bari Aldo Moro), Angelo Sante Varvara (University of Bari Aldo Moro and University of Milan) and Graziano Pesole (University of Bari Aldo Moro and CNR-IBIOM). *MitSorter: a standalone resource for methylation-based discrimination of mitochondrial DNA ONT reads.*

Abstract. Background: The discrimination between mitochondrial DNA (mtDNA) and nuclear mitochondrial DNA segments (NUMTs) is crucial for variant calling, as NUMTs can confound the analysis of mtDNA mutations. As Oxford Nanopore Technologies (ONT) enables direct methylation detection, offering the opportunity to classify these sequences based on the lack of CpG methylation in human mtDNA, we developed MitSorter, a bioinformatic tool that distinguishes mtDNA from NUMTs.

Methods: MitSorter processes ONT reads by integrating high-accuracy basecalling with Dorado and methylation-aware models to generate modification-aware BAM (modBAM) files. The tool assesses 5mCpG methylation at the read level, partitioning the dataset into two BAM files: unmethylated mtDNA reads and highly methylated NUMT reads. The tool was validated using the GIAB human ONT dataset.

Results: Mitochondrial-aligned reads showed a bimodal distribution. Considering all reads across three tested samples (N=99,897), the majority (69%,N=68,947) exhibited low CpG methylation levels (<5%), consistent with true mtDNA. A secondary peak at higher methylation levels (>60%) was observed, likely corresponding to NUMTs. The unsorted modBAM exhibited a mean 5mC fraction of 0.37±0.13. MitSorter generated an unmethylated modBAM with a mean 5mC fraction of 0.0022±0.0003, while the methylated modBAM had a 5mC fraction of0.63±0.28. Similar results were found for normal pancreatic tissue and for 12 private samples tested.

Conclusion: MitSorter is a reliable and efficient tool for distinguishing mtDNA from NUMTs using raw ONT data. This tool may greatly improve mtDNA variant analysis and facilitate the identification of novel NUMTs, as mtDNA-to-nuclear transfer is now considered an ongoing event, particularly in cancers.

Keywords: human mitochndrial DNA, variant calling, methylation, NUMTs

John Hawkins (European Molecular Biology Laboratory (EMBL)), Ilia Kats (German Cancer Research Center (DKFZ)), Rebecca Wagner (German Cancer Research Center (DKFZ)), Micheala Behm (German Cancer Research Center (DKFZ)), Dominik Lindenhofer (European Molecular Biology Laboratory (EMBL)), Lars Steinmetz (European Molecular Biology Laboratory (EMBL)) and Oliver Stegle (European Molecular Biology Laboratory (EMBL)). *Higher throughput and fidelity screens with flexible, higher accuracy barcode decoding.*

Abstract. DNA barcodes and their decoding are at the core of high-throughput molecular biology. Large-scale assays such as pooled CRISPR screens and their single-cell manifestations are pushing the boundaries of what is accurately feasible with current barcode decoding technology, as we ever demand the capability to identify more perturbations, more cells, more omics, more insights in the same assay. Despite the critical nature of the decoding task, few methods exist that permit accurate, flexible and versatile decoding in the more demanding of these applications. Current state-of-the-art is many tools for individual tasks, often with severe limitations in their applicability, and often with inflexible vendor-specific solutions. Here, we propose a principled information theory-based barcode decoding software solution that addresses a wide range of decoding tasks in a unified framework, including demanding applications. Our approach can handle flexible error models, including those encountered in long-read sequencing such as Nanopore, and supports decoding arbitrary and non-trivial constructions of multiple barcodes, linkers, UMIs, and logical combinations thereof. Finally, we present a graphical user interface for clear, user-friendly specification of each barcode decoding task.

Keywords: Barcodes, High throughput sequencing, Long-read sequencing, Information theory, Demultiplexing, UMIs, Pooled assays

Dohun Yi (Hanyang University) and Jin-Wu Nam (Hanyang Univ.). *Global cell line misidentification: regional variations and escalating risks.*

Abstract. Human cell line misidentification has posed a longstanding challenge to biomedical research, yet the full scale remains inadequately characterized. In this study, we re-authenticated a global dataset of 79,696 human cell line samples using ultrafast variant profiling and a cell type identification tool, revealing a misidentification rate of 5.12%. Striking regional differences were observed: misidentification in China has reached 16.7%, largely driven by HeLa infiltration at the cell bank level, whereas in other countries, misidentification originated at the laboratory level. Alarmingly, misidentification rates show a rising trend, suggesting an increased risk of compromised research outcomes. The accumulation of misidentified biospecimens in public repositories could mislead both experimental and data-driven biomedical sciences, underscoring the urgent need for heightened vigilance in cell line authentication.

Keywords: Rapid variant profiling, Cell line misidentification, Large-scale sequencing data analysis, Public data surveillance

Xinzhu Jiang (UM-SJTU Joint Institute, Shanghai Jiao Tong University, Shanghai), Cheng Wang (Global Institute of Future Technology, Shanghai Jiao Tong University, Shanghai), Jinpu Cai (UM-SJTU Joint Institute, Shanghai Jiao Tong University, Shanghai), Yuxuan Hu (School of Computer Science and Technology, Xidian University, Xi'an), Jin Gu (Department of Automation, Tsinghua University, Beijing), Qiuyu Lian (Gurdon Institute, University of Cambridge) and Hongyi Xin (Global Institute of Future Technology, Shanghai Jiao Tong University, Shanghai).
Harmonization, integration, and annotation of single-cell protein and transcriptomic atlases via unified cell ontology learning.
Abstract. Integrating and consistently annotating single-cell multi-omic data across diverse studies is essential for fully characterizing cellular heterogeneity. Technologies that concurrently measure surface proteins and transcriptomes at the single-cell level provide a unique opportunity to organize cells into unified ontologies, leveraging the well-established "cluster of differentiation" (CD) marker systems. However, existing integration methods often fail to preserve such ontologies due to technical biases, annotation inconsistencies, and mosaic surface markers across datasets. These methods either ignore cell ontologies by projecting cells into "flat" latent spaces, or rely on predefined ontologies—which are typically unavailable or unreliable in pathological contexts.

To address these challenges, we present CITE-pool, an unsupervised ontology-preserving framework for integrating mosaic single-cell protein and transcriptomic datasets. Inspired by CD-based annotation systems, CITE-pool builds a protein-anchored cell ontology to hierarchically integrate and organize cells across studies. When surface markers are missing or insufficient, pseudo-markers are imputed from transcriptomic features to continue ontology construction at deeper levels.

Benchmarking on both simulated and real datasets demonstrates CITE-pool's strengths in transparent integration, fine-grained annotation, and preservation of condition-specific variation. The resulting ontology enhances low-dimensional visualizations and supports flexible query-to-reference mapping with interpretable justification workflows analogous to flow cytometry gating. The pseudo-marker imputation further generalizes to spatial transcriptomics and enables identification of lymphoid structures in tumor microenvironments. Applied to COVID-19 datasets, CITE-pool reveals both shared and condition-specific cell ontologies, intuitively highlighting disease-associated cell states and markers. These results demonstrate the power and interpretability of CITE-pool for integrative atlas construction and full-resolution cellular annotation.

Keywords: Cell ontology, Single-cell multi-omics, Unsupervised integration, Surface protein inference

David Wragg (University of Aberdeen), Eunchai Kang (University of Aberdeen) and Mike Morgan (University of Aberdeen). *Harvesting more reads from single-cell combinatorial barcoding data with Scarecrow.*

Abstract. Combinatorial barcoding technologies such as split-pool-ligation sequencing (SPLiT-seq) involve sequential rounds of cell barcoding with potential to capture millions of single-cell transcriptomes in a single experiment. Accurate assignment of reads to cell barcode combinations depends on faithful sequencing, and existing bioinformatic pipelines assume barcodes are in identical positions in every sequencing read. However, small shifts in barcode position (1-2bp) can be caused by technical sequencing artifacts such as polymerase slippage, adapter misalignment, or rare indels introduced during adaptor ligation and/or tagmentation. While existing cell and sample demultiplexing tools can handle a small number of single nucleotide mismatches, the consequence of ignoring positional shifts can lead to loss of data for downstream analysis.

Here, we introduce scarecrow, which addresses these challenges by first screening a subset of reads against barcode whitelists to generate position-specific barcode distributions. These barcode profiles are then used to flexibly identify barcode positions in each read whilst accounting for jitter, prioritising barcode matches with the fewest mismatches and jitter. Our benchmarks on Parse Evercode and Scale Biosciences data show a 10-64% increase in usable reads compared to using a fixed position. Our approach requires no prior knowledge of barcode positions within the library structure, overcoming the limitations of existing tools. The outputs from scarecrow are structured for compatibility with standard tools (e.g., kallisto, STAR, umi-tools), offering a robust solution for maximizing data yields across diverse single-cell experiments.

Keywords: single cell RNA-sequencing, combinatorial barcoding, sequence analysis & processing

Vanessza Fentor (University of Edinburgh), Nadine Bestard Cuche (University of Edinburgh), Kellie Horan (University of Edinburgh), Meryam Beniazza (University of Edinburgh), Andrea Corsinotti (University of Edinburgh), Barry McColl (University of Edinburgh) and Josef Priller (University of Edinburgh). *Comparative Analysis of Xenium and Visium HD Spatial Transcriptomics Platforms Using a Modular Spatial Transcriptomics Pipeline.*

Abstract. Emerging spatial transcriptomics platforms offer unique advantages, but their comparative strengths and limitations in complex neuropathological contexts remain incompletely characterized. We employ a highly customizable analysis pipeline to perform a comprehensive comparison of two 10x Genomics technologies: Xenium (subcellular resolution) and Visium HD (whole-transcriptome), using matched cortical demyelination samples from a multiple sclerosis model, with accompanying single-nucleus RNA-seq data for validation. We aim to highlight strengths and weaknesses of these technologies that should be considered in the context of brain biology, which is a notoriously difficult site especially for cell segmentation. Which technology is better at answering which questions?

The pipeline's modular design allows high customization of analysis tools of each platform's data while enabling direct comparison through standardized metrics. Preliminary observations suggest platform-dependent detection patterns, though analysis is still undergoing to fully characterize their respective capabilities in neuroinflammatory contexts.

By maintaining identical sample preparation and analysis workflows where possible, this study may provide insights into how technological differences in resolution, sensitivity, and transcriptome coverage could influence biological interpretation in neurodegenerative disease research. The flexible framework could prove particularly valuable for researchers working with multiple spatial transcriptomics technologies, potentially facilitating more informed experimental design decisions for future investigations of complex neuroinflammatory processes.

Keywords: spatial transcriptomics, transcriptomics, Xenium, Visium HD, benchmarking, snRNAseq

Serghei Mangul (Stefan cel Mare University of Suceava), Mihai Covasa (Stefan cel Mare University of Suceava), Mihai Dimian (Stefan cel Mare University of Suceava), Viorel Munteanu (Technical University of Moldova) and Andrei Lobiuc (Stefan cel Mare University of Suceava). *Collecting, Processing, and Managing Personal Medical Data with PHGL-COVID.*

Abstract. The traditional provider-centric healthcare model, heavily reliant on physician expertise and the Healthcare Delivery Value Chain (HDVC), is being challenged by the rise of patient-centered Personal Health Records (PHRs). These digital tools empower patients to actively engage in their healthcare journey, influencing diagnosis, treatment, and outcomes. PHRs foster a proactive healthcare approach, shifting the clinical paradigm towards a more participatory model that enhances physician-patient interactions and improves patient health.

Integrating diverse and complex health data, including Electronic Health Records (EHRs), environmental, and genomic information, into comprehensive PHRs, coupled with Artificial Intelligence (AI) and Natural Language Processing (NLP) powered analytics, can further refine this proactive stance. This review explores the critical need for this data expansion, emphasizing its potential to revolutionize physician-patient dynamics. We examine how PHRs act as a catalyst for proactive care and the synergistic relationship between PHRs and AI, proposing a user-centric digital health data platform architecture and outlining general functional requirements.

Furthermore, we identify and analyze the challenges hindering the full potential of PHRs within the HDVC. Ultimately, this review underscores the transformative impact of PHRs in fostering a patient-centric and data-driven healthcare landscape, promising significant advancements in physician-patient relationships, continuity of care, and overall patient well-being.

Keywords: Personal Medical Data, ML, genomics

Nuo Xu (University of Toronto), Katherine Rynard (University of Toronto), Shreejoy Tripathy (University of Toronto), Maahil Arshad (University of Toronto), Elizabeth Radley (University of Toronto), Hua Luo (University of Toronto), Harry Smith (University of Toronto), Ellie Hogan (University of Toronto), Maria Fafouti (University of Toronto), Melanie Davie (University of Toronto), Ai Tian (University of Toronto), Brett Trost (University of Toronto), Julien Muffat (University of Toronto), John Calarco (University of Toronto), Hyun Lee (University of Toronto), Craig Smibert (University of Toronto) and Howard Lipshitz (University of Toronto). *Uncovering novel protein-coding isoforms in human iPSC-derived cortical neurons and their relevance to ASD.*

Abstract. Precise gene regulation is critical for human cortical neuron development, yet existing reference databases incompletely capture the diversity of transcripts and proteins expressed in these cells. To address this gap, we used a multifaceted functional genomics approach to analyze human induced pluripotent stem cells (iPSCs) differentiated into cortical neurons. We performed long-read RNA sequencing to capture full-length transcripts, short-read RNA sequencing to validate splice junctions, and proteomics to confirm translation of novel open reading frames.

Our data reveal a rich landscape of 200,000 splice isoforms (~100,000 novel). These mRNA isoforms code for 65,000 protein isoforms (40,000 novel), many of which are directly validated by proteomics, underscoring their biological relevance. Notably, mRNA and protein expression dynamics often differ across genes, in part, explained by extensive mRNA isoform switching across neuronal development, including changes in untranslated regions and splicing events that may influence mRNA stability and translation.

Importantly, whole-genome sequencing from autism spectrum disorder (ASD) cohorts suggests that some variants previously deemed noncoding may disrupt newly identified protein-coding exons. Reclassifying these variants as potentially pathogenic could refine ASD risk assessments and implicate novel disease mechanisms. In sum, our work provides a high-resolution view of transcript and protein complexity in human cortical neurons and underscores the importance of isoform-centric analyses for understanding neurodevelopment.

Keywords: alternative splicing, long-read sequencing, autism spectrum disorder, nextflow, developmental biology, iPSC-derived neurons

Marta Benegas Coll (BioBam Bioinformatics), Stefan Götz (BioBam Bioinformatics) and Ana Conesa Cegarra (Spanish National Research Council (CSIC)). *The Influence of Clustering Quality on Cell Type Prediction Accuracy.*

Abstract. Accurate cell type prediction is a critical step in the interpretation of single-cell RNA-seq data, as downstream biological insights strongly depend on the reliability of these predictions. Most annotation strategies rely on an initial unsupervised clustering step to identify groups of transcriptionally similar cells. However, clustering is sensitive to parameter choices, which can lead to substantial variation in cell grouping.

While it is widely acknowledged that clustering quality influences downstream analyses, the importance of conducting in-depth quality assessment remains underappreciated. Additionally, the extent to which suboptimal clustering impacts cell type prediction remains insufficiently characterized. To investigate this, an analysis is designed to assess the consistency and accuracy of cell type predictions at varying clustering qualities.

To conduct this analysis, a range of clustering outputs was generated using the Seurat package by varying parameters such as resolution and the number of dimensions. Clustering performance was evaluated using a variety of established metrics that categorize the outputs from high-quality to suboptimal. Cell type annotations were then obtained using both cluster-level (CellKb) and cell-level (SingleR) prediction tools on a well-annotated dataset. This setup enabled the comparison of predicted annotations against ground truth labels to estimate prediction accuracy.

By analyzing the correlation between clustering quality and annotation performance, this study examined the robustness of reference-based prediction tools and evaluated whether strong annotation methods can compensate for poor clustering. The findings aim to guide best practices in single-cell workflows, evaluate the importance of clustering assessment, and propose a set of quality metrics to conduct it.

Keywords: scRNA-Seq, clustering assessment, cell type prediction, quality check, single-cell

Fatemeh Kordevani (San Raffaele Scientific Institute), Francesca Tortorelli (San Raffaele Scientific Institute), Cristina Toffalori (San Raffaele Scientific Institute), Fabio Ciceri (San Raffaele Scientific Institute; Vita-Salute San Raffaele University), Luca Vago (San Raffaele Scientific Institute; Vita-Salute San Raffaele University) and Marco Punta (San Raffaele Scientific Institute). *Cell-Type Deconvolution Under Stress: The Impact of Noise, Evenness, and Biological Complexity.*

Abstract. Bulk RNA-seq is a widely adopted and cost-effective technology that provides quantification of RNA transcripts within biological samples. It is relatively common for bulk samples to be constituted of unsorted mixtures of different cell-types. Estimating the cellular composition of these samples, especially in disease contexts such as acute myeloid leukemia (AML), can offer valuable prognostic insights. To address this need, numerous tools have been developed that try to deconvolve cell-type signals from bulk data. In this study, we systematically evaluated 12 such deconvolution methods using the DeconBenchmark package.

We constructed cell type-specific expression profiles (including B, T, NK, myeloid, and AML cells) from an in-house single-cell RNA-seq dataset derived from post-transplant relapse AML bone marrow (BM-AML) samples. For benchmarking, we used two datasets: pseudo-bulk mixtures with known composition and 32 bulk BM-AML samples annotated with FACS-derived cell fractions. Performance was assessed using Pearson correlation and root mean square error between predicted and ground-truth cell fractions.

Most methods performed well on pseudo-bulks but showed decreased performance on real-case scenario bulk samples. Differences between single-cell and bulk technologies along with the fact that bulk samples are typically more noisy than pseudo-bulk ones (e.g., due to the presence of unprofiled cell types) can help explain some of these differences. Interestingly, we noted that most methods performed best on skewed cell profiles, with their accuracy instead decreasing when predicting more evenly balanced samples.

Our tests suggest that cell-type deconvolution of bulk RNA-seq cancer samples remains a challenging problem.

Keywords: Transcriptomics, Deconvolution, Benchmarking, Acute Myeloid Leukemia, Shannon Index

Enrique Vidal Beneyto (Universitat Politècnica de València (UPV), Genomics of Gene Expression lab, (I2SysBio-CISC), Paterna, Spain), Loris Mannino (Neuronal and Tissue Regeneration Lab, Centro Investigación Príncipe Felipe (CIPF), Valencia, Spain), Guillem Paniagua-Soriano (Neuronal and Tissue Regeneration Lab, Centro Investigación Príncipe Felipe (CIPF), Valencia, Spain), Eric López-Mocholi (Neuronal and Tissue Regeneration Lab, Centro Investigación Príncipe Felipe (CIPF), Valencia, Spain), Victoria Moreno-Manzano (Neuronal and Tissue Regeneration Lab, Centro Investigación Príncipe Felipe (CIPF), Valencia, Spain), Javier Buceta (The.Si.M.Bio.Sys. Lab, (I2SysBio-CISC-UV), Paterna, Spain), Sonia Tarazona-Campos (BiostatOmics Group, Universitat Politècnica de València (UPV), Valencia, Spain) and Ana Conesa (Genomics of Gene Expression lab, (I2SysBio-CISC), Paterna, Spain). *Cells2Spine: Spatial Transcriptomics and Single-Cell RNA-Seq in Spinal Cord Injury.*

Abstract. Spinal cord injury (SCI) remains a leading cause of paralysis and currently, no effective cure exists to restore neurological function following complete injury, though limited recovery is possible in incomplete cases. The cellular and molecular mechanisms underlying this recovery remain incompletely characterized.

This study investigated transcriptional and cellular composition changes across anatomical regions following SCI in rats, attributing gene expression alterations to specific cell populations. We employed 10X Visium Spatial Transcriptomics and Chromium Single-Cell technologies alongisde computational tools such as Scanpy for data exploration and visualization, EdgeR for differential expression analysis, RCTD and Cell2location for spatial deconvolution, and C-SIDE for linking spinal cord gene signatures to specific cellular populations.

Our spatial transcriptomic analysis revealed a distinct injury-specific cluster characterized by peripheral immune cell infiltration, as well as an increase of oligodendrocyte populations in the gray matter. Differential gene expression primarily highlighted upregulation of remyelination pathways and repair mechanisms at the injury site. Integration of spatial and single-cell data enabled precise mapping of transcriptional changes to specific cell types involved in the post-injury response cascade.

This methodological framework successfully addresses our research objectives and provides valuable insights into the molecular landscape following SCI. The approach can be extended to additional experimental conditions and temporal points to further investigate similar biological questions, potentially advancing our understanding of recovery mechanisms following incomplete spinal cord injury. These initial findings lay the groundwork for future studies that may contribute to the development of targeted therapeutic strategies to enhance functional recovery after SCI.

Keywords: Single-Cell, Spatial Transcriptomics, Spinal Cord, Gene Expression

Alessandro Brandulas Cammarata (University of Lausanne). *Robust Data-driven gene expression inference for RNA-seq using intergenic regions as estimation of background noise.*

Abstract. Bulk and single-cell RNA-Seq are powerful and widely used techniques that provide quantitative information on gene expression. While the primary focus of many applications is to estimate gene expression levels, a crucial first step in assessing gene activity is to distinguish technical or biological transcriptional noise from actively expressed genes. Typically, this is accomplished by setting an arbitrary abundance threshold (e.g., TPM>2) for calling a gene expressed. However, because of the substantial variation in technical and biological noise levels across RNA-Seq experiments, the usage of such a fixed threshold can lead to errors. To overcome these limitations, we propose an updated dynamic approach. First, we infer the amount of noise in a library by selecting reads mapped to intergenic regions. Using those intergenic regions and a fitted hurdle model, we can define a gene as expressed if its abundance is significantly higher than the distribution of this background noise. The accuracy of this approach outperforms other existing methods in determining the true state of genes, based on our evaluation compared to epigenetic markers, ribo-seq data, and genetic evidence in three animal species. It is also computationally efficient, allowing it to scale to a large database, such as Bgee. To further evaluate the quality of our prediction, we show that the proportion of expressed genes is biologically meaningful and stable between libraries originating from the same tissue (blood, brain), in both model and non-model organisms, whereas the standard fixed threshold leads to significant differences between samples on the 52 species tested.

Keywords: transcriptomics, expression state, Biological noise, gene filtering, single-cell

Francesca Longhin (Department of Information Engineering, University of Padova), Giacomo Baruzzo (Department of Information Engineering, University of Padova), Enidia Hazizaj (AB ANALITICA, Srl), Diego Boscarino (AB ANALITICA, Srl), Dino Paladin (AB ANALITICA, Srl) and Barbara Di Camillo (Department of Information Engineering, University of Padova). *Realistic simulation of NGS reads from tumoral samples with MOV&RSim.*

Abstract. Advances in bioinformatics pipelines for variant calling have accelerated with the decreasing cost of next-generation sequencing (NGS). Accurate detection of somatic mutations is critical for precision oncology, particularly for guiding therapy decisions. However, somatic variant calling remains challenging due to cancer heterogeneity, diverse mutational landscapes, and sequencing noise. A comprehensive dataset of fully characterised tumoral genomes, representing the variability across cancer types, is still lacking, even among synthetic data, limiting systematic evaluation and optimization of variant calling tools.

In this work, we evaluated nine existing somatic simulators (Synggen, BAMSurgeon, SVEngine, VarSim, Xome-Blender, tHapMix, Pysim-sv, SCNVSim, HeteroGenesis) for their ability to control biological (variant type, number, position, length, content, zygosity; sample clonality and contamination) and technical parameters (sequencing errors, coverage, base quality). None provided full control over both domains, nor guidance for cancer-specific parameter tuning. To address this, we developed MOV&RSim, a novel simulator that leverages data-driven information to set variants and reads characteristics, generating realistic tumoral samples, and providing complete control on both biological and technical parameters. Additionally, we leveraged well-annotated variant databases (COSMIC and TGCA) to create cancer-specific presets that inform the simulator's parameters for 21 cancer types.

MOV&RSim, packaged in Docker and freely available for academic use, enables users to simulate biologically realistic and technically nuanced tumoral samples. It represents the most flexible and comprehensive simulation framework currently available for benchmarking and optimizing somatic variant calling pipelines across diverse cancer types.

Keywords: Realistic Simulation, Cancer-specific Presets, Genetic Variants, Sequencing Reads, Variants characteristics, Reads characteristics, Somatic sample simulator, Gold-standard sample

Juan Francisco Cervilla (Earlham Institute), Carlos Blanco (Institute of Integrative Systems Biology (i2sysbio) - CSIC), Carolina Monzó (Institute of Integrative Systems Biology (i2sysbio) - CSIC), Wilfried Haerty (Earlham Institute) and Ana Conesa (Institute of Integrative Systems Biology (i2sysbio) - CSIC). *SQANTI-single cell: Quality control and curation of long-read sequencing data and transcriptome assemblies at single cell resolution.*

Abstract. Long read RNA sequencing (lrRNA-seq) has enabled the study of full-length isoforms and their implications, as there is growing evidence that alternative splicing is involved in many regulatory processes and diseases. Moreover, recent advances in single cell technologies (sclrRNA-seq) have allowed combining both technologies to study isoforms expression and diversity within a complex population of cells. However, technical errors stemming from library preparation protocols and sequencing platforms, along with the different strategies for transcriptome reconstruction, introduce additional sources of variability and potential biases in isoform detection, quantification, and interpretation at the cell level.

Here, we present SQANTI-single cell (SQANTI-sc), a computational tool designed to assess the quality and curation of single cell, long read data. Leveraging SQANTI3 and SQANTI-reads, originally developed for bulk long-read sequencing, SQANTI-sc adapts and expands their functionalities to the single cell context. Through a series of modules, SQANTI-sc evaluates sample quality at read and isoform level using SQANTI3 structural categories, detects potential technical biases from single cell library preparation and sequencing protocols, and facilitates transcript models and cells curation, providing in-depth visualization of read and isoform metrics at single cell level.

SQANTI-sc has been tested using publicly available single cell, long read RNA-seq datasets from Oxford Nanopore Technologies and Pacific Biosciences to ensure its effectiveness regardless of the platform used. SQANTI-sc offers an efficient and robust solution for quality control, curation, and comparison of single cell, long read data; improving reliability and interpretability from a novel perspective.

Keywords: long-reads, single cell, quality control

Anna-Lena Katzke (Deparment of Human Genetics, Hannover Medical School), Marvin Döbel (Institute of Medical Genetics and Applied Genomics, University Hospital Tübingen), Jan Hauke (Center for Familial Breast and Ovarian Cancer, Center for Integrated Oncology (CIO), University Hospital Cologne), Benedikt Schnur (Deparment of Human Genetics, Hannover Medical School), Gunnar Schmidt (Deparment of Human Genetics, Hannover Medical School) and Marc Sturm (Institute of Medical Genetics and Applied Genomics, University Hospital Tübingen). *HerediClassify: Automated variant classification in hereditary breast and ovarian cancer.*

Abstract. Background:
Since the introduction of next-generation sequencing, variant classification has become the bottleneck in genetic diagnostics. Progressively more detailed variant classification guidelines and the need for regular variant reinterpretation based on new data exacerbate the bottleneck. Many of the rules used in variant interpretation can be automated to support variant classification efforts.
Methods:
HerediClassify is a Python-based tool that automates the variant classification guidelines by the American College of Medical Genetics (ACMG). 19/28 ACMG criteria are automated, though data availability limits application of rules depending on functional and segregation data. Additionally, gene-specific guidelines for ATM, BRCA1, BRCA2, CDH1, PALB2, PTEN, and TP53 are implemented.
Results:
HerediClassify is designed for modularity and extensibility by drawing inspiration from functional programming and package management. For instance, users can choose between different implementations of one criterion through a configuration file.
On a validation dataset consisting of 721 variants from ClinGen, HerediClassify achieves an average F1-Score of 93% across all criteria. HerediClassify outperforms other state-of-the-art classification tools on a hereditary breast and ovarian cancer dataset of 130 variants with a F1-Score of 68% for final variant classification, including vaRHC (59%) and VarSome (63%).
Conclusion:
HerediClassify is a powerful tool that can support variant classification. It is currently used by experts of the German Consortium for Hereditary Breast and Ovarian Cancer. The main application of HerediClassify is the preselection of variant classification criteria and the prioritization of variants for classification and reclassification.

Keywords: Variant classification, Automated variant classification, ACMG, Human genetics, Genetic diagnostics, Hereditary Breast and Ovarian Cancer

Daniel León-Periñán (Max Delbrück Center), Nikos Karaiskos (Max Delbrück Center) and Nikolaus Rajewsky (Max Delbrück Center). *Malva: Real-time Sequence Search across Billions of Cells.*

Abstract. Modern single-cell and spatial sequencing technologies create increasingly large repositories of nucleotide data from organisms across planet Earth. Current analysis methods require extensive preprocessing, limiting real-time exploration of the sequence-space by researchers and AI agents. Here, we present Malva, a universal search index that unifies multi-terabyte repositories into an instantly queryable, dynamic resource — hundreds of millions of cells across thousands of experiments. Malva enables immediate analysis of any nucleotide pattern with single-cell and spatial resolution. enables immediate, single-cell and spatial analyses of any nucleotide pattern, for instance identifying circular RNAs, tracking 3′ UTR dynamics during embryonic development, and detecting alternative splicing events across cell types. Importantly, Malva can discover cell types directly from unaligned reads without reference genomes, allowing cross-species analysis independent of orthology models. The indexed sequences also eliminate preprocessing barriers for AI, directly feeding neural networks that predict spatial expression patterns from primary sequence. By enabling ultrarapid sequence-to-phenotype mapping, Malva opens new avenues for biological discovery in the era of massive sequencing.

Keywords: single-cell, spatial transcriptomics, reference-free, sequence search, databases, language models

Hélène Collinot (AP-HP), Maryline Favier (Inserm), Rachel Onifarasoaniaina (Inserm), Alicia Gouge (Inserm), Djihane Djeridane (Inserm), Isabelle Lagoutte (Inserm), Sébastien Jacques (Inserm), Daniel Vaiman (Inserm), Céline Méhats (Inserm) and Christophe Le Priol (Inserm). *Improving the quality of spatial transcriptomics data clustering through recursivity.*

Abstract. A classical clustering workflow of spatial transcriptomics data consists of several ordered steps, i.e. normalization of the raw sequencing counts, selection of variable genes, dimensionality reduction of the dataset, integration of data from different samples, before the clustering process itself. Each of these steps is highly parametrizable, and some parameters can drastically affect the final clustering result. Besides, clustering algorithms are affected by randomness, which prevents reproducibility. However, the choice of parameters and the lack of reproducibility are rarely addressed in biological studies.

We have developed a recursive clustering workflow that generates clusterings using multiple parameter settings and identifies the best one in a data-driven manner. We applied it recursively in multiple independent rounds of sub-clustering to two Visium datasets in reproductive biology studies and to published Visium datasets. We compared the obtained recursive clusterings with those obtained with the classical approach, which consists in analyzing a whole dataset at once. By analyzing mouse brain datasets, we obtained more refined spatially defined clusterings. In a term parturition study, we were able to identify clusters mainly driven by specific cell types that are crucial to address to the biological issue of the study.

Thus, our recursive approach significantly improves the quality of spatial transcriptomics data clustering. Furthermore, by combining the selection of the best clustering according to a quality metric and the recursive application of the whole clustering process, our approach may provide an answer to the lack of reproducibility of clustering results in biological studies.

Keywords: clustering, recursivity, reproducibility, spatial, transcriptomics

Nadja Nolte (National Insitute of Biology, Jožef Stefan International Postgraduate School), Kristina Gruden (National Insitute of Biology), Lauren McIntyre (Department of Molecular Genetics and Microbiology, University of Florida Genetics Institute) and Marko Petek (National Insitute of Biology). *A pipeline for allele specific expression analysis in polyploid plants using long-read RNA-seq* .

Abstract. Long-read RNA sequencing enables identification of novel and complex isoforms and the separation of transcripts from highly similar loci, such as alleles. Measuring expression at the allele level allows investigation of allelic imbalance, which can indicate cis-regulatory differences between alleles. However, polyploid plant genomes produce highly similar transcripts that cannot be sufficiently assigned to their allele of origin using short-read RNA sequencing.

We developed an end-to-end pipeline for allele-specific expression analysis using long-read RNA-seq in polyploid plants. Using tetraploid potato as a model system, we demonstrate examples of genes with allelic imbalance and differential allele usage between tissues and conditions. By comparing short-read and long-read RNA-seq data from identical samples, we show the significant advantages of long-read RNA-seq for accurate allele identification and quantification in complex polyploid genomes.

Keywords: long-read transcriptomics, allele-specific expression, polyploid plants

Fabian Jetzinger (BioBam Bioinformatics S.L.), Stanley Cormack (Institute for Integrative Systems Biology (I2SysBio), Spanish National Research Council (CSIC)), Jorge Mestre-Tomás (Institute for Integrative Systems Biology (I2SysBio), Spanish National Research Council (CSIC)), José Manuel Morante-Redolat (University of Valencia), Isabel Fariñas (University of Valencia), Luis Ferrández (Institute for Integrative Systems Biology (I2SysBio), Spanish National Research Council (CSIC)), Stefan Götz (BioBam Bioinformatics S.L.), Carolina Monzó (Institute for Integrative Systems Biology (I2SysBio), Spanish National Research Council (CSIC)), Alejandro Paniagua (Institute for Integrative Systems Biology (I2SysBio), Spanish National Research Council (CSIC)) and Ana Conesa (Institute for Integrative Systems Biology (I2SysBio), Spanish National Research Council (CSIC)). *Join & Call vs. Call & Join – Identifying Transcripts Across Biological Replicates.*

Abstract. In the rapidly advancing field of long-read transcriptomics, the choice of how to combine data from multiple replicates can greatly influence transcript isoform discovery, yet has received little attention. Here, we compare two strategies—"Call & Join," in which transcripts are identified separately in each biological replicate then combined with TAMA Merge, and "Join & Call," in which reads are pooled before identifying isoforms—using a novel data set of PacBio and Oxford Nanopore data from five mouse brain and kidney samples spiked with Lexogen SIRVs. Six popular tools (IsoQuant, FLAIR, Bambu, TALON, Mandalorion, IsoSeq) were evaluated via SQANTI3 against RefSeq annotations and by SIRV recovery metrics. We find that permissive tools like FLAIR uncover more novel, rare isoforms under the "Join & Call" strategy, whereas conservative tools like IsoQuant yield greater novelty under the "Call & Join" strategy, likely due to sample-specific calls and shifting read-to-isoform assignments. Furthermore, while a substantial number of shared transcripts are consistently found across all individual samples and are recovered by both strategies, we also observe a large number of sample-specific isoforms present in only one sample. This suggests that even a small number of biological replicates may be sufficient to separate commonly shared isoforms from sample-specific ones. Our results further indicate that optimal performance in isoform identification may require tailoring both strategy and tool selection to the specific aims of a study — whether prioritizing known transcript recovery or the discovery of novel isoforms.

Keywords: Long-read sequencing, Transcriptome reconstruction, transcript identification, Join & Call, Call & Join, SIRVs, IsoQuant, FLAIR, Bambu, TALON, Mandalorion, IsoSeq, TAMA Merge, SQANTI3, Oxford Nanopore, PacBio, biological replicates

Ilya Levantis (Laverock Therapeutics), Kevin Gillinder (Laverock Therapeutics), Mariano Olivera Fedi (Laverock Therapeutics) and Lauren Overend (Laverock Therapeutics). *Rapid turn-around CRISPR outcome analysis using ONT Minion Sequencing.*

Abstract. We show that characterisation and validation of CRISPR experiment outcomes can be achieved using ONT Minion sequencing enabling cheaper and more rapid genome-editing workflows.

Generating clonal cell lines using CRISPR can be a time-consuming task, requiring multiple stages of sequencing and analysis for QC and characterisation before a final clonal cell line is produced. Existing software for quantifying allele composition in gene-editing experiments, and characterising genotypes of clonal lines is limited to using highly accurate short-read sequencing as input. The resulting need to use Illumina short read sequencing during this process can significantly extend the time taken to produce clonal cell lines with validated edits due to the long turn-around times this sequencing technology entails.

By tailoring our bioinformatics workflow to use ONT long read sequencing, we were able to leverage the agility of ONT Minion-based targeted amplicon sequencing (easy to carry out in-house sequencing to lower costs and reduce turn-around time). This allowed us to carry out sequencing and analysis of ~2000 CRISPR experiment outcomes within 24 hours with comparable accuracy to standard short read-based workflows.

Additionally, the use of long read sequencing enables the accurate characterisation of experiments involving large knock-ins or generation of structural variants, making the analysis and validation of more complex gene-editing experiments easier to conduct.

Keywords: CRISPR, genome editing, gene editing, oxford nanopore, minion, long read, genotyping, allele quantification

Hsueh-Ting Chu (Department of Computer Science and Information Engineering, Asia University). *Trio Genome Phasing from High-throughput Sequencing with Large Language Models.*

Abstract. Accurate phasing of diploid genomes into maternal and paternal haplotypes is crucial for understanding genetic inheritance, disease susceptibility, and population genetics. This study explores the potential of Large Language Models to enhance trio-based genome phasing. The goal of this research is to develop the scheme of word-embedding for sequencing reads and then train large language models with parental genome sequencing data for phasing offspring genome.

The core objective of this work is to develop and evaluate an LLM-based framework trained on parental genome sequencing data to accurately identify offspring genomic reads to the paternal and maternal chromosomes. We explored different embedding strategies and LLM architectures to optimize performance and assess the model's ability to handle high-throughput sequencing data. This work opens new possibilities for integrating advanced natural language processing techniques into the field of genomics.

Keywords: Genome Phasing, Haplotype Analysis, Large Language Models, Sequencing Reads, Word Embeddings

Lennart Bartels (Biomolecular Data Science in Pneumology, Research Center Borstel, Leibniz Lung Center, Germany), Sebastien Boutin (Department of Infectious Diseases and Microbiology, University Hospital Schleswig-Holstein Campus Lübeck), Christian Utpatel (Molecular and Experimental Mycobacteriology, Research Center Borstel, Leibniz Lung Center, Germany), Stefan Niemann (Molecular and Experimental Mycobacteriology, Research Center Borstel, Leibniz Lung Center, Germany), Dennis Nurjadi (Department of Infectious Diseases and Microbiology, University Hospital Schleswig-Holstein Campus Lübeck) and Inken Wohlers (Biomolecular Data Science in Pneumology, Research Center Borstel, Leibniz Lung Center, Germany). *High-quality Klebsiella pneumoniae genomes from consensus of various, deeply characterized PacBio, ONT and Illumina-based assemblies.*

Abstract. Klebsiella pneumoniae is a bacterial pathogen that can cause pneumonia, urinary tract and bloodstream infection as well as meningitis, typically in immune-compromised individuals. It is
the leading cause of infections acquired in health-care settings. Rising resistances to antimicrobials and hypervirulent strains increase health concerns. Klebsiella pneumoniae resistant to last-line carbapenem antibiotics leads the bacterial priority pathogens list of the World Health Organization (WHO) in 2024[1]. High-quality genomes are essential to identify known and new resistance and virulence factors, to provide reference sequences for mapping-based approaches and as knowledge base for research on pathogen biology, resistance mechanisms and novel antibacterial medicines. Due to horizontal gene transfer, bacterial genomes of the same species can differ considerably with respect to sequence and protein content. Further, circular chromosomes are complemented by plasmid sequences that pose specific challenges for assembly, but are clinically important, since they represent a key mechanism for acquiring resistance factors. According to the European Nucleotide Archive, more than thousands of Klebsiella pneumoniae isolates have been sequenced with all major technologies, typically in the larger context of clinical surveillance.

We generated high-quality assemblies of the genomes of 105 carbapenem-resistant Klebsiella pneumoniae clinical isolates collected in a hospital in Vietnam in 2021[2]. These assemblies are based on published Illumina and Oxford Nanopore Technologies (ONT) data (R.9.4.1 flow cells; guppy version 6.3.9 basecaller) complemented with Pacific Biosciences (PacBio) HiFi sequencing data (Sequel II) for a subset of 56 isolates of the same cohort. Quality control of sequencing data was performed using fastp[3] and NanoPlot[4], showing diverse characteristics in terms of sequencing depth (Illumina 14-249x, median 61x; ONT 49-388x, median 158x; PacBio 27-95x, median 60x), read N50 (ONT 2,653-11,214x, median 5,334x; PacBio 7,847-15,937x, median 11,933x) and base accuracy (ONT Q15 reads 15-39%, PacBio Q30 reads 77-83%). The 300 bp paired-end Illumina reads were quality trimmed and ONT reads pre-processed with Porechop_abi[5] and filtlong[6]. Using ONT data, an end-to-end microbial hybrid assembly workflow, hybracter[7], was applied, which includes dedicated plasmid assembly and annotation using plassembler[8] and internally the assembler flye[9]. For 81 of 105 isolates, hybracter assemblies were complete (77%). Circular plasmid numbers varied between 4 and 11 across these assemblies (median 5). Assembly errors with respect to HiFi data were assessed for 46 of the complete assemblies (one outlier removed) using Inspector[10] and show few base errors (0-141, median 9), only three assemblies with a structural error, from 1 to 142 small-scale assembly errors (median 15) and a HiFi-based QV score that ranges from 44 to 67 (median 56). As an alternative approach, short-read first assembly was computed using Unicycler[12]. Manual inspection of assembly graphs of unfinished assemblies revealed the presence of typically one long repeat sequence that hinders the unambiguous resolution of the bacterial chromosome. We used Autocycler[13] to create consensus assemblies from
different read subsets assembled with a range of different assemblers to account for assembler specific assembly errors. Autocycler clusters contigs from different assemblies of the same read sets based on sequence similarity and tries to resolve these clusters into one consensus contig. Therefore ONT read sets were subsampled to four independent read subsets and were assembled with six assemblers (flye[9], raven[14], canu[15], miniasm[16], necat[17] and nextdenovo[18]). If available, four HiFi assemblies were added (flye[9], raven[14], hicanu[19], hifiasm[20]). Overall, 24 (28 including HiFi) assemblies were used as input for autocycler. In 57 isolates, both hybracter and autocycler report completely resolved assemblies. In 37 (64%) of these cases the number of replicons obtained using both methods matched exactly. Autocycler further reports metrics that indicate how well the clustering of contigs from the input assemblies worked. Thereby, the cluster balance score describes how balanced the clusters are in terms of the number of clustered contigs from the input assemblies and ranged between 0.175 and 0.846 (median 0.510). The cluster tightness score varied between 0.655 and 0.908 (median 0.811) and quantifies the sequence similarity within clusters. Theses scores reflect the support of the consensus assembly by the input assemblies. Overall, for 66 isolates the consensus assembly could fully resolve the contig clusters from the input assemblies.

For a cohort of carbapenem-resistant Klebsiella pneumoniae, a WHO top-priority bacterial pathogen, we have performed assembly, detected potential errors and generated consensus assemblies. Towards this, three different types of sequencing data, each covering a range of read characteristics, have been used together with nine different assembly tools and an end-to-end microbial assembly workflow. This has been followed by the application of a recently released tool that generates a consensus from a large number of bacterial assemblies. As a result, 66 assemblies of exceptional high quality with respect to the chromosome as well as plasmids were generated and their

support based on concordance of input assemblies quantified. These assemblies serve as Klebsiella pneumoniae reference genomes, e.g. for variant identification and accurate and complete protein sequence annotation.

References

1. WHO bacterial priority pathogens list, 2024: Bacterial pathogens of public health importance to guide research, development and strategies to prevent and control antimicrobial resistance. https://www.who.int/publications/i/item/9789240093461.
2. Sy, B. T. et al. Heterogeneity of colistin resistance mechanism in clonal populations of carbapenem-resistant Klebsiella pneumoniae in Vietnam. Lancet Reg. Health - West. Pac. 51, 101204 (2024).
3. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics 34, i884–i890 (2018).
4. De Coster, W. & Rademakers, R. NanoPack2: population-scale evaluation of long-read sequencing data. Bioinformatics 39, btad311 (2023).
5. Bonenfant, Q., Noé, L. & Touzet, H. Porechop_ABI: discovering unknown adapters in ONT sequencing reads for downstream trimming. 2022.07.07.499093 Preprint at https://doi.org/10.1101/2022.07.07.499093 (2022).
6. Wick, R. rrwick/Filtlong. (2025).
7. Bouras, G. et al. Hybracter: Enabling Scalable, Automated, Complete and Accurate Bacterial Genome Assemblies. 2023.12.12.571215 Preprint at https://doi.org/10.1101/2023.12.12.571215 (2024).
8. Bouras, G., Sheppard, A. E., Mallawaarachchi, V. & Vreugde, S. Plassembler: an automated bacterial plasmid assembly tool. Bioinformatics 39, btad409 (2023).
9. Kolmogorov, M., Yuan, J., Lin, Y. & Pevzner, P. A. Assembly of long, error-prone reads using repeat graphs. Nat. Biotechnol. 37, 540–546 (2019).
10. Chen, Y., Zhang, Y., Wang, A. Y., Gao, M. & Chong, Z. Accurate long-read de novo assembly evaluation with Inspector. Genome Biol. 22, 312 (2021).
11. Gurevich, A., Saveliev, V., Vyahhi, N. & Tesler, G. QUAST: quality assessment tool for genome assemblies. Bioinformatics 29, 1072–1075 (2013).
12. Wick, R. R., Judd, L. M., Gorrie, C. L. & Holt, K. E. Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. PLOS Comput. Biol. 13, e1005595 (2017).
13. Wick, R. rrwick/Autocycler. (2025).
14. Vaser, R. & Šikić, M. Time- and memory-efficient genome assembly with Raven. Nat. Comput. Sci. 1, 332–336 (2021).
15. Koren, S. et al. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res. 27, 722–736 (2017).
16. Li, H. Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. Bioinformatics 32, 2103–2110 (2016).
17. Chen, Y. et al. Efficient assembly of nanopore reads via highly accurate and intact error correction. Nat. Commun. 12, 60 (2021).
18. Hu, J. et al. NextDenovo: an efficient error correction and accurate assembly tool for noisy long reads. Genome Biol. 25, 1–19 (2024).
19. Nurk, S. et al. HiCanu: accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads. Genome Res. 30, 1291–1305 (2020).
20. Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. Nat. Methods 18, 170–175 (2021).

Keywords: Genome Assembly, De Novo Assembly, Long Read Sequencing, Bacterial Pathogens, Whole Genome Sequencing, Klebsiella pneumoniae

Dominika Galová (Institute of Molecular Biology of the Slovak Academy of Sciences), Nikola Klištincová (Institute of Molecular Biology of the Slovak Academy of Sciences), Francesca Maisto (Institute of Molecular Biology of the Slovak Academy of Sciences), Andrea Puškárová (Institute of Molecular Biology of the Slovak Academy of Sciences), Mária Bučková (Institute of Molecular Biology of the Slovak Academy of Sciences), Jelena Pavlović (Institute of Molecular Biology of the Slovak Academy of Sciences) and Domenico Pangallo (Institute of Molecular Biology of the Slovak Academy of Sciences). *Metagenomic Exploration of Microbial Biosorption Potential for Rare Earth Element Recovery from Slovak Urban Waste.*

Abstract. Rare earth elements (REEs) are critical components in the production of modern electronic devices and play an essential role in the development of low-carbon and clean energy technologies. Given the increasing demand and environmental costs of traditional mining, urban mining emerges as a sustainable solution by recovering REEs from urban waste. One innovative branch of urban mining is biomining, which utilizes microorganisms to extract valuable metals in an eco-friendly manner. Biomining includes two main processes: bioleaching and biosorption. This study aims to explore the biosorption potential of microbial communities present in environmental samples (soil and wastewater) using a metagenomic approach. Whole metagenome sequencing was performed using the MinION platform from Oxford Nanopore Technologies (ONT), enabling comprehensive analysis of environmental DNA (eDNA). The main objectives included identifying microbial taxa involved in REE interactions, detecting genes related to metal resistance, with a focus on metallothioneins. To support gene identification and annotation, curated databases was constructed by integrating data from the BacMet database, which provides a wide array of sequences related to metal metabolism and resistance. This integrative approach enhanced our understanding of microbial roles in REE recovery and support the development of bio-based recovery strategies. The findings may contribute to more sustainable practices in metal recovery and circular economy efforts.

Keywords: Rare earth elements, Microbial communities, Metagenomics, Oxford Nanopore Technologies, MinION sequencing, Metal resistance genes

Nikola Klištincová (Institute of Molecular Biology, Slovak Academy of Sciences), Dominika Galova (Institute of Molecular Biology of the Slovak Academy of Sciences), Francesca Maisto (Institute of Molecular Biology, Slovak Academy of Sciences), Mária Bučková (Institute of Molecular Biology, Slovak Academy of Sciences), Andrea Puškárová (Institute of Molecular Biology, Slovak Academy of Sciences), Lucia Kraková (Institute of Molecular Biology, Slovak Academy of Sciences) and Domenico Pangallo (Institute of Molecular Biology, Slovak Academy of Sciences). *Microbial Communities Involved in Gas Oil Tank Corrosion Revealed by Oxford Nanopore Sequencing.*

Abstract. We investigated corrosion events in two 10,000 m³ gas oil tanks using Oxford Nanopore Technology (ONT)-based high-throughput sequencing. The microbial communities potentially involved in corrosion, and genes related to metal resistance and biocorrosion were characterized.

Both tanks were constructed from carbon steel (STN 11 378). Tank 1 (T1) rested on a concrete base, while Tank 2 (T2) included a sand layer between the steel sheet and concrete. We analysed two samples from T1: T1_S1 (corroded concrete) and T1_S2 (clean concrete, control). From T2, five samples were studied: T2_CS (clean sand, control), T2_S25_1 (sand in contact with corroded steel), T2_S25_2 (corroded steel sheet), T2_S46_1 (sand in contact with corroded steel), and T2_S46_2 (corroded steel sheet).

DNA was extracted using the Qiagen DNeasy PowerSoil Pro Kit. Sequencing was performed on the MinION platform with the Ligation Sequencing Kit (SQK-LSK114) and PCR barcoding. Data processing included basecalling and adapter trimming using Dorado (ONT), read assembly with Flye, and taxonomic classification via Kraken2 and the PFPlus database. Genes associated with metal resistance were identified using the BacMet database. Visualizations were produced in RStudio.

Taxonomic analysis revealed significant diversity across samples. Corroded sites harboured members of the phyla Pseudomonadota, Bacillota, and Actinomycetota. Using BacMet, we also identified specific genes linked to metal resistance.

This study highlights the utility of Oxford Nanopore sequencing for investigating microbial-driven corrosion and supports its application in refinery-related biocorrosion research.

Acknowledgements: This study was funded by project SAS-NSTC-JRP-2024-06_EMERGE.

Keywords: Biocorrosion, Oxford Nanopore Sequencing, Metal Resistance Genes, Microbial Communities, Taxonomic Classification, Refinery Corrosion Research

Vinzenz May (Core unit bioinformatics, Berlin Institute of Health @ Charité Berlin), Manuel Holtgrewe (Core unit bioinformatics, Berlin Institute of Health @ Charité Berlin) and Dieter Beule (Core unit bioinformatics, Berlin Institute of Health @ Charité Berlin). *SvirlPool: Structural Variant Calling with Local Assemblies fom Nanopore Long Reads.*

Abstract. Structural variant (SV) detection is integral to rare disease diagnostics and cancer genomics, as well as in studying genome variation in populations, where a great challenge is the identification of shared variants among multiple samples. Existing methods of SV-detection with long reads usually use read-to-reference alignments to detect SVs on abstractions of SV signals or rely on computationally extremely expensive de novo assembly of the whole genome if feasible.

We propose a novel method to combine alignment-based and assembly methods. We identify candidate regions in read-to-reference alignments and assemble clusters of parts of aligned reads. The resulting polished virtual fragments are then aligned to a reference sequence of choice and used for SV-calling. The comparison of those virtual fragments of sequencing samples allows us to very exactly match SVs that are shared between several samples, since each virtual fragment contains sequence information about its original reads.

We have developed the first local assembly-based method to detect SVs specifically with Nanopore long reads. In the most popular benchmark on deletions and insertions we achieved a f1 score of more than 95%. We tested the Mendelian consistency in a family of six trios in ten members and found a maximum discordance of only 6%, which is much better than any other method we tested.

We are currently working on the detection of further types of structural variants and general fine-tuning. In the future, we will test the feasibility of this approach to cancer research cases.

Keywords: structural variants, long reads, nanopore, rare diseases, genomics

Anna Diamant (Université Côte d'Azur, Institut de Pharmacologie Moléculaire et Cellulaire), Eamon Mcandrew (Université Côte d'Azur, Institut de Pharmacologie Moléculaire et Cellulaire), Pascal Barbry (Université Côte d'Azur, Institut de Pharmacologie Moléculaire et Cellulaire), Georges Vassaux (Université Côte d'Azur, Institut de Pharmacologie Moléculaire et Cellulaire) and Kevin Lebrigand (Université Côte d'Azur, Institut de Pharmacologie Moléculaire et Cellulaire). *Allos: python package to explore single-cell and spatial isoform-level transcriptomics.*

Abstract. Alternative splicing is a fundamental process in gene expression that enables a single gene to produce multiple mRNA transcripts giving rise to proteins with distinct structures and functions. This mechanism is a major contributor to the diversity of proteins in human cells, with more than 90% of human genes undergoing alternative splicing. When dysregulated, alternative splicing can lead to the production of abnormal proteins potentially associated with disease. However, most single-cell RNA sequencing methods focus on gene-level quantification missing the diversity of gene isoforms.

Allos is a Python package, based upon scverse ecosystem, developed to streamline the statistical analysis and exploration of single-cell long-read transcriptomics data. It allows researchers to easily handle and explore isoform level gene expression, providing deeper insights into cellular heterogeneity and underlying biological processes. Allos is agnostic to sequencing platforms and quantification tools.

The package introduces the AnnDataIso class that inherits from the AnnData class, extending its functionality to support isoform-level data exploration such as differential isoform usage and expression. It integrates as a part of the scverse (Foundational tools for single-cell omics data analysis) ecosystem.

Allos includes intuitive visualization tools for known and novel transcripts, automated report generation for differential isoform usage, and clear, interpretable plots to highlight transcript-level differences across cell populations or experimental conditions.

With its user-friendly design and integration with existing tools, Allos helps fill the current gap in single-cell and spatial transcriptomics by enabling broader use of isoform-level analysis to the academic community.

Keywords: single-cell transcriptomics, spatial transcriptomics, long-read RNA seq, alternative splicing, isoform profiling, python package

Desmond Singh (University of Waterloo), Yifan Yang (University of Waterloo) and Brendan McConkey (University of Waterloo). *A Weighted Resampling Approach to Differential Expression Analysis.*

Abstract. Differential expression analysis (DEA) allows researchers to relate changes in gene expression to specific phenotypes. DEA relies on RNA-seq datasets that measure expression in thousands of genes, often across a few samples per condition. These high-dimensional data are then typically analyzed with parametric statistical models employing the negative binomial distribution for inference. Variance stabilization is typically required for such datasets, often based on the mean-variance trend across all genes. Some potential data-dependent issues with analyzing RNA-seq data using an assumed distribution include poor model fit and/or insufficient variance stabilization, potentially compromising nominal false discovery rate control. As an alternative, we propose a weighted resampling approach that combines variance stabilization using priors with a resampling procedure to estimate the statistical significance of differential expression. The presented method is distribution agnostic, effective for any sample size, and develops gene-by-condition-specific priors to estimate stabilized variances. Priors are built as cumulative distribution functions (CDFs) by sampling normalized expression data from a pool of reference genes with expression levels similar to those of a target gene. The CDF for a target gene-by-condition is calculated as a combination of the prior CDF and observed experimental data, weighted by the number of experimental data points. Output CDFs are then repeatedly sampled to generate a distribution of Log2FC estimates to compare differential expression between experimental conditions. We demonstrate the validity of this approach using simulation experiments and its efficacy by analyzing human and plant RNA-seq datasets.

Keywords: Differential Expression Analysis, RNA-Seq, High-Throughput Sequencing, Transcriptomics, Resampling, Computational, R Programming, Cumulative Distribution Functions, Simulations, Statistical

Peiying Cai (Department of Molecular Life Sciences and Swiss Institute of Bioinformatics, University of Zurich), Mark D Robinson (Department of Molecular Life Sciences and Swiss Institute of Bioinformatics, University of Zurich) and Simone Tiberi (Department of Statistical Sciences, University of Bologna). *Detection of differential spatial patterns in spatial omics data.*

Abstract. Spatially resolved transcriptomics (SRT) technologies measure gene expression while preserving spatial context. Several methods have been developed to identify spatially variable genes (SVGs), i.e., genes whose expression profiles vary across tissue. However, at present, no method allows comparing spatial patterns between experimental conditions with multiple samples.

Here we present an approach for identifying genes with differential spatial patterns (DSP) across conditions in multi-sample SRT datasets. The method is based on a negative binomial model of gene expression data and uses a likelihood ratio test, to identify genes that exhibit changes in spatial expression structure. We designed simulation studies, where our framework exhibits strong performance, in terms of sensitivity, specificity and runtime.

Key strengths of our framework include: (i) modeling multiple samples across multiple conditions, reducing uncertainty from individual samples and identifying genes with DSP across experimental conditions; (ii) region-specific testing, allowing investigation of the mRNA abundance changes between conditions in areas of particular interest; and (iii) computational efficiency and compatibility with diverse SRT platforms.

The method is freely available as part of the DESpace Bioconductor R package.

Keywords: Spatial omics, Bioinformatics, Statistical software

Furkan Eris (ETH Zurich), Ulysse McConnell (ETH Zurich), Can Firtina (ETH Zurich) and Onur Mutlu (ETH Zurich). *RawBench: A comprehensive benchmarking framework for raw nanopore signal analysis.*

Abstract. Nanopore sequencing technologies continue to advance rapidly, but conventional basecalling pipelines struggle to keep pace with the computational demands of processing increasingly complex signals and pore chemistries. Raw nanopore signal analysis has emerged as a promising alternative to these resource-intensive approaches. While attempts have been made to benchmark these methods, existing evaluation frameworks 1) fail to include the latest improvements in nanopore datasets, 2) overlook techniques that bypass basecalling entirely, and 3) lack the flexibility to accommodate new tools in the fast-evolving field of raw nanopore signal analysis.

Our goal is to provide an extensible benchmarking suite addressing the above issues that enables designing and comparing new methods for raw signal analysis using up-to-date and representative datasets. To this end, we introduce RawBench, a flexible framework for evaluating raw nanopore signal analysis across organisms with varying genome complexities. We observe that all raw signal pipelines share a common structure: 1) encoding signals, 2) encoding reference genomes, and 3) comparing these representations. RawBench enables modular integration and comparison of different techniques used in the different stages of raw signal analysis pipelines. By doing so, RawBench addresses our goal of enabling researchers to easily benchmark new methods as well as design effective raw signal analysis pipelines for many applications.

Decision: (conflict)
Keywords: raw nanopore signal analysis, nanopore sequencing, multimodal representation, benchmarking suite, basecalling

Samuel Coleman IV (University of Utah, Biomedical Informatics), Luca Pinello (Massachusetts General Hospital, Molecular Pathology Unit) and Kendell Clement (University of Utah, Biomedical Informatics). *Analysis of Single-Cell DNA Readouts from CRISPR Screens with CRISPR SCope.*

Abstract. CRISPR perturbation screens are an effective tool for a high-throughput functional assessment of the genome. Recent technologies allow readout of CRISPR perturbations at single-cell resolution using targeted amplicon sequencing. However, analyzing these data presents unique challenges due to cell-specific CRISPR-induced mutations that complicate traditional analysis methods which rely on shared mutations across multiple cells.

We introduce CRISPR SCope, a software for high-throughput analysis and visualization of single-cell DNA readouts from CRISPR screens. Our method precisely links genomic perturbations with phenotypes by reporting allele sequences in individual cells at CRISPR targets. CRISPR SCope offers three key innovations: (1) a novel amplicon quality score statistic that identifies high-quality cells with sufficient read depth across targeted sites for downstream analysis; (2) biologically informed read alignment utilizing CRISPResso2 to increase accuracy by incorporating CRISPR editing outcome characteristics to reduce sequencing artifacts; and (3) multi-level analysis granularity, from cell-specific editing outcomes to experiment-wide summaries, with direct DNA sequence readouts for assigning alleles and genotypes at the single-cell level.

Results are summarized and presented in an automatically generated HTML report with an accompanying HDF5 file for further analysis. Overall, CRISPR SCope represents a significant advancement in single-cell CRISPR screen analysis by offering direct readouts of DNA perturbations and empowering the exploration of cellular perturbations with greater resolution. CRISPR SCope is available at https://github.com/clementlab/CRISPR-SCope

Keywords: CRISPR, CRISPR Screen, DNA, Genomics, Single-cell sequencing

Sara Baldinelli (Genome Biology Unit, European Molecular Biology Laboratory (EMBL), Heidelberg, Germany.), Vaishali Grewal (Division Signaling and Functional Genomics, German Cancer Research Center (DKFZ), Heidelberg, Germany.), Michael Dorrity (Molecular Systems Biology Unit, European Molecular Biology Laboratory (EMBL), Heidelberg, Germany.), Stefan Peidli (Genome Biology Unit, European Molecular Biology Laboratory (EMBL), Heidelberg, Germany.), Michael Boutros (Division Signaling and Functional Genomics, German Cancer Research Center (DKFZ), Heidelberg, Germany.) and Wolfgang Huber (Genome Biology Unit, European Molecular Biology Laboratory (EMBL), Heidelberg, Germany.). *Combinatorial Barcoding Meets In Vivo CRISPR: Decoding Context-Dependent Networks via sci-RNA-seq.*

Abstract. Through the use of a heat-inducible CRISPR/Cas9 system, we can perturb genes otherwise essential during development in an adult context. Scaling up this experimental approach is the central tool of the DECODE project, founded with the aim of systematically mapping context-dependent genetic networks across dynamic tissues in vivo. Drosophila melanogaster serves as our model system, in which we specifically perturb only cells expressing intestinal stem cell markers. By integrating single-cell transcriptomics and high-resolution imaging of thousands of conditional knockouts, DECODE aims to resolve how genetic networks dynamically adapt their topology across cell types and external stimuli.

To achieve organism-scale profiling, we utilize single-cell combinatorial indexing RNA sequencing (sci-RNA-seq), a split-pool barcoding strategy that exponentially scales throughput while reducing costs to ~1 cent per cell. With the transition from whole-cell to single-nucleus profiling, this method is able to avoid challenges in cell dissociation and minimize tissue-specific biases. Even though nuclear RNA lacks cytoplasmic transcripts, nuclei isolation enables universal application across tissues, including RNase-rich adult samples, and reduces dissociation-induced stress artifacts. snRNA-seq retains sufficient biological signal to characterize population heterogeneity and often achieves a higher signal-to-noise ratio.

A careful optimization of sci-RNA-seq is required to mitigate artifacts like index hopping and ensure high-quality data. QC metrics are lower compared to droplet-based methods but are sufficient for robust population characterization when improved. Understanding the challenges and limitations of sci-RNA-seq remains an active area of investigation and we aim to further evaluate its performance and benchmark it against gold-standard methods.

Keywords: scRNA-seq, In-vivo CRISPR-Cas9, Perturb-seq, Combinatorial barcoding, sci-RNA-seq

Félix-Antoine Trifiro (Université de Sherbrooke) and Marie Brunet (Université de Sherbrooke). *Demystifying structural variation in neuroblastoma using data-driven personalized annotations.*

Abstract. Cancer is the first cause of death among Canadian children. Pediatric cancers differ from adults' by their genomic signatures, enriched in structural alterations and copy number variations. Neuroblastoma (NB) is the most common extracranial tumor in children, and the survival rate of high-risk NB remains at 50%. Despite the likelihood of genomic structural variation, sequencing analyses are still aligned to a reference genome. Here, we suggest leveraging commonly discarded reads (DRP) due to their disagreement with the reference architecture to build personalized genomic profiles.

We developed an analytical pipeline to build personalized genomic architectures and highlight structural alterations. We applied it on a cohort of 292 children with various cancer types. Whole exome sequencing from paired healthy and tumor tissues is aligned to the reference genome using the standard analytical pipeline (BWA). 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 specific Smith-Waterman algorithm. Confident DRPs are then used to build a personalized genomic architecture supported by the sequencing data. Here, we focused on NB. Three patients (18%) presented a significant increase of DRPs between normal and tumoral tissues. One patient presented genomic remodelling of the ASIC2 locus, whose disruption has been linked with NB tumor type and stage. Another patient presented a novel retrocopy of CUEDC2 whose overexpression has been correlated NB tumor development. Our approach highlighted overlooked structural and copy number variations relevant to the patients' diagnosis.

Keywords: Genomic, Structural alteration, Personnalized architecture, Pediatric cancers

Marjan Hosseini (University of Connecticut), Thomas Bergendahl (University Of Connecticut), Ella Veiner (University Of Connecticut), Zane Smith (University of Tennessee, Knoxville), Tala Yasenpoor (University Of Connecticut), Jill Wegrzyn (University Of Connecticut), Sorin Istrail (Brown University), Margaret Staton (University of Tennessee, Knoxville) and Derek Aguiar (University Of Connecticut). *pHapCompass: Probabilistic Assembly and Uncertainty Quantification of Polyploid Haplotype Phase.*

Abstract. Haplotype assembly is an important component of foundational molecular and population genetics problems, with uses including interpreting the effects of genetic variation on complex traits and reconstructing genealogical relationships. Assembling the haplotypes of polyploid genomes is challenging due to the exponential search space of haplotype phasings and read assignment ambiguity. We present pHapCompass, a probabilistic polyploid haplotype assembler that models the joint distribution over haplotype phasings with a Markov random field whose conditional independence structure is defined by observed haplotype segments. Haplotypes are assembled using algorithms that maintain uncertainty across multiple plausible haplotype phasings (forward filtering backward sampling) or select the haplotype assembly with highest likelihood (max-product). Ambiguities in polyploid phase introduce a sequential matching problem, which is resolved by a variant selection algorithm that generates candidate phasings by enforcing consistency with partial phasing obtained in earlier iterations. We benchmark pHapCompass through comparisons with existing haplotype assembly approaches using experimental and simulated data that reflect the realistic genomic complexity of polyploidy organisms, i.e., autopolyploidy and allopolyploidy. Further, we generalize the vector error rate and minimum error correction evaluation criteria for partially phased haplotypes. Results show that pHapCompass yields higher overall phasing accuracy, longer phased blocks, and robust performance across varying genomic complexities and polyploid structures, in addition to its interpretability and quantification of uncertainty.

Keywords: haplotype assembly, polyploid haplotype phase, probabilistic graphical models, Markov random field, forward filtering backward sampling, max-product

Richard A. Schäfer (Northwestern University Feinberg School of Medicine) and Rendong Yang (Northwestern University Feinberg School of Medicine). *Efficient Genome Indexing for Large-Scale Linked Interval Data.*

Abstract. Efficiently querying specific genomic regions is fundamental in bioinformatics, which allows for extracting relevant feature information from large genomic datasets. While existing tools provide query capabilities, they are limited to basic interval manipulation and do not natively support linked interval data or complex relationships between genomic features. We introduce genogrove, a hybrid graph data structure that facilitates scalable interval queries. Distributed as a command-line tool and a general-purpose library with bindings to multiple programming languages, genogrove integrates seamlessly into diverse bioinformatics workflows. It supports advanced queries involving linked genomic intervals, making it useful for applications where the relationships between intervals need to be considered. We demonstrate how it serves as an efficient interval search structure for large-scale datasets and lays the foundation for more advanced genomic analyses. We applied it to detecting transcript isoforms from long-read data and the genotyping of HLA alleles, demonstrating its usability in a wide range of applications in bioinformatics.

Keywords: interval, query, data structure, library, bioinformatics

David Requena (Johns Hopkins Bloomberg School of Public Health), Daniel F. Guevara-Diaz (Johns Hopkins Bloomberg School of Public Health), Daniel Garbozo-Santillan (Johns Hopkins Bloomberg School of Public Health) and Robert H. Gilman (Johns Hopkins Bloomberg School of Public Health). *Blood transcriptomic profiling identifies CHIT1 as a potential biomarker of early Chronic Chagas Cardiomyopathy.*

Abstract. Chagas disease is caused by Trypanosoma cruzi, being the parasitic infection with highest morbidity and mortality in the Americas. Up to 45% of infected individuals develop Chronic Chagas Cardiomyopathy (CCC), a deadly condition if not treated early. However, diagnosis is challenging as it affects heart tissue. To uncover potential indirect biomarkers, we screened whole-blood transcriptomes of 21 Bolivian patients using single-end sequencing and state-of-the-art bioinformatic methods. We compared early-CCC patients (B+) to patients with Chagas infection without cardiac issues (A+), and to Chagas-negative patients having cardiac symptoms (B-) or healthy controls (A-). Unsupervised clustering revealed sex as a confounding factor and was included as a covariate. Consistent with early disease stages, few differentially expressed genes were found in B+ vs A+. Among these, CHIT1 showed upregulation from A+ to B+ in our dataset, which was validated in two external cohorts. Using it, we developed a prognostic model using sex and CHIT1 expression, which showed a good performance (AUC: 0.89, ANOVA p-value: 0.02). Moreover, competitive gene-set tests (CAMERA and GSEA) revealed enrichment of macrophage- and cardiomyocyte-related signatures in B+; whereas signatures for megakaryocytes, platelets, NK cells, and neutrophils were reduced. This pilot study highlights key genes and signatures potentially involved in CCC progression and provides candidates that, upon future testing, may serve for CCC detection.

Keywords: RNA-seq, Early detection, Trypanosoma cruzi, Cardiomyopathy, Prognosis

Shir Liya Dadon (Hebrew University of Jerusalem) and Oren Ram (Hebrew University of Jerusalem). *Investigating Resistance Associated Mutations in Lung Cancer at Single-Cell Resolution.*

Abstract. Tumor cells accumulate mutations rapidly and become genetically different from one another, this process gives the cells an advantage of survival in an environment of competition over space and resources. Both chemotherapy and targeted cancer therapies eliminate most tumor cells but leave behind treatment-resistant clones, which are much more difficult to treat successfully. PC9 cells originated from non-small cell lung cancer adenocarcinoma, from the leading cause of cancer death worldwide. Using sc-rDSeq, a novel drop based full length single cell RNAseq technology, we profiled PC9 lung cancer cells under 3 treatment conditions - DMSO (control), Gefitinib, and Osimertinib resulting in drug-tolerant persister cells. We developed a computational pipeline to profile single nucleotide variations (SNVs) at the single-cell level. Since the treatment was short-lived, we found that most SNVs with resistance potential were already present in the control group. To identify oncogenic drivers, we focused on SNVs that were enriched compared to the control and uncovered genes such as MUC16 and TNKS1BP1. Interestingly, we also identified genes with depleted SNVs like MDH2 and IGFBP6, suggesting a disadvantageous effect of the SNV. Together with gene expression, we can pinpoint drug resistant essential genes, which can serve as candidates for targeted therapy. These findings provide insight into SNV heterogeneity in tumor cells, and our pipeline offers a valuable tool for detecting early signs of resistance in cancer samples.

Keywords: Single-cell, Full-length RNAseq, NSCLC, SNV detection, Resistance

Rosanna Smith (University of Southampton), Sarah Frampton (University of Southampton), Benjamin Stevens (University of Southampton), Jade Forster (University of Southampton), Jane Gibson (University of Southampton), Mark Cragg (University of Southampton) and Jonathan Strefford (University of Southampton). *Assembly of the high homology immune FCGR2/3 locus with Oxford Nanopore long-reads.*

Abstract. Monoclonal antibody therapy has transformed cancer treatment, but success is influenced by the engagement of Fc-gamma receptors on tumour immune cells. The human FCGR2/3 genetic locus encodes three low-affinity activating Fc-gamma receptors (FCGR2A, FCGR2C, FCGR3A) and a single inhibitory one (FCGR2B), where the ratio of activating to inhibiting receptors impacts antibody therapy resistance. Modulation of this ratio requires understanding of FCGR gene regulatory mechanisms, for which genetic characterization of the FCGR2/3 locus is a critical step. However, the locus contains a ~98% homologous, ~85kb segmental duplication which confounds short-read sequencing approaches and leads to loss and misattribution of experimental signal. The locus is also known to exhibit structural variation on the same scale as the segmental duplication, leading to further complexity.

We have deployed Oxford Nanopore long-read genomic DNA sequencing of primary cell material to assess this complex region and assemble haplotype-resolved FCGR2/3 loci across a cohort of 22 individuals. Both Cas9-targeting enrichment and adaptive sampling methods were employed to enrich locus-specific information, resulting in high, but heterogenous alignment coverage across the region. When combined with technology version limitations, high homology, and structural variation, this results in significant challenges for region-specific genome assembly.

We report the computational approaches taken, limitations, and cases of successful assembly for our cohort of 22 individuals. Together with assessment of structural variation within the Human Pangenome Reference Consortium cohort, we aim to determine genetic variability within potential gene regulatory regions and dissect regulatory mechanisms for a complex set of genes which impact cancer treatment outcome.

Keywords: Long-read, Nanopore, Immune, FCGR, Assembly, Segmental duplication, Homology

Ferdinand Popp (Division of Applied Bioinformatics, German Cancer Research Center (DKFZ), 69120, Heidelberg, Germany), Chen Hong (German Cancer Research Center (DKFZ), 69120, Heidelberg, Germany), Nicholas Abad (Division of Applied Bioinformatics, German Cancer Research Center (DKFZ), 69120, Heidelberg, Germany), Nicola Biondi (Division of Applied Bioinformatics, German Cancer Research Center (DKFZ), 69120, Heidelberg, Germany), Yoann Pageaud (Division of Applied Bioinformatics, German Cancer Research Center (DKFZ), 69120, Heidelberg, Germany), Benedikt Brors (Division of Applied Bioinformatics, German Cancer Research Center (DKFZ), 69120, Heidelberg, Germany) and Lars Feuerbach (Division of Applied Bioinformatics, German Cancer Research Center (DKFZ), 69120, Heidelberg, Germany). *TelomereHunter2: Enhanced Tool for in-silico Telomere Analysis.*

Abstract. TelomerHunter extracts telomeric reads from sequencing data to measure telomere content and composition. It is used to study the cancer hallmark replicative immortality, which is tightly linked to the activation of telomere maintenance mechanisms (TMM). In precision oncology programs, the software is applied to differentiate tumors with telomerase reactivation (TERT) from those which employ the alternative lengthening of telomeres (ALT) pathway. Increasing sequencing depth and novel data formats in medical sequencing programs, as well as, a raising interest from non-human fundamental research programs in sequencing based telomere analysis, are new challenges in the field.

Here, we present TelomereHunter2, a comprehensive update addressing these requests. Key improvements include:

- Support for compressed sequencing formats such as CRAM, which are used to attenuate the data deluge in biomedical research.
- Compatibility with non-human genomes (e.g. mouse, dog) for cross-species studies.
- Containerization (Docker/Singularity/Apptainer) to enhance the applicability in diverse IT environments, such as medical centers and hospitals.
- Optimized algorithms with improved runtime to reduce turnaround time in precision oncology programs.
- Additional interactive visualizations to support patient diagnosis, prognosis and therapy selection by interdisciplinary molecular tumor boards.

TelomereHunter2 is currently in beta testing and will be released as open-source software implemented in Python3.

Keywords: Cancer informatics, Telomere biology, High-throughput sequencing, Alternative lengthening of telomeres (ALT), Telomerase activation

Lars Feuerbach (Division of Applied Bioinformatics, German Cancer Research Center (DKFZ)), Niklas Engel (Division of Applied Bioinformatics, German Cancer Research Center (DKFZ)), Ferdinand Popp (Division of Applied Bioinformatics, German Cancer Research Center (DKFZ)), Lea Herzel (Division of Applied Bioinformatics, German Cancer Research Center (DKFZ)), Julie Surmely (Division of Applied Bioinformatics, German Cancer Research Center (DKFZ)), Hanna Frieß (Division of Applied Bioinformatics, German Cancer Research Center (DKFZ)), Malte Simon (Division of Applied Bioinformatics, German Cancer Research Center (DKFZ)), Charles Imbusch (Division of Applied Bioinformatics, German Cancer Research Center (DKFZ)) and Benedikt Brors (Division of Applied Bioinformatics, German Cancer Research Center (DKFZ)). *Quantifying immune cell telomere content at single-cell resolution in context of PD-1 checkpoint immunotherapy.*

Abstract. Introduction: Biological processes such as aging, carcinogenesis, and immune response rely on the ability to maintain or rapidly expand cell populations. The fitness of the involved cells is constrained by their replicative potential, which is reflected in the cellular telomere content. To study the impact of telomere length on immunotherapy, we describe a novel workflow for the inference of telomere content from scATAC-seq data. We then characterize the response of the T-cell compartment to programmed cell death protein 1 (PD1) blockade in basal cell carcinoma on the telomerome-level.

Method: We apply the TelomereHunter software to scATAC-seq data to determine telomere content on single-cell level in a publicly-available hematopoietic dataset consisting of 35,139 cells. Integrating information from open-chromatin-based signatures to assess cell identity, we characterize the heterogeneity of telomere length for individual cell populations pre- and post-immunotherapy.

Results: The extracted telomeric reads from the scATAC-seq data reflect the expected telomereome to genome fraction. Telomere content distributions differ significantly between cell populations. We observe that the median telomere content in intermediate and terminal exhausted CD8+ T-cells prior to treatment onset is significantly correlated to response to PD-1 checkpoint blockade. Likewise, telomere content correlates with post-treatment cell proliferation in terminally exhausted and T follicular helper cells from patients responding to immunotherapy.

Conclusion: Telomere content measurement from scATAC-seq data has a sufficiently high signal-to-noise ratio to detect significant differences between cell types and states. Furthermore, the telomere content of CD8+ exhausted T-cells at treatment onset is a putative biomarker for successful PD-1-based immunotherapy.

Keywords: telomere, immunotherapy, single-cell analysis, biomarker, computational workflow

Yan Gao (Department of Data Science, Dana-Farber Cancer Institute) and Heng Li (Department of Data Science, Dana-Farber Cancer Institute; Department of Biomedical Informatics, Harvard Medical School). *LongcallD: joint calling of small variants and large structural variants from long reads.*

Abstract. Accurate detection of genetic variants across all size ranges is critical for understanding genomic diversity and disease mechanisms. Existing tools often handle small variants, i.e., single-nucleotide polymorphisms (SNPs) and small insertions/deletions (indels), and large structural variants (SVs) separately, leading to fragmented analyses. We present longcallD, a unified algorithm for joint calling of small variants and SVs from long-read sequencing data. Leveraging the phasing information inherent in long-read data, longcallD phases long reads into haplotypes using germline SNPs and small indels prior to SV detection. This approach also enables sensitive detection of low-allele-fraction mosaic variants, including single-nucleotide variants and transposable element insertions. Benchmarking against the Genome in a Bottle HG002 T2T v1.1 germline variant reference demonstrates that longcallD achieves comparable or superior small variant calling and consistently outperforms existing methods in SV detection, particularly in repetitive regions. Additional evaluations using COLO829 and H2009 synthetic datasets indicate that longcallD significantly reduces false positive mosaic variant calls originating from germline events when compared to current approaches. Implemented in C, longcallD offers an efficient and streamlined solution for comprehensive variant analysis in genomic research. The software is available at https://github.com/yangao07/longcallD.

Keywords: long reads, variant calling, haplotype phasing, structural variants, mosaic variants

Daniela E. Kirwan (Institute for Infection & Immunity, City St. George's, University of London. London, SW17 0RE, United Kingdom.), Daniel Garbozo-Santillan (Bioinformatics Group in Multiomics and Immunology. New York, NY, 10016, USA.), Daniel F. Guevara-Díaz (Bioinformatics Group in Multiomics and Immunology. New York, NY, 10016, USA.), Carla Apaza (Laboratorio de Bioinformática, Biología Molecular y Desarrollos Tecnológicos, Universidad Peruana Cayetano Heredia), Diego Taquiri-Diaz (Laboratorio de Bioinformática, Biología Molecular y Desarrollos Tecnológicos, Universidad Peruana Cayetano Heredia), Deborah L.W. Chong (Institute for Infection & Immunity, City St. George's, University of London. London, SW17 0RE, United Kingdom.), Mirko Zimic (Laboratorio de Bioinformática, Biología Molecular y Desarrollos Tecnológicos, Universidad Peruana Cayetano Heredia), Jon S. Friedland (Institute for Infection & Immunity, City St. George's, University of London. London, SW17 0RE, United Kingdom.) and David Requena (Bioinformatics Group in Multiomics and Immunology. New York, NY, 10016, USA). *Platelet transcriptomic profiling provides novel insights into Tuberculosis disease.*

Abstract. Tuberculosis (TB), caused by Mycobacterium tuberculosis, remains one of the leading infectious diseases worldwide. Although elevated platelet counts are frequently observed in TB patients, the molecular mechanisms underlying their role in the innate immune response and disease progression remain poorly understood. To uncover potential biological processes, we performed bulk transcriptomic profiling of platelets from 20 Peruvian patients using paired-end sequencing and state-of-the-art bioinformatics methods. Transcript quantification was performed with Salmon, followed by removal of ribosomal and low count genes (mean normalized counts <2). Differential expression analysis of TB patients over controls revealed 269 up- and 78 downregulated genes (FDR < 0.05). Eleven upregulated genes were also present in the transcriptome of three external whole-blood datasets; including interferon-stimulated genes such as STAT1, GBP1, RSAD2, CMPK2, and IFIT3. Also, FCGR1A, which has the strongest differential signal and it is associated with intracellular mycobacterial response and proinflammatory signaling. Moreover, using three different competitive gene-set enrichment analysis methods (CAMERA, PADOG, and GSEA) identified 21 gene sets highly enriched in infected patients. Notably, several converged to innate immune processes, including pathogen sensing, inflammatory responses, glycolipid binding, and antigen presentation via MHC class I. Altogether, this study provides new insights into the gene expression changes and biological pathways in platelets during TB infection. This may contribute to a better understanding of their role in immune modulation and pathogenesis, potentially leading to new ways to attack the disease.

Keywords: RNA-seq, Platelets, Tuberculosis, Gene set enrichment analysis, Innate Immunity

Nika Mansouri Ghiasi (ETH Zurich), Talu Güloglu (ETH Zurich), Harun Mustafa (ETH Zurich), Can Firtina (ETH Zurich), Konstantina Koliogeorgi (ETH Zurich), Konstantinos Kanellopoulos (ETH Zurich), Haiyu Mao (King's College London), Rakesh Nadig (ETH Zurich), Mohammad Sadrosadati (ETH Zurich), Jisung Park (Pohang University of Science and Technology (POSTECH)) and Onur Mutlu (ETH Zurich). *SAGe: A Lightweight Algorithm-Architecture Co-Design for Mitigating the Data Preparation Bottleneck in Large-Scale Genome Analysis.*

Abstract. Given the exponentially growing volumes of genomic data, there are extensive efforts to accelerate genome analysis. We demonstrate a major bottleneck that greatly limits and diminishes the benefits of state-of-the-art genome analysis accelerators: the data preparation bottleneck, where genomic data is stored in compressed form and needs to be decompressed and formatted first before an accelerator can operate on it. To mitigate this bottleneck, we propose SAGe, an algorithm-architecture co-design for highly-compressed storage and high-performance access of large-scale genomic data. SAGe overcomes the challenges of mitigating the data preparation bot- tleneck while maintaining high compression ratios (comparable to genomic-specific compression algorithms) at low hardware cost. This is enabled by leveraging key features of genomic datasets to co-design (i) a new (de)compression algorithm, (ii) hardware, (iii) storage data layout, and (iv) interface commands to access storage. SAGe stores data in structures that can be rapidly interpreted and decompressed by efficient streaming accesses and lightweight hardware. To achieve high compression ratios using only these lightweight structures, SAGe exploits unique features of genomic data. We show that SAGe can be seamlessly integrated with a broad range of genome analysis hardware accelerators to mitigate their data preparation bottlenecks. Our results demonstrate that SAGe improves the average end-to-end performance and energy efficiency of two state-of-the-art genome analysis accelerators by 3.0x–32.1x and 18.8x–49.6x, respectively, compared to when the accelerators rely on state-of-the-art decompression tools.

Keywords: Genome analysis hardware acceleration, DNA sequence data compression, Algorithm-hardware co-design

Sara Knaack (University of Wisconsin Madison). *A bridge to precision knowledge: Experimentally motivated quantitative analysis approaches bring de novo power to RNA-seq interpretation.*

Abstract. A historical trope places physics in the sphere of precision knowledge, but biology in a sphere of lower precision. This belies the increasingly precise nature of cellular and molecular biology experimental techniques. Current data analysis approaches in biology don't fully address the challenges of experimental reproducibility. Take the task of differential gene expression (DEG) analysis in RNA-seq analysis. Differing methods typically produce distinct results on even identical data. Differing methods are commonly used for different platforms/modalities, e.g., EBSeq and DESeq2 on bulk RNA-seq data, or Wilcoxon rank and MAST on snRNA-seq data. Choices of fold-change and p-value or adjusted-p-value thresholds and multiple hypothesis correction method all impact conclusions. This shifts analysis away from unbiased, de novo interpretation towards tuning and curation from prior knowledge. In short, the essential reproducibility and precision of knowledge generated is not as well understood as hoped in this paradigm. I present exploratory analysis of published bulk RNA-seq data from yeast following experimentally motivated statistical principles used in particle physics to assign uncertainty values to observed counts. These principles enable intuitive understanding of the reproducibility of observations and high-confidence DE gene calling. These principles can further extend to clustering and (co-association) network construction through choices of distance metrics, facilitating an effective combination of experimental reasoning and confidence in de novo statistical power. More broadly, such approaches enable objective interpretation and integration of data across experiments (and modalities), crucial for establishing precision knowledge and maximizing benefit from resource-intensive experiments, and support progress towards precision medicine.

Keywords: RNA-seq, differentially expressed gene, DE gene analysis, precision, measurement, precision medicine, quantitative biology, experimental data analysis

Rebecca Chen (University of Calgary), Libby Redman (University of Calgary), Lynsey Melville (Moredun Research Institute), Dave Bartley (Moredun Research Institute) and John Gilleard (University of Calgary). *Screening drug resistance mutations in parasitic nematodes using Oxford Nanopore long-read amplicon sequencing.*

Abstract. Illumina targeted amplicon sequencing of β-tubulin, benzimidazole drug target, is commonly applied to screen for benzimidazole resistance (BZ-R) SNPs in human and animal parasitic nematode populations. However, a major limitation is the short sequenced fragment (<600bp), covering only a small region of the complete β-tubulin gene, which restricts our ability to detect other potential BZ-R conferring mutations. Long-read sequencing from Oxford Nanopore Technologies (ONT) has become increasingly affordable but is still less accurate than Illumina.

To assess the accuracy and sensitivity of BZ-R SNP detection in parasite populations using ONT long-read amplicon sequencing, frequencies of canonical BZ-R SNPs were compared to Illumina short-read sequencing as ground truth. Parasitic nematode (Nematodirus battus and Haemonchus contortus) β-tubulin gene was amplified and sequenced from 30 farm populations using ONT MinION and Illumina MiSeq in multiple PCR replicates.

A pipeline for non-synonymous SNP detection from amplicon sequencing was developed. Briefly, adapter sequences and low quality reads were trimmed and removed with Cutadapt. Reads were aligned to reference sequence using Minimap2, and variants called with Clair3 (ONT) or GATK (Illumina). SNPs resulting in synonymous and non-synonymous mutations in exon regions were annotated and visualized with a custom R script.

Canonical BZ-R SNP frequencies were very consistent across PCR replicates and comparable to Illumina results. Additional β-tubulin non-synonymous SNPs were also detected in regions not covered by Illumina short-read sequencing. These results show ONT long-read amplicon sequencing is reliable for SNP detection in parasite populations and uncovers additional potential BZ-R SNPs with our pipeline.

Keywords: oxford nanopore, amplicon sequencing, variant calling, drug resistance

Shay Golan (University of Haifa and Recihman University), Ido Tziony (Bar-Ilan University), Matan Kraus (Bar-Ilan University), Yaron Orenstein (Bar-Ilan University) and Arseny Shur (Bar-Ilan University). *GreedyMini: Generating low-density DNA minimizers.*

Abstract. Minimizers is the most popular k-mer selection scheme in algorithms and data structures analyzing high-throughput sequencing (HTS) data. In a minimizers scheme, the smallest k-mer by some predefined order is selected as the representative of a sequence window containing w consecutive k-mers, which results in overlapping windows often selecting the same k-mer. Minimizers that achieve the lowest frequency of selected k-mers over a random DNA sequence, termed the expected density, are desired for improved performance of HTS analyses. Yet, no method to date exists to generate minimizers that achieve minimum expected density. Moreover, for k and w values used by common HTS algorithms and data structures there is a gap between the densities achieved by existing selection schemes and a recent theoretical lower bound. Here, we present GreedyMini, a toolkit of methods to generate minimizers with low expected or particular density, to improve minimizers, to extend minimizers to larger alphabets, k, and w, and to measure the expected density of a given minimizer efficiently. We demonstrate over various combinations of k and w values, including those of popular HTS methods, that GreedyMini can generate DNA minimizers that achieve expected densities very close to the lower bound, and both expected and particular densities much lower compared to existing selection schemes. Additionally, we show that the k-mer rank-retrieval time by GreedyMini is comparable to that of common k-mer hash functions. We expect GreedyMini to improve the performance of many HTS algorithms and data structures and advance the research of k-mer selection schemes.

Keywords: minimizers, high-throughput sequencing, de Bruin graphs