High Throughput Sequencing
Algorithms and Applications
A special track of the ISMB-ECCB 2023 meeting
Lyon, France, July 25-26, 2023

ISMB-ECCB 2023 HiTSeq Track Proceedings

Lyon, France
July 25-26, 2023
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

Kuan-Hao Chao (Johns Hopkins University), Aleksey V Zimin (Johns Hopkins University), Mihaela Pertea (Johns Hopkins University) and Steven L. Salzberg (Johns Hopkins University). *The first gapless, reference-quality, fully annotated genome from a Southern Han Chinese individual.*

Abstract. We used long-read DNA sequencing to assemble the genome of a Southern Han Chinese male. We organized the sequence into chromosomes and filled in gaps using the recently completed T2T-CHM13 genome as a guide, yielding a gap-free genome, Han1, containing 3,099,707,698 bases. Using the T2T-CHM13 annotation as a reference, we mapped all genes onto the Han1 genome and identified additional gene copies, generating a total of 60,708 putative genes, of which 20,003 are protein coding. A comprehensive comparison between the genes revealed that 235 protein-coding genes were substantially different between the individuals, with frameshifts or truncations affecting the protein-coding sequence. Most of these were heterozygous variants in which one gene copy was unaffected. This represents the first gene-level comparison between two finished, annotated individual human genomes.

Keywords: genome assembly, annotation, DNA sequencing, reference genome, variant calling

Nisha Hemandhar Kumar (University Medical Center Goettingen), Eugenio Fornasiero (University medical center goettingen), Mattia Pelizzola (Fondazione Istituto Italiano di Tecnologia (IIT)), Mattia Furlan (Fondazione Istituto Italiano di Tecnologia (IIT)), Verena Kluever (University Medical Center Göttingen), Emanuel Barth (Friedrich Schiller University Jena), Manja Marz (Friedrich Schiller University Jena), Sebastian Krautwurst (Friedrich Schiller University Jena), Cristina Cheroni (European Institute of Oncology, University of Milan) and Giuseppe Testa (European Institute of Oncology, University of Milan, Human Technopole). *Detailed analysis of the aged brain transcriptome reveals differences in mRNA properties and impaired mRNA turnover.*

Abstract. Brain aging is characterized by a progressive loss of tissue integrity and increased cellular heterogeneity, leading to impaired function, increased susceptibility to disease and death. In this work, we collected and analyzed two large datasets including mRNA levels from Illumina and Oxford Nanopore technologies in young adult and aged adult mice. We report the first transcriptome-wide differential transcript usage study of brain aging. We provide the community with a large resource of whole brain transcriptomes and comprehensive analyses that identify widespread diversity in RNAs during aging. Specifically, we observed that the mRNAs encoding for neuronal synaptic proteins are upregulated with age and this is conserved in the human context. We also observed that a subset of the genes that are upregulated in the aged brain is associated with neurodegenerative diseases. In addition, we report that the RNA molecules that are longer and have a shorter 3'UTR are abundant in the aged brain and a subset of these are alternatively spliced at the 3'UTR. Finally, we observed a difference in the turnover of mRNAs at different ages. Overall, based on these observations, we speculate that alterations at the 3'UTR may play an active role in the aging process.

Keywords: Physiological aging, proteogenomics, differential transcript usage, neurodegeneration, Illumina, Oxford Nanopore technology

Zhengyu An (The Institute of Science and Technology for Brain-inspired Intelligence, Fudan University), Jie Zhang (The Institute of Science and Technology for Brain-inspired Intelligence, Fudan University), Jijun Wang (Shanghai Key Laboratory of Psychotic Disorders, Shanghai Jiao Tong University School of Medicine, Shanghai, China), Jingqi Chen (The Institute of Science and Technology for Brain-inspired Intelligence, Fudan University) and Xing-Ming Zhao (The Institute of Science and Technology for Brain-inspired Intelligence, Fudan University). *Long-read WGS revealed potentially pathogenic structural variants in repeat regions for a Chinese schizophrenia cohort.*

Abstract. We used Long-read sequencing (LRS) to identify structural variants (SVs) in 141 schizophrenic cases and obtained a median of 14,392 high-confidence SVs per individual using an alignment-based pipeline. We compared these SVs with those detected by short-read sequencing (SRS) of the same samples and found that 31.5% of the SVs detected by LRS were consistent with 64.5% of those detected by SRS. LRS-specific SVs were enriched in segmental duplications (SD) and simple repeats (SR) regions, demonstrating the advantages of LRS in these regions. After filtering through the public population-scale SV sets, we identified 618 potential pathogenic SVs that were enriched in SR regions and carried by multiple cases, detected only by LRS-based methods. We also identified previously reported SCZ-associated SVs, such as TRA in DISC1 and VNTRs in SLC6A4, among this set. These potential pathogenic SVs affected 551 genes, which were significantly enriched in developmental and synaptic-related pathways that tended to be associated with SCZ. Our study highlights the effectiveness of LRS in identifying SVs, particularly in repeat regions, and provides new insights into the genetic mechanisms of diseases.

Keywords: Long-read sequencing, structural variants, schizophrenia, repeat regions

Alaina Shumate (Johns Hopkins University), Brandon Wong (Johns Hopkins University), Geo Pertea (Lieber Institute for Brain Development) and Mihaela Pertea (Johns Hopkins University). *Efficient and robust transcriptome reconstruction from hybrid RNA-seq data.*

Abstract. Short-read and long-read RNA sequencing technologies each have their strengths and weaknesses for transcriptome assembly. While short reads are highly accurate, they are rarely able to span multiple exons. Long-read technology can capture full-length transcripts, but its relatively high error rate often leads to mis-identified splice sites. The initial version of StringTie2, our guided transcriptome assembler, was able to handle either short or long read data but would not be able to handle both data types at the same time. We improved on StringTie2 and released a new version of StringTie that is capable of handling mixed transcriptomic data that includes both short and long RNA-seq reads sequenced from the same sample. By taking advantage of the strengths of both long and short reads, hybrid-read assembly with StringTie is more accurate than long-read only or short-read only assembly, and on some datasets it can more than double the number of correctly assembled transcripts, while obtaining substantially higher precision than the long-read data assembly alone. Using real and simulated data, we show that hybrid-read assemblies achieve greater precision and sensitivity than both the corrected or uncorrected long-read only, and short-read only assemblies as well as better estimates of gene expression levels.

Keywords: transcriptome assembly and quantification, short-read RNA-sequencing, long-read RNA-sequencing, long-read error correction

Ruibin Xi (Peking University), Zijie Jin (Peking University), Wenjian Huang (Peking University), Ning Shen (Zhejiang University Medical Center), Juan Li (Peking University), Xiaochen Wang (Peking University), Jiqiao Dong (GeneX health Co. Ltd. Beijing) and Peter Park (Harvard University).
Single-cell gene fusion detection by scFusion.

Abstract. Gene fusions can play important roles in tumor initiation and progression. While fusion detection so far has been from bulk samples, full-length single-cell RNA sequencing (scRNA-seq) offers the possibility of detecting gene fusions at the single-cell level. However, scRNA-seq data have a high noise level and contain various technical artifacts that can lead to spurious fusion discoveries. Here, we present a computational tool, scFusion, for gene fusion detection based on scRNA-seq. We evaluate the performance of scFusion using simulated and five real scRNA-seq datasets and find that scFusion can efficiently and sensitively detect fusions with a low false discovery rate. In a T cell dataset, scFusion detects the invariant TCR gene recombinations in mucosal-associated invariant T cells that many methods developed for bulk data fail to detect; in a multiple myeloma dataset, scFusion detects the known recurrent fusion IgH-WHSC1, which is associated with overexpression of the WHSC1 oncogene. Our results demonstrate that scFusion can be used to investigate cellular heterogeneity of gene fusions and their transcriptional impact at the single-cell level.

Keywords: single-cell sequencing, gene fusion, RNA sequencing, structural variation, deep learning, breakpoints

Ram Ayyala (USC), Dottie Yu (USC) and Serghei Mangul (USC). *Rigorous Benchmarking of HLA Callers for RNA-seq Data.*

Abstract. Precise identification of alleles in the human leukocyte antigen (HLA) region of the human genome is crucial for many clinical and research applications. However, HLA typing remains challenging due to the highly polymorphic nature of the HLA loci. With Next-Generation Sequencing (NGS) data becoming widely accessible, many computational tools have been developed to predict HLA types from RNA sequencing (RNA-seq) data. Despite this development, there remains a lack of comprehensive and systematic benchmarking of RNA-seq based HLA callers. To address this limitation, we rigorously compared the performance of all 9 HLA callers currently published on six gold standard datasets spanning 652 RNA-seq samples. In each case, we produced evaluation metrics of accuracy for each caller that is the percentage of correctly predicted alleles. We then reported the HLA genes and alleles most prone to misprediction. Furthermore, we evaluated the performance of each caller through their runtime and usage of CPU and memory. Our study is also the first to evaluate effect of read length on prediction quality using each tool, and the effect of ancestral origin (African vs. European) on accuracy. This study offers crucial information for researchers and clinicians regarding appropriate choices of methods for HLA typing.

Keywords: Human Leukocyte Antigen, HLA typing, RNA-seq, short read, HLA callers, immunogenomics, benchmarking

Mikaela Koutrouli (Novo Nordisk Foundation Center of Protein Research), Radha Swaminathan (Department of Chemistry and Chemical Biology, University of New Mexico, Albuquerque), Jeremy Edwards (Department of Chemistry and Chemical Biology, University of New Mexico, Albuquerque) and Lars Juhl Jensen (Novo Nordisk Foundation Center of Protein Research). *Visualizing Spatial Transcriptomics with U-CIE Color Encoding.*

Abstract. Spatial transcriptomics is a cutting-edge technique that enables the analysis of gene expression patterns within specific regions of tissue or organs. However, analyzing the large and complex datasets generated from spatial transcriptomics experiments remains a challenge. Here we propose U-CIE, a method for visualizing high-dimensional data by encoding it as colors using a combination of dimensionality reduction and the CIELAB color space. U-CIE allows genome-wide expression patterns within tissue or organ sections to be visualized and highlights the distribution of different cell types across the spatial transcriptomics data. U-CIE first uses UMAP to reduce high-dimensional gene expression data to three dimensions while preserving the spatial information. Next, the resulting three-dimensional representation is embedded within the CIELAB color space, generating a color encoding that captures much of the original structure of the data. U-CIE has been successfully applied to a mouse brain section dataset to highlight the distribution of different cell types across the spatial transcriptomics data and provide insights into the organization of these cells within brain regions. U-CIE has the potential to be a powerful tool for exploring spatial transcriptomics data and gaining new insights into cellular organization and function.

Keywords: spatial transcriptomics, visualization, color encoding, cell type composition, single cells

Michael P Lynch (University of Limerick), Yufei Wang (Dana-Farber Cancer Institute), Laurent Gatto (UCLouvain) and Aedin C Culhane (University of Limerick). *demuxSNP: supervised demultiplexing of scRNAseq data using cell hashing and SNPs.*

Abstract. Single-cell sequencing allows unprecedented understanding of biologically relevant differences between individual cells. Multiplexing, that is loading multiple biological samples into each sequencing lane, is widely used to further reduce sequencing costs. The sequencing reads must then be demultiplexed or identified as being from a particular biological sample. Methods to date have either used cell hashing labels (tags) or SNPs. We present our approach and its corresponding R package 'demuxSNP' which overcomes current technical challenges in demultiplexing scRNAseq reads which can be applied to genetically distinct biological samples.
demuxSNP uses data from both tags and SNPs. demuxSNP performs SNP feature selection then trains a doublet-aware knn classifier on the SNP profiles of singlet cells called with high confidence using cell tagging methods. Low confidence cells (cells which we could not confidently call using cell tagging methods) are then assigned based on their SNP profiles. demuxSNP is a computationally efficient and cell-type unbiased algorithm for demultiplexing genetically distinct biological samples.

Keywords: scRNAseq, single-cell, demultiplexing, cell hashing, SNPs

Tim Dunn (University of Michigan) and Satish Narayanasamy (University of Michigan). *vcfdist: Accurately benchmarking phased small variant calls in human genomes.*

Abstract. Accurately benchmarking small variant calling accuracy is critical for the continued improvement of human whole genome sequencing (WGS). Discovering true positive variants during WGS is critical for the detection of cancer and many other genetic diseases. In this work, we show that current variant calling evaluations are biased towards certain variant representations and may misrepresent the relative performance of different variant calling pipelines. In particular, this leads to inconsistent evaluations near low-complexity repetitive regions.

We propose solutions, first exploring the affine gap alignment parameter design space for complex variant representation and suggesting a standard. Next, we present our tool "vcfdist" and demonstrate the importance of enforcing local phasing for evaluation accuracy. We then introduce the notion of partial credit for mostly-correct calls and present an algorithm for clustering dependent variants. Lastly, we motivate using alignment distance metrics to supplement precision-recall curves for understanding variant calling performance.

The performance of vcfdist is compared against vcfeval, the currently accepted standard for VCF benchmarking. We evaluate 64 phased VCF submissions to the precisionFDA's variant calling "Truth Challenge V2" and show that vcfdist improves measured (SNP, INDEL) performance consistency across variant representations from R^2 = (0.96888, 0.97243) for vcfeval to (0.99999, 0.99996) for vcfdist.

Keywords: variant calling, small variants, germline variants, complex variants, benchmarking, VCF evaluation, alignment

Lauren Mak (Cornell Tri-Institutional Computational Biology & Medicine Program, Weill Cornell Medicine of Cornell University), Braden Tierney (Institute for Computational Biomedicine, Weill Cornell Medicine, New York, NY, 10065, USA), Cynthia Ronkowski (University of Southern California), Michael Toomey (Tri-Institutional Computational Biology & Medicine Program, Weill Cornell Medicine of Cornell University), Juan Sebastián Andrade Martínez (Universidad de los Andes), Sam Zimmerman (Section on Pathophysiology and Molecular Pharmacology, Joslin Diabetes Center, Boston, MA, USA), Chenlian Fu (Tri-Institutional Computational Biology & Medicine Program, Weill Cornell Medicine of Cornell University), Malika Kopbayeva (School of Molecular and Theoretical Biology, Tartu, Estonia), Anna Noyvert (School of Molecular and Theoretical Biology, Tartu, Estonia), Brett Farthing (Zymo Research, 17062 Murphy Ave, Irvine, CA 92614), Shuiquan Tang (Zymo Research, 17062 Murphy Ave, Irvine, CA 92614), Christopher Mason (Cornell University) and Iman Hajirasouliha (Cornell University). *A modular metagenomics analysis system for integrated multi-step data exploration.*

Abstract. Motivation: Computational analysis of large-scale metagenomics sequencing datasets has proved to be both incredibly valuable for extracting isolate-level taxonomic and functional insights from complex microbial communities. However, thanks to an ever-expanding ecosystem of metagenomics-specific algorithms and file formats, designing studies, implementing seamless and scalable end-to-end workflows, and exploring the massive amounts of output data have become studies unto themselves. Furthermore, there is little inter-communication between output data of different analytic purposes, such as short-read classification and metagenome assembled genomes (MAG) reconstruction. One-click pipelines have helped to organize these tools into targeted workflows, but they suffer from general compatibility and maintainability issues.
Results: To address the gap in easily extensible yet robustly distributable metagenomics workflows, we have developed a module-based metagenomics analysis system written in Snakemake, a popular workflow management system, along with a standardized module and working directory architecture. Each module can be run independently or conjointly with a series of others to produce the target data format (ex. short-read preprocessing alone, or short-read preprocessing followed by de novo assembly), and outputs aggregated summary statistics reports and semi-guided Jupyter notebook-based visualizations, The module system is a bioinformatics-optimized scaffold designed to be rapidly iterated upon by the research community at large.

Keywords: Metagenomics, Microbiomes, Big data analysis, Pipeline design, Benchmarking, Metagenome-assembled genomes, Taxonomic classification

Yeremia Gunawan Adhisantoso (Leibniz University Hannover), Jan Voges (Leibniz University Hannover) and Jörn Ostermann (Leibniz University Hannover). *PEKORA: High-Performance 3D Genome Reconstruction Using K-th Order Spearman's Rank Correlation Approximation.*

Abstract. Advances in high-throughput sequencing technologies have enabled the use of genomic information to better understand biological processes through studies such as genome- wide association studies, polygenic risk score estimation and chromosome conformation capture. The study of spatial chromosome organization of the human genome plays an important role in understanding gene regulation. Chromosome conformation capture techniques, such as Hi-C, can capture long-range interactions between all pairs of loci on all chromosomes. These techniques have revealed structures of genome organization, such as A/B compartments, topologically associated domains, chromatin loops and frequently interacting regions.
Although the advancement of Hi-C techniques enables the generation of massive amounts of high-resolution data, we face several challenges such as a high proportion of missing data and noisy observed interaction frequencies. Therefore, it is currently unfeasible to reconstruct high-resolution genome structures efficient at high accuracy using existing state-of-the-art methods. To remedy this situation, we present PEKORA, a high-performance 3D genome reconstruction method using k-th order Spearman's rank correlation approximation. PEKORA outperforms the state of the art by a huge margin of 35% on average.

Keywords: Hi-C, contact matrix, chromosome conformation capture, 3D reconstruction, optimization, machine learning

David Pellow (Tel-Aviv University), Lianrong Pu (Tel-Aviv University), Baris Ekim (Massachusetts Institute of Technology), Lior Kotlar (Ben-Gurion University), Ron Shamir (Tel-Aviv University) and Yaron Orenstein (Bar-Ilan University). *Efficient minimizer orders for large values of k using minimum decycling sets.*

Abstract. Minimizers are ubiquitously used in algorithms for efficient searching, mapping, and indexing of high-throughput DNA sequencing data.
Minimizer schemes select a minimum k-mer with respect to a predefined k-mer order in every L-long sub-sequence of the target sequence.
Commonly used minimizer orders select more k-mers
than necessary and therefore provide limited improvement in runtime and memory usage of downstream analysis tasks. The recently introduced universal k-mer hitting sets produce minimizer orders with fewer selected k-mers. Unfortunately, generating compact universal k-mer hitting sets is currently infeasible for k>13.
Here, we close the gap of efficient minimizer orders for large values of k by introducing decycling-set-based minimizer orders. We show that in practice these new minimizer orders select a number of k-mers comparable to that of minimizer orders based on universal k-mer hitting sets, and can also scale to larger k.
Furthermore, we developed a method that computes the minimizers in a sequence on the fly without keeping the k-mers of a decycling set in memory.
This enables the use of these minimizer orders for any value of k. We expect the new orders to improve the runtime and memory usage of algorithms and data structures in high-throughput DNA sequencing analysis.

Keywords: Minimizers, K-mers, Decycling set, de Bruijn graph

Justin Moy (Bioinformatics Program, Boston University, Boston, MA, USA), Irzam Sarfraz (Bioinformatics Program, Boston University, Boston, MA, USA), Vishal Shah (Bioinformatics Program, Boston University, Boston, MA, USA), Edward Ruiz (Section of Hematology and Medical Oncology, School of Medicine, Boston University, Boston, MA, USA), Guo-Cheng Yuan (Precision Immunology Institute, Icahn School of Medicine at Mount Sinai, New York, NY, USA) and Ruben Dries (Section of Hematology and Medical Oncology, School of Medicine, Boston University, Boston, MA, USA). *Developing a unified framework to identify the layers and dynamics of spatial co-expression programs in ultra-resolved spatial data.*

Abstract. The Mouse Organogenesis Spatiotemporal Transcriptomic Atlas (MOSTA) contains spatial gene expression measurements for ~25,000 genes, captured by Stereo-Seq at 500nm resolution, for the entire mouse embryo, across eight developmental time points from E9.5 to E16.5. Data on this scale (terabytes) provide the opportunity to develop methods that untangle spatial organization, and its association with biological function, at the level of the whole organism.
We created a pipeline to identify spatial programs at the level of gene expression, niche organization, and cellular morphology. Using MOSTA expression and nuclei imaging data from timepoints E12.5, E14.5, and E16.5, we optimized a Hidden Markov Random Field model that differentially weights feature similarity with neighborhood effect to classify individual spatial units into larger spatially coherent domains. It is compatible with different types of spatial information: expression levels, tissue organization, and cell morphology. We also identified spatial co-expression modules (genes displaying spatial co-expression in a smoothed spatial k-nearest neighbor network) and evaluated the stability of these modules across adjacent tissue sections and consecutive time points of development. These analyses allow scientists to assess the biological role and organization of multiple spatial programs and how they change in a spatio-temporal manner.

Keywords: spatial transcriptomics, hidden markov random field, mouse embryogenesis, hmrf, development, markov, MOSTA, cell segmention, co-expression modules, niche organization, machine learning, spatiotemporal

Jin-Wu Nam (Department of Life Science, Hanyang University). *Ultrafast prediction of somatic structural variations by filtering out reads matched to pan-genome k-mer sets.*

Abstract. Variant callers typically produce massive numbers of false positives for structural variations, such as cancer-relevant copy-number alterations and fusion genes resulting from genome rearrangements. Here we describe an ultrafast and accurate detector of somatic structural variations that reduces read-mapping costs by filtering out reads matched to pan-genome k-mer sets. The detector, which we named ETCHING (for efficient detection of chromosomal rearrangements and fusion genes), reduces the number of false positives by leveraging machine-learning classifiers trained with six breakend-related features (clipped-read count, split-reads count, supporting paired-end read count, average mapping quality, depth difference and total length of clipped bases). When benchmarked against six callers on reference cell-free DNA, validated biomarkers of structural variants, matched tumour and normal whole genomes, and tumour-only targeted sequencing datasets, ETCHING was 11-fold faster than the second-fastest structural-variant caller at comparable performance and memory use. The speed and accuracy of ETCHING may aid large-scale genome projects and facilitate practical implementations in precision medicine.
Citation: https://doi.org/10.1038/s41551-022-00980-5

Keywords: Pan-genome K-mer sets, Somatic Structural Variations, Fusion Genes, Machine Leanrning, Enhancer Hijacking, Whole Genome Sequencing, Cancer Panel Sequencing

George Howitt (Peter MacCallum Cancer Centre), Yuzhou Feng (Peter MacCallum Cancer Centre), Lucas Tobar (Peter MacCallum Cancer Centre), Dane Vassiliadis (Peter MacCallum Cancer Centre), Peter Hickey (Walter and Eliza Hall Institute of Medical Research), Mark A. Dawson (Peter MacCallum Cancer Centre), Sarath Ranganathan (Royal Children's Hospital), Shivanthan Shanthikumar (Royal Children's Hospital, The University of Melbourne), Melanie Neeland (Murdoch Children's Research Institute), Jovana Maksimovic (Peter MacCallum Cancer Centre, The University of Melbourne) and Alicia Oshlack (Peter MacCallum Cancer Centre). *Benchmarking single-cell hashtag oligo demultiplexing methods.*

Abstract. As single cell RNA-sequencing becomes more accessible and reliable, research groups can design replicated experiments with multiple biological samples. For large experiments, sample multiplexing is often used to reduce cost and limit batch effects.
A commonly used multiplexing technique involves tagging cells prior to pooling with a hashtag oligo (HTO) that can be sequenced along with the cells' RNA to determine their sample of origin. Several tools have been developed to demultiplex HTO sequencing data and assign cells to samples, but these tools are often tested using data with high-quality HTO labelling and low contamination. Using experimental data sets with both good and poor labelling of samples, we critically assess the performance of seven HTO demultiplexing tools: hashedDrops, HTODemux, GMM-Demux, demuxmix, deMULTIplex, BFF and HashSolo. Each sample in our data sets has also been demultiplexed using genetic variants from the RNA, enabling comparison of HTO demultiplexing techniques against complementary data from the genetic "ground truth". We find that all methods perform similarly where HTO labelling is of high quality, but methods that assume a bimodal counts distribution perform poorly on lower quality data. We also provide heuristic approaches for assessing the quality of HTO counts in a scRNA-seq experiment.

Keywords: Single-cell, Benchmarking, Demultiplexing

Jack Fraser-Govil (Wellcome Sanger Institute) and Zemin Ning (Wellcome Sanger Institute).
Reconstructing Sampling Biases from Coverage Data.

Abstract. In biological applications it is common for the underlying biochemical theory to indicate that some observable should follow a Poisson distribution, often the result of assuming that processes occur with a constant mean rate: an example being the base coverage of a DNA sequencing effort. However, it is well known that the observed distributions often demonstrate significant over-dispersion compared to a Poisson. In this work we assume this is the result of a marginalisation ('blurring') of the sampling rate, and use advanced demarginalization techniques to recover the functional form of the underlying sampling bias which can result from features in the genome, or inherently from the sequencing platform. Using Bayesian methods, we infer the statistical significance of multiple underlying models. In doing so, we discover a highly multi-modal sampling bias, with some cases demonstrating that no part of the genome was sampled to the naively recovered mean sampling rate. We suggest that this result could have important uses in the future of fields of cancer detection and sequence data quality control, and that this methodology generalises to multiple other processes within the field of genomics.

Keywords: Genome Sequencing, Statistical Analysis, Data Quality Control, Read Coverage

Laura Masatti (Department of Biology, University of Padova), Stefania Pirrotta (Department of Biology, University of Padova), Robert Fruscio (Department of Obstetrics and Gynecology, Università degli Studi Milano-Bicocca, SanGerardo Hospital, Monza), Lorenzo Ceppi (Dipartimento Ostetricia e Ginecologia, Grande Ospedale Metropolitano Niguarda, Milano), Nicolò Gnoato (Department of Biology, University of Padova), Laura Mannarino (Department of Oncology, Istituto di Ricerche Farmacologiche Mario Negri IRCCS, Milan), Luca Beltrame (Department of Oncology, Istituto di Ricerche Farmacologiche Mario Negri IRCCS, Milan), Chiara Romualdi (Department of Biology, University of Padova), Maurizio D'Incalci (Department of Oncology, Istituto di Ricerche Farmacologiche Mario Negri IRCCS, Milan), Sergio Marchini (Department of Oncology, Istituto di Ricerche Farmacologiche Mario Negri IRCCS, Milan) and Enrica Calura (Department of Biology, University of Padova). *From RNA-seq to Single-cell sequencing, the riddle of epithelial ovarian cancer heterogeneity.*

Abstract. Epithelial ovarian cancer (EOC) is one of the most frequently diagnosed cancers in women and the major cause of mortality. High-grade serous EOC (HGS-EOC) represents its most aggressive subtype. Effective therapy includes a primary surgical cytoreduction, followed by chemotherapy based on platinum analogues, but, despite an initial positive response, the majority relapse and have a poor outcome. RNA-Seq data gave us the power to investigate differential gene expression, immune gene signatures and pathways to define tumor microenvironment (TME) and immune profile of this EOC. Otherwise single-cell sequencing technology (scRNA-seq) nowadays has become a powerful method to investigate cell-to-cell transcriptomic variation, revealing new cell types and providing insights into developmental processes of this heterogeneous disease. A first cohort of patients consisting in RNA-seq of longitudinal biopsies of HGS-EOC collected at primary surgery (naïve tumor) and relapse after chemotherapy, have been compared with a scRNA-seq patient dataset composed of four sites: the ovary sampled at the exploratory laparoscopy naïve to chemotherapy, and three metastatic sites at the debulking surgery after NACT. Through the analysis and comparison of these two dataset, this study aims to identify transcriptional changes and cell composition changes occuring during chemotherapy treatment and relapses.

Keywords: Ovarian Cancer, Tumor Microenvironment, RNA sequencing, Single-cell sequencing, Resistance

Stefania Pirrotta (Biology Department, University of Padova), Laura Masatti (Biology Department, University of Padova), Nicolò Gnoato (Biology Department, University of Padova), Paolo Martini (Department of Molecular and Translational Medicine, University of Brescia), Massimo Bonora (Department of Medical Sciences, Section of Experimental Medicine, LTTA, University of Ferrara) and Enrica Calura (Biology Department, University of Padova). *mitology: a new tool to dissect mitochondrial activity from transcriptome.*

Abstract. Mitochondria are a main control center for metabolism and OXPHOS and the phenotypic manifestations of an impaired mitochondrial function may be highly heterogeneous. Further, with the new technologies of single-cell and spatial transcriptomics it is now possible to explore alterations and dissect heterogeneity at a single-cell resolution.
With the aim to provide a tool to explore mitochondrial activity in different types of transcriptomic profiles, we developed the mitology R package. A list of genes was obtained from mitochondrial-specific databases and from Gene Ontology database. Then, from the Reactome pathway database and from the Gene Ontology database, pathways and terms enriched in our list were selected and reorganized in categories used to determine processes associated with well-defined gene sets. Leveraging these categories, we can now dissect mitochondrial processes at different specificity levels. Our tool uses this information to perform a single-transcriptome analysis from gene expression input of samples, cells or spots.
Here, we provide a new tool that helps in inspecting and dissecting mitochondrial activity. It represents a strong instrument for mitochondrial studies and their impact in disease onset and progression as transcriptomes can now be studied from the mitochondrial point of view and provide a powerful contribution in clinical studies.

Keywords: Mitochondrial activity, R package, transcriptome, single-cell, spatial

Dehan Cai (City University of Hong Kong), Jiayu Shang (City University of Hong Kong) and Yanni Sun (City University of Hong Kong). *HaploDMF: viral haplotype reconstruction from long reads via deep matrix factorization.*

Abstract. Lacking strict proofreading mechanisms, many RNA viruses can generate progeny with slightly changed genomes. Being able to characterize highly similar genomes (i.e., haplotypes) in one virus population helps study the viruses' evolution and their interactions with the host/other microbes. High-throughput sequencing data has become the major source for characterizing viral populations. However, the inherent limitation on read length by next-generation sequencing makes complete haplotype reconstruction difficult. In this work, we present a new tool named HaploDMF that can construct complete haplotypes using third-generation sequencing (TGS) data. HaploDMF utilizes a deep matrix factorization model with an adapted loss function to automatically learn latent features from aligned reads. The latent features are then used to cluster reads of the same haplotype. Unlike existing tools whose performance can be affected by the overlap size between reads, HaploDMF can achieve highly robust performance on data with different coverage, haplotype number, and error rates. In particular, it can generate more complete haplotypes even when the sequencing coverage drops in the middle. We benchmark HaploDMF against the state-of-the-art tools on simulated and real sequencing TGS data on different viruses. The results show that HaploDMF competes favorably against all others.

Keywords: Viral haplotypes, Third generation sequencing, Long reads, Deep Matrix Factorization

Serghei Mangul (Department of Clinical Pharmacy, School of Pharmacy, University of Southern California) and Viorel Munteanu (Technical University of Moldova). *A rigorous benchmarking of methods for SARS-CoV-2 lineage abundance estimation in wastewater.*

Abstract. While wastewater-based genomic (WBG) surveillance is promising for viral propagation and dynamics at a population level, there is a need for comprehensive benchmarking of methods for SARS-CoV-2 lineage abundance estimation in wastewater samples. To fully unlock the potential of wastewater-based genomic surveillance we created extensive benchmarks to measure the accuracy of bioinformatics methods aimed to estimate the relative abundance of SARS-CoV-2 lineages in the wastewater samples.

We benchmarked 20 tools by exploring the dependence of the accuracy on several parameters, including different sequencing technology, length of sequencing fragments, read length, error rate, and coverage. We simulated in-silico a series of 42 mixtures mimicking samples with various lineages relative profiles and generated 16 in-vitro benchmarking real data that allowed measuring the sensitivity and specificity of bioinformatics tools.

Our results on 42 simulated samples mixed with different abundances Kallisto outperform other tools, showing a higher percentage (25%) of estimations for abundance below the absolute error of 0.1 and can detect the lower frequency of 1% below the relative error of 0.2 compared to other tools for the same experimental settings.

Our research will inform the broad biomedical community about feasible bioinformatics methods for quantifying SARS-CoV-2 strains abundances in wastewater samples.

Keywords: wastewater based-surveillance, bioinformatics, SARS-CoV-2, COVID-19

Andrew Zheng (University of Toronto), Jim Shaw (University of Toronto) and Yun William Yu (University of Toronto). *Mora: abundance aware metagenomic read re-assignment for disentangling similar strains.*

Abstract. Taxonomic classification of reads obtained by metagenomic sequencing is often a first step for understanding a microbial community, but correctly assigning sequencing reads to the strain or sub-species level has remained a challenging computational problem. We introduce Mora, a MetagenOmic read Re-Assignment algorithm capable of assigning short and long metagenomic reads with high precision, even at the strain level. Mora is able to accurately re-assign reads by first estimating abundances through an expectation-maximization algorithm and then utilizing abundance information to re-assign query reads. The key idea behind Mora is to maximize read re-assignment qualities while simultaneously minimizing the difference from estimated abundance levels, allowing Mora to avoid over assigning reads to the same genomes. On simulated diverse reads, this allows Mora to achieve F1 scores comparable to other algorithms while having less runtime. However, Mora significantly outshines other algorithms on very similar reads. We show that the high penalty of over assigning reads to a common reference genome allows Mora to accurately infer correct strains for real data in the form of short E. coli reads and long Covid-19 reads.

Keywords: Metagenomics, Read Re-Assignment, Abundance Quantification

Laura Covill (Karolinska Institute), Tim Holmes (University of Bergen), Tessa Campbell (Karolinska Institute), Ram Vinay Pandey (Karolinska Institute), Marie Meeths (Karolinska Institute) and Yenan Bryceson (Karolinska Institute). *Multi-omics workflow from diseased cell-types can be used to robustly prioritize known disease-causing non-coding variants in proof-of-concept patients .*

Abstract. Although the advent of next-generation sequencing has increased diagnostic success in instances of monogenic disease, analysis of exonic sequences can still only provide a molecular diagnosis in ~20-40% of cases. The functional impact of variants in non-coding regions are more difficult to predict. We have created and evaluated a multi-omics pipeline to increase diagnostic yield, by identifying regions of aberrant allelic imbalance in ATACseq and RNAseq from cell populations best displaying the phenotype of an individual patient. As proof-of-concept we tested the workflow on two patients with a known non-coding variant associated with hemophagocytic lymphohistiocytosis type 3 (FHL3). By filtering variants based on differential accessibility and allelic imbalance of ATACseq from patient cells vs control subsets, in addition to minor allele frequency (MAF) in a population cohort and genomic region conservation, candidate variants in non-coding regions could be identified and ranked for functional analysis.

Keywords: Diagnostics, Monogenic disease, ATAC-seq, RNA-seq, Whole genome sequencing

Wim L. Cuypers (University of Antwerp), Sandra Van Puyvelde (University of Antwerp), Kris Laukens (University of Antwerp) and Pieter Meysman (University of Antwerp). *Transcriptional rewiring in Salmonella Typhimurium from sub-Saharan Africa .*

Abstract. Transcriptional rewiring is a fundamental process in the adaptation of pathogens, yet the role of coding regions in Salmonella has been disproportionately highlighted in previous studies while neglecting the potential role of non-coding mutations and transcriptional rewiring. In this study, we re-implemented the iterative comparison of gene co-expression method in R and made it available as a package for future research (https://github.com/Cuypers-Wim/gccR). This package enables the calculation of gene co-expression conservation between two bacterial strains, taking into account technical variation in the dataset, and identifies genes with significantly diverged or conserved co-expression profiles. To illustrate the efficacy of this approach, we investigated the differences in gene co-expression between two Salmonella strains: the commonly used lab strain S. Typhimurium 14028s and the clinical isolate S. Typhimurium D23580, representative of ST313 strains causing bloodstream infections in sub-Saharan Africa. Our results indicate little overall divergence in gene co-expression between both strains, with high conservation in genes linked to translation processes. However, we observed lower conservation in genes linked to colonising the gut compared to genes required for intracellular replication and survival, which is potentially linked to the more host-adapted properties of S. Typhimurium D23580.

Keywords: Transcriptional rewiring, Gene co-expression, Gene expression, Salmonella

Angelo Velle (Department of Biology, University of Padova), Nicolò Gnoato (Department of Biology, University of Padova), Ilaria Billato (Department of Biology, University of Padova), Stefania Pirrotta (Department of Biology, University of Padova), Enrica Calura (Department of Biology, University of Padova) and Chiara Romualdi (Department of Biology, University of Padova). *Multi-omics integration: a regression based approach.*

Abstract. Large datasets containing different omics are increasingly available in public databases. However, capturing all the information contained in these data is a major challenge. Since the different omics are biologically related, it is fundamental to statistically detect their interplay with models that take into account all the omics and try to detect the key molecular players involved. To solve this need, we provide an easy-to-use R package for omics data integration.
Our package is designed to detect the association between the expression of a target and its regulators while taking into account their genomics modifications such as Copy Number Variations and methylation. In some cases, the number of regulators for a given target could be very high. To handle this eventuality, we provide a penalized model that will automatically keep only the most important regulators. We are also evaluating the possibility of adding more models for integration and expanding it to single-cell data. The package also provides functions for visualizing results to make model interpretation straightforward.
Our package provides a solid and easy-to-use way to solve the problem of multi-omics integration while allowing the detection and visualization of their interplay.

Keywords: multi-omics, integration, regression

Xiao Luo (Bielefeld University), Xiongbin Kang (Bielefeld University) and Alexander Schoenhuth (Bielefeld University). *VeChat: Correcting errors in long reads using variation graphs.*

Abstract. Error correction is the canonical first step in long-read sequencing data analysis. Current self-correction methods, however, are affected by consensus sequence induced biases that mask true variants in haplotypes of lower frequency showing in mixed samples. Unlike consensus sequence templates, graph based reference systems are not affected by such biases, so do not mistakenly mask true variants as errors. We present VeChat, as an approach to implement this idea: VeChat is based on variation graphs, as a popular type of data structure for pangenome reference systems. Extensive benchmarking experiments demonstrate that long reads corrected by VeChat contain 4 to 15 (Pacific Biosciences) and 1 to 10 times (Oxford Nanopore Technologies) less errors than when being corrected by state of the art approaches. Further, using VeChat prior to long-read assembly significantly improves the haplotype awareness of the assemblies. VeChat is an easy-to-use open-source tool and publicly available at https://github.com/HaploKit/vechat.

Keywords: Error correction, Third-generation sequencing, Genome Assembly, Variation graphs, Computational pan-genomics, Mixed samples, Metagenomes

Alessandro Brandulas Cammarata (University of Lausanne), Frederic B. Bastian (SIB Swiss Institute of Bioinformatics), Marc Robinson-Rechavi (University of Lausanne), Sara S. Fonseca Costa (University of Lausanne), Marta Rosikiewicz (SOPHiA GENETICS), Julien Roux (University of Lausanne) and Julien Wollbrett (University of Lausanne). *Using curated intergenic regions and a weighted merging function to identify active transcription: a novel approach for improving gene expression analysis.*

Abstract. While the primary focus of many RNA-seq 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. However, the usage of a fixed abundance threshold often leads to either a loss of information or an increase in false positives. To overcome these limitations, we propose an updated approach for the Bgee database. We identify a set of genes that are confidently non-expressed in each library using reads mapped to curated intergenic regions; and from their distribution, we compute a p-value. Additionally, we introduce a weighted merging function to aggregate per-gene expression calls signal from multiple libraries into a single presence/absence call. Its accuracy outperforms other existing methods in determining the true state of genes, when compared to reference sets in three distinct species. This approach can be applied to bulk as well as single-cell RNA-Seq. We also show that our method yields considerably fewer false positives when classifying genes carried by sex chromosomes. Overall, this novel approach has the potential to enhance the accuracy of genes accessible to differential expression analysis across conditions.

Keywords: bulk RNA-Seq, single-cell RNA-Seq, gene expression, transcriptional noise, RNA-seq pre-processing

Kevin Berg (Institute for Virology and Immunobiology, University of Würzburg), Manivel Lodha (Institute for Virology and Immunobiology, University of Würzburg), Yilliam Cruz Garcia (Cancer Systems Biology Group, Theodor Boveri Institute, University of Würzburg), Thomas Hennig (Institute for Virology and Immunobiology, University of Würzburg), Elmar Wolf (Cancer Systems Biology Group, Theodor Boveri Institute, University of Würzburg), Bhupesh Prusty (Institute for Virology and Immunobiology, University of Würzburg) and Florian Erhard (Chair of Computational Immunology, University of Regensburg). *Correcting 4sU induced quantification bias in nucleotide conversion RNA-seq data.*

Abstract. Nucleoside analogues like 4-thiouridine (4sU) are used to metabolically label newly synthesized RNA. Chemical conversion of 4sU before sequencing induces T-to-C mismatches in reads sequenced from labelled RNA, allowing to obtain total and labelled RNA expression profiles from a single sequencing library. Cytotoxicity due to extended periods of labeling or high 4sU concentrations has been described, but the effects of extensive 4sU labeling on expression estimates from nucleotide conversion RNA-seq have not been studied. Here, we performed nucleotide conversion RNA-seq with escalating doses of 4sU with short-term labeling (1h) and over a progressive time course (up to 2h) in different cell lines. With high concentrations or at later time points, expression estimates were biased in an RNA half-life dependent manner. We show that bias arose by a combination of reduced mappability of reads carrying multiple conversions, and a global, unspecific underrepresentation of labelled RNA due to impaired reverse transcription efficiency and potentially global reduction of RNA synthesis. We developed a computational tool to rescue unmappable reads, which performed favourably compared to previous read mappers, and a statistical method, which could fully remove remaining bias.

Keywords: SLAM-seq, Nucleotide conversion RNA-seq, 4sU labelling

Marco Oliva (Department of Computer and Information Science and Engineering, University of Florida), Travis Gagie (Diego Portales University) and Christina Boucher (University of Florida).
Building a Pangenome Alignment Index via Recursive Prefix-Free Parsing.

Abstract. Pangenomics alignment has emerged as an opportunity to reduce bias in biomedical research. Traditionally, short read aligners---such as Bowtie and BWA---were used to index a single reference genome, which was then used to find approximate alignments of reads to that genome. Unfortunately, these methods can only index a small number of genomes. Moni, an emerging pangenomic aligner, uses a preprocessing technique called prefix-free parsing to build a dictionary and parse from the input---these, in turn, are used to build the main run-length encoded BWT, and suffix array of the input. This is accomplished in linear space in the size of the dictionary and parse. Therein lies the open problem that we tackle in this paper. Although the dictionary scales sub-linearly with the size of the input, the parse becomes orders of magnitude larger than the dictionary. To scale the construction of Moni, we need to remove the parse from the construction of the RLBWT and suffix array. We solve this problem, and demonstrate that this improves the construction time and memory requirement allowing us to build the RLBWT and suffix array for 1000 diploid human haplotypes from the 1000 genomes project using less than 600GB of memory.

Keywords: Pangenomics, Alignment, r-index

Elena Weiß (Ludwig Maximilian University of Munich) and Caroline C. Friedel (Ludwig Maximilian University of Munich). *RegCFinder: de novo discovery of genomic regions with differential read density.*

Abstract. To date, no methods are available for the targeted identification of genomic regions with differences in sequencing read distributions between two conditions. Existing approaches either cannot be targeted to genomic regions of interest, only determine changes in total read numbers, require a predefined subdivision of input regions, or average across multiple input regions. Here, we present RegCFinder, which automatically identifies subregions of input windows with differences in read density between two conditions. For this purpose, the problem is defined as an instance of the well-established all maximum scoring subsequences problem, which can be solved in linear time. Subsequently, statistical significance and relative use of these regions within the input windows are determined with DEXSeq. RegCFinder allows flexible definition of input windows, making it possible to target the analysis to any regions of interests, e.g. promoter regions, gene bodies, peak regions, and more. Furthermore, any type of sequencing data can be used as input, thus, RegCFinder lends itself to a wide range of applications and biological questions. We illustrate the usefulness of RegCFinder on two applications. In both cases, we can both confirm previous observations regarding changes in read distributions, but also identify interesting novel subgroups of genes with distinctive changes.

Keywords: sequencing data, differential analysis, ChIP-seq, PRO-seq

Madeleine Duran (University of Washington), Brent Ewing (University of Washington), Eliza Barkan (University of Washington), David Kimelman (University of Washington), Jennifer Franks (University of Washington), Amy Tresenreider (University of Washington) and Cole Trapnell (University of Washington). *Hooke: A tool for differential analysis of cellular composition in single-cell perturbation experiments.*

Abstract. Advancements in multiplexing techniques have enabled the application of single-cell genomic methods to comprehensively study the effects of high-throughput perturbation experiments at whole-embryo scale. Such analyses aim to pinpoint key genes, cell types, and signaling pathways that control cell fate decisions during development. However, there is a lack of statistically principled tools for measuring how cell types shift after perturbations (genetic, chemical, or environmental) and identifying which genes regulate those transitions. Hooke is a new software package that uses Poisson-Lognormal models to perform differential analysis of cell abundances for perturbation experiments read out by single-cell RNA-seq. This versatile framework allows users to both 1) perform multivariate statistical regression to describe how perturbations alter the relative abundances of each cell state and 2) describe how all pairs of states co-vary as a parsimonious network of partial correlations. To demonstrate Hooke's utility, we analyzed a single-cell atlas of zebrafish organogenesis that includes wild-type and genetic perturbations at whole-embryo scale across multiple time points. With this method, we identified novel genetic requirements for relatively rare cell types in the embryonic kidney. Hooke will be available as an open-source R package and will enable users to dissect genetic dependencies in single-cell perturbation experiments.

Keywords: Single Cell, Developmental Biology, Gene Regulation

Dmitrii Meleshko (Weill Cornell Medical College), Rui Yang (Memorial Sloan Kettering Cancer Center), David Danko (Cornell University), Salil Maharjan (Weill Cornell Medicine), Anton Korobeynikov (Saint Petersburg State University) and Iman Hajirasouliha (Cornell University).
Blackbird: structural variant detection using synthetic and low-coverage long-reads.

Abstract. Recent benchmarks of structural variant (SV) detection tools revealed that the most medium-range (50-10,000 bp) SVs cannot be resolved with short-read sequencing, but long-read SV callers achieve great results. However, long-read sequencing has a higher cost and requires higher input DNA. Lowering sequence coverage reduces cost, but long-read SV callers perform poorly with coverage below 10$\times$. Synthetic long-read (SLR) technologies have great potential for SV detection, though their long-range information has been hard to utilize for events shorter than 50 kbp.

We propose a novel integrated alignment- and local-assembly-based algorithm, Blackbird, that uses SLR together with low-coverage long reads to improve the detection of challenging medium-size events. Without the need for a whole genome assembly, Blackbird uses barcode information encoded in SLR to accurately assemble small segments and use long reads for an improved assembly.

We evaluated Blackbird on simulated and real human genome datasets. Using the HG002 GIAB callset, we demonstrated that in hybrid mode, Blackbird demonstrated results comparable to state-of-the-art long-read tools using significantly lower long-read coverage. Blackbird requires only 5$\times$ to achieve F1 scores similar to PBSV and Sniffles2 using 10$\times$ long-read coverage. Additionally, Blackbird in the SLR-only is more sensitive than popular short-read methods.

Keywords: Synthetic long reads, SV calling, long reads, linked-reads, local assembly, variant discovery

Vinzenz May (Core unit bioinformatics, Berlin Institute of Health @ Charité Berlin), Dieter Beule (Core unit bioinformatics, Berlin Institute of Health @ Charité Berlin) and Manuel Holtgrewe (Core unit bioinformatics, Berlin Institute of Health @ Charité Berlin). *Structural Variant Calling from Long Read-based Local Assemblies.*

Abstract. Correct structural variant (SV) detection is integral to rare disease diagnostics and cancer genomics, as well as genome variation in populations.
Existing methods of structural variant detection with long reads usually use read-to-reference alignments to detect SVs on abstractions of SV signals. Typical problems are the detection of very long or complex SVs as well as SVs within difficult regions, such as repeats or low complexity regions.
We propose a novel method, in which we assemble parts of reads that are aligned to SV candidate regions to resolve such difficult loci or re-construct complex SVs. The great advantage we try to exploit is that the depth of coverage can be significantly lower than it would be necessary for de novo assembly, while difficult regions can still be resolved better than with the initial alignments. This has become a feasible solution since the per base error rates of available long read technologies have decreased dramatically in the past four years. The challenges are to identify the regions correctly and to determine the correct number of haplotypes.

Keywords: Long reads, structural variants, structural variant calling, structural variant detection, 3rd generation sequencing, local sequence assembly

Sébastien Gradit (Institut Pasteur) and Axel Cournac (Institut Pasteur). *Statistical inference of repeated sequence contacts in Hi-C maps (Hi-C BERG).*

Abstract. Increasingly detailed investigations of the spatial organization of genomes reveal that chromosome folding influences or regulates dynamic processes such as transcription, DNA repair and segregation. Hi-C approach is commonly used to characterize genome architecture by quantifying physical contacts' frequency between pairs of loci through high-throughput sequencing. These sequences cause challenges during the analysis' alignment step, due to the multiplicity of plausible positions to assign sequencing reads. These unknown parts of the genome architecture, that may contain biological information, remains hidden throughout downstream functional analysis. To overcome these limitations, we have developed HiC-BERG, a method combining statistical inference with input from DNA polymer behavior characteristics and features of the Hi-C protocol to assign with robust confidence repeated reads in a genome and "fill-in" empty vectors in contact maps. HiC-BERG is intended to be applicable to different types of organisms. We will present the program and key validation tests, before applying it to unveil hidden parts of the genomes of E.coli, S.cerevisiae and P.falciparum. HiC-BERG shows that repeated sequences may be involved in singular genomic architectures. Our method can provide an alternative visualization of genomic contacts under a wide variety of biological conditions allowing a more complete view of genome plasticity.

Keywords: Hi-C, Contact map, Genome architecture, Genome conformation, Genomics, Genome contacts, Statistical inference

Johannes Ostner (Helmholtz Munich, Munich, Germany), Christian L. Müller (Helmholtz Munich, Munich, Germany) and Hongzhe Li (University of Pennsylvania, Philadelphia, PA). *Flexible and efficient modeling of compositional count data from high-throughput sequencing.*

Abstract. Population count data derived from high-throughput sequencing experiments provides valuable information about the composition of biological environments such as cell populations or microbial communities. However, accurate statistical modeling of such data is challenging due to its high dimensionality, excessive number of zeros, correlation of features, and compositional constraints. The recent class of compositional power interaction models (PIMs; Yu et al., 2021) accommodates these properties and can be optimized efficiently through score matching methods. In our work, we use PIMs to model covariate influence on the data composition and subsequently perform differential testing. We extend the score matching estimator to include latent effect variables and derive an extended parameter optimization scheme that selects the ideal power transform for accurate data representation without needing zero replacement strategies. We demonstrate that PIMs are better suited than other popular distributional methods to describe real and simulated high-throughput sequencing data with correlated features while requiring very low computational resources.
We further evaluate the model's ability to discover significant and pairwise interactions between features and showcase its flexibility through applications to blood cell compositions obtained through single-cell RNA sequencing, as well as amplicon sequencing data of the human gut microbiome.

Keywords: compositional data, high-throughput sequencing, differential abundance testing, generative model, scRNA-seq, microbiome, score matching, optimization

Mara Stadler (Helmholtz Munich) and Christian L. Müller (Helmholtz Munich). *A Workflow for Stable Interaction Detection in High-Throughput Biological Data.*

Abstract. The advent of large-scale data (e.g., from biotechnology) has made the development of suitable statistical techniques a cornerstone of modern interdisciplinary research. These data often contain many features but limited sample size, and are accompanied by experimental noise. A common research question in data-driven observational studies is to determine how features impact a readout of interest. Typically, only a subset of features is relevant, and they may interact in a concerted fashion. Thus, a major concern is to identify these relevant effects from a large number of possible combinations of features. To address this, we propose a robust statistical workflow to recover interactions in the data-scarce regime. Our multi-stage approach uses a lasso model for hierarchical interactions combined with stability-based model selection in a replicate consistent workflow. We demonstrate its superior performance compared to state-of-the-art techniques using synthetic data and show its wide applicability in a number of different biological applications including histone modification-protein interactions and combinatorial drug effects on cell morphological features.

Keywords: Robust interaction detection, High-throughput biological data, high-dimensional statistics, lasso, stability selection, histone modifications, drug interactions

Laura Martens (Technical University Munich), David Fischer (Broad Institute), Vicente Yépez (Technical University Munich), Fabian Theis (Helmholtz Center Munich) and Julien Gagneur (Technical University Munich). *Modeling fragment counts improves single-cell ATAC-seq analysis.*

Abstract. Single-cell ATAC-sequencing (scATAC-seq) is a powerful technique for studying chromatin regulation at the single-cell level. Typically, scATAC-seq data is binarized to indicate open chromatin regions, but the implications of this binarization are not well-understood. In this study, we demonstrate that a quantitative treatment of scATAC-seq data improves the goodness-of-fit of existing models and their applications, including clustering, cell type identification, and batch integration. Our contribution is twofold. First, we show that fragment counts, but not read counts, can be modeled using standard count distribution. Second, we compare the effects of binarization versus a count-based model (PoissonVAE) on scATAC-seq data using publicly available datasets, and highlight the biological effects that are missed by a binary treatment. We show that high count peaks in scATAC-seq data correspond to important regulatory regions such as super enhancers and highly transcribed promoters, similar to observations in bulk ATAC-seq data. Furthermore, we demonstrate that fragment counts in promoter regions correlate with gene expression, emphasizing a quantitative signal in promoter accessibility. Our results have significant implications for scATAC-seq analysis, suggesting that handling the data quantitatively can improve the accuracy of machine learning models used for investigating single-cell regulation.

Keywords: single-cell genomics, chromatin regulation, single-cell ATAC-seq analysis, quantitative treatment

Lea Vandamme (CNRS, University of Lille), Bastien Cazaux (University of Lille) and Antoine Limasset (CNRS, University of Lille). *Kmer2Reads an associative index for Third Generation Sequencing data.*

Abstract. Studying biological sequences typically involves using a reference genome, but obtaining accurate assemblies from sequencing data can be challenging due to genomic repeats, errors, and biases.
Hence, working directly with raw data output by sequencers, without pre-processing, can be preferable. Our objective is to develop multifaceted indexes able to identify reads containing a specific k-mer in a given dataset. Popular indexes, dubbed colored de Bruijn graphs associate the kmer origin among thousand of datasets. However they are not able to index each reads separately. To address this challenge, we present K2R, which leverages redundancy in the data to limit memory usage. Specifically, we use super-k-mers to reduce the number of entries in our structures and employ the concept of color to minimize memory impact of repetitive k-mer data. We present the main results obtained by comparing K2R with state-of-the-art methods such as hashing methods (e.g., read connector) and full-text indexing (e.g., r-index), in terms of memory impact, throughput, and time consumption for creation and query. We compare the performance of the tools encompassing varying coverage levels and error rates, to evaluate their advantages, disadvantages and respective comfort zone.

Keywords: Indexing, Data structure, Third generation sequencing

Ajita Shree (Indian Institute of Technology Kanpur), Musale Krushna Pavan (Indian Institute of Technology Kanpur) and Hamim Zafar (Indian Institute of Technology Kanpur). *Supervised integration of single-cell datasets using hierarchical deep-generative model paired with cell-type classifier.*

Abstract. Integration of heterogeneous datasets generated through advanced single-cell sequencing techniques enables improved understanding of the cellular states and expression programs underlying complex biological systems. Hence, integration methods focus on the removal of complex-nested batch effects arising due to the heterogeneity in samples generated across tissue locations, time and conditions. However, it is equally important to conserve the biology while improving batch-correction. The goal is to leverage any available cell type annotation for cells for an improved integration. Here we present a supervised data integration framework called scDREAMER-Sup that employs a novel adversarial hierarchical variational autoencoder with two neural network classifiers for improved bio-conservation and batch-correction respectively. We evaluated the performance of scDREAMER-Sup on 5 challenging real datasets consisting of ~1 million cells and multiple cell sub-types. We further introduced semi-supervised setting to address the challenge of missing cell type annotations in integration tasks and experimented with {10%, 20%, 50%} of missing cell type annotations categories. We compared scDREAMER-Sup's performance against that of scANVI and scGEN, top performing state-of-the-art methods that utilize cell type labels under both supervised as well as semi-supervised settings and demonstrated that scDREAMER-Sup significantly outperformed other methods with an overall improvement of 36%-48% in combined composite score.

Keywords: data-integration, adversarial training, variational auto-encoder, single-cell sequencing data

Jonathan Ogata (Virginia Commonwealth University), Wancen Mu (University of North Carolina-Chapel Hill), Eric Davis (University of North Carolina-Chapel Hill), Bingjie Xue (University of Virginia), Chuck Harrell (Virginia Commonwealth University), Nathan Sheffield (University of Virginia), Douglas Phanstiel (University of North Carolina at Chapel Hill), Michael Love (University of North Carolina at Chapel Hill) and Mikhail Dozmorov (Virginia Commonwealth University).
excluderanges: exclusion sets for T2T-CHM13, GRCm39, and other genome assemblies.

Abstract. Exclusion regions are sections of reference genomes with abnormal pileups of short sequencing reads. Removing reads overlapping them improves biological signal, and these benefits are most pronounced in differential analysis settings. Several labs created exclusion region sets, available primarily through ENCODE and Github. However, the variety of exclusion sets creates uncertainty which sets to use. Furthermore, gap regions (e.g., centromeres, telomeres, short arms) create additional considerations in generating exclusion sets. We generated exclusion sets for the latest human T2T-CHM13 and mouse GRCm39 genomes and systematically assembled and annotated these and other sets in the 'excluderanges' R/Bioconductor data package, also accessible via the BEDbase.org API. The package provides unified access to systematically annotated 82 GenomicRanges objects covering six organisms, multiple genome assemblies and types of exclusion regions. For human hg38 genome assembly, we recommend 'hg38.Kundaje.GRCh38_unified_blacklist' as the most well-curated and annotated, and sets generated by the Blacklist tool for other organisms. Package website: https://bioconductor.org/packages/excluderanges/, https://dozmorovlab.github.io/excluderanges/

Keywords: exclusion regions, blacklist, gap regions, genome assembly, T2T-CHM13

Lianbo Yu (The Ohio State University), Yue Zhao (University of Michigan) and Lang Li (The Ohio State University). *CEDA: integrating gene expression data with CRISPR-pooled screen data identifies essential genes with higher expression.*

Abstract. Clustered regularly interspaced short palindromic repeats (CRISPR)-based genetic perturbation screen is a powerful tool to probe gene function. However, experimental noises, especially for the lowly expressed genes, need to be accounted for to maintain proper control of false positive rate. We develop a statistical method, named CRISPR screen with Expression Data Analysis (CEDA), to integrate gene expression profiles and CRISPR screen data for identifying essential genes. CEDA stratifies genes based on expression level and adopts a three-component mixture model for the log-fold change of single-guide RNAs (sgRNAs). Empirical Bayesian prior and expectation–maximization algorithm are used for parameter estimation and false discovery rate inference. Taking advantage of gene expression data, CEDA identifies essential genes with higher expression. Compared to existing methods, CEDA shows comparable reliability but higher sensitivity in detecting essential genes with moderate sgRNA fold change. Therefore, using the same CRISPR data, CEDA generates an additional hit gene list.

Keywords: CRISPR Screening, Empirical Bayes, Normal Mixture Model, EM Algorithm

David Gómez-Sánchez (Spanish National Cancer Research Centre CNIO, Computational Oncology, Madrid, Spain.), Bárbara Hernando (Spanish National Cancer Research Centre CNIO, Computational Oncology Group, Madrid, Spain.), Joe Sneath Thompson (Spanish National Cancer Research Centre CNIO, Computational Oncology Group, Madrid, Spain.), Diego García-López (Tailor Bio, Cambridge UK), Alice Cádiz (Spanish National Cancer Research Centre CNIO, Computational Oncology Group, Madrid, Spain.), Luis Paz-Ares (H12O-CNIO Lung Cancer Clinical Research Unit, Madrid, Spain.) and Geoff Macintyre (Spanish National Cancer Research Centre CNIO, Computational Oncology Group, Madrid, Spain.). *Genome-wide copy number profiling using DNA sequencing targeted panels.*

Abstract. Next Generation Sequencing (NGS) panels are routinely used for small variant detection in cancer clinical care. The capture efficiency of these panels is not optimal, resulting in off-target DNA being sequenced alongside on-target DNA. These off-target reads are distributed across the genome, therefore allowing for a whole genome to be derived, although we see significant biases in the number of reads in these regions. Here we present a method to correct these biases and generate robust genome-wide copy number profiles.
A variety of capture based NGS protocols were analysed using our novel computational method and were compared with the current gold standard: Shallow Whole Genome Sequencing (sWGS), demonstrating that comparable results could be obtained from either sWGS or NGS targeted samples depending on sample purity, preservation method, and read depth.
We then benchmarked our algorithm with other NGS-based copy number extraction methods, showing an improvement in performance without the need of a matched normal tissue or filtering out noisy data, usual drawbacks for other methods in the cancer field. As most clinical sequencing workflows rely on targeted capture gene panels, our method has the potential for new biomarker discovery through the addition of robust copy number profiles to these assays.

Keywords: copy-number, NGS, capture based sequencing, sequencing panels, CIN, cancer, copy number, chromosomal instability, tumor, method

Timothé Rouzé (CNRS, Univ Lille), Camille Marchet (CNRS, Univ Lille) and Antoine Limasset (CNRS, Univ Lille). *Scalable analysis of sequencing data analysis using novel Fractional Hitting Set method.*

Abstract. A challenge for Bioinformatics is to keep up with the amount of data generated by high throughput sequencing.
Being able to compare such volume of data remains a scalability challenge which is the focus of many methodological papers.
To achieve drastic memory cost reduction, a possibility is to transform documents into "sketches" of highly reduced sizes that can be quickly compared to compute the documents similarity with bounded error.
The most used tools rely on fixed sized sketches using techniques such as Minhash or HyperLogLog. However, those techniques have a relatively poor accuracy when the compared datasets are very dissimilar in size or content.
To cope with this problem, novel methods proposed to construct adaptive sketches, scaling linearly with the size of the input, by selecting a fraction of the documents' k-mers.
Several techniques were proposed to perform uniform sub-sampling with theoretical guarantees such as modimizer/modminhash, scaled minhash/FracMinHash.
With SuperSampler, we improve such schemes by combining them with the concept of super-k-mers thus drastically reducing resources usage (CPU, memory, disk).
In this poster, we show that SuperSampler can use an order of magnitude less resources than state of the art with equivalent results.

Keywords: Local sensitive hashing, sketching, large scale, genome comparison

Xinyi Lin (The University of Hong Kong), Chuen Chau (The University of Hong Kong), Kun Ma (The University of Hong Kong), Yuanhua Huang (The University of Hong Kong) and Joshua Ho (The University of Hong Kong). *DCATS: differential composition analysis for flexible single-cell experimental designs.*

Abstract. Differential composition analysis – the identification of cell types that have statistically significantly change in abundance between multiple experimental conditions – is one of the most common tasks in single cell omic data analysis. However, it remains challenging to perform differential composition analysis in the presence of flexible experimental designs and uncertainty in cell type assignment. Here, we introduce a statistical model and an open source R package, DCATS, for differential composition analysis based on a beta-binomial regression framework that addresses these challenges. Our empirical evaluation shows that DCATS consistently maintains high sensitivity and specificity compared to state-of-the-art methods.

Keywords: differential composition analysis, single-cell RNA sequencing, beta-binomial generalized linear model

Dmitry S. Shcherbo (Cancer Research UK Cambridge Institute, University of Cambridge and Cancer Research UK Cambridge Centre, Cambridge), Matthew D. Eldridge (Cancer Research UK Cambridge Institute, University of Cambridge and Cancer Research UK Cambridge Centre, Cambridge), Maria Neofytou (Cancer Research UK Cambridge Institute, University of Cambridge and Cancer Research UK Cambridge Centre, Cambridge), Wendy N. Cooper (Cancer Research UK Cambridge Institute, University of Cambridge and Cancer Research UK Cambridge Centre, Cambridge), James D. Brenton (Cancer Research UK Cambridge Institute, University of Cambridge and Cancer Research UK Cambridge Centre, Cambridge), Hui Zhao (Cancer Research UK Cambridge Institute, University of Cambridge and Cancer Research UK Cambridge Centre, Cambridge) and Nitzan Rosenfeld (Cancer Research UK Cambridge Institute, University of Cambridge and Cancer Research UK Cambridge Centre, Cambridge). *Evaluation of panels of normal samples for somatic copy number alteration detection tools.*

Abstract. Short DNA molecules found in blood plasma originate mostly from dying cells including tumour cells in cancer patients. Information on the genomes can be extracted from cell-free DNA (cfDNA) in several ways including presence of point mutations, cytosine methylation and others. Chromosomal rearrangements and somatic copy number alterations (SCNAs) are key events in many cancer types and can help in the diagnosis of cancer and often associated with treatment response. Bioinformatic tools for SCNA detection from shallow whole genome sequencing (sWGS) data are widely used in cancer research. Some of the algorithms are specifically tailored to detect SCNAs in cfDNA rather than in genomic DNA from tumour tissues and most of them require baseline estimation from a panel of normal samples (PoN) with a priori absence of SCNAs. PoNs are often distributed within software packages, however they are generated on datasets that are not always based on cfDNA. Given specific properties of cfDNA (short fragment length, biased coverage profiles, etc) we investigate the influence of PoNs generated with various parameters and on different data types. We use samples from 1000 Genomes project and cfDNA WGS data to assess performance of CNAclinic and ichorCNA on sWGS of cfDNA of cancer patients.

Keywords: cancer, somatic copy number alteration detection, cell-free DNA

Nuria Sánchez de la Blanca Carrero (Germans Trias i Pujol Research Institut), Pablo Sacristán Gómez (Instituto de Investigación Sanitaria del Hospital Universitario de La Princesa (IIS-HUP)), Ana Serrano Somavilla (Instituto de Investigación Sanitaria del Hospital Universitario de La Princesa (IIS-HUP)), Santiago Guerra Cantera (Instituto de Investigación Sanitaria del Hospital Universitario de La Princesa (IIS-HUP)), Raúl Fernández Contreras (Instituto de Investigación Sanitaria - Hospital Universitario de la Princesa (IIS-HUP)), Cristina Sánchez Guerrero (Universidad Complutense de Madrid), Miguel Antonio Sampedro Núñez (Instituto de Investigación Sanitaria - Hospital Universitario de la Princesa (IIS-HUP)), José Luis Muñoz de Nova (Instituto de Investigación Sanitaria - Hospital Universitario de la Princesa (IIS-HUP)), Mónica Marazuela Azpiroz (Instituto de Investigación Sanitaria - Hospital Universitario de la Princesa (IIS-HUP)) and Rebeca Martínez Hernández (Instituto de Investigación Sanitaria - Hospital Universitario de la Princesa (IIS-HUP)). *Digging into the molecular differences among conditions by the integration of spatial transcriptomics samples: Beyond gene plotting.*

Abstract. CONTEXT: RNA-seq allows us to uncover general molecular differences. However, it has not enough depth to identify more subtle disparities between different cellular subpopulations. Spatial transcriptomics(ST) is a novel method whose strong point is the gene spatial location, but, can we compare different conditions to get disease-associated signatures and susceptibility pathways within cell populations using ST? We afford it in autoimmune thyroid diseases/AITD (Hashimoto's thyroiditis/HT and Graves' disease/GD) vs controls.
METHODOLOGY: 3 HT, 3 GD and 2 controls were sequenced using Visium ST (10XGenomics). We compared the isolation of cell populations by pathology-based and unsupervised clustering. Then, we proceeded with a sample integration using harmony followed by a re-clustering strategy and pseudobulks differential expression analysis of: thyrocytes, connective tissue(CT) and vessels separately. Finally, validation using public single cell(SC) repositories and immunostaining.
RESULTS: We obtained a significant correlation between pathology-based and unsupervised clustering. We revealed damaged epithelial cells in AITD close to infiltration; molecular signatures which correlate to fibroblast subpopulations in CT from HT and GD samples and differential angiogenic processes guessed from their vessels. SC, immunostaining and literature validated ST results.
CONCLUSIONS: ST is also useful to integrate samples and to infer molecular distinctions in a cellular context among conditions.

Keywords: Spatial transcriptomics, Data integration, Enrichment analysis, Clustering, Translational bioinformatics, Autoimmune thyroid diseases

Solène Brohard (Centre National de Recherche en Génomique Humaine (CNRGH), Institut de Biologie François Jacob, CEA), Florence Glibert (Centre National de Recherche en Génomique Humaine (CNRGH), Institut de Biologie François Jacob, CEA), Cédric Fund (Centre National de Recherche en Génomique Humaine (CNRGH), Institut de Biologie François Jacob, CEA), Olivier Alibert (Centre National de Recherche en Génomique Humaine (CNRGH), Institut de Biologie François Jacob, CEA), Jean-François Deleuze (Centre National de Recherche en Génomique Humaine (CNRGH), Institut de Biologie François Jacob, CEA), Eric Bonnet (Centre National de Recherche en Génomique Humaine (CNRGH), Institut de Biologie François Jacob, CEA) and Sophie Chantalat (Centre National de Recherche en Génomique Humaine (CNRGH), Institut de Biologie François Jacob, CEA). *High-order chromatin contact in regulatory regions with the Pore-C approach.*

Abstract. In humans, gene transcription is under the control of the promoter, which is located immediately upstream of the genes, and the enhancers, which are distal regulatory regions. Recent studies suggest the existence of hubs of interactions involving multiple enhancers and promoters. In order to define the structure of the interacting hubs of chromatin regions at high resolution, we have implemented the Pore-C approach based on long-read multiway contact nanopore sequencing. Deshpande et al. (Nature Biotechnology, 2022), proposed dedicated bioinformatics tools to identify multi-way contacts and reveal significant cooperativities. We used the tools and the results of this study to validate our Pore-C experiments obtained on reference cell lines. We first mapped fragments located in multiway contact reads. Then, we generated virtual pairwise contacts and verified that we reproduced 3D genomic features classically observed in HIC maps. We also showed that the range of inter- and intrachromosomal interactions was very similar. We finally identify significant intra-chromosomal high-order contacts in regulatory regions using the Chromunity tool, described in the paper. We will now generate new Pore-C datasets on different cell types and conditions to explore the role of several proteins involved in chromatin remodeling in the formation of these hubs.

Keywords: Pore-C, Multi-way contacts, Interaction hubs

Bailey Francis (European Molecular Biology Laboratory - European Bioinformatics Institute), Mohab Helmy (European Molecular Biology Laboratory - European Bioinformatics Institute), Jingtao Lilue (Gulbenkian Institute of Science), Laura Reinholdt (The Jackson Laboratory), Anne Czechanski (The Jackson Laboratory), Emma Betteridge (Wellcome Sanger Institute), Iraad Bronner (Wellcome Sanger Institute), David Adams (Wellcome Sanger Institute) and Thomas Keane (European Molecular Biology Laboratory - European Bioinformatics Institute). *A Telomere-to-Telomere (T2T) complete mouse genome.*

Abstract. The mouse reference genome, based on the laboratory strain C57BL/6J, has served as a foundation for improving our understanding of human health and genetics for over twenty years. However, the current mouse reference genome (GRCm39) contains over 170 known gaps and issues, and is missing key features such as telomere and centromere sequences. Here, we combine novel high-molecular-weight DNA extraction methodologies and ultra-long sequencing technologies on mESCs from a C57BL/6J x CAST/EiJ F1 animal to generate two of the most complete reference-quality mouse genome sequences to date. These new T2T mouse assemblies add significant amounts of novel sequence when compared to their respective current reference genomes (over 150Mbp and 250Mbp for C57BL/6J and CAST/EiJ respectively, of which 100Mbp and 150Mbp constitutes newly placed telocentric sequences). Our C57BL/6J assembly closes over 95% of the previously unresolved autosomal gaps in GRCm39 with over 12Mbp of novel sequences. Additionally, we have shown that our new T2T assemblies significantly improve the representation of previously hard-to-assemble regions when compared to the current reference genomes (e.g. PAR, KZFPs). As a result, these assemblies represent a major milestone in the journey towards a fully complete mouse reference genome.

Keywords: genome assembly, telomere-to-telomere, DNA sequencing, oxford nanopore, pacbio hifi, comparative genomics, mouse genomics, reference genome

Alister D'Costa (Ontario Institute for Cancer Research; University of Toronto, Dept of Computer Science), Philip Zuzarte (Ontario Institute for Cancer Research), Michael Molnar (Ontario Institute for Cancer Research), Tracy Murphy (University Health Network), Mark Minden (University Health Network), Yun William Yu (University of Toronto, Dept of Mathematics) and Jared Simpson (Ontario Institute for Cancer Research; University of Toronto, Dept of Computer Science, Dept of Molecular Genetics). *Detecting Chromosomal Translocations using Augmented Genome Sequence Graphs.*

Abstract. Chromosomal translocations have the potential to generate fusion proteins and disrupt gene expression. Current methods to identify translocations from long DNA sequencing reads typically rely on alignments to a linear reference, with a read requiring a high-quality alignment to two different genomic positions. While effective, reference-based translocation detection may generate false positive calls tied to polymorphic insertions. In this work we show that augmented genome sequence graphs, that contain known variation not found in the linear reference genome, can be used to effectively detect chromosomal translocations with far fewer false positive calls and in less time than existing state of the art methods. Demonstrating our method on a set of leukemia samples with known translocations, we show large decrease in the number of false positive translocation calls using a graph-based translocation detection approach.

Keywords: Structural Variations, Sequence Graphs, Translocation Detection, Reference Free Approaches

Pablo Garcia-Nieto (Chan Zuckerberg Initiative), Signe Chambers (Chan Zuckerberg Initiative), Harley Thomas (Chan Zuckerberg Initiative), Maximilian Lombardo (Chan Zuckerberg Initiative), Brian Aevermann (Chan Zuckerberg Initiative), Karen Liang (Chan Zuckerberg Initiative), Seve Badajoz (Chan Zuckerberg Initiative), Colin Megill (Chan Zuckerberg Initiative), Trent Smith (Chan Zuckerberg Initiative), Daniel Hegeman (Chan Zuckerberg Initiative), Arathi Mani (Chan Zuckerberg Initiative), Timmy Huang (Chan Zuckerberg Initiative), Kuni Katsuya (Chan Zuckerberg Initiative), Emanuele Bezzi (Chan Zuckerberg Initiative), Bruce Martin (Chan Zuckerberg Initiative), Andrew Tolopko (Chan Zuckerberg Initiative), Alexander Tarashansky (Chan Zuckerberg Initiative), Nayib Gloria (Chan Zuckerberg Initiative), Fran McDae (Clever Canary), Dave Rogers (Clever Canary), Mim Hastie (Clever Canary), Amanda Infeld (Chan Zuckerberg Initiative), Meghan Urisko (Chan Zuckerberg Initiative), Norbert Tavares (Chan Zuckerberg Initiative), Garabet Yeretssian (Chan Zuckerberg Initiative), Bailey Marshall (Chan Zuckerberg Initiative), Ivana Jelic (Chan Zuckerberg Initiative), Jonah Cool (Chan Zuckerberg Initiative), Ambrose Carr (Chan Zuckerberg Initiative), Michael Cherry (Stanford), Jason Hilton (Stanford), Jennifer Zamanian (Stanford), Jennifer Chien (Stanford), Erica Rutherford (Stanford), Lian Morales (Stanford), Jim Chaffer (Stanford) and Dana Sadgat (Chan Zuckerberg Initiative). *The Census of CZ CELLxGENE Discover is an API for efficient and low-latency access to the largest standardized single-cell data repository.*

Abstract. CZ CELLxGENE Discover has released all of its human and mouse single-cell data through its Census (cellxgene-census.readthedocs.io) – a free-to-use service with an API and data that allows for querying its single-cell data corpus directly from Python or R. The API uses a new technology that allows for efficient and low-latency querying. The data are fully standardized and hosted publicly for free access, and they are composed by a count matrix of 50 mi cells (observations) by >60 k genes (features) accompanied by 11 cell metadata variables (e.g. cell type, tissue, sequencing technology, donor id, etc) and gene metadata that includes GENCODE-based IDs and gene names. While these data are built from more than 500 datasets, the APIs enable convenient cell- and gene-based filtering to obtain any slice of interest in a matter of seconds. All data can be quickly transformed to numpy, pandas, anndata, Seurat, or R base objects.

Keywords: Cell Census, Single Cell, Metadata, Gene-based, Free access

Jens-Uwe Ulrich (Hasso Plattner Institute) and Bernhard Renard (Hasso Plattner Institute). *Taxor: Fast and space-efficient taxonomic classification of long reads.*

Abstract. Correctly identifying all organisms in an environmental or clinical sample is fundamental in many metagenomic sequencing projects. Over the last years, many tools have been developed that classify short and long sequencing reads by comparing their nucleotide sequences to a predefined set of references. Although those methods already utilize flexible data structures with low memory requirements, the constantly increasing number of reference genomes in the databases poses a major computational challenge to the profilers regarding memory usage, index construction and query time. Here, we present Taxor as a fast and space-efficient tool for taxonomic profiling by utilizing hierarchical interleaved XOR filters. Taxor shows a precision of 99.9% for read classification on the species level while retaining a recall of 96.7%, outperforming tools like Kraken2 and Centrifuge in terms of precision by 3-9%. Our benchmarking based on simulated and real data indicates that Taxor accurately performs taxonomic read classification while reducing the index size of the reference database and memory requirements for querying by a factor of 2-12x when compared to other profiling tools.

Keywords: metagenomics, taxonomic profiling, long-read classification, species identification, XOR filter, k-mer, syncmer, nanopore sequencing

Navonil De Sarkar (Fred Hutchinson Cancer Center), Robert Patton (Fred Hutchinson Cancer Center), Peter Nelson (Fred Hutchinson Cancer Center) and Gavin Ha (Fred Hutchinson Cancer Center). *Nucleosome Patterns in Circulating Tumor DNA Reveal Transcriptional Regulation of Advanced Prostate Cancer Phenotypes.*

Abstract. Advanced prostate cancers comprise distinct phenotypes, but tumor classification remains clinically challenging. Here, we harnessed circulating tumor DNA (ctDNA) to study tumor phenotypes by ascertaining nucleosome positioning patterns associated with transcription regulation. We sequenced plasma ctDNA whole genomes from patient-derived xenografts representing a spectrum of androgen receptor active (ARPC) and neuroendocrine (NEPC) prostate cancers. Nucleosome patterns associated with transcriptional activity were reflected in ctDNA at regions of genes, promoters, histone modifications, transcription factor binding, and accessible chromatin. We identified the activity of key phenotype-defining transcriptional regulators from ctDNA, including AR, ASCL1, HOXB13, HNF4G, and GATA2. To distinguish NEPC and ARPC in patient plasma samples, we developed prediction models that achieved accuracies of 97% for dominant phenotypes and 87% for mixed clinical phenotypes. Although phenotype classification is typically assessed by IHC or transcriptome profiling from tumor biopsies, we demonstrate that ctDNA provides comparable results with diagnostic advantages for precision oncology.

Keywords: cfDNA, ctDNA, liquid biopsies, phenotype modeling, nucleosome positioning

Meghan Violette (University of Tampa), Michael Middlebrooks (University of Tampa) and Padmanabhan Mahadevan (University of Tampa). *Evaluation of various short read genome assemblers on sea slug genomic data.*

Abstract. Elysia crispata is a photosynthetic sea slug that consumes green algae and then sequesters chloroplasts from the algae in special cells lining the digestive tubules, in a process called kleptoplasty. This sea slug can photosynthesize using stolen chloroplasts for 3-4 months after feeding. We used Illumina short read sequencing to determine the genome of this sea slug and evaluated the performance of various short read genome assemblers on this sea slug genomic data. The genome assemblers evaluated were MEGAHIT, SPAdes, ABySS, MaSuRCA, Clover, Platanus allee, Mini SR, Wengan, IDBA, SOAPdenovo, Geneious, GATB-Minia and Discovar denovo. The top 2 genome assembly programs based on BUSCO completeness and total number of contigs were MaSuRCA and MEGAHIT. MaSuRCA reference assisted, MaSuRCA and MEGAHIT produced the assemblies with the highest percent of complete and partial core genes with the sea slug data. However, they are still highly fragmented given the total number of contigs in the final assemblies. The MaSuRCA assemblies took 15 hours, but produced the best assembly compared to MEGAHIT which took 3 hours. MaSuRCA produced the best assembly with reference assisted mode, suggesting that reference assisted assembly may be the best option when a suitable reference genome is available.

Keywords: sea slug, short read, assembler, genome

Satoshi Nomura (Division of Systems Biology, Nagoya University Graduate School of Medicine), Yasuhiro Kojima (Laboratory of Computational Life Science, National Cancer Center Research Institute), Kodai Minoura (Japanese Red Cross Aichi Medical Center Nagoya Daichi Hospital) and Teppei Shimamura (Division of Systems Biology, Nagoya University Graduate School of Medicine).
Deep generative modeling for multimodal velocity estimation from single-cell multiomics data.

Abstract. Single-cell multiomics provides an opportunity to comprehend the regulatory relationships across modalities, including transcriptome and regulome. However, this approach is experimentally limited to revealing static snapshots at the time of observation, which hinders our understanding of dynamic state changes orchestrated across modalities. Although RNA velocity addresses this issue by estimating temporal changes in transcriptome, inferring dynamics in other modalities remains challenging.
To overcome this limitation, we develop a deep generative model named mmVelo, which estimates cell-state dependent dynamics across multiple modalities. mmVelo learns the dynamics of cellular states based on spliced and unspliced mRNA counts and projects them onto other modalities, thereby inferring cross-modal dynamics, such as RNA-chromatin accessibility dynamics.
We applied mmVelo to single-cell multiomics data from developing mouse brain and validated the accuracy of the estimated chromatin accessibility dynamics. Furthermore, we discover that known lineage-determining transcription factors play a crucial role in regulating chromatin accessibility in mouse skin. Finally, we demonstrate in human brain development that by using multiomics data as a bridge, the dynamics of other modalities can be inferred from single-modal data through cross-modal generation.
Overall, mmVelo offers a unique advantage in understanding the dynamic interactions between modalities, providing insights into regulatory relationships across molecular layers.

Keywords: single-cell multiomics, deep generative model, multimodal prediction, RNA velocity, cell differentiation, gene regulation

Koichiro Majima (Division of Systems Biology, Nagoya University Graduate School of Medicine), Kodai Minoura (Japanese Red Cross Nagoya Daiichi Hospital), Yasuhiro Kojima (Laboratory of Computational Life Science, National Cancer Center Research Institute) and Teppei Shimamura (Division of Systems Biology, Nagoya University Graduate School of Medicine). *Variational Inference for Single-Cell Transcriptome with DNA Barcoding Reconstructs Unobserved Cell States and Differentiation Trajectories.*

Abstract. Single-cell RNA sequencing (scRNA-seq) is a powerful tool for characterizing cell types and states. However, it has limitations in measuring changes in gene expression during dynamic biological processes such as differentiation due to the destruction of cells during analysis. Recent studies combining scRNA-seq with lineage tracing have provided clonal information but still face challenges such as observations at discrete time points and difficulty in tracking cells within a certain lineage over the time course, since early observations are not direct ancestors of cells in the same lineage observed later time point. To address these issues, we developed Lineage Variational Inference (LineageVI), a model based on the framework of variational autoencoder (VAE), to convert single-cell transcriptome observation with DNA barcoding into the latent state dynamics consistent with the clonal relationship by assuming a common ancestor. This model enables us to quantitatively capture the cell state transitions. We demonstrate how our model can recapitulate differentiation trajectories in hematopoiesis and learn potential dynamics and estimated backward transitions from later to earlier observations in the latent space. Restoring transcriptomes at each time point in each lineage showed an increase in undifferentiated marker expression and a decrease in differentiation marker expression according to ancestors.

Keywords: Single-cell RNA sequencing (scRNA-seq), Lineage tracing, neural network, variational autoencoder (VAE), differentiation, Hematopoiesis

Jaebeom Kim (Interdisciplinary Program in Bioinformatics, Seoul National University) and Martin Steinegger (School of Biological Sciences, Seoul National University). *Metabuli: sensitive and specific metagenomic classification through a novel joint analysis of amino-acid and DNA sequences.*

Abstract. Assigning taxonomic labels to metagenomic reads involves a trade-off between specificity and sensitivity, depending on the sequence type employed. DNA-based metagenomic classifiers offer higher specificity by capitalizing on mutations to differentiate closely related taxa. Conversely, AA-based classifiers provide higher sensitivity in detecting homology due to the increased conservation of AA.

To solve the trade-off, we developed Metabuli based on a novel k-mer structure, metamer, that simultaneously stores AA and DNA. Metabuli compares metamers first using AA for sensitivity and subsequently with DNA for specificity. We compared Metabuli to DNA-based (Kraken2, KrakenUniq, Centrifuge) and AA-based (Kraken2X, Kaiju, MMseqs2 Taxonomy) tools. In an inclusion test, where 2382 query subspecies were present in databases, DNA-based tools classified up to twice as many reads as AA-based tools to correct (sub)species. However, in an exclusion test, where 367 query species were excluded from databases, AA-based tools showed about twice higher sensitivity in genus-level classification.

Only Metabuli showed state-of-art level performance in both, achieving species-level precision of ~99% and sensitivity of ~97% in the inclusion test, and precision of ~65% and sensitivity of ~48% in the exclusion test. It demonstrates the robustness of Metabuli in diverse contexts of metagenomic studies. (github.com/steineggerlab/Metabuli)

Keywords: Metagenomics, Metagenomic classification, Next-generation sequencing, Taxonomic classification

Dominik Hadzega (Medirex Group Academy n. p.o.), Michaela Hyblova (Medirex Group Academy n. p.o.), Petra Hirjakova (Medirex Group Academy n. p.o.), Katarina Kalavska (Translational Research Unit, Faculty of Medicine, Comenius University and National Cancer Institute), Lucia Kucerova (Translational Research Unit, Faculty of Medicine, Comenius University and National Cancer Institute), Andrea Soltysova (Medirex Group Academy n. p.o.), Lubos Klucar (Institute of Molecular Biology SAS), Gabriel Minarik (Medirex Group Academy n. p.o.) and Michal Mego (Translational Research Unit; 2nd Department of Oncology, Comenius University and National Cancer Institute).
Transcritomic study on cisplatin resistance in testicular germ cell tumours.

Abstract. Resistance for therapy against Germ cell tumours (GCTs) brings serious complications. Although, these processes have been studied, exact mechanism is still in need of further clarification. Here, we were observing changes in trancriptome profile between parental cell lines and those resistant on cisplatin. We studied 7 samples from 6 different parental cell lines and same number of derived resistant cell lines. Using standard procedure of differentially expressed genes analysis, using statistical test by DESeq2, we identified 2 protein coding genes DAZ1 and DAZ2 as differentially expressed and apart from that 2 non-coding RNA genes. Gene enrichment analysis by gProfiler2 on wider set of potentially differentially expressed genes showed upregulation of genes involved in cancers, and cancer related processes (drug metabolism, constitutive signalling by aberrant PIK3). In this study, we simultaneously continue with sequencing more samples and DNA-based variant analysis. This research aims to help us understand mechanism behind cisplatin resistance of GCTs and potentially improve the treatment.This work was supported by the Slovak Research and Development Agency under the Contract no. APVV-20-0158 and with the support of the OP Integrated Infrastructure for the project with the code ITMS: 313011AVH7, co-financed by the European Regional Development Fund.

Keywords: Germ cell tumours, cisplatin resistance, RNA-seq, Differentially expressed genes, Gene enrichment

Özlem Muslu (TRON -Translationale Onkologie an der Universitätsmedizin der Johannes Gutenberg-Universität Mainz gemeinnützige GmbH), Jonas Ibn-Salem (TRON -Translationale Onkologie an der Universitätsmedizin der Johannes Gutenberg-Universität Mainz gemeinnützige GmbH), Thomas Bukur (TRON -Translationale Onkologie an der Universitätsmedizin der Johannes Gutenberg-Universität Mainz gemeinnützige GmbH), Nathalie Buchholz (TRON -Translationale Onkologie an der Universitätsmedizin der Johannes Gutenberg-Universität Mainz gemeinnützige GmbH), Stefania Gangi Maurici (TRON -Translationale Onkologie an der Universitätsmedizin der Johannes Gutenberg-Universität Mainz gemeinnützige GmbH), Alina Henrich (TRON -Translationale Onkologie an der Universitätsmedizin der Johannes Gutenberg-Universität Mainz gemeinnützige GmbH), Martin Löwer (TRON -Translationale Onkologie an der Universitätsmedizin der Johannes Gutenberg-Universität Mainz gemeinnützige GmbH) and Ugur Sahin (BioNTech Mainz). *A benchmarking dataset for somatic variant calling by orthogonal deep sequencing confirmation.*

Abstract. The influence of somatic mutations in cancer led to various somatic mutation callers, yet benchmarking such callers need to be standardized further by providing extensive ground truth data, especially for low frequency mutations. While consortium efforts such as TCGA-MC3, PCAWG-Pilot63, and SEQC2 achieve this by providing orthogonal deep sequencing data for a subset of called mutations, they do not cover the entire mutational spectrum and for patient derived cohorts it is not possible to perform additional orthogonal sequencing due to lack of sufficient DNA material. Here we present a cell line based data set that contains matched whole-exome sequencing data with two technical replicates for tumor and normal DNA of three cell lines of melanoma and pancreas cancer. All mutations called by Mutect2, Strelka2, and our AI-based variant caller, VariantMedium were deep sequenced using Illumina MiSeq (mean coverage=34870X). The presented data set contains a total of 1395 variants consisting of 1222 SNVs and 173 indels, with 894 confirmed somatic, 152 confirmed germline mutations, and 349 variants that were categorized as no mutation. Notably, 21% of all variant candidates have variant allele frequencies below 0.1. With this data set, more extensive benchmarking studies can be performed, specifically for low frequency mutations.

Keywords: Somatic mutation, Whole exome sequencing, Targeted deep sequencing, Benchmarking

Ilaria Billato (Department of Biology, University of Padova), Chiara Romualdi (Department of Biology, University of Padova), Gabriele Sales (Department of Biology, University of Padova) and Davide Risso (Department of Statistical Sciences, University of Padova). *ScalablePCA: Benchmarking principal component analysis for large-scale single-cell RNA-sequencing data.*

Abstract. The size and complexity of single-cell RNA-seq data are increasing, making standard workflows too computationally demanding. Existing tools for single cells do not scale efficiently to such large datasets and operate out-of-memory. To address this problem, the use of more efficient algorithms and out-of-memory data representations becomes essential for the analysis.
This work compares SVD algorithms, applied to the computation of the first 50 principal components of single-cell RNA-seq data, to IRLBA and randomized SVD algorithms as implemented in the R/Bioconductor BiocSingular, Python Scanpy and ScikitLearn packages. We tested the above methods on a real single-cell RNA-seq dataset from 10X Genomics that contains approximately 1.3 million cells and 30,000 genes isolated from the mouse brain. We found that randomized PCA was the most memory efficient, taking 7.05 GB of RAM, while IRLBA PCA took 7.48 minutes to compute the top 50 principal components, using a single core. This benchmark will represent a useful guideline to find out the best trade-off regarding time and memory consumption and to observe how computational times and costs change using e GPU-based rather than CPU-based pipelines.

Keywords: benchmark, transcriptomic, principal component analysis, scalable, CPU, GPU, computational, workflow

Haichao Wang (Cancer Research UK Cambridge Institute, University of Cambridge and Cancer Research UK Cambridge Centre, Cambridge), Elkie Chan (LKS Faculty of Medicine, The University of Hong Kong, Hong Kong SAR, China), Aadhitthya Vijayaraghavan (Cancer Research UK Cambridge Institute, University of Cambridge and Cancer Research UK Cambridge Centre, Cambridge), Paulius Mennea (Cancer Research UK Cambridge Institute, University of Cambridge and Cancer Research UK Cambridge Centre, Cambridge), Emma-Jane Ditter (Cancer Research UK Cambridge Institute, University of Cambridge and Cancer Research UK Cambridge Centre, Cambridge), Wendy N Cooper (Cancer Research UK Cambridge Institute, University of Cambridge and Cancer Research UK Cambridge Centre, Cambridge), Arif Surani (Cancer Research UK Cambridge Institute, University of Cambridge and Cancer Research UK Cambridge Centre, Cambridge), Maria Neofytou (Cancer Research UK Cambridge Institute, University of Cambridge and Cancer Research UK Cambridge Centre, Cambridge), Tommy Kaplan (School of Computer Science and Engineering, The Hebrew University of Jerusalem, Israel), Nitzan Rosenfeld (Cancer Research UK Cambridge Institute, University of Cambridge and Cancer Research UK Cambridge Centre, Cambridge) and Hui Zhao (Cancer Research UK Cambridge Institute, University of Cambridge and Cancer Research UK Cambridge Centre, Cambridge). *Accurate fragment-length estimation of cell-free DNA.*

Abstract. Upon cell death, short cell-free DNA fragments are released into the circulatory system, allowing a non-invasive view of various cellular processes in tissues. Of particular interest are tumor-derived cfDNA fragments, which appear at very low abundance. Recently, “fragmentomic” features, including cfDNA fragment length distribution, end motifs, and genomic location, were utilized to distinguish between ctDNA and normal cfDNA. As shown, different methods for library preparation and alignment often bias these features. In this study, we investigated how to properly analyze Illumina sequenced cfDNA data for optimal accuracy.
Five library preparation methods were applied to 10 healthy individuals’ cfDNA. sWGS was performed and multiple aligners were compared to quantify the variations in the results. We also aligned both trimmed and untrimmed fastq files to see if soft-clipping could bias the feature call. The biases in features inferred using different tools will be reported.
Trimgalore and trimmomatic do not trim 5’ adapters and may not be suitable for library preparations that included 5’ manipulation. BWA-MEM assumes fragment lengths are normally distributed whereas in cfDNA both mono- and di-nucleosomal fragments are expected, requiring specific parameter settings. We investigated and recommend a set of tools that give the expected and robust fragment length distribution.

Keywords: cancer, liquid biopsy, cfDNA, fragmentomics

Daniel López López (Fundación progreso y salud), Gema Roldan (Fundacion Progreso y Salud), María Peña Chillet (Fundación Progreso y Salud) and Joaquin Dopazo (Fundacion Progreso y Salud).
The Spanish Polygenic Risk Score Reference Distribution: A Resource for Personalized Medicine.

Abstract. We present the Spanish polygenic risk score (PRS) reference distribution, a database for the Spanish population consisting of 3124 PRS distributions for common diseases and quantitative traits. The reference includes PRS for various types of cancer, disorders associated with the digestive, cardiovascular, neuronal, and immune systems, as well as quantitative traits such as hematological measurements, and anthropometric levels. The distributions can be explored at http://csvs.clinbioinfosspa.es/?tab=prs.
We released the pipeline we utilized to preprocess, phase and impute samples in our reference cohort. This makes it possible to compute PRS for external genomes and exomes, which can then be compared to a specific reference distribution in a standardized manner. Our pipeline is designed to handle large cohorts in parallel, and can be run on local or cloud-based infrastructures.
The use of these resources can assist in selecting the most suitable PRS, determining the relative risk for patient stratification, and calibrating absolute risk values of PRS for the Spanish population. This can aid in the incorporation of PRS in the Spanish healthcare system. Furthermore, this approach can be applied to establish population-specific PRS distributions for other populations, facilitating the adoption of PRS in healthcare systems worldwide.

Keywords: polygenic risk score, Spanish population, personalized medicine, disease prediction, quantitative traits

Yrjö Koski (Institute for Molecular Medicine Finland, University of Helsinki), Biswajyoti Sahu (Applied Tumor Genomics Research Program, Faculty of Medicine, University of Helsinki), Kimmo Palin (Applied Tumor Genomics Research Program, Faculty of Medicine, University of Helsinki) and Esa Pitkänen (Institute for Molecular Medicine Finland (FIMM), University of Helsinki). *NanoDS: Simulating nanopore sequencing data with distribution approximations.*

Abstract. Oxford nanopore sequencing allows simultaneous capture of long sequencing reads and rich signal data, facilitating epigenetic modification detection and genome assembly.
Data simulation methods can be used to create large datasets for developing and benchmarking computational methods, and augment existing datasets to train robust machine learning models. The use of simulated data can alleviate the need to generate costly real-world data and avoid data privacy issues. However, simulated data has to accurately resemble real data in order for it to be useful.
Here we propose NanoDS (Nanopore distribution simulator), a probabilistic deep learning method for simulating nanopore sequencing data. NanoDS predicts the parameters that describe the signal level and event length, i.e. number of signal measurements, distributions for each k-mer in a DNA sequence. We model the signal levels and event lengths with Gaussian and k-inflated negative binomial distributions, respectively.
NanoDS was trained using two datasets from PCR-amplified DNA samples. We validated NanoDS by basecalling the generated signals with a state-of-the-art basecaller. We also compared NanoDS to other simulation methods proposed in the literature. These experiments showed that NanoDS can accurately simulate nanopore sequencing signals and can be used to generate datasets for developing and benchmarking computational methods.

Keywords: nanopore sequencing, deep learning, simulation, probabilistic modeling

Thea Fennell (MRC Laboratory of Molecular Biology, University of Cambridge) and Ieva Berzanskyte (MRC Laboratory of Molecular Biology). *Single-cell sequencing and cross-species mapping – conserved immunological markers revealed in bone marrow.*

Abstract. Background: Traditionally non-model species are increasingly emerging as the focus of research; most recently, the Golden Hamster (Mesocricetus auratus), with its human-like COVID-19 lung pathology. In this analysis, we use a mouse cell atlas to characterise bone marrow from M. auratus, using self-assembling manifolds (SAM). The most immediate application of this profile is to pathology. However, for the single-cell community, there was a deeper, more methodological motive: to amplify the power of reference-based mapping.
Results: Bone marrow was extracted from hamster legs and sequenced using 10X. Hamster genes were annotated via reciprocal BLAST against mouse. This facilitated integration with the Tabula Muris Senis, using SAM for reference-based mapping of cell identity (SAMap). Assignments were consistent with and exceeded the predictive power of marker-based mapping, conducted in parallel, supporting marker conservation from mouse.
Conclusions: Cellular atlases can improve significantly on the assignment of identity in single-cell datasets. Our study recapitulates this finding. Critically, we succeeded in extending this improvement across species, through the use of innovative SAM algorithms. Thus, we spotlight a novel approach for the community to exploit curated cell atlases and transcriptomes, beyond the constraints of their original, model organism; and beyond this coincidental case of the Golden Hamster.

Keywords: single-cell sequencing, cross-species, bone marrow, immunology, cell identity, cell atlas, reference-based mapping, marker-based mapping, scRNA-seq, RNA-seq, transcriptomics, bioinformatics, self-assembling manifolds, homology, hamster, golden hamster, mouse, Tabula Muris Senis, COVID-19

Eleftherios Zormpas (Biosciences Institute, Faculty of Medical Sciences, Newcastle University, NE2 4HH, UK.), Rachel Queen (Bioinformatics Support Unit, Faculty of Medical Sciences, Newcastle University, NE2 4HH, UK.), Adrienne Unsworth (Bioinformatics Support Unit, Faculty of Medical Sciences, Newcastle University, NE2 4HH, UK.), Quentin Anstee (Translational & Clinical Research Institute, Faculty of Medical Sciences, Newcastle University, NE2 4HH, UK.), Alexis Comber (School of Geography and Leeds Institute for Data Analytics, University of Leeds, LS2 9NL, UK.) and Simon Cockell (School of Biomedical, Nutritional and Sport Sciences, Faculty of Medical Sciences, NE2 4HH, UK.). *Geographically weighted methods for Spatial Transcriptomics data analysis.*

Abstract. Advancements in technology have made it possible to use RNA sequencing in situ. This enables the comprehensive analysis of the entire transcriptome with almost single-cell accuracy while preserving the spatial information of the tissue. The resulting spatially-resolved 'omics data presents a new analytical challenge for molecular biology – how to leverage the spatial aspect of the data effectively? Since tissues are congregations of intercommunicating cells, identifying local and global patterns of spatial association is imperative to elucidate the processes which underlie tissue function. Performing spatial data analysis requires particular considerations of the distinct properties of data with a spatial dimension, which gives rise to an association with a different set of statistical and inferential considerations. By their nature, the geographical sciences primarily use spatially oriented data and over many years have developed the necessary tools and methods to analyse them robustly. Here we discuss the application of a selection of such methods in the biological context, examining a publicly available dataset from the 10X Visium platform.

Keywords: spatial data, spatial analysis, spatial transcriptomics, geography, geographically weighted methods

Igor Martayan (ENS Rennes). *Fractional Hitting Sets.*

Abstract. Minimizers (m-mers where m < k) play a key role in modern methods for efficient searching, mapping and indexing of long genomic sequences. Grouping k-mers based on their minimizers is useful for distributed and parallel processing, as well as compact encoding using super-k-mers (consecutive k-mers sharing the same minimizer). One way to choose minimizers is to use Universal Hitting Sets (UHS), which select a small set of minimizers such that every k-mer is guaranteed to contain at least one of them. While standard minimizer selection usually achieve a density of 2 selected minimizers per k-mer, UHS-based approach typically achieve a lower density. We introduce Fractional Hitting Sets (FHS), which select a fraction of the minimizers uniformly at random, without having to cover every k-mer. By relaxing this constraint of universality, we can reduce the density even further than with UHS. We derive a theoretical model for FHS allowing us to predict the expected density and the size of the super-k-mers based on the chosen fraction. We then show that FHS are suitable for both high coverage with reduced density, and small unbiased coverage with a minimal footprint.

Keywords: k-mer, minimizer, subsampling, sketching, density

Clara Inserte (Institute of Medical Informatics, University of Münster, Münster, 48149, Germany), Kornelius Kerl (Department of Pediatric Hematology and Oncology, University Children's Hospital Muenster, 48149 Münster, Germany), Julian Varghese (Institute of Medical Informatics, University of Münster, Münster, 48149, Germany) and Sarah Sandmann (Institute of Medical Informatics, University of Münster, Münster, 48149, Germany). *Optimized selection of marker genes for spatial transcriptomics deconvolution.*

Abstract. Different from bulk and common single-cell RNA-seq analyses, spatial transcriptomics provide the option to analyze the positional context of cells in a tissue. However, common approaches lack single-cell resolution. As a consequence, the analysis of these data requires a deconvolution step to infer the cell type composition of each spot. Several tools and algorithms are available for this purpose, most of which require a matching single-cell dataset and a list of specific marker genes for each cell type, usually defined via the top differentially expressed genes. However, this approach only considers changes in expression levels and their significance.
We performed a detailed analysis of genes and their potential eligibility as markers, considering changes in expression level and specificity of gene expression. The ideal marker gene will have significantly increased expression and will be expressed in all cells belonging to a specific cell type, while not being expressed in any of the others. Comparing different approaches on the basis of real spatial transcriptomics data from six cases of hepatoblastoma, we developed a filtration strategy resulting in a better and more specific list of marker genes that can be used for the deconvolution of spatial transcriptomics.

Keywords: deconvolution, spatial transcriptomics, marker genes, single-cell

Ruichao Wang (Joint Research Center Computational Biomedicine, University Hospital RWTH Aachen, Aachen, Germany) and Kjong-Van Lehmann (Cancer Research Center Cologne-Essen, University Hospital Cologne, Cologne, Germany). *Distance analysis of mutational signatures .*

Abstract. Mutational signatures are context-specific frequency patterns of somatic mutations that result from endogenous and/or exogenous factors and the presence of distinct types of mutational signatures can provide information on the aetiology of tumours. Reference catalogues of mutational signatures have been created that are used to determine the presence or absence of the according patterns in a tumour sample.
However, the differences observed between two signatures based on the mutational profile alone can be very subtle. Typically, the cosine distance is used to define a formal distance between two mutational signatures. This however can provide misleading results when trying to identify signature similarities or the presence of signatures in existing data.
Therefore, we have been comparing different distance metrics to determine the most robust metric to distinguish and in return quantify the presence of mutational signatures in given samples. To compare different distance metrics, we simulated mutational signatures based on the existing COSMIC reference catalogues. We considered several performance criteria to provide a better understanding on how to assess mutational signatures in the future.

Keywords: Somatic mutations, Mutational signatures, Distance metrics

Sara Terzoli (Humanitas Research Hospital), Laura Mannarino (Humanitas Research Hospital), Valentina Cazzetta (Università degli studi di Milano), Lara Paracchini (Humanitas Research Hospital), Rosalba Portuesi (Humanitas Research Hospital), Domenico Vitobello (Humanitas Research Hospital), Maurizio D'Incalci (Humanitas Research Hospital), Joanna Mikulak (Humanitas Research Hospital), Sergio Marchini (Humanitas Research Hospital) and Domenico Mavilio (Università degli Studi di Milano). *Single-cell RNA sequencing reveals heterogeneity within different histological subtypes of epithelial ovarian cancer.*

Abstract. Epithelial ovarian cancer (EOC) is a gynecological malignancy that develops within the ovary or the fallopian tubes. It is classified into different histotypes including high-grade serous, low-grade serous, endometrioid, clear cells, and mucinous with heterogeneous profiles and clinical outcomes. However, information regarding the heterogeneity of tumor-infiltrating immune cells (TIICs) and blood-associated immune cells as well as the role of the surrounding tumor microenvironment (TME) in tumor progression, is lacking.
We performed 3' single-cell RNA sequencing on both peripheral blood mononuclear cells (PBMCs), sorted TIICs, and tumor cells in a cohort of EOC patients encompassing four different histotypes: high-grade serous, endometrioid, clear cells, and mucinous. We investigated the heterogeneity of TME by using different unsupervised learning algorithms including a holistic clustering approach, functional enrichment analysis (Reactome), differentiation and activation trajectory analysis (RNA velocity, CellRank), cell-cell communication (NicheNet), cytokine profiling (CytoSig database), and transcription factor inference (SCENIC).
Overall, our analysis revealed that different histotypes are characterized by high heterogeneity in the immune cell subpopulations in terms of ratio and phenotype. Moreover, it provided insights into the comprehension of tumor-immune cell interactions, outlining a correlation between the immune system and the clinical outcome in EOC patients.

Keywords: Epithelial ovarian cancer, tumor-infiltrating immune cells, Single-cell RNA sequencing, RNA velocity, transcription factor

Paolo Marzano (Università degli Studi di Milano), Sara Terzoli (Humanitas University), Simone Balin (Università degli Studi di Milano), Silvia Della Bella (Università degli Studi di Milano), Valentina Cazzetta (Università degli Studi di Milano), Rocco Piazza (Università di Milano Bicocca), Inga Sandrock (Hannover Medical School), Sarina Ravens (Hannover Medical School), Likai Tan (Hannover Medical School), Immo Prinz (Hannover Medical School), Antonio Voza (Humanitas University), Joanna Mikulak (Humanitas Research Hospital) and Domenico Mavilio (Università degli Studi di Milano). *Transcriptomic propensity of TNFhigh MAIT cells to provide B cell help following SARS-CoV-2 vaccination.*

Abstract. Recent findings indicated an association between MAIT cells and the immune response to the BNT162b2 vaccine, where MAIT cell frequency was associated with an increased adaptive immune response.
Herein, to investigate the effect of repeated SARS-CoV-2 vaccinations on MAIT cells, we performed a longitudinal 5' scRNA-seq coupled with scTCR-seq analysis on the peripheral blood samples of six healthy adults naïve for the SARS-CoV-2 infection and immunized with the two doses of the mRNA-based vaccine BNT162b2. Taking advantages of computational approaches, including functional pathway enrichment analyses and the gene expression–effector cell-polarization's fate probabilities correlation (RNA Velocity and CellRank), we identified MAIT cells as the major source of TNF-α across circulating lymphocytes, and this TNFhigh signature increased upon the second administration of the vaccine. Notably, the increased TNF-α expression correlated with SARS-CoV-2 specific antibody titers. Therefore, by modeling the intercellular communication with the NicheNet algorithm, we observed that the TNF-α-profile predicts the transcriptional changes of conventional switched memory B cells, deputed to high-affinity long-term memory.
Overall, our results indicate that MAIT cells promote B cell functionality in response to the vaccine, favoring effective and long-term protection against SARS-CoV-2 infection, suggesting the use of MAIT cells as cellular adjuvants in mRNA-based vaccines.

Keywords: single-cell RNA sequencing, TCR sequencing, RNA velocity, NicheNet, SARS-CoV-2, BNT162b2 mRNA vaccine, MAIT cells, Unconventional T cells, Immune response, TNF-α, B cells

Carolin Walter (Institute of Medical Informatics), Sarah Sandmann (Institute of Medical Informatics), Luisa Klotz (Department of Neurology with Institute of Translational Neurology) and Julian Varghese (Institute of Medical Informatics). *Comparison of deconvolution algorithms for GeoMx spatial transcriptomics immune cell data.*

Abstract. Spatial Transcriptomics (ST) is a powerful Next Generation Sequencing-based technique that combines expression information and spatial context for a given sample, and thus allows new insights into the cellular composition and spatial organization of tissues. Several ST techniques with different structures exist, e.g. grid-based spot data, or marker-based regions of interest (ROI), but most current approaches are limited to multi-cell resolution. A precise estimation of the abundance of cell types in a chosen region of interest is therefore essential for accurate ST data interpretation. We compared the performance of deconvolution algorithms regarding the inference of immune cell types on original and published NanoString GeoMx ST data from multiple sclerosis (MS) lesions and lung tumor tissue. Both published and single cell-based custom immune cell profiles were used for the cell type deconvolution, and effects of raw data quality and algorithm parameter settings on the estimated immune cell populations were assessed. With preprocessed immune signature matrices as basis for computational deconvolution, SpatialDecon, FARDEEP, and EPIC consistently identified core immune cell populations in a subset of seven MS regions of interest. Detection of other cell types was more variable, and algorithm-dependent. Lower sample quality generally impeded accurate cell type deconvolution.

Keywords: spatial transcriptomics, Next Generation Sequencing, NanoString GeoMx, deconvolution, benchmarking

Shaolei Teng (Howard University). *Computational mutagenesis and transcriptome analysis of Curly Su protein in transgenic flies.*

Abstract. The Curly Su (dMPO) protein, a homolog of the human myeloperoxidase (hMPO), is involved in wing development in Drosophila melanogaster. The dMPO contributes to various cellular and physiological processes through its production of reactive oxygen species (ROS). As the sequences of dMPO and hMPO are similar, dMPO is an excellent candidate for experimental validation for the development and immunity studies. To investigate the effects of specific mutations on dMPO and hMPO, we performed saturated computational mutagenesis to identify the target mutations based on predicted folding energy changes and proximity to Post-Translational Modification (PTM) sites. We constructed transgenic fruit flies with G378W, Del 305-687, S590A, K552R, and W621R mutations using genome editing. We observed wing phenotypes and the overall lifespan of the samples during husbandry. The transcriptome analysis was conducted for both transgenic and wild-type samples using RNAseq. We utilized the R Bioconductor package, BigPint, to visualize differentially expressed genes (DEGs) between treatment samples and conducted gene ontology analysis of the DEGs to provide functional attributes to down-regulated and up-regulated genes. The combination of computational tools and genetics experiments enabled rapid insights into the novel functional effects of missense mutations in target proteins.

Keywords: Computational mutagenesis, Curly Su, Myeloperoxidase, RNAseq, transgenic fly

Lilian Marchand (CRIStAL - CNRS UMR 9189 - Université de Lille), Hélène Touzet (CRIStAL - CNRS UMR 9189 - Université de Lille) and Jean-Stéphane Varré (CRIStAL - CNRS UMR 9189 - Université de Lille). *Assessing alternatively spliced transcript diversity with long reads.*

Abstract. We introduce RNA-tailor, a novel tool designed to precisely inventory the repertoire of alternatively spliced transcripts of a target gene from third-generation sequencing data. Alternative splicing (AS) is a regulation mechanism that enables the production of various RNA isoforms. Abnormally spliced RNAs can be responsible for various diseases and cancers. For the study of AS and despite their higher error sequencing rate, long read sequencing technologies are preferred to short reads as they are able to capture full length transcripts, allowing to grasp the combinatorics of exons. Thanks to the usage of splice-aware alignment tools and fine refinement steps, RNA-tailor aims to provide a nucleotide-level precise picture of alternative transcripts of a given target gene using only a reference sequence. It makes the method usable for analysis on both model and non-model species, and provides unbiased and accurate results that better reflect the transcript diversity of a sample with a great potential for novel isoform discovery. We will present results on an ONT mouse transcriptome dataset and a subset of 19 genes for which annotated isoforms differs in number or AS events. RNA-tailor preliminary results show a better level of prediction than recent tools (Freddie, FLAIR), without prior knowledge.

Keywords: alternative splicing, third generation sequencing, transcript isoform identification, long read

Satria Kautsar (DOE Joint Genome Institute, Lawrence Berkeley National Lab, US), Harrison Ho (School of Natural Sciences, University of California at Merced, US) and Zhong Wang (DOE Joint Genome Institute, Lawrence Berkeley National Lab, US). *Axolotl: A Scalable Apache Spark-based Library for High-troughput Genomic Data Analysis.*

Abstract. Next-generation sequencing has substantially increased genomic data volume and complexity, often exceeding terabytes in size. Traditional bioinformatic tools, designed for single computer operations, struggle to cope with these datasets. Despite the emergence of parallel frameworks like Apache Spark, Dask, Polars, and Ray, their application to genomic problems remains limited.

We introduce Axolotl, a scalable library built on Apache Spark, specifically designed for large-scale genomic data analysis. Axolotl creates genomics-specific function modules, enabling biologists to utilize Python for distributed computing environments. Users can harness Spark's built-in SQL and Machine Learning libraries for scalable bioinformatics analysis. We present two distinct use cases: a global-scale examination of over 1.5 million biosynthetic gene clusters (BGCs) and a distributed batch computation of polygenic risk scores (PRS). Axolotl efficiently processes these datasets in parallel using 32+ 16-core compute nodes, virtually combining the power of a 512-core, 4TB RAM machine, with entire analysis pipelines implemented in fewer than 50 lines of code.

Our findings highlight Axolotl's potential to revolutionize how researchers tackle large genomic data sets, enabling swift, scalable, and accurate analyses across a broad spectrum of omics applications.

Keywords: Axolotl, Apache Spark, Large-scale genomics analysis

Mikele Milia (University of Padova), Nicola Ferro (University of Padova), Barbara Di Camillo (University of Padova) and Giacomo Baruzzo (University of Padova). *mopo16Sweb: A webapp for multi-objective optimization of 16S rRNA primers sequences on the cloud.*

Abstract. Targeted amplicon sequencing of the 16S ribosomal RNA gene is a common approach to study microbial communities in a site. However, the accuracy of this methodology strongly depends on the choice of primer pairs.
In our previous work, we developed mopo16S [DOI:10.1186/s12859-018-2360-6], a multi-objective optimization framework to simultaneously maximize primers efficiency, specificity and coverage. Here we present mopo16Sweb, a powerful tool to easily design optimal 16S primers that further improve mopo16S functionalities.
First, we have extended the multi-objective optimization framework by designing new fitness functions to include user-specified constraints in the optimization process.
Second, we have simplified the specification of the required input by including built-in presets of known bacteria sequences and primers pairs, querying the most widely used metagenomic databases (GreenGenes, Ribosomal Database Project, SILVA and probeBase).
Third, we have improved the analysis of the output-optimized primers by adding interactive plots/tables that simplify the selection of the best primers among the ones in the Pareto front output.
We have included all the above updates in mopo16Sweb, a novel interactive (containerized) webapp for 16S primers optimization on the cloud (https://mopo16sweb.dei.unipd.it/) or in user server.

Keywords: optimization, multi-objective optimization, webapp, 16S sequencing, PCR primer, metagenomics

Shunhua Han (Illumina, Inc.), Vitor Ounchic (Illumina, Inc.), Varun Jain (Illumina, Inc.), Pavana Anur (Tempus Labs, Inc.), Francisco De La Vega (Tempus Labs, Inc.) and James Han (Illumina, Inc.).
Overcoming the Challenges of Variant Calling in PMS2 High Homology Regions for Improved Lynch Syndrome Diagnosis using Whole-Genome Sequencing.

Abstract. Detecting pathogenic variants in PMS2 is crucial for diagnosing Lynch Syndrome, an autosomal dominant cancer predisposition syndrome. However, the presence of a highly homologous pseudogene, PMS2CL, complicates variant calling using short-read sequencing. In this study, we introduce a computational method, Multi-Region Joint Detection (MRJD), to address these challenges and improve the reliability of small variant calling in the ~11kb PMS2 homology region. MRJD detects variants in paralogous regions by jointly genotyping all paralogous regions, including reads with ambiguous alignment. The method offers a default mode that balances precision and recall, and a high sensitivity mode that maximizes the ability to identify all potential variants. These methods are implemented as part of the DRAGEN software suite v4.2, allowing users to choose the best fit for their needs.
We benchmarked MRJD on 150 samples from the Illumina Polaris diversity panel and found it outperforms default germline small variant calling, particularly for INDELs. The high sensitivity mode achieves 96% aggregated recall for SNPs and INDELs, while maintaining acceptable precision. MRJD contributes to a more reliable diagnosis of Lynch Syndrome, improving risk assessment for affected individuals. This method also paves the way for further research on variant calling in genes with high homology challenges.

Keywords: PMS2, Lynch Syndrome, Whole-Genome Sequencing, Variant Calling, Bioinformatics

Or Lazarescu (Department of Clinical Biochemistry & Pharmacology, Ben-Gurion University of the Negev, Beer Sheva, Israel), Yulia Haim (Department of Clinical Biochemistry & Pharmacology, Ben-Gurion University of the Negev, Beer Sheva, Israel), Maya Ziv-Agam (Department of Clinical Biochemistry & Pharmacology, Ben-Gurion University of the Negev, Beer Sheva, Israel), Idan Hekselman (Department of Clinical Biochemistry & Pharmacology, Ben-Gurion University of the Negev, Beer Sheva, Israel), Juman Jubran (Department of Clinical Biochemistry & Pharmacology, Ben-Gurion University of the Negev, Beer Sheva, Israel), Danny Kitsberg (Edmond & Lily Safra Center for Brain Sciences, The Hebrew University of Jerusalem, Jerusalem, Israel), Antje Korner (Department of Women and Child Health, Centre of Pediatric Research (CPL), Leipzig University, Leipzig, Germany), Rinki Murphy (Department of Medicine, Faculty of Medical and Health Sciences, University of Auckland, Auckland, New Zealand), Matthias Bluher (Medical Department III – Endocrinology, Nephrology, Rheumatology, University of Leipzig, Leipzig, Germany), Naomi Habib (Edmond & Lily Safra Center for Brain Sciences, The Hebrew University of Jerusalem, Jerusalem, Israel), Assaf Rudich (Department of Clinical Biochemistry & Pharmacology, Ben-Gurion University of the Negev, Beer Sheva, Israel) and Esti Yeger-Lotem (Department of Clinical Biochemistry & Pharmacology, Ben-Gurion University of the Negev, Beer Sheva, Israel). *Comparison of the human visceral and subcutaneous single-nuclei atlases uncovers common and depot-selective cell types and communication networks.*

Abstract. Adipose tissues are major endocrine organs that are crucial for human health. Visceral and subcutaneous adipose tissues have overlapping and distinct functions, but a thorough comparison at the single-cell level has been lacking. Here we report single-nuclei RNA-seq analysis of human visceral and subcutaneous adipose tissues collected from 10 donors of both sexes and a range of BMI. Analysis of 83,731 visceral and 37,879 subcutaneous nuclei revealed over 20 cell types. Several cell types showed marked differences in prevalence between the two tissues which were not shown before, including adipocytes, adipose progenitor and stem cells, immune cells, vascular smooth muscle cells and endothelial cells. Among adipocytes, we identified 'classical' adipocytes characterized by enrichment in lipid metabolism pathways. Notably, we identified eight novel 'specialized' adipocyte subtypes, four of which were depot-specific. In addition to expressing known adipocyte marker genes like classical adipocytes, these specialized adipocyte subtypes were enriched in immune-related, extracellular matrix deposition (fibrosis) or vascularization pathways. Analysis of intercellular communication patterns revealed interactions between adipocyte subtypes and immune cells that included common and depot-selective routes. Our analyses extend current understanding of the different biology of human adipose tissues, and propose new depot-selective avenues to manipulate adipose tissues to promote health.

Keywords: single-nuclei RNA-seq, adipose tissue, human cell atlas

Franziska Lang (TRON Translational Oncology Mainz), Patrick Sorn (TRON Translational Oncology Mainz), Martin Suchan (TRON Translational Oncology Mainz), Alina Henrich (TRON Translational Oncology Mainz), Christian Albrecht (TRON Translational Oncology Mainz), Nina Koehl (TRON Translational Oncology Mainz), Aline Beicht (TRON Translational Oncology Mainz), Pablo Riesgo-Ferreiro (TRON Translational Oncology Mainz), Christoph Holtsträter (TRON Translational Oncology Mainz), Barbara Schrörs (TRON Translational Oncology Mainz), David Weber (TRON Translational Oncology Mainz), Martin Löwer (TRON Translational Oncology Mainz), Ugur Sahin (BioNTech SE; Johannes Gutenberg University Mainz; TRON Translational Oncology Mainz) and Jonas Ibn-Salem (TRON Translational Oncology Mainz). *splice2neo combines the effect of somatic mutations on splicing with RNA-seq support to predict tumor-specific splice junctions as neoantigen candidates.*

Abstract. Splicing is dysregulated in many tumors, and thereof resulting tumor-specific transcript isoforms may encode neoantigens. Detecting tumor-specific splicing is challenging because splice junctions identified in tumor transcriptomes can also appear in healthy tissues. However, somatic mutations can disrupt or create canonical splicing motifs leading to individual tumor-specific targets.

Splice2neo integrates the predicted splice effects from somatic mutations with splice junctions detected in tumor RNA-seq. We excluded canonical splice junctions from healthy tissue samples, annotated the resulting transcript and peptide sequences, and developed a stringent detection rule to predict splice junctions as tumor-specific targets. In a verification cohort of melanoma samples, we identified 1.7 target splice junctions per tumor and estimated a false discovery rate of 0.04. For individual examples of exon-skipping events, we confirmed the expression in tumor-derived RNA by quantitative real-time PCR experiments. Most splice junctions encoded at least one neoepitope candidate with predicted strong MHC I or MHC II binding. Compared to neoepitope candidates derived from non-synonymous point mutations, the splicing-derived neoepitope candidates had a lower self-similarity to corresponding wild-type peptides.

In summary, splice2neo helps to identify tumor-specific splice junctions as neoantigen candidates to expand the target repertoire for personalized cancer immunotherapies.

https://github.com/TRON-Bioinformatics/splice2neo

Keywords: Cancer, Splicing, Tumor-specificity, Neoantigen candidates, Personalized immune therapies

Anima Sutradhar (University of Edinburgh), Giovanni Stracquadanio (University of Edinburgh), Jonathan Pointon (FUJIFILM Diosynth Biotechnologies UK) and Christopher Lennon (FUJIFILM Diosynth Biotechnologies UK). *Transcriptome-wide meta-analysis of codon usage in Escherichia coli.*

Abstract. Synonymous codons display an inherent non-random distribution, called codon usage bias, and can differ across species from the level of genes to genomes. This has been exploited by the biotechnology industry, where recombinant proteins are back-translated to DNA by selecting codons to maximise transcription and yield. However, obtaining accurate and representative codon bias estimates requires the identification of highly expressed genes and their codon variation across samples and conditions. To address this, we developed Codon Usage Bias from RNA-sequencing (CUBseq), a fully automatic meta-analysis pipeline to build highly expressed gene panels and estimate codon usage bias from RNA sequencing experiments.
Here, we use CUBseq to estimate codon usage bias in Escherichia coli using RNA sequencing data from 6,763 samples across 72 strains. We found a set of 115 highly expressed genes in our dataset, with negligible variation across strains, suggesting codon usage to be stable across different strains. We then compared our codon usage bias estimates to the widely used genome-derived Kazusa and CoCoPUTs codon usage tables, where we found significant variations across several codons, suggesting that the transcriptome plays an important role in influencing codon preference. Overall, CUBseq provides a novel and robust method for transcriptome-based codon usage analysis.

Keywords: codon usage, transcriptomics, RNA-sequencing, codon optimisation, biostatistics, synthetic biology, computational biology

Christophe Bécavin (Université Côte d'Azur, CNRS, Nice, France), Antoine Collin (3IA, CNRS, Sophia Antipolis, France) and Pascal Barbry (3IA, CNRS, Sophia Antipolis, France). *Checkatlas: One liner quality control tool for your single-cell atlases.*

Abstract. The development of single-cell atlases has become a significant area of research in the last decade. However, there is currently no global quality control tool to evaluate the final reconstructed atlas. To address this issue, the Checkatlas bioinformatic tool was created to provide a user-friendly way to assess the overall quality of single-cell atlases. Checkatlas screens all relevant files in the working directory and produces quality control report for every atlas in one single html file. The tool generates a summary table of all the atlases in the working directory, quality control metrics in table and figure formats, and metadata regarding cell annotation and experimental design. Checkatlas also includes a catalog of essential metrics for evaluating clustering, annotation, and visualization, making it a valuable resource for researchers working with diverse single-cell atlas datasets. The tool is multi-threaded and can be deployed on a computing cluster for efficient and speedy analysis. In this presentation, we will demonstrate Checkatlas's use cases and its successful application in evaluating the quality of 80 COVID and healthy donor atlases.

Keywords: single-cell, Atlas, Human Cell Atlas, Quality control, Clustering metrics, Cell annotation metrics, scanpy, seurat, multiqc, nextflow

Hanna Slowik (Silesian University of Technology), Joanna Zyla (Silesian University of Technology), Anna Papiez (Silesian University of Technology), Joanna Polanska (Silesian University of Technology) and Michal Marczyk (Silesian University of Technology). *Understanding zero counts in single-cell RNA-sequencing data to develop a score for evaluation of data imputation methods.*

Abstract. Single-cell RNA sequencing (scRNA-seq) technology, which enables parallel profiling of hundreds of thousands of cells, has already been successfully applied to search for new cell types or to understand different cellular states. A well-known feature of scRNA-seq is data sparsity, i.e. a high percentage of zero counts in the data matrix. Such an event could occur due to: (i) technical reasons when a cell's transcript is present but undetected because of inefficient cDNA polymerization, amplification error, or low sequencing depth; (ii) biological reasons when zeros reflect a real lack of expression in a cell. We investigated several technical factors that can contribute to expression shift between bulk and scRNA-seq platforms and found that a low level of bulk gene expression representing true expression is the main factor, however, RNA integrity, gene or UTR3 length, and the number of transcripts could also be important. Next, we developed a true biological zero (TBZ) score by calculating the ratio between the distribution of genes not expressed in scRNA-seq but observed in bulk and normally expressed genes. Finally, we used the TBZ score to test existing data imputation methods that can preserve true zeros: ALRA, DrImpute, SAVER, and scImpute.

Keywords: single-cell sequencing, data imputation, zeros

Dario Simionato (Department of Information Engineering, University of Padua, Italy), Antonio Collesei (Venetian Oncology Institute (IOV-IRCSS)), Federica Miglietta (Veneto Institute of Oncology, IOV-IRCCS, Padua, Italy) and Fabio Vandin (Department of Information Engineering, University of Padua, Italy). *ALLSTAR: Inference of ReliAble CausaL RuLes between Somatic MuTAtions and CanceR Phenotypes.*

Abstract. Recent advances in DNA sequencing technologies have allowed the detailed characterization of whole-exomes and whole-genomes in large cohorts of tumors. These studies have highlighted the extreme heterogeneity of somatic mutations between tumors. Such heterogeneity hinders out our ability to identify alterations important for the disease. Several tools have been developed to identify somatic mutations related to cancer phenotypes. However, such tools identify only correlations, with no guarantee of highlighting causal relations.
We describe ALLSTAR, a novel tool to infer reliable causal relations between somatic mutations and cancer phenotypes. In particular, our tool identifies reliable causal rules highlighting combinations of somatic mutations with the highest impact in terms of average effect on the phenotype. While we prove that the underlying computational problem is NP-hard, we develop a branch-and-bound approach that employs PPI networks and novel bounds for pruning the search space, while correcting for multiple hypothesis testing.
Our extensive experimental evaluation on synthetic data shows that ALLSTAR is able to identify reliable causal relations in large cancer cohorts. Moreover, the reliable causal rules identified by our tool in cancer data show that ALLSTAR identifies several somatic mutations known to be relevant for cancer phenotypes as well as novel biologically meaningful relations.

Keywords: causal-inference, bioinformatics, NP-hard, causal-rules, observational-dataset, cancer

Nico Alavi (Max Planck Institute for Molecular Genetics), M-Hossein Moeinzadeh (Max Planck Institute for Molecular Genetics), Jakob Hertzberg (Max Planck Institute for Molecular Genetics), Uirá Souto Melo (Max Planck Institute for Molecular Genetics), Maryam Ghareghani (Max Planck Institute for Molecular Genetics), Anton Kriese (Max Planck Institute for Molecular Genetics), Julika Wenzel (Max Planck Institute for Molecular Genetics), Eldar Abdullaev (Max Planck Institute for Molecular Genetics), Robert Schöpflin (Max Planck Institute for Molecular Genetics), Malte Spielmann (Universiy of Lübeck, Institute of Human Genetics), Stefan Mundlos (Max Planck Institute for Molecular Genetics, Charité-Universitätsmedizin Berlin) and Martin Vingron (Max Planck Institute for Molecular Genetics). *Multi-Platform Comparison of Structural Variant Detection Methods in the Context of Clinical Diagnostics.*

Abstract. Structural variants (SVs) have been associated with many monogenic and complex disorders. However, their accurate and reliable detection remains challenging, often resulting in diverging results from current SV detection methods. As a guide for SV calling in the context of clinical diagnostics, we compared SV detection methods based on Illumina short-reads, PacBio long-reads, 10x Genomics and TELL-Seq linked-reads, as well as Bionano optical maps in a cohort of 20 patients. This comparison was enabled by a novel approach for constructing manually curated benchmark SV callsets for each patient. Our results indicate caller-specific performance differences across categories of SV types, sizes, and genomic contexts.
Furthermore, we present an unsupervised learning approach to systematically identify the failure modes of current short-read-based SV detection methods. For that, each SV was annotated with a number of features. These include features about the variant itself, the genomic context, and the alignment of the reads supporting an SV. These features serve as the basis to construct a kNN-graph for clustering SVs. Next, groups of clusters that are not well detected by short-read sequencing are inferred. Identifying such biases can help improve short-read-based SV calling.

Keywords: Structural Variants, Sequencing, Clinical Diagnostics

Ivan Tolstoganov (Department of Mathematics, Science for Life Laboratory, Stockholm University, Stockholm, Sweden), Zhoutao Chen (Universal Sequencing Technology Corporation, Carlsbad, California 92011, USA), Pavel Pevzner (University of California, San Diego, San Diego CA, USA) and Anton Korobeynikov (Saint Petersburg State University, Saint Petersburg, Russia). *SpLitteR: Diploid genome assembly using linked TELL-Seq reads and assembly graphs.*

Abstract. Abstract
Recent advances in long-read sequencing technologies enabled accurate and contiguous de novo assemblies of large genomes and metagenomes. However, even long and accurate high-fidelity (HiFi) reads do not resolve repeats that are longer than the read lengths. This limitation negatively affects the contiguity of diploid human genome assemblies since two haplomes share many long identical regions. To generate the telomere-to-telomere assemblies of diploid genomes, biologists now construct their HiFi-based phased assemblies and use additional experimental technologies to transform these phased assemblies into more contiguous diploid assemblies. The barcoded linked-reads, generated using an inexpensive TELL-Seq technology, provide an attractive way to bridge unresolved repeats in phased assemblies of diploid genomes.
Here, we present a SpLitteR tool for haplotype phasing and scaffolding in an assembly graph using barcoded linked-reads. We benchmark SpLitteR on assembly graphs produced by various long-read assemblers and show how TELL-Seq reads facilitate phasing and scaffolding in these graphs. This benchmarking demonstrates that SpLitteR improves upon the state-of-the-art linked-read scaffolders in the accuracy and contiguity metrics. SpLitteR is implemented in C++ as a part of the freely available SPAdes package and is available at https://cab.spbu.ru/software/splitter.

Keywords: long-read assembly, linked-reads, haplotype phasing, scaffolding, repeat resolution

Arda Söylev (Heinrich Heine University Düsseldorf), Samarendra Pani (Heinrich Heine University Düsseldorf), Tobias Rausch (European Molecular Biology Laboratory, Genome Biology Unit, Heidelberg, Germany), Jan Korbel (European Molecular Biology Laboratory, Genome Biology Unit, Heidelberg, Germany) and Tobias Marschall (Heinrich Heine University Düsseldorf). *SVarp: Structural Variation Discovery in Pangenomes.*

Abstract. With the introduction of the first draft human pangenome reference, it has been shown that there is a significant improvement in SV genotyping using short-reads with pangenomes compared to a linear reference. However, there is still a lack of tools able to discover novel SVs from read alignments to a pangenomic reference. Here we present SVarp that closes this gap by calling novel phased variation sequences on graph genomes using third generation long sequencing reads. In order to assess the performance of SVarp, we randomly generated SVs of small (<1 kb) and large (>1 kb) size for various SV types, embedded them into the T2T reference genome and mapped them into the HPRC Minigraph pangenome. In order to find specific SV types, we used the sequences (SVtigs) outputted by SVarp as input to SVIM-asm assembly-based SV caller. Based on our comparison results against the true call sets; SVarp can find deletions and duplications with >90% recall and ~99% precision (average of small and large events). On the other hand, although small insertions and large inversions have ~90% recall and precision, large insertions and small inversions suffer from low recall (~30% and ~50% respectively) but high precision (>90%).

Keywords: pangenomes, genomics, structural variation

Corentin Thuilliez (INSERM U1015, Gustave Roussy Cancer Campus), Maria Eugenia Marques Da Costa (INSERM U1015, Gustave Roussy Cancer Campus), Nathalie Gaspar (INSERM U1015 & Department of Pediatric and Adolescent Oncology, Gustave Roussy Cancer Campus), Pierre Khneisser (Department of Medical Biology and Pathology, Gustave Roussy Cancer Campus), Gael Moquin-Beaudry (INSERM U1015, Gustave Roussy Cancer Campus), Jean-Yves Scoazec (Department of Medical Biology and Pathology, Gustave Roussy Cancer Campus) and Antonin Marchais (INSERM U1015 & Department of Pediatric and Adolescent Oncology, Gustave Roussy Cancer Campus).
CellFromSpace: A versatile tool for spatial transcriptomic data analysis through reference-free deconvolution and guided cell type/activity annotation.

Abstract. Spatial transcriptomic is one of the most promising technologies to analyze spatial distribution and interaction. Spatially barcoded next generation (NGS) sequencing-based methods enable the detection of transcripts on tissue sections. Several of these technologies, are near single cell with spots encompassing 1-20 cells. Therefore, a deconvolution step is required to gain insight into the mixture of cells.
Here, we propose a new method named CellFromSpace (CFS), based on the independent component analysis, a blind signal separation method, to deconvolute, without reference single cell data, spatial transcriptomic data. We developed an R package and a shiny interface to accelerate the annotation of the signal.
Visium fresh frozen and FFPE samples of adult mouse brain and human tumors from 10x genomics were analyzed. We were able to recapitulate the structure of the mouse brain using our method with high fidelity. Furthermore, we quickly identified cell types and activities within heterogeneous cancer tissues. The method also enables to subset the signal and the spot corresponding to specific cell, to drive further analysis usually performed for scRNA-seq such as trajectory inference.
In conclusion, CFS provides a full workflow to analyze and quickly interpret results from NGS-based spatial transcriptomics analysis without reference single cell dataset.

Keywords: Spatial transcriptomics, deconvolution, ICA, tool

Samuel Hamilton (Northwestern University), Gaurav Gadhvi (University of Michigan), Script Investigators (N/A) and Deborah Winter (Northwestern University). *Machine Learning Informed Guidelines for Optimal RNA-seq Quality Control.*

Abstract. Every RNA-seq experiment demands rigorous quality control (QC) to remove poor quality samples which may mask meaningful results. While the specifics of RNA-seq quality control depend on the goals of an experiment, thus far there has been minimal investigation into which metrics best identify poor quality samples. In addition, there is a lack of tools which support context-specific quality control analysis. Together, this creates obstacles for researchers looking for a data-informed justification on whether to exclude a potentially problematic sample.
To address this, we investigated the utility of different RNA-seq QC metrics using 252 RNA-seq samples from human patients. We examined the relationships between these QC metrics and 3 distinct endpoint metrics which together characterize RNA-seq sample quality holistically. Then, we trained and interrogated a random forest model to identify the most useful QC metrics and key inflection points in their effect on quality that inform guidelines for sample exclusion. Lastly, we developed an open source software tool, QC Doctor, which generates holistic visual summaries of RNA-seq sample quality that can be customized to a user's experimental goals. Together, this work provides data-informed guidelines and tools to aid researchers in improving the rigor of their RNA-seq experiments.

Keywords: RNA-seq, Quality Control, Machine Learning, Genomics, Software

Gryte Satas (Memorial Sloan Kettering Cancer Center), Matthew A. Myers (Memorial Sloan Kettering Cancer Center), Seongmin Choi (Memorial Sloan Kettering Cancer Center) and Sohrab Shah (Memorial Sloan Kettering Cancer Center). *Leveraging Evolutionary Constraints to Refine Somatic Variant Calls from Single-Cell Sequencing Data.*

Abstract. Single-cell DNA sequencing (scDNA-seq) technologies enable scaled measurements of tumor cell genomes. However, low per-cell coverage and technical biases present analytical challenges for identifying nucleotide resolution mutations. Accurate calling of small variants (SNVs and indels) is a critical prerequisite for many downstream analyses but call sets from scDNA-seq data often contain many false positives ('artifacts'). We introduce ArtiCull, a variant call refinement algorithm that exploits evolutionary constraints to identify artifacts in scDNA-seq data. ArtiCull requires no external training data, manual inspection, or prior knowledge of artifact profiles. Instead, ArtiCull uses somatic evolutionary models to identify a subset of high-confidence artifactual and true variants; these labeled variants are then used to train a feature-based classifier. This enables researchers to train patient-, cohort-, or technology-specific classifiers attuned to the specific profile of technical biases in their dataset. Validation with matched bulk sequencing data shows that ArtiCull greatly improves SNV calling precision with minimal loss of recall. We demonstrate that ArtiCull improves the identification of clones in scDNA-seq data, and increases sensitivity of mutational signature analyses to identify processes active in a small number of cells.

Keywords: scDNA, variant calling, cancer genomics, algorithm, evolution

Yifan Zhao (Department of Biomedical Informatics, Harvard Medical School), Hong Wei Yang (Department of Neurological Surgery, University of Massachusetts Medical School), Rona S. Carroll (Department of Neurological Surgery, University of Massachusetts Medical School), Bethany C. Berry (Department of Neurological Surgery, University of Massachusetts Medical School), Xiaochen Wang (School of Mathematical Sciences, Peking University), Ruibin Xi (School of Mathematical Sciences, Peking University), Mark D. Johnson (Department of Neurological Surgery, University of Massachusetts Medical School) and Peter J. Park (Department of Biomedical Informatics, Harvard Medical School). *Allele-specific identification of somatic copy numbers variants in single-cell whole-genome sequencing data.*

Abstract. Accurate detection of somatic copy number variants (CNVs) in single-cell whole-genome sequencing (scWGS) is challenging due to sequencing depth limitations and amplification artifacts. Existing single-cell CNV callers are primarily designed for large CNVs arising in neoplastic cells, and often have low sensitivity for non-clonal CNVs and small (<1Mb) CNVs. Here, we propose scBIC-seq, a novel CNV detection algorithm that utilizes read count, B-allele frequency, and haplotype signals to accurately infer both clonal and non-clonal CNVs. Benchmarking experiments demonstrated superior performance compared to existing methods, especially at the sub-megabase scale. Remarkably, scBIC-seq enabled the detection of subtle genomic changes in minute cell populations in a longitudinal meningioma case, tracing the origin of the second clonal expansion back to a time even before the first surgical resection. Furthermore, applying scBIC-seq to scWGS data from eleven neurotypical human brains revealed an increase in somatic CNV burden in post-mitotic neurons during normal aging.

Keywords: single cell copy number variant detection, single cell whole-genome sequencing, brain somatic mosaicism

Shulan Tian (Mayo Clinic, Rochester) and Eric Klee (Mayo Clinic, Rochester). *An integrative data mining workbench for disease association and risk prediction in large patient sequencing projects.*

Abstract. Mayo Clinic Tapestry study is a population-scale sequencing initiative. With a goal of sequencing 100,000 consented patient participants, To accelerate genomic findings, we developed a data mining workbench that integrates genetic variants with phenotypic data in EMRs. The workbench is designed to enable the selection of appropriate analysis workflows based on the prevalence of a disease trait. For rare disease studies, the resources such as lists of curated genes and the prior knowledge of gene inheritance pattern are critical. While for common disease studies, we implemented multiple software packages for data QC, population stratification estimation, GWAS and PRS analysis. To demonstrate its broad applications, we applied the workbench to two common diseases, obesity and nonalcoholic fatty liver disease (NAFLD). Through rare variant analysis, the workflow prioritized key genes in obesity that were previously found to be significantly associated with BMI in multiple studies. While in NAFLD cohort, the PRS in the 95% quantile confers increased risk (OR=4.653, 95% CI, 3.88-5.58) compared to the control cohort. Finally, applied to rare disease study, we identified rare pathogenic germline variants in a cohort of 146 Cholangiocarcinoma subjects, which are among key variants previously known to be involved in Cholangiocarcinoma pathogenesis.

Keywords: Population genomics, Genomic data mining, Disease association studies, Risk prediction

Logan Blaine (Harvard Medical School, Massachusetts General Hospital, BROAD institute), Jayoung Ryu (Harvard Medical School, Massachusetts General Hospital, BROAD institute), Lucas Ferreira (Harvard Medical School, Massachusetts General Hospital, BROAD institute), Martin Jankowiak (BROAD institute), Matthias Heinig (Helmholtz Zentrum München Deutsches Forschungszentrum für Gesundheit und Umwelt (GmbH)) and Luca Pinello (Harvard Medical School, Massachusetts General Hospital, BROAD institute). *PerTurbo: A scalable Bayesian analysis framework for CRISPR screens with single cell RNA readouts.*

Abstract. Single-cell CRISPR screens combine CRISPR-Cas9 genetic perturbations with single cell RNA sequencing to directly test how genetic elements or variants impact gene expression. As the data collected from individual screens continues to grow exponentially, a need has emerged for computational methods that can efficiently analyze millions of single-cell transcriptomic profiles while remaining robust to technical confounders. To address this gap, we developed PerTurbo: a fast, fully Bayesian tool for estimating perturbation effects in single cell CRISPR screens. Like previous CRISPR-specific analysis methods, PerTurbo identifies differentially expressed genes (DEGs) between perturbed and unperturbed cells, while treating the observed guide RNA (gRNA) counts as noisy measurements of the true perturbation(s) received by each cell. But thanks to its efficient, GPU-accelerated implementation using stochastic variational inference, PerTurbo runs thousands of times faster, making it feasible to run transcriptome-wide DEG tests for each gRNA. Additionally, unlike other methods, PerTurbo tests for effects on both the mean and variance of transcript counts. We highlight PerTurbo's superior scalability and performance by performing transcriptome wide tests on several datasets to investigate perturbation effects on enhancers, genes, or disease-associated variants in modulating gene expressions at single-cell resolution.

Keywords: scRNA-seq, CRISPR, Perturb-seq, Bayesian modeling, Probabilistic programming

Matteo Lepur (BC Cancer Research Center, Department of Molecular Oncology), Kevin Yang (BC Cancer Research Center, Department of Molecular Oncology), Felix Fu (BC Cancer Research Center, Department of Molecular Oncology), Patricia Galipeau (Fred Hutchinson Cancer Research Center, FHIND Cancer), Adam Kreitzman (Fred Hutchinson Cancer Research Center, FHIND Cancer), Minjeong Ko (Fred Hutchinson Cancer Research Center, FHIND Cancer), Amy Paguirigan (Fred Hutchinson Cancer Research Center, FHIND Cancer), Vinci Au (BC Cancer Research Center, Department of Molecular Oncology), Viviana Cerda (BC Cancer Research Center, Department of Molecular Oncology), Esther Kong (BC Cancer Research Center, Department of Molecular Oncology), Daniel Lai (BC Cancer Research Center, Department of Molecular Oncology), Michael Van Vliet (BC Cancer Research Center, Department of Molecular Oncology), Elena Zaikova (BC Cancer Research Center, Department of Molecular Oncology), Alexandre Bouchard-Côté (University of British Columbia, Department of Statistics), Sam Apricio (BC Cancer Research Center, Department of Molecular Oncology), Gavin Ha (BC Cancer Research Center, Department of Molecular Oncology) and Andrew Roth (BC Cancer Research Center, Department of Molecular Oncology). *LiquidBayes: Integrated analysis of single whole genome sequencing and ctDNA.*

Abstract. Cancer is uncontrolled growth of a collection of cells. These cells can be parsed into distinct subgroups, called clones, characterized by a unique set of genetic mutations. Cancer treatments' success hinges on its ability to destroy each of these clones, however genetic variation across clones often leaves treatment ineffective. Hence, it is of clinical importance to understand the clonal structure of cancer.

Circulating tumour DNA (ctDNA) are fragments of tumour DNA found in blood. Serial samples of ctDNA can be obtained non-invasively via blood samples and provide a lens into the changing tumour abundance and clonal structure. Existing methods estimate tumour abundance using ctDNA but ignore its underlying clonal structure.

We introduce LiquidBayes, a Bayesian statistical model, that infers tumour abundance and clonal structure by integrating ctDNA with single-cell whole-genome sequencing (scWGS). LiquidBayes uses scWGS to infer clone specific copy number profiles and leverage this as prior information on clonal structure.

LiquidBayes outperformed two state of the art approaches, ichorCNA and MRDetect, on semi-synthetic data. We performed scWGS and ctDNA sequencing on pre- and post-treatment samples from a patient with triple negative breast cancer. LiquidBayes shows both a reduction in overall tumour abundance and shift in clonal structure after treatment.

Keywords: High-throughput sequencing, Machine learning and computational biology, Medical informatics

Zsolt Balazs (University of Zurich), Todor Gitchev (University of Zurich), Ivna Ivankovic (University of Zurich) and Michael Krauthammer (University of Zurich). *Fragmentstein - Facilitating data-reuse for cell-free DNA fragment analysis.*

Abstract. Motivation: Cell-free DNA (cfDNA) sequencing is a promising diagnostic and monitoring tool in cancer care. The development of novel computational and analytic workflows in the field is, however, impeded by limited data sharing due to the strict control of genomic data. While the analysis of copy number variants (CNV), nucleosome footprints, and fragmentation patterns from cfDNA sequencing data do not theoretically require actual sequence information, current bioinformatics software is generally developed to process alignment files containing sensitive sequence data.
Results: We present Fragmentstein, a lightweight command line tool for converting non-sensitive cfDNA fragmentation data, consisting only of cfDNA fragment coordinates, into sequence alignment mapping (SAM/BAM) files which most contemporary cfDNA sequencing analysis tools require as input. Fragmentstein merges fragment coordinates and mapping quality scores with sequence information from a reference genome to create SAM/BAM files that contain fragment coordinates from the sample but no sensitive genome sequence. To demonstrate the utility of Fragmentstein, we analyze a publicly available dataset and show that CNVs, nucleosome footprints, and fragment length features can be fully recovered from non-sensitive fragment data.

Keywords: Cell-free DNA, Data sharing, Fragmentomics, Data reuse

Dmytro Horyslavets (Institute of Molecular Biology and Genetics of NASU), Harun Mustafa (ETH Zurich), Mikhail Karasikov (ETH Zurich), Andre Kahles (ETH Zurich) and Alina Frolova (Institute of Molecular Biology and Genetics of NASU). *Sequence-read extraction from Counting de Bruijn graphs.*

Abstract. As the availability of biological sequencing data continues to grow at an exponential rate, efficient methods for storing, indexing, and analyzing this data have become increasingly important. Annotated de Bruijn graphs have emerged as a popular method for representing large sets of sequencing data, as they enable efficient storage of k-mer sets and their annotations in a compressed form. In turn, the Counting de Bruijn graph was developed as a generalization of the annotated de Bruijn graph, which allows supplementing each node-label relation with one or more attributes such as k-mer count or coordinates. The concept of the Counting de Bruijn graphs is utilized in the MetaGraph framework, which offers a unique approach to indexing global coordinates by utilizing a number of compression techniques for both the graph and the annotations. In this work, we present an algorithm for extracting read sequences from a Counting de Bruijn graph, which was implemented within the MetaGraph framework. This task is of critical importance as getting the read sequences of interest from which the graph was built would open new opportunities for downstream analysis after sequence search.

Keywords: k-mer sets, de Bruijn graphs, Counting de Bruijn graphs, MetaGraph, graph traversal, sequence search

Timothé Rouzé (CNRS, Univ Lille), Camille Marchet (CNRS) and Antoine Limasset (CNRS).
SuperSampler: efficient scaled sketches for metagenomics and extensive genomics compositional analysis.

Abstract. A challenge for Bioinformatics is to keep up with the amount of data generated by high throughput sequencing.
Being able to compare such volume of data remains a scalability challenge which is the focus of many methodological papers.
To achieve drastic memory cost reduction, a possibility is to transform documents into "sketches" of highly reduced sizes that can be quickly compared to compute the documents similarity with bounded error.
The most used tools rely on fixed sized sketches using techniques such as Minhash or HyperLogLog. However, those techniques have a relatively poor accuracy when the compared datasets are very dissimilar in size or content.
To cope with this problem, novel methods proposed to construct adaptive sketches, scaling linearly with the size of the input, by selecting a fraction of the documents' k-mers.
Several techniques were proposed to perform uniform sub-sampling with theoretical guarantees such as modimizer/modminhash, scaled minhash/FracMinHash.
With SuperSampler, we improve such schemes by combining them with the concept of super-k-mers thus drastically reducing resources usage (CPU, memory, disk).
In this poster, we show that SuperSampler can use an order of magnitude less resources than state of the art with equivalent results.

Keywords: Local sensitive hashing, Sketching, Large scale, Genome comparison

Arnab Chakrabarti (Centrum für Integrierte Onkologie (CIO) Köln, RWTH Aachen University), Hiroshi Hamano (RWTH Aachen), Lancelot Seillier (University Hospital RWTH Aachen) and Kjong-Van Lehmann (Centrum für Integrierte Onkologie (CIO) Köln, Uniklinik Köln). *Estimate mutational signature exposure from sparse clinical sequencing data.*

Abstract. A typical analysis estimates the presence of known mutational signatures in each sample. However, current approaches rely on a large number of mutations to accurately estimate mutational signature exposure. Making this analysis possible when only sparse mutation data are available, such as data generated from panel sequencing or samples with low mutational burden, requires novel developments in the current methodologies for estimating mutational signature exposures. Here we present our work of assessing signature exposures using a novel predictive modeling approach. Our strategy follows two main steps. First, using a statistical model, we identify relevant signals from cancer mutations based on a mutational signature reference catalog (e.g., COSMIC [2]). Second, we use these mutational signals to train a predictive model. The model aims to estimate informative regions with respect to mutational signatures from the cancer genome sequence that are being considered when estimating the mutational signature exposure on a single sample.

Keywords: cancer genomics, mutational signatures, rna sequencing

Lin Zhang (Tohoku University), Hafumi Nishi (Tohoku University; Ochanomizu University) and Kengo Kinoshita (Tohoku University). *Identification and validation of heterogeneous neutrophils by integrated analysis of single-cell and bulk RNA-sequencing in COVID-19.*

Abstract. The coronavirus disease (COVID-19) can alter leukocyte phenotype in terms of the disease severity, including neutrophil activation signatures in severe cases. In recent years, accumulating evidence has revealed unexpected phenotypic heterogeneity and diverse functions of neutrophils in many other diseases. However, the complexity of neutrophil phenotypes and their relative impacts on COVID-19 pathogenesis have not been well addressed. Here, we integrated public single-cell and bulk RNA-sequencing data from healthy donors and COVID-19 patients to investigate neutrophil heterogeneity and uncover how they contribute to disease pathogenesis. We identified and described eight neutrophil subtypes (namely C1-C8). The eight subtypes exhibited different activation signatures, subtype compositions, and enriched pathways according to COVID-19 severity. C4 neutrophil subtype was associated with severe and fatal patients. Cell-cell communication analysis revealed different neutrophil phenotypes of the eight subtypes, such as the transmembrane receptor expression of CD45 and secretion of PPBP in the C4 fraction. In addition, the bulk RNA-seq datasets analysis using a cellular deconvolution approach validated the relative abundances of neutrophils and expansion of the C4 fraction in severe COVID-19 patients. Our work provides a framework to understand the functional heterogeneity of neutrophils and sheds light on the prevention and treatment of COVID-19.

Keywords: Single-cell RNA-seq, Bulk RNA-seq, Neutrophil heterogeneity, COVID-19, Cellular deconvolution

Davide Cozzi (Università degli studi di Milano Bicocca), Massimiliano Rossi (University of Florida), Simone Rubinacci (University of Lausanne), Travis Gagie (Dalhousie University), Dominik Köppl (Tokyo Medical and Dental University), Christina Boucher (University of Florida) and Paola Bonizzoni (Università degli Studi di Milano-Bicocca). *μ-PBWT: a lightweight r-indexing of the PBWT for storing and querying UK Biobank Data.*

Abstract. Motivation: The positional Burrows-Wheeler Transform (PBWT) is a data structure that indexes haplotype sequences in a manner that enables finding maximal haplotype matches in h sequences containing w variation sites in O(hw)-time, by Durbin's Algorithm 5. This represents a significant improvement over classical quadratic-time approaches. However, the original PBWT data structure does not allow for queries over biobank panels that consist of several millions of haplotypes, if an index of the haplotypes must be kept entirely in memory.
Results: We leverage the notion of r-index proposed for the BWT to present a memory efficient method for constructing and storing the run-length encoded PBWT, and computing one-vs-all set maximal matches (SMEMs) queries in haplotype sequences. We implement our method, which we refer to as μ-PBWT, and evaluate it on datasets of 1000 Genome Project and UK Biobank data. Our experiments demonstrate that the μ-PBWT reduces the memory usage up to a factor of 20% compared to the best current PBWT-based indexing. In particular, μ-PBWT produces an index that stores high-coverage whole genome sequencing data of chromosome 20 in about a third of the space of its BCF file.
Availability: https://github.com/dlcgold/muPBWT and https://bioconda.github.io/recipes/mupbwt/README.html

Keywords: Succinct data structures, Burrows–Wheeler transform, Positional Burrows–Wheeler transform, Pattern matching

Jim Shaw (Department of Mathematics, University of Toronto), K. Jun Gao (Department of Computer Science, University of Toronto), Jared Simpson (Ontario Institute for Cancer Research; Department of Molecular Genetics, University of Toronto) and Yun William Yu (Department of Mathematics, University of Toronto). *ChromMiniGraph: Space-Efficient Minimizer-based Pangenome Reference Graph and Haplotype Mapping Tool.*

Abstract. Advances in sequencing technologies have enabled construction of individualized references representing genetic variations across populations. However, existing graph genome software has the disadvantage of being memory and storage-intensive, as it often stores complete reference sequences along the graph. We introduce ChromMiniGraph, a tool for constructing space-efficient pangenome reference that uses k-mer sampling to reduce memory and storage requirements while maintaining accuracy. The tool constructs a directed acyclic graph through iterative chaining with colored nodes representing haplotypes, which can be then used to map sequencing reads. ChromMiniGraph maps reads by subsampling them and performing colinear chaining onto a topologically linearized coordinate of the graph using banded dynamic programming. To further improve chaining accuracy, ChromMiniGraph identifies superbubbles in the graph to augment the linearized coordinate with an alternative sparse distance matrix to score anchors that straddle superbubbles. ChromMiniGraph can correctly assign haplotypes to simulated reads with short indels and complex structural variations and successfully map PacBio reads from HG01243 to a reference graph constructed using Human Chromosome 20 reference sequences obtained from the Genome Reference Consortium and 1000 Genomes Project with accuracy and efficiency. Overall, ChromMiniGraph offers a streamlined workflow for creating and visualizing pangenome references, read phasing, and identifying structural variations.

Keywords: pangenome, graph genome, read phasing, pangenome reference graph, data structure, graph algorithms

Elise Amblard (CNRS, UMR 5525, VetAgro Sup, Grenoble INP, TIMC, University Grenoble Alpes, 38000 Grenoble, France), Vadim Bertrand (CNRS, UMR 5525, VetAgro Sup, Grenoble INP, TIMC, University Grenoble Alpes, 38000 Grenoble, France) and Magali Richard (CNRS, UMR 5525, VetAgro Sup, Grenoble INP, TIMC, University Grenoble Alpes, 38000 Grenoble, France). *A multi-block approach to improve deconvolution of cancer omic data.*

Abstract. Bulk transcriptomes and epigenomes are routinely measured in the clinic to diagnose and classify cancer patients. However, these classifications do not account for intra-tumor heterogeneity, i.e. proportions of the cell types mixed in a sample. This information is critical as it has an impact on the tumor behavior with respect to its evolution and treatment response.
The current state-of-the-art procedure to de-mix samples is to apply deconvolution algorithms on one block of data, either the transcriptome or the methylome. Nevertheless, there is no consensus on the best deconvolution method. In our project, we propose to combine both blocks. Our work hypothesis is that joint deconvolution should perform better. Indeed, we hope that more information, and from different nature, would improve the de-mixing task.
We face the following challenges: how to do joint deconvolution? How to compare multi-block versus simple-block approaches?
We first collected several deconvolution tools in the literature and tested different timings for the block integration step within the deconvolution pipeline. We then compared multi-block strategies with simple-block ones. Our first results showed that the multi-block approach exhibited superior performances.
Finally, we will use the output of multi-block deconvolution to build a more refined stratification of pancreatic cancer patients.

Keywords: deconvolution, multi-block, omic, cancer, benchmark, classification

Alejandro Paniagua (Institute for Integrative Systems Biology (CSIC-UV), Spanish National Research Council, Paterna, Valencia, Spain.), Francisco Pardo-Palacios (Institute for Integrative Systems Biology (CSIC-UV), Spanish National Research Council) and Ana Conesa (Institute for Integrative Systems Biology (CSIC-UV), Spanish National Research Council). *Evidence-driven annotation of the Trichechus manatus latirostris genome using long-reads.*

Abstract. Third-generation sequencing (TGS) has made draft genome production more accessible, however, the structural annotation of those genomes has not progressed at the same pace. This study explores the potential of using long-read RNA sequencing (lrRNA-seq) to improve gene annotation. The accuracy of lrRNA-seq supported gene prediction was evaluated for different sequencing platforms (PacBio or Nanopore), pipelines, and annotation approaches besides the gene prediction tool AUGUSTUS.
We used Nanopore and PacBio sequencing data from the WTC11 human cell line, processed independently and in combination using IsoSeq3 or FLAIR pipelines. Incorporating lrRNA-seq data during the gene prediction step significantly improved gene prediction accuracy, particularly with transcript models generated from PacBio long-reads.
This approach was then applied to annotate the Trichechus manatus latirostris genome, resulting in 25% and 12.3% more BUSCO genes than using only experimental data or ab initio predictions, respectively. The findings suggest that lrRNA-seq is a valuable source of experimental data for supporting gene annotation in mammalian species.

Keywords: Genome annotation, Third-generation sequencing, lrRNA-Seq, Evidence-driven annotation, Florida manatee

Maryna Chepeleva (Luxembourg Institute of Health), Tijana Randic (University of Luxembourg), Demetra Philippidou (University of Luxembourg), Reka Toth (Luxembourg Institute of Health), Stephanie Kreis (University of Luxembourg) and Petr Nazarov (Luxembourg Institute of Health).
Deconvolution of scRNA-seq data of melanoma cell lines links molecular profiles with drug resistance.

Abstract. Decrypting molecular processes in oncology is essential for precise diagnostics and deployment of the most effective targeted therapy that leads to improved treatment outcomes and reduced risks of adverse effects. However, inter- and intra-tumor heterogeneity hides underlying mechanisms of tumor drug response and adaptation or resistance. Here we linked molecular profiles of melanoma with drug resistance.
We analyzed single-cell RNAseq data of four NRAS-mutant melanoma cell lines (MelJuso, Sklmel30, M20, IPC298) in four stages of treatment with MEK1/2 plus CDK4/6 inhibitors, applying the previously developed R/Bioconductor package consICA. This package implements a reference-free deconvolution method that separates mixed molecular profiles into statistically independent signals. We were able to map single cells on a cell cycle and observed a strong linkage between the proportion of proliferating cells and adaptation to the treatment, which occurred with individual speed for each cell line. In three of four cell lines, we observed increased motility in resistant samples. We also observed several signals linked to ATP synthesis and metabolism that were modulated by the treatment and resistance. Interestingly, gene signals involved in mRNA processing and chromatin remodeling were mainly down-regulated in resistant cells, suggesting potential changes at the epigenetic level.

Keywords: deconvolution, transcriptomics, melanoma, drug resistance

Hai C. T. Nguyen (UNIST), Bukyung Baik (UNIST) and Dougu Nam (UNIST). *Benchmarking integration of single-cell differential expression.*

Abstract. Integration of single-cell RNA sequencing data between different samples has been a major challenge for analyzing cell populations. However, strategies to integrate differential expression analysis of single-cell data remain underinvestigated. Here, we benchmark 46 workflows for differential expression analysis of single-cell data with multiple batches. We show that batch effects, sequencing depth and data sparsity substantially impact their performances. Notably, we find that the use of batch-corrected data rarely improves the analysis for sparse data, whereas batch covariate modeling improves the analysis for substantial batch effects. We show that for low depth data, single-cell techniques based on zero-inflation model deteriorate the performance, whereas the analysis of uncorrected data using limmatrend, Wilcoxon test and fixed effects model performs well. We suggest several high-performance methods under different conditions based on various simulation and real data analyses. Additionally, we demonstrate that differential expression analysis for a specific cell type outperforms that of large-scale bulk sample data in prioritizing disease-related genes.

Keywords: Single cell RNA-seq, Transcriptomics, Batch integration, Differential expression analysis

Karoliina Salenius (Tampere University), Reija Autio (Tampere University), Jake Lin (Tampere University), Christophe Roos (Euformatics Oy), Jussi Volanen (Euformatics Oy) and Matti Nykter (Tampere University). *Improving Genomic Variant Detection efficiency: Insights from Alignment Tool and Variant Caller Comparisons.*

Abstract. Advances in next-generation sequencing have increased the usage of whole genome sequencing (WGS) for studying disease-related polymorphisms. Accurate detection of genomic variants is imperative, and selecting the appropriate bioinformatic pipeline is non-trivial. We assessed three alignment tools (bwa-mem, minimap2 and dragmap-os) and three variant callers (GATK, GATK-DRAGEN and DeepVariant) using Genome in a Bottle consortium datasets and real data.

GATK showed lowest accuracy for indels, whereas SNV results were similar across pipelines with slight improvements in DeepVariant. Base quality score recalibration significantly increased computational time and had adverse effects on DeepVariant's accuracy. Filtering difficult genomic regions reduced variant calling time for GATK tools, but had no effect on DeepVariant, which is the fastest tool when GPU is available.

All aligners produced equally good results with minimal differences. Minimap2 was the fastest of the three, while bwa-mem and dragmap-os had similar runtimes. Notably, aligning data with bwa-mem is still recommended for some structural variant detection methods.

Overall, DeepVariant consistently performed well across aligners with high sensitivity and precision. It required no region filtering and had compatible resource requirements. The choice of preprocessing pipeline depends on the study's requirements, emphasizing the need for careful consideration of downstream analysis right from the start.

Keywords: Next-generation sequencing (NGS), Whole genome sequencing (WGS), Genomic variants, Benchmarking, Variant calling, Alignment, Data preprocessing, Short read sequencing

Risa K Kawaguchi (Kyoto University). *Large-scale integration analysis of epigenetic data for inferring the dosage-dependent binding of reprogramming transcription factors.*

Abstract. Transcription factors (TFs), which can bind to specific DNA regions to control epigenetic and transcriptomic regulation, have been widely studied. High-throughput epigenetic analyses such as ChIP-seq and CUT&Tag can identify DNA binding sites of a number of TFs, followed by the prediction of their binding motifs. However, the detected binding sites of the same protein has a variety across studies, potentially due to the conditional differences and/or batch effects. Recently, the dosage-dependent binding of TFs gains the attention to explain the variability of those binding sites. For example, BANC-seq revealed the change of binding profiles of several TFs in nanomolar range, suggesting the important role of chromatin context and DNA sequence for TF binding.
In this study, we performed the large-scale comparison of ChIP-seq datasets of TFs and their family genes. By focusing on the reprogramming factors, our ongoing analysis suggests that the integration of binding profiles of different experiments can be a clue for clarifying the impacts of biological contexts to determine the binding probability of TFs at each site. Comparing the co-expression patterns of the TF family genes, it is aimed at predicting the pioneering ability and dosage dependency of each TF and its interacting proteins.

Keywords: transcription factor, reprogramming factor, ChIP-seq, epigenomics

Katrin Frauenknecht (National Center of Pathology (NCP), Laboratoire national de santé), Michel Valtey (Luxembourg Centre of Neuropathology (LCNP) , Laboratoire national de santé), Camille Cialini (Luxembourg Centre of Neuropathology (LCNP), Luxembourg Institute of Health), Arnaud Muller (Bioinfo Platform, Luxembourg Institute of Health (LIH)), Lorraine Richart (Luxembourg Centre of Neuropathology (LCNP), Luxembourg Institute of Health), Michel Mittelbronn (National Center of Pathology (NCP), Laboratoire national de santé), Petr Nazarov (Multiomics Data Science Group, Bioinfo Platform, Luxembourg Institute of Health (LIH)) and Reka Toth (Multiomics Data Science Group, Bioinfo Platform, Luxembourg Institute of Health (LIH)). *Estimation of originating cell composition from cfDNA methylomes.*

Abstract. Nanopore sequencing has emerged as a significant technique for DNA methylation analysis, as it enables the detection of base modifications without the need for additional conversion steps. This allows for the measurement of methylation using a small amount of input material, making it increasingly useful in liquid biopsies to detect cell-free DNA (cfDNA) methylation patterns. Such patterns hold promise for the early diagnosis and monitoring of tumors originating from various sources. Identifying the originating cell composition (OCC) is a key step; however, its accurate prediction poses a challenge due to data sparsity and limited overlapping CpG sites across samples. To address this challenge, we have developed a method that leverages the methrix R package's speed and efficacy to fit individual models based on the non-missing sites in each sample, thereby estimating OCC efficiently. This method can utilize both array and sequencing-based references and works with any sequencing-based cfDNA methylomes. Although our estimates demonstrate that even 1 million reads suffice for accurate OCC prediction, we have incorporated quality control measures to assess the discriminatory ability of covered CpGs in the reference dataset. Overall, our approach provides a simple way of OCC estimation based on the DNA methylation patterns of cfDNA.

Keywords: Nanopore, methylation, prediction, Originating cell composition

Ivna Ivanković (University of Zurich), Zsolt Balázs (University of Zurich), Todor Gitchev (University of Zurich), Norbert Moldovan (Cancer Center Amsterdam, Amsterdam UMC), Florent Moulière (Cancer Center Amsterdam, Amsterdam UMC) and Michael Krauthammer (University of Zurich).
The effects of bioinformatics preprocessing on cell-free DNA analysis.

Abstract. Cell-free DNA (cfDNA) is a valuable liquid biopsy biomarker for cancer diagnosis and monitoring, carrying genetic and epigenetic information released into the bloodstream from normal and cancerous cells. While cfDNA analysis often relies on DNA sequencing and subsequent bioinformatics processing, the effects of bioinformatics preprocessing on cfDNA measurements remain understudied. To investigate whether preprocessing choices affect cfDNA analysis outputs, we built a modular bioinformatics pipeline that evaluates a range of commonly used preprocessing settings. We evaluated the effect of preprocessing on low-pass whole-genome sequencing of plasma cfDNA (median coverage 2.38x). cfDNA fragment size, coverage, copy number changes and differential coverage analysis over DNase hypersensitivity sites were recovered in a cohort of 20 lung cancer and 20 healthy cfDNA samples. We found that the analyzed features remain robust to preprocessing such as read trimming and reference genome builds. However, we observed that strict alignment filtering improves the differentiation between cancer and healthy samples for coverage-related features, such as differential coverage analysis over DNase hypersensitivity sites. In conclusion, our findings indicate that bioinformatic preprocessing choices in cfDNA analysis have minimal impact on distinguishing cancer and healthy samples with only few features benefitting from strict alignment filtering.

Keywords: Cell-free DNA, Bioinformatics pipeline, NGS

Dario Righelli (Department of Statistical Sciences, University of Padova), Kelly Eckenrode (Graduate School of Public Health and Health Policy, City University of New York, NY, NY, United States), Marcel Ramos (Graduate School of Public Health and Health Policy, City University of New York, NY, NY, United States), Ricard Argelaguet (European Bioinformatics Institute (EMBL-EBI), Hinxton, Cambridgeshire, CB10 1SD, UK), Christophe Vanderaa (de Duve Institute, Université catholique de Louvain, Avenue Hippocrate 75, Brussels, 1200, Belgium), Ludwig Geistlinger (Graduate School of Public Health and Health Policy, City University of New York, NY, NY, United States), Aedin Culhane (School of Medicine, University of Limerick, Limerick, Ireland), Laurent Gatto (de Duve Institute, Université catholique de Louvain, Avenue Hippocrate 75, Brussels, 1200, Belgium), Vincent Carey (Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts, United States), Martin Morgan (Roswell Park Comprehensive Cancer Institute, Buffalo, New York, United States), Davide Risso (Department of Statistical Sciences, University of Padova, Padova, Italy.) and Levi Waldron (Graduate School of Public Health and Health Policy, City University of New York, NY, NY, United States). *Curated Single Cell Multimodal Landmark Datasets for R/Bioconductor.*

Abstract. Most high-throughput sequencing methods for single-cells focus on quantifying gene expression. However, recent advancements in multimodal profiling techniques have enabled the simultaneous measurement of multiple -omics within the same cells. To facilitate the development of new statistical and computational methods for analyzing such data, it is important to have readily available landmark datasets that adhere to standard data classes.

We have gathered, processed, and compiled publicly available landmark datasets from several single-cell multimodal protocols, including CITE-Seq, ECCITE-Seq, SCoPE2, scNMT, 10X Multiome, seqFISH, and G&T-seq. These datasets are released as Bioconductor classes and are documented and distributed as the SingleCellMultiModal package through Bioconductor's ExperimentHub. This allows for the retrieval of landmark datasets from seven different single-cell multimodal data generation technologies using a single command, eliminating the need for additional data processing or manipulation, and allowing to analyze and develop methods within Bioconductor's extensive ecosystem.

We present two illustrative examples of integrative analyses that are greatly simplified by the use of SingleCellMultiModal. This package will facilitate the advancement of bioinformatic and statistical methods within the Bioconductor framework, enabling researchers to address the challenges associated with integrating multiple molecular layers and analyzing phenotypic outcomes, such as cell differentiation, activity, and disease.

Keywords: Single cell multimodal, Bioconductor, Data analysis, Genomics, Transcriptomics, Proteomics, Spatial transcriptomics, Bioinformatics

Chao-Jung Wu (UQAM), Hui-Wen Liu (Biogen), Lauren Tereshko (Biogen), Dongdong Lin (Biogen), Svetlana Bergelson (Biogen), Baohong Zhang (Biogen), Abdoulaye Baniré Diallo (Université du Québec à Montréal) and Cullen Mason (Biogen). *Assessing functional impurities in rAAV production platforms by long-read sequencing.*

Abstract. Recombinant adeno-associated virus (rAAV)-mediated gene therapy has been applied for human diseases. However, the rAAV capsids contain heterogeneous mixtures of full-length and truncated genomes and residual host cell and plasmid DNA, depending on the manufacturing process. Therefore, a method is needed to characterize the encapsidated DNA of rAAV in order to support process development and batch release. The emerging long-read sequencing (LRS) has achieved AAV single-genome resolution. Here we propose a Python-based LRS profiling framework to classify and quantitate residual DNA species in rAAV products. We designed a reference that contains universal genetic components that are commonly used in rAAV production, including AmpR, KanR, Rep and Cap genes along with HPV18, Ad5 and hg38 genomes. We accessed the impurities of rAAV production from public and in-house LRS datasets. Analyzing the lambda fragments supplemented in these datasets showed that sequencing introduced size biases, which couldn't be fully rescued by regression but is improvable within library preparation. Functional potential of impurities were assessed through indicators derived from long-read alignments, which enabled us to quantitatively compare impurities between manufacturing batches. We demonstrated that LRS provides informative metrics for rAAV production and can facilitate process development to ensure therapeutic product safety and quality.

Keywords: rAAV genotyping, residual DNA, long-read sequencing, PacBio, Nanopore, gene therapy vector quality

Ammarah Anwar (Universität Klinikum Düsseldorf (Department of pediatrics Oncology and Hematology)), Triantafyllia Brozou (Universität Klinikum Düsseldorf (Department of pediatrics Oncology and Hematology)), Layal Yasin (Universität Klinikum Düsseldorf (Department of pediatrics Oncology and Hematology)), Ute Fischer (Universität Klinikum Düsseldorf (Department of pediatrics Oncology and Hematology)), Arndt Borkhardt (Universität Klinikum Düsseldorf (Department of pediatrics Oncology and Hematology)) and Carolin Walter (Institut für Medizinische Informatik - Universität Münster). *Unveiling the Landscape of De Novo Mutations in Pediatric Cancer: A Bioinformatics Approach with Trio Exome Sequencing and Haplotype Phasing.*

Abstract. In this study, we conducted whole-exome sequencing to investigate the origins and inheritance patterns of de-novo mutations in pediatric cancer patients. We analyzed a cohort of 280 patients using Illumina short-read trio exome sequencing and identified variants using GATK/Varscan pipelines. Among the high impact SNP/indel variants, 0.36% were identified as de-novo. Our analysis revealed 22.5% of the patients had three de-novo mutations per patient, with some individuals having 28. Interestingly, 5.71% of patients did not exhibit any de-novo mutations. The estimated rate of exonic de novo sequence variants in our study was 6.67x10-8 per generation. We observed a positive correlation between paternal age and the number of de-novo mutations, while maternal age did not show a significant effect. Further investigation focused on Tier1-3 genes, encompassing 364 genes. These mutations were identified in 6.78% of the cohort, and pathogenic variants were found in TP53, SOS1, PTPN11 and MSH6 genes, potentially associated with leukemia (BCP-ALL) and glioblastoma. Moreover, our analysis revealed higher prevalence of de-novo mutations within CpG regions compared to non-CpG regions. By performing read-based haplotype phasing, we determined that 74.28% of phased de-novo mutations originated from the paternal lineage, while 25.7% originated from the maternal lineage.

Keywords: Pediatric cancers, De novo mutations, Read-based haplotype phasing, Linear regression analysis, CpG regions, Illumina short-read trio exome sequencing