High Throughput Sequencing
Algorithms and Applications
A special track of the ISMB 2022 meeting
Madison, WI, July 10-14, 2022

ISMB 2022 HiTSeq Track Proceedings

Madison, WI, United States
July 12-13, 2022
http://www.hitseq.org

Organizers:

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

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

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

Birte Kehr, Ph.D.
Regensburg Center for Interventional Immunology, Regensburg, Germany
E-mail: Birte.Kehr@klinik.uni-regensburg.de

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

Sebastian Schmidt (Helsinki University), Shahbaz Khan (Indian Institute of Technology Roorkee), Jarno Alanko (University of Helsinki) and Alexandru I. Tomescu (University of Helsinki). *Matchtigs: minimum plain text representation of kmer sets.*

Abstract. Kmer-based methods are widely used in bioinformatics, which raises the question of what is the smallest practically usable representation (i.e. plain text) of a set of kmers.

We propose a polynomial algorithm computing a minimum such representation (which was previously posed as a potentially NP-hard open problem), as well as an efficient near-minimum greedy heuristic.

When compressing genomes of large model organisms, read sets (Illumina short reads) thereof or bacterial pangenomes, with only a minor runtime increase, we decrease the size of the representation by up to 60% over unitigs and 27% over previous work.

Additionally, the number of strings is decreased by up to 97% over unitigs and 91% over previous work.

Finally, our small representation has advantages in downstream applications, as it speeds up queries on the popular kmer indexing tool Bifrost by 1.66x over unitigs and 1.29xover previous work.

Keywords: kmer sets, plain text compression, graph algorithm, sequence analysis, genomic sequences, minimum-cost flow, Chinese postman problem

Mingze Gao (The University of Hong Kong), Yuanhua Huang (The University of Hong Kong) and Chen Qiao (The University of Hong Kong). *UniTVelo: temporally unified RNA velocity reinforces single-cell trajectory inference.*

Abstract. The recent breakthrough of single-cell RNA velocity methods brings attractive promises to automatically identifying directed trajectory on cell differentiation, states transition and response to perturbations, which is uniquely demanded in in-vivo applications and abnormal conditions. However, the existing RNA velocity methods, including scVelo, are often found to return erroneous results, partly due to model violation of complex expression profiles or lack of temporal regularization. Here, we present UniTVelo, a statistical framework of RNA velocity that models the flexible transcription dynamics of spliced and unspliced RNAs via a spliced RNA oriented framework. Uniquely, it also supports the effective inference of unified latent time across genes and orders cells on individual genes in the phase portrait, especially for multiple-rate kinetics genes and those with stable and monotonic changes across the transcriptome. With ten datasets, we demonstrate that UniTVelo returns the expected trajectory in different biological systems, including hematopoietic differentiation and those even with weak kinetics or complex branches. Specifically, UniTVelo correctly identifies the differentiation trajectories of the human bone marrow development, from hematopoietic stem cells to three distinct branches. This system is complex and cannot be fully resolved by other currently available RNA velocity methods.

Keywords: RNA velocity, transcriptional dynamics, trajectory inference

Mohammed Alser (ETH Zurich), Jeremy Rotman (University of Southern California), Dhrithi Deshpande (University of Southern California), Kodi Taraszka (UCLA), Huwenbo Shi (Harvard University), Pelin Burcak Icer (ETH-Zurich), Harry Taegyun Yang (UCLA Department of Computer Science, Zarlab), Victor Xue (University of California Los Angeles), Sergey Knyazev (University of California, Los Angeles), Benjamin D. Singer (Northwestern University Feinberg School of Medicine), Brunilda Balliu (Leiden University Medical Center), David Koslicki (Penn State University), Pavel Skums (Georgia State University), Alex Zelikovsky (Georgia State University), Can Alkan (Bilkent University, Department of Computer Engineering), Onur Mutlu (Carnegie Mellon University, ETH Zurich) and Serghei Mangul (USC).
Technology dictates algorithms: recent developments in read alignment.

Abstract. Aligning sequencing reads onto a reference is an essential step in the majority of genomic analysis pipelines. Computational algorithms for read alignment have evolved in accordance with technological advances, leading to today's diverse array of alignment methods. We survey algorithmic foundations and methodologies across 107 alignment methods published between 1988 and 2021, for both short and long reads. We discuss the weakness and strengths of the algorithms using our rigorous experimental evaluation. We separately discuss how longer read lengths produce unique advantages and limitations to read alignment techniques.

Our review focuses on the interplay between technological development and algorithm development. It can explain the success behind popular read aligners, guide the choice of the most appropriate read alignment tools for particular problems, and identify new algorithmic research directions in response to the advancement of long-read technologies and novel sequencing protocols. It also discusses how general alignment algorithms have been tailored to the specific needs of various domains in biology, including whole transcriptome, adaptive immune repertoires of T and B cells receptors, and human microbiome studies.

Keywords: Genome analysis, Read mapping, Genome sequencing, Genome sequencing technologies, Transcriptome, Adaptive immune repertoire, RNA-Seq alignment, Metagenomic alignment, DNA Sequence alignment, Approximate string matching, Pre-alignment filtering, Read alignment

Sohta Nishida (Graduate School of Information Science and Technology, Osaka University), Junko Yoshida (Department of Physiology II, Nara Medical University), Shigeto Seno (Graduate School of Information Science and Technology, Osaka University), Kyoji Horie (Department of Physiology II, Nara Medical University) and Hideo Matsuda (Graduate School of Information Science and Technology, Osaka University). *A clustering method based on cell type identification for single-cell RNA-seq data.*

Abstract. Recent advances in single-cell RNA-seq (scRNA-seq) technology have made it possible to perform high-throughput, large-scale transcriptome profiling at single-cell resolution. Unsupervised learning, such as data clustering, is central to identifying and characterizing novel cell types and gene expression patterns. Clustering is used to computationally identify groups of cells by comparing the gene-expression profiles of the groups. The result of the clustering enables us to summarize complex scRNA-seq data into a digestible format for human interpretation. This allows us to describe population heterogeneity with discrete labels that are easier to understand, rather than trying to understand the higher-dimensional manifolds in which cells exist.

Clustering includes graph-based clustering and k-means clustering, in which the genes of each cell are clustered as features. However, these methods do not reflect the cell types in each cluster. Therefore, we have proposed a clustering method using the results of cell type identification as features. Cell type identification results in a score for each cell type. By using the scores, the method utilizes the proportion for clustering. The performance of the method will be demonstrated by applying the method to several benchmarking single-cell RNA-seq datasets.

Keywords: clustering, cell type identification, single-cell RNA-seq

Yu Chen (University of Alabama at Birmingham), Yiqing Wang (University of Alabama at Birmingham), Weisheng Chen (University of Alabama at Birmingham), Yuwei Song (University of Alabama at Birmingham), Yixin Zhang (University of Alabama at Birmingham), Herbert Chen (University of Alabama at Birmingham) and Zechen Chong (University of Alabama at Birmingham). *Gene fusion detection and characterization in long-read cancer transcriptomes with FusionSeeker*.

Abstract. Gene fusions are prevalent in a wide array of cancer types with different frequencies. They often play a critical role in tumorigenesis and progression, and some are serving as therapeutic targets. A large number of tools have been developed and applied to short-read cancer transcriptome sequencing data for gene fusion detection. However, it's always challenging to identify chimeric reads or discordant read pairs that represent gene fusions from short reads especially given the innate splicing structures of isoforms. Long-read RNA sequencing technologies, such as PacBio Iso-Seq and Nanopore direct RNA sequencing, can generate full-length transcript sequencing reads and may alleviate these issues, therefore showing great potential in gene fusion detection. However, to our knowledge, there are only two long-read gene fusion detection tools available, and their performance is limited in terms of sensitivity and precision. To take full advantage of long-read sequencing, we developed a novel method, FusionSeeker, to comprehensively characterize gene fusions in long-read cancer transcriptome data and reconstruct accurate fused transcripts from raw reads.

FusionSeeker consists of three major steps: raw signal detection, candidate event clustering and filtering, and transcript sequence reconstruction. FusionSeeker first scans the read alignments for split-read patterns and records raw signals of gene fusions if two alignments from the same read reside in two distinct genes. All raw signals are then clustered with a density-based clustering algorithm into candidate fusions. Candidates that cannot reach the minimum number of supporting reads are considered as noises and will be discarded, with only confident gene fusions left. For each event, FusionSeeker collects fusion-containing reads to generate an accurate consensus sequence with partial order alignment. The final output of FusionSeeker includes a list of confident gene fusion events and their transcript sequences. Benchmarked on three replicate simulated datasets, FusionSeeker achieved accuracy over 93% for gene fusion detection using both PacBio Iso-Seq and Nanopore direct RNA sequencing-like reads, which was consistently higher than the other two long-read gene fusion callers, JAFFAL and LongGF. FusionSeeker successfully reconstructed full-length transcript sequences for more than 99.5% of simulated gene fusion events, with an average sequence identity over 99%. The accurate transcript sequences also improved breakpoint accuracy, with 94% and 74.36% of the detected gene fusions having exact breakpoints on simulated Iso-Seq and Nanopore reads, respectively. We then applied the three long-read gene fusion detection tools on two cancer cell lines, SKBR-3 and MCF-7. In SKBR-3 cell line, FusionSeeker identified 15 previously validated gene fusion events, while JAFFAL and LongGF detected 13 and 10 events, respectively. When comparing gene fusion calls from the three tools, 19 events were reported only by FusionSeeker. 17 out of these 19 FusionSeeker-unique events were cross-validated in DNA sequencing data with a validation rate of 89.47%, which was higher than JAFFAL-unique (27.27%) and LongGF-unique events (60%). In particular, we observed a four-hop gene fusion event, CSNK2A1:NCOA3:MMP24OS:TSHZ2 , in SKBR-3 cell line, with all breakpoints located in intronic regions. In MCF-7 cell line, we designed PCR primers and validated 7 novel gene fusion events that have not been previously reported. In both SKBR-3 and MCF-7 cell lines, FusionSeeker transcript sequences showed significantly higher identity comparing to the reference gene sequences than the raw reads, indicating that most sequencing errors have been corrected by our transcript reconstruction module.

Our long-read gene fusion detection algorithm, FusionSeeker, can accurately identify gene fusions in both in silico and real long-read cancer transcriptome datasets, allowing comprehensive characterization of gene fusions and prediction of exact breakpoints. The accurate transcript sequences reported by FusionSeeker may facilitate the inference of the amino acid sequences and even the structures of the fused proteins.

Keywords: Gene fusion, Cancer genomics, Long-read sequencing, Transcriptome

Apurva Gopisetty (German Cancer Research Center (DKFZ) and German Cancer Consortium (DKTK), Heidelberg, Germany), Aniello Federico (German Cancer Research Center (DKFZ) and German Cancer Consortium (DKTK), Heidelberg, Germany), Didier Surdez (Institut Curie Research Centre, Paris, France), Elnaz Saberi-Ansari (Institut Curie Research Centre, Paris, France), Yasmine Iddir (Institut Curie Research Centre, Paris, France), Alexandra Saint-Charles (Institut Curie Research Centre, Paris, France), Anna-Lisa Böttcher (German Cancer Research Center (DKFZ) and German Cancer Consortium (DKTK), Heidelberg, Germany), Norman Mack (German Cancer Research Center (DKFZ) and German Cancer Consortium (DKTK), Heidelberg, Germany), Benjamin Schwalm (German Cancer Research Center (DKFZ) and German Cancer Consortium (DKTK), Heidelberg, Germany), Justyna Wierzbinska (Bayer AG, Pharmaceuticals, Research and Development, Berlin, Germany), Joshua Waterfall (Institut Curie Research Centre, Paris, France), Andreas Schlicker (Bayer AG, Pharmaceuticals, Research and Development, Berlin, Germany), Louis F Stancato (Eli Lilly and Company, Indianapolis, IN, USA), Gilles Vassal (Department of Clinical Research, Gustave Roussy, Villejuif, France), Natalie Jäger (German Cancer Research Center (DKFZ) and German Cancer Consortium (DKTK), Heidelberg, Germany), Gudrun Schleiermacher (Institut Curie Research Centre, Paris, France), Jan Koster (Department of Oncogenomics, Amsterdam University Medical Centre, Amsterdam, the Netherlands), Stefan M Pfister (German Cancer Research Center (DKFZ) and German Cancer Consortium (DKTK), Heidelberg, Germany) and Marcel Kool (German Cancer Research Center (DKFZ) and German Cancer Consortium (DKTK), Heidelberg, Germany). *ITCC-P4: Genomic profiling and analyses of pediatric patient tumor and patient-derived xenograft (PDX) models for high throughput in vivo testing*.

Abstract. Advancements in state-of-the-art molecular profiling techniques have resulted in better understanding of pediatric cancers and their drivers. Many new types and subtypes of pediatric cancers have been identified with distinct molecular and clinical characteristics.The ITCC-P4 consortium is a preclinical collaboration between academic centers across Europe and several pharmaceutical companies, with the overall aim to establish a sustainable platform of >400 molecularly well-characterized PDX models of high-risk pediatric cancers and use them for in vivo testing of novel mechanism-of-action based treatments.Currently, 340 models are fully established, including 87 brain and 253 non-brain tumor models, together representing different tumor types both from primary (113) and relapsed (92)/metastatic disease (42). 252 of these models have been fully molecularly characterized, representing 18 pediatric cancer entities and 43 different subtypes.Using low coverage whole-genome and whole exome sequencing, somatic mutation calling, DNA copy number, transcriptome analysis and methylation profiling we have observed that the molecular profile of most PDX models closely mimics their original tumors. Clonal evolution of somatic variants was only observed in some PDX-tumor pairs or between disease states. Somatic copy number variant analysis highlights specific alterations; for instance, MYB, MYC, MYCN, NTRK3, PTEN loss differently distributed between PDX-patient tumor pairs in high-grade gliomas.

Keywords: Cancer genomics, pediatric solid tumors, whole exome sequencing, patient-derived xenograft models, transcriptome analysis, methylation profiling

Renuka Gattu (Kakatiya University) and Shamitha Gangupanthula (Kakatiya University). *Next-Generation Sequencing (NGS) in populations of Indian Tropical Tasar Silkworm, Antheraea mylitta .*

Abstract. The tropical tasar silkworm, a semi-domesticated wild sericigenous insect,found in the form of 44 ecoraces In India, with variations in phenotypic traits.The wide range of distribution of the species has encountered diverse geographic and climatic variations of the distinct are as,leading to marked differences in not only phenotypical and physiological traits but also in the commercial and technological aspects. A.mylitta Drury,which is an exclusive ecorace of the states of Andhra Pradesh and Telangana, is well known for its superior commercial characters,but, is on the verge of extinction due to its weaknesses involtinism,emergence,hatching,lowyieldetc.Theecoraceconservationis essential to utilize their valuable genes in enhancing productivity and to build variation in new population through hybridization. Modern sequencing methods like NGS technologies and Insilco analysis are used in population genetic studies to investigate the evolutionary forces affecting genetic variation.In the present studies, the genomic DNA of parental ecoraces - Andhra local and Daba TV of A. mylitta and their hybrid populations were sequenced independently using the Illumina NextSeq500 in order to analyze their genetic relationship.The sequencing library revealed that the fragment size ranged between 200bp to 700bp and identified 35877 sites in 8 samples.Further, the phylogenetic tree showed closely and distantly related taxa among the populations.

Keywords: Antheraea mylitta, Ecoraces, Next-Generation Sequencing

Lu Pan (Karolinska Institutet), Trung Nghia Vu (Karolinska Institutet), Yudi Pawitan (Karolinska Institutet) and Huy Dinh (McCardle Laboratory for Cancer Research, Department of Oncology, University of Wisconsin). *Isoform-level quantification for single-cell RNA sequencing*.

Abstract. RNA expression at isoform level can potentially reveal cellular subsets and corresponding biomarkers that are not visible at gene level. However, due to the strong 3′ bias sequencing protocol, mRNA quantification for high-throughput single-cell RNA sequencing such as Chromium Single Cell 3′ 10× Genomics is currently performed at the gene level. We have developed an isoform-level quantification method for high-throughput single-cell RNA sequencing by exploiting the concepts of transcription clusters and isoform paralogs. The method, called Scasa, compares well in simulations against competing approaches including Alevin, Cellranger, Kallisto, Salmon, Terminus and STARsolo at both isoform- and gene-level expression. The reanalysis of a CITE-Seq dataset with isoform-based Scasa reveals a subgroup of CD14 monocytes missed by gene-based methods.

Keywords: Computational biology, Bioinformatics, Isoform quantification, Single-Cell RNA-Sequencing, scRNA-Seq

Valérie Marot-Lassauzaie (Berlin Institute for Medical Systems Biology, Max Delbrück Center in the Helmholtz Association, Berlin, Germany), Brigitte Joanne Bouman (Berlin Institute for Medical Systems Biology, Max Delbrück Center in the Helmholtz Association, Berlin, Germany), Fearghal Declan Donaghy (Berlin Institute for Medical Systems Biology, Max Delbrück Center in the Helmholtz Association, Berlin, Germany) and Laleh Haghverdi (Berlin Institute for Medical Systems Biology, Max Delbrück Center in the Helmholtz Association, Berlin, Germany). *κ-velo improves single-cell RNA-velocity estimation.*

Abstract. Single-cell transcriptomics has been used to study dynamical processes such as cell differentiation. RNA velocity (La Manno et. al. 2020) was a breakthrough towards obtaining a more complete description of the dynamics of such processes. Here, simultaneous measurement of new unspliced and old spliced mRNA adds a temporal dimension to the data. The change in mRNA abundance, called RNA velocity, is used to infer the progression of cells through the dynamical process. However, reliable velocity analysis is still impeded by multiple computational issues. State-of-the-art methods for velocity inference (Bergen et. al. 2020) have issues in velocity inference as well as visualisation. Moreover, there are inconsistencies in current processing pipelines and the single-cell specific (stochastic) part of the dynamic is lost through multiple layers of data smoothing.

We introduce a new method for RNA velocity analysis that addresses some of the issues in velocity estimation. We also propose that visualisation of the velocities based on the Nystroem projection method represents the single-cell stochasticity better than current practices. Finally, we adjust the processing pipeline for consistency with downstream velocity estimation. We validate our model on simulation and on real data, and compare it to current state-of-the-art.

Keywords: RNA velocity, single-cell sequencing, trajectory inference, transcriptional dynamics

Jana Musilova (Brno University of Technology, FEEC, Department of Biomedical Engineering), Xenie Kourilova (Brno University of Technology, FCH, Department of Food Chemistry and Biotechnology), Matej Bezdicek (University Hospital Brno, Department of Internal Medicine—Hematology and Oncology), Stanislav Obruca (Brno University of Technology, FCH, Department of Food Chemistry and Biotechnology) and Karel Sedlar (Brno University of Technology, FEEC, Department of Biomedical Engineering). *Comparison of short and long reads in structural and functional annotation of non-model bacteria*.

Abstract. DNA sequencing is a unique way to gain insight into the structure of the genome and the functions of an organism. In this study, we compared the widely used Illumina short reads and Oxford Nanopore long reads sequencing technologies in structural and functional annotation of non-model bacteria. We examined Schlegelella thermodepolymerans subspecies DSM 15264, LMG 21645, and CCUG 50061, non-model Gram-negative industrially utilizable representatives. Although these bacteria have a significant potential for the production of polyhydroxyalkanoates - degradable bioplastics by utilizing waste from the agro-food industry, assemblies of their genomes are not available.

The results revealed the Nanopore as the more efficient approach for initial genome characterization. Compared to Illumina, Nanopore revealed more structural genomic features and assigned more genes to the Clusters of Orthologous Groups (COGs). Moreover, Nanopore resulted in the largest contig and N50 many times higher and the number of contigs many times lower than Illumina assemblies. On the other hand, Nanopore sequencing has been shown to be error-prone. Consequently, assemblies of Nanopore's individual genomic features are less accurate, resulting in incomplete structural annotation and incorrect functional annotation in several cases. Illumina sequencing is, therefore, more applicable for detailed studies of specific genomic regions.

Keywords: Schlegelella thermodepolymerans polyhydroxyalkanoates, PHA, whole-genome assembly, Clusters of Orthologous Groups, Open Reading Frames

Jim Shaw (University of Toronto) and Yun William Yu (University of Toronto). *Theory of local k-mer selection with applications to long-read alignment* .

Abstract. Motivation:

Selecting a subset of k-mers in a string in a local manner is a common task in bioinformatics tools for speeding up computation. Arguably the most well-known and common method is the minimizer technique, which selects the 'lowest-ordered' k-mer in a sliding window. Recently, it has been shown that minimizers may be a sub-optimal method for selecting subsets of k-mers when mutations are present. There is however a lack of understanding behind the theory of why certain methods perform well.

Results:

We first theoretically investigate the conservation metric for k-mer selection methods. We derive an exact expression for calculating the conservation of a k-mer selection method. This turns out to be tractable enough for us to prove closed-form expressions for a variety of methods, including (open and closed) syncmers, (a, b, n)-words, and an upper bound for minimizers. As a demonstration of our results, we modified the minimap2 read aligner to use a more conserved k-mer selection method and performed a long-read transcriptome mapping experiment. Our results give new insights into how new k-mer selection strategies offer new parameterizations that can be used for optimizing speed and alignment quality.

Keywords: syncmers, minimizers, k-mers, long-read, mapping, sequence alignment, theory

Aarti Venkat (Tempus Labs Inc.), Daniel Cook (Google Health), Yannick Pouliot (Tempus Labs, Inc.), Pi-Chuan Chang (Google Health), Andrew Carroll (Google Health) and Francisco De La Vega (Tempus Labs, Inc.). *Accurate germline variant calling from RNA-Seq data using deep learning and Genome-in-a-Bottle reference cell-line data.*

Abstract. RNA-Seq is the leading technology for genome-wide transcript quantification and characterization. RNA-Seq data may contain useful information about transcribed genetic variants. However, accurate variant calling from RNA-Seq data is challenging due to the huge variation in depth of coverage. We leveraged DeepVariant to call variants by retraining a previous whole-exome sequencing CNN model with RNA-Seq alignments. We trained on data from RNA-Seq data from three cell lines used as sources of the Genome-in-a-Bottle (GiaB) reference materials. To represent assay variability, we sequenced HG002 and HG005 in triplicate and HG001 in 10 replicates. Benchmarking shows that our training improves the F1 score for all coding regions from 0.08 for the initial whole exome sequencing starting model to 0.64 after the training cycle. With a genotype quality score threshold set to provide a ≤1.5% false discovery rate, we obtained a sensitivity of 37% for all coding regions and 92% for coding regions of highly expressed genes. Our results show that DeepVariant models trained with RNA-Seq data with high quality truth sets can deliver accurate germline variant calls.

Keywords: Variant calling, RNAseq, Deep Learning

Carolin Walter (Westfälische Wilhelms-Universität Münster) and Julian Varghese (Westfälische Wilhelms-Universität Münster). *Basic4CVis: an R/Shiny app for 4C-seq quality control and interaction visualization*.

Abstract. Circular chromosome conformation capture with high-throughput sequencing (4C-seq) is a next-generation sequencing technique that offers detailed insights into the three-dimensional structure of the genome around a chosen viewpoint. Since 4C-seq is characterized by a semi-quantitative, fragmented data structure and technical biases that distort the actual signal, the analysis strategies have to be adapted accordingly. 4C-seq replicate experiments are invaluable for the detection of regular and differential interactions between conditions, but add additional challenges to the analysis.

We present Basic4CVis, an R/Shiny app for the analysis and visualization of 4C-seq data. The R package offers routines for the filtering and quality control of a 4C-seq experiment's virtual fragment library data, related statistical overviews per sample, functionality for both near-cis and far-cis visualization of single samples and replicate interactions, and a user-friendly graphical user interface that allows to display overlaps and specific interactions for settings with multiple conditions and differential interactions. While standard data preprocessing and virtual fragment library generation are conducted with the R/Bioconductor package Basic4Cseq, sets of interacting regions can be imported from other 4C-seq analysis algorithms or text files for visualization purposes. Thus, Basic4CVis is a flexible addition to 4C-seq analyses with replicates or multiple conditions.

Keywords: 4C-seq, quality control, visualization, R/Shiny

Kerui Peng (University of Southern California) and Serghei Mangul (University of Southern California).
Rigorous benchmarking of T cell receptor repertoire profiling methods for cancer RNA sequencing.

Abstract. The ability to identify and track T cell receptor (TCR) sequences from patient samples becomes central to the field of cancer research. The available high-throughput method to profile T cell receptor repertoires is TCR sequencing. However, the available TCR-Seq data is limited compared to RNA sequencing. We have benchmarked the ability of RNA-Seq-based methods to profile TCR repertoires by examining 19 bulk RNA-Seq samples across four cancer cohorts including both T cell rich and poor tissues. We have performed a comprehensive evaluation of the existing RNA-Seq-based repertoire profiling methods using targeted TCR-Seq as the gold standard. We also highlighted scenarios under which the RNA-Seq approach is suitable and can provide comparable accuracy to the TCR-Seq approach. Results show that these methods are able to effectively capture the clonotypes and estimate the diversity of TCR repertoires, as well as provide relative frequencies of clonotypes in T cell rich tissues and monoclonal repertoires. However, these methods have limited power in T cell poor tissues, especially in polyclonal repertoires. The results of our benchmarking provide an appealing argument to incorporate RNA-Seq into immune repertoire screening of cancer patients as it offers knowledge into transcriptomic changes that exceed the limited information provided by TCR-Seq.

Keywords: T cell receptor repertoire sequencing, RNA sequencing, High-throughput technologies

Kristen Beck (IBM), Edward Seabolt (IBM), Akshay Agarwal (IBM Corp), Gowri Nayar (IBM Research), Simone Bianco (IBM), Harsha Krishnareddy (IBM), Timothy Ngo (IBM), Mark Kunitomi (IBM), Vandana Mukherjee (IBM Research, Almaden) and James Kaufman (IBM). *Semi-Supervised Pipeline for Autonomous Annotation of SARS-CoV-2 Genomes.*

Abstract. SARS-CoV-2 sequencing has scaled dramatically, yet existing genome annotation methods can result in missing or incorrect gene/protein sequences. To overcome this limitation, we developed a novel semi-supervised pipeline for automated gene, protein, and functional domain annotation of SARS-CoV-2 genomes that is reference-free and overcomes atypical genomic traits. With this, we analyzed 66,000 genomes and identified the comprehensive set of known proteins with 98.5% set membership accuracy and 99.1% accuracy in length prediction, compared to proteome references, including Replicase polyprotein 1ab (with its transcriptional slippage site). Compared to Prokka (base) and VAPiD, we yielded 6.4- and 1.8-fold increase in protein annotations. Our method generated 13,000,000 gene, protein, and domain sequences—some conserved spatiotemporally and others representing emerging mutations e.g. D614G and N501Y. For spike glycoprotein domains, we achieved >97.9% reference sequence identity and characterized RBD variants. We demonstrated robustness and extensibility on an additional 4,000 genomes spanning eight variants of concern and interest. In this cohort, we successfully identified all keystone spike glycoprotein mutations with >99% accuracy and demonstrated high protein and domain annotation accuracy. This work comprehensively presents the molecular targets to refine biomedical interventions for SARS-CoV-2 with a scalable, high-accuracy method to analyze newly sequenced infections as they arise.

Keywords: high throughput method, COVID-19, viral genomics, semi-supervised, genome annotation, autonomous

Yan Yang (Tempus Labs, Inc.), Len Trigg (Real Time Genomics, Inc.), Kurt Gaastra (Real Time Genomics, Inc.), Sean Irvine (Real Time Genomics, Inc.), Gene Selkov (Tempus Labs, Inc.), Kyung Choi (Tempus Labs, Inc.), Robert Huether (Tempus Labs, Inc.) and Francisco De La Vega (Tempus Labs, Inc.). *Accurate genotyping of UGT1A1 dinucleotide repeat polymorphism from targeted NGS data for the assessment of irinotecan chemotherapy adverse events.*

Abstract. The gene UGT1A1 encodes the enzyme responsible for the glucuronidation of SN-38, the active metabolite of IRI. Wild-type UGT1A1 contains six TA repeats [A(TA)6TAA] in its promoter region. Polymorphic UGT1A1 alleles with a higher number of TA repeats, such as UGT1A1 *28 /(TA)7 and *37/(TA)8 alleles, cause decreased enzyme activity and are associated with adverse events of irinotecan, a chemotherapy drug. Genotyping of UGT1A1 polymorphisms from NGS data is challenging due to artifacts from DNA polymerase slippage. We developed a novel method, BayeSTR, to call accurate UGT1A1 repeat genotypes from target capture NGS data. BayeSTR analyzes read alignments to a graph-based model representing the possible repeat alleles, applies an empirically derived “stutter” denoising model, and then performs genotype calling by a Bayesian model. We validated our method with germline data from the Tempus xT tumor-normal matched NGS test, which targets 648 cancer related genes including the UGT1A1 promoter. We observed 100% accuracy through analysis of sequencing data from a collection of 54 Coriell cell-line DNA samples whose UGT1A1 genotypes were established orthogonally. BayeSTR allows for automated, accurate UGT1A1 promoter genotyping from targeted NGS data.

Keywords: UGT1A1 promotor, repeat regions, BayeSTR, IRI-induced adverse events

Yukun Tan (UT MD Anderson Cancer Center), Vakul Mohanty (UT MD Anderson Cancer Center), Shaoheng Liang (UT MD Anderson Cancer Center), Kun Hee Kim (UT MD Anderson Cancer Center), Jun Ma (UT MD Anderson Cancer Center), Marc Jan Bonder (German Cancer Research Center), Xinghua Shi (Temple University), Zechen Chong (University of Alabama at Birmingham) and Ken Chen (UT MD Anderson Cancer Center). *novoBreak-rna: local assembly for novel splice junction detection from RNA-seq data*.

Abstract. Splice junction, govern the process of removing introns by the RNA splicing machinery, is a vital component of eukaryotic genes. Identification of splice junction provides valuable insights of alternative splicing and fusion transcripts events, which have been found in most of the hallmarks of cancer and can potentially apply to cancer diagnosis, prognosis, and therapy. However, most of the available tools for splice junction detection directly align paired-end short reads to the genomic reference and identify the splice junctions from the discordant read pairs. Although computationally efficient, alignment-based approaches are fundamentally limited in detecting sequences that are substantially different from the reference, as such are most likely containing splice junctions due to the challenges in accurately splitting and aligning short fragments. On the other hand, the de novo whole transcriptome assembly approach, attempting to assemble all reads into a single consensus transcriptome, is computationally intensive. In this study, we proposed a local assembly-based framework, called novoBreak-rna, which modify our well-attested genomic structural variation breakpoint assembly tool novoBreak to assemble novel splice junctions in RNA-seq data. The results using real data of prostate cancer from TCGA demonstrate that our method can achieve higher sensitivity to detect the novel splice junctions.

Keywords: Splice junction, Local assembly, RNA-seq

Yeonghun Lee (Gwangju Institute of Science and Technology) and Hyunju Lee (Gwangju Institute of Science and Technology). *Integrative reconstruction of cancer genome karyotypes using InfoGenomeR*.

Abstract. Annotation of structural variations (SVs) and base-level karyotyping in cancer cells remains challenging. Here, we present Integrative Framework for Genome Reconstruction (InfoGenomeR)-a graph-based framework that can reconstruct individual SVs into karyotypes based on whole-genome sequencing data, by integrating SVs, total copy number alterations, allele-specific copy numbers, and haplotype information. Using whole-genome sequencing data sets of patients with breast cancer, glioblastoma multiforme, and ovarian cancer, we demonstrate the analytical potential of InfoGenomeR. We identify recurrent derivative chromosomes derived from chromosomes 11 and 17 in breast cancer samples, with homogeneously staining regions for CCND1 and ERBB2, and double minutes and breakage-fusion-bridge cycles in glioblastoma multiforme and ovarian cancer samples, respectively. Moreover, we show that InfoGenomeR can discriminate private and shared SVs between primary and metastatic cancer sites that could contribute to tumour evolution. These findings indicate that InfoGenomeR can guide targeted therapies by unravelling cancer-specific SVs on a genome-wide scale. This paper was published in Nat Commun 12, 2467 (2021) https://doi.org/10.1038/s41467-021-22671-6.

Keywords: Whole-genome sequencing, Genome reconstruction, Structural variations

Yashna Paul (AbbVie Deutschland GmbH & Co. KG, Genomics Research Center, Knollstrasse, 67061 Ludwigshafen), Gen Lin (AbbVie Deutschland GmbH & Co. KG, Genomics Research Center, Knollstrasse, 67061 Ludwigshafen), Maya Woodbury (AbbVie, Cambridge Research Center, 200 Sidney Street Cambridge, MA 02139), Robert Talanian (AbbVie, Cambridge Research Center, 200 Sidney Street Cambridge, MA 02139), Knut Biber (AbbVie Deutschland GmbH & Co. KG, Genomics Research Center, Knollstrasse, 67061 Ludwigshafen), Janina S. Ried (AbbVie Deutschland GmbH & Co. KG, Genomics Research Center, Knollstrasse, 67061 Ludwigshafen) and Astrid Wachter (AbbVie Deutschland GmbH & Co. KG, Genomics Research Center, Knollstrasse, 67061 Ludwigshafen). *snRNA-seq resolved glial and neuronal communication changes during Alzheimer's disease progression.*

Abstract. In Alzheimer's disease (AD), reactive microglia and astrocytes are suggested to disrupt neuronal functions potentially leading to neurodegeneration and cognitive decline. As microglia and astrocytes may act together in the disease process, we aimed to identify ligand-receptor interaction pairs involved in AD pathology by using single nuclei RNA sequencing (snRNA-seq) NeuN-negative profiles of glial cells from brain tissue of 18 AD and control donors.

Six permutation-based approaches implemented in the LIANA framework (CellChat, Connectome, iTALK, CellPhoneDB, NATMI and SCA) that assigned cell-cell interaction scores were used in combination to identify astrocyte-microglia subtype interactions specific to AD. A public snRNA-seq study including 24 AD and 24 control donors (Mathys et al., 2019) was not only used to validate these interactions but was further utilized to infer glial-neuronal interactions. Interactions associated with progression of AD were identified by correlating interaction scores to pathological determinants such as APOE status, Braak stage, total tangle and total plaque. Interactions uniquely occurring in early and late pathology AD indicated involvement of specific biological processes in different disease stages.

This study highlights human glial-neuronal interactions with known AD GWAS hits, drug targets or association to AD pathology.

Keywords: snRNA-seq, glial-neuronal interactions, Alzheimer's Disease

Taylor Reiter (University of Colorado Anschutz Medical Campus) and Casey Greene (University of Colorado Anschutz Medical Campus). *Building a compendium of publicly available microbial isolate RNA-seq data*.

Abstract. Researchers who focus on model organisms greatly benefit from compendia like recount, GTEx, and The Cancer Genome Atlas, which improve the findability and decrease the analysis burden for gene expression data from different experiments. While gene expression compendia exist for some bacterial model organisms like Escherichia coli and Pseudomonas aeruginosa, no compendium exists that unites gene expression profiles across all bacterial and archaeal species. We produced a compendium that integrates the 59,239 publicly available isolate bacterial and archaeal RNA-seq samples, creating a community resource that stands to improve data access and decrease time-to-insight for researchers interested in microbial gene expression. The main product of the pipeline is a normalized ortholog count table that includes all processed samples. Additionally, to support cross-domain and domain-specific inquiries the pipeline allows flexible data outputs. These include strain- or species-specific count tables and interconversion between annotation formats (e.g. ortholog to reference genome). All research products, including strain profiles, reference pangenomes, raw and normalized compendia, annotation maps, and analysis code will be made publicly available. Our pipeline is encoded in Snakemake and is available at github.com/greenelab/2022-microberna.

Keywords: gene expression, compendium, automated workflow

Lixing Yang (University of Chicago). *Mutational signatures of complex genomic rearrangements in human cancer*.

Abstract. Complex genomic rearrangements (CGRs) are common in cancer and are known to form via two aberrant cellular structures—micronuclei and chromatin bridge. However, which mechanism is more relevant to CGR formation in cancer and whether there are other undiscovered mechanisms remain unknown. Here we developed a computational algorithm 'Starfish' to analyze 2,014 CGRs from 2,428 whole-genome-sequenced tumors and discover six CGR signatures based on their copy number and breakpoint patterns. Through extensive benchmarking, we show that our CGR signatures are highly accurate and biologically meaningful. Three signatures can be attributed to known biological processes—micronuclei- and chromatin-bridge-induced chromothripsis and circular extrachromosomal DNA. More than half of the CGRs belong to the remaining three signatures not been reported previously. A unique signature, we named "hourglass chromothripsis", with localized breakpoints and small amount of DNA loss is abundant in prostate cancer. We find SPOP is associated with hourglass chromothripsis and may play an important role in maintaining genome integrity.

Keywords: Structural variations, Cancer genomics, Complex genomic rearrangements, Chromothripsis, Mutational signatures, Genome instability

Nicholas Abad (Division of Applied Bioinformatics, German Cancer Research Center (DKFZ), Heidelberg, Germany), Cindy Körner (Division of Molecular Genome Analysis, German Cancer Research Center (DKFZ), Heidelberg, Germany) and Lars Feuerbach (Division of Applied Bioinformatics, German Cancer Research Center (DKFZ), Heidelberg, Germany). *Identifying Functional, Non-Coding Somatic Single Nucleotide Variants through the REMIND-Cancer Bioinformatics Pipeline.*

Abstract. Although current personalized cancer treatment approaches primarily target mutations in protein-coding regions, the relevance of non-coding regulatory regions has been previously demonstrated. However, the ability to detect these mutations through statistical methods is limited due to their low recurrence and missing statistical power. We overcome this by applying the REMIND-Cancer Pipeline, which is an integrative computational pipeline that combines genomic, transcriptomic and chromatin accessibility information to identify functional promoter mutations. The pipeline consists of three major steps: (1) exclude all mutations that show no potential for increasing gene expression by modifying promoter sequences, (2) rank the remaining candidates by a multivariate scoring function, and (3) allow for the in-depth analysis of the top scoring mutations through a multi-functional visualization tool. We analyzed the publicly-available PCAWG dataset, which consists of 2,583 patients and 43,639,986 SNVs, and the pipeline along with the manual inspection of these candidates highlighted 8 candidate promoter mutations. In validation experiments, 7 of these mutations exhibited an increase in promoter activity when comparing the mutant to its wild type. With a specificity of 87.5% and a 3-week lab validation turnover, our method represents a substantial improvement over existing workflows and the pipeline approaches applicability in precision oncology programs.

Keywords: non-coding regulatory regions, promoter mutations, REMIND-Cancer Pipeline, identify functional promoter mutations, bioinformatics pipeline

Yu Chen (UAB), Yixin Zhang (UAB), Amy Wang (UAB), Min Gao (UAB) and Zechen Chong (UAB). *Accurate long-read de novo assembly evaluation with Inspector*.

Abstract. Long-read de novo genome assembly continues to advance rapidly. However, there is a lack of effective tools to accurately evaluate the assembly results, especially for structural errors. We present Inspector, a reference-free long-read de novo assembly evaluator which faithfully reports types of errors and their precise locations. Notably, Inspector can correct the assembly errors based on consensus sequences derived from raw reads covering erroneous regions. Based on in silico and long-read assembly results from multiple long-read data and assemblers, we demonstrate that in addition to providing generic metrics, Inspector can accurately identify both large-scale and small-scale assembly errors.

Keywords: De novo assembly, Long reads, Assembly evaluation, Assembly error, Genome assembly

Marina Yurieva (The Jackson Laboratory for Genomic Medicine), Dan Skelly (The Jackson Laboratory for Genomic Medicine), Candice Baker (The Jackson Laboratory for Genomic Medicine), Will Schott (The Jackson Laboratory for Genomic Medicine), Sandy Diagle (The Jackson Laboratory for Genomic Medicine) and Joshy George (The Jackson Laboratory for Genomic Medicine). *Quantifying the accuracy of genetic demultiplexing of pooled single cell genomics in the mouse across multiple tissues and data types.*

Abstract. Single cell genomics is a rapidly growing and widely used technology that helps to understand cellular heterogeneity and to elucidate the cell type-specific mechanisms mediating disease susceptibility. Nevertheless, the costs of single cell genomic assays remain relatively high and sample throughput low. Genetic demultiplexing is a method that can be used to identify cells from individuals based on natural genetic variation in single cell datasets. It has been used in several human studies but have not been applied to data from non-human systems. A detailed examination of the factors influencing the power and accuracy of labelled sample assignments is not available.

Here we examine the parameters affecting the success of a single cell genetic demultiplexing study using demuxlet1, a tool used to separate samples using genetic variation. We find that sequencing depth is the main factor of demultiplexing success, suggesting that independent genetic variants are the key quantity powering genetic demultiplexing. We provide a pipeline that can be used to split a BAM file by individual cell barcodes and downsample each individual cell's reads to determine the robustness of sample assignments as a function of sequencing depth.

Keywords: Single cell genomics, scRNAseq, scATACseq, Demultiplexing, Computational tools

Agata Muszyńska (Institute of Automatic Control, Electonics and Computer Science, Silesian University of Technology, Gliwice, Poland), Ryszard Przewłocki (Department of Molecular Neuropharmacology, Institute of Pharmacology Polish Academy of Sciences, Kraków, Poland) and Paweł P. Łabaj (Małopolska Centre of Biotechnology UJ, Kraków, Poland). *Exploring mouse transcriptomic landscape.*

Abstract. The mouse is a widely studied animal, as in many aspects its biology is conserved to ours, and it constitutes a valuable model organism. However, the complexity of its transcriptome is not yet fully elucidated. One of the mechanisms responsible for this is alternative splicing (AS), which has been reported to be characteristic of almost all genes in mammals and may be one of the most widely exploited mechanisms responsible for increased transcriptomic and proteomic complexity. We present the results of studying the splicing events in data from an experiment focused on neuropathic pain in mice. In our study, we focused on finding commonalities in the collection of fairly diverse samples to expand the currently known mouse transcriptomic landscape. We hypothesize that the inclusion of different pathological factors should allow detection of the spinal cord novel alternative splicing events (nASEs) characteristic under all conditions. We found that the vast majority of nASEs are common among all samples in our study. Furthermore, the results of the functional analysis showed a clear connection to the nervous system. This result might indicate that the mouse reference model lacks information for brain tissue, but also reflect expected neuroplasticity.

Keywords: RNA-seq, alternative splicing, transcriptomics

Qimin Zhang (The Pennsylvania State University), Qian Shi (The Pennsylvania State University) and Mingfu Shao (The Pennsylvania State University). *Accurate assembly of multi-end RNA-seq data with Scallop2*.

Abstract. Modern RNA-sequencing protocols can produce multi-end data, where multiple reads originating from the same transcript are attached to the same barcode. The long-range information in the multi-end reads is beneficial in phasing complicated spliced isoforms, but assembly algorithms that leverage such information are lacking. Here we introduce Scallop2, a reference-based assembler optimized for multi-end RNA-seq data. The algorithmic core consists of three steps: (1) using an algorithm to 'bridge' multi-end reads into single-end phasing paths in the context of splice graph, (2) employing a method to refine erroneous splice graphs by utilizing multi-end reads that fail to bridge, and (3) piping the refined splice graph and bridged phasing paths into an algorithm that integrates multiple phase-preserving decompositions. Tested on 561 cells in two Smart-seq3 datasets and on ten Illumina paired-end RNA-seq samples, Scallop2 substantially improves the assembly accuracy compared with two popular assemblers StringTie2 and Scallop. Scallop2 represents a significant leap forward for transcript assembly and therefore enables further improvement of the identification of novel transcripts and the downstream isoform-level expression analysis. More importantly, Scallop2 enables accurate construction of transcriptomes at single-cell resolution, which benefits a broader use and advances biological and biomedical research in the era of single-cell omics.

Keywords: transcriptome assembly, multi-end RNA-seq, single-cell RNA-seq, Transcriptomics, Smart-seq3

Timothy Collingsworth (Vor Biopharma), Kit Cummins (Vor Biopharma), Azita Ghodssi (Vor Biopharma), Michael Pettiglio (Vor Biopharma), Japan Mehta (Vor Biopharma), Nipul Patel (Vor Biopharma), Caroline McGowan (Vor Biopharma), Jeff Pimentel (Vor Biopharma), Anjali Kapuria (Vor Biopharma), Meltem Isik (Vor Biopharma), Alejandra Falla (Vor Biopharma), Ruijia Wang (Vor Biopharma), Shu Wang (Vor Biopharma), Dane Hazelbaker (Vor Biopharma), Elizabeth Paik (Vor Biopharma), Michelle Lin (Vor Biopharma), John Lydeard (Vor Biopharma), Gary Ge (Vor Biopharma) and Tirtha Chakraborty (Vor Biopharma). *TransACT provides enhanced detection and characterization of translocation events from high-throughput sequencing data at base-pair resolution for gene editing products.*

Abstract. Gene editing is a powerful approach to improve our ability to treat specific diseases with an unmet medical need. CRISPR-Cas-based gene editing has broad therapeutic applications but also has the potential to increase the possibility of chromosomal translocations after introducing genomic cuts, especially when introducing multiple edits (multiplex editing), nullifying or diminishing the benefits of the therapy by precipitating additional disorders. The development of computational tools to support translocation detection and quantification methods therefore represents a necessary and impactful contribution to the field.

Here, we enhance tools designed for unidirectional sequencing to improve scalable detection and characterization of on-on, on-off, and off-off target translocation events in edited genomes. Our bioinformatics package, TransACT (Translocation Analysis Computational Toolkit), can detect translocation in unidirectional as well as targeted amplicon next generation sequencing data. In addition, we implement advanced false positive filtering to increase the confidence level and generate summary statistics with translocation visualizations at single base-pair resolution. Finally, we demonstrate the accuracy and limit of detection using spike-in translocation datasets.

TransACT is a sophisticated translocation detection and quantification method especially useful for the evaluation of multiplex editing techniques to assess the pre-clinical and clinical safety of gene editing drug products.

Keywords: Gene editing, Translocation, High-throughput sequencing, Visualization

James Denvir (Marshall University), Vinícius Magalhães Borges (Marshall University), Alejandro Q. Nato Jr. (Marshall University) and Adeoluwa Adeluola (Marshall University). *Detecting multiple SARS-CoV-2 variants from short-read sequencing reads of community wastewater samples.*

Abstract. Tracking the spread of SARS-CoV-2 variants has been an essential tool in the public health response to the COVID-19 pandemic. The inflow to public wastewater treatment facilities is a source of SARS-CoV-2 viruses from the community served by the facility. Short-read sequencing of these viral samples has the potential to identify variants present in the sample. However, the combination of the short read length and the heterogeneity of the sample pose challenges to the analysis. We demonstrate a novel graph-theory based analytical approach to the analysis of sequencing data from heterogeneous SARS-CoV-2 samples. Briefly, we identify sites in the viral genome which are polymorphic in the sample, and then identify subsets of these, which we term “discriminating mutation sets,” which segregate with reads. We applied this analysis to data from sequencing of wastewater sampled in January 2022 and, by counting reads consistent with each of the discriminating mutation sets, were able to provide estimates of the relative abundance of Delta and Omicron variants in the samples. This technique also shows potential for identification and relative quantification of variants at a more fine-grained phylogenic level.

Keywords: Sequence analysis, High-throughput sequencing, SARS-CoV-2, Wastewater sample analysis

Arjun Srivatsa (Carnegie Mellon University), Haoyun Lei (Carnegie Mellon University) and Russell Schwartz (Carnegie Mellon University). *A Clonal Evolution Sequence Simulator for Planning Somatic Evolution Studies*.

Abstract. Somatic evolution plays a key role in development and aging as well as in disease processes, notably cancer. The importance of understanding mechanisms of somatic mutability has promoted a proliferation of new sequencing technologies, each with distinctive capabilities and limitations. The enormous space of possible combinations of sequencing modalities poses a substantial challenge for selecting optimal technologies for any particular scientific questions. Versatile simulation tools are thus needed to make it possible to explore and optimize potential study designs. We present a clonal evolution and sequencing simulator allowing for generating synthetic data from a wide range of clonal lineages, variant classes, and sequencing technologies designed for evaluating study designs for assessing somatic mutation mechanisms. Users can define properties of the somatic evolutionary process, mutation classes (e.g., single nucleotide polymorphisms, copy number changes, and classes of structural variation), and biotechnology options (e.g., coverage, bulk vs single cell, whole genome vs exome, error rate, number of samples). The simulator then generates synthetic sequence reads and their corresponding ground-truth parameters for the given study design. We demonstrate its utility in evaluating and optimizing study designs to detect differences in somatic mutation mechanisms between sequence samples.

Keywords: Somatic Evolution, Simulation, Sequencing

Lauren Coombe (BC Cancer, Genome Sciences Centre), Rene Warren (BC Cancer, Genome Sciences Centre), Vladimir Nikolic (BC Cancer, Genome Sciences Centre), Johnathan Wong (BC Cancer, Genome Sciences Centre) and Inanc Birol (BC Cancer, Genome Sciences Centre). *GoldRush-Link: Integrating minimizer-based overlap detection and gap-filling to the ntLink long read scaffolder*.

Abstract. Generating high-quality de novo genome assemblies for model and non-model organisms opens the door to a plethora of important downstream studies. To leverage the repeat-spanning evidence from long-read sequencing technologies, we previously developed ntLink, a minimizer-based long-read scaffolding tool. However most scaffolders, including ntLink, introduce gap sequences ("N"s) between joined sequences, leaving large stretches of unresolved assembly bases, and naively join overlapping sequences. To address these limitations, we added two new features to ntLink: overlap detection and gap-filling. These features are crucial to our new de novo long read assembly tool, GoldRush, and are integrated in the GoldRush-Link stage of the pipeline. Both the overlap detection and gap-filling features are alignment-free, relying on lightweight minimizer mappings. As demonstrated by tests on assemblies from human individuals NA24385 and NA19240, these new features increase the contig NGA50 lengths 502-fold and 7-fold, respectively, while maintaining the high scaffold NGA50 lengths achieved through scaffolding with ntLink. With these two functionalities, >99% gaps were filled for each individual, leaving fewer than 55 N's per 100 kbp in the final assemblies. These modular improvements in ntLink would benefit a wide variety of assembly workflows, including but not limited to GoldRush.

Keywords: de novo genome assembly, long reads, minimizers, gap-filling, assembly finishing

Armaghan Sarvar (Genome Sciences Centre, BC Cancer Agency), Lauren Coombe (Genome Sciences Centre, BC Cancer Agency), René Warren (Genome Sciences Centre, BC Cancer Agency) and Inanc Birol (Genome Sciences Centre, BC Cancer Agency). *Stash: A data structure based on stochastic tile hashing*.

Abstract. Storing and analyzing large sequencing datasets is computationally expensive and developing scalable data structures and algorithms is essential for analyzing their information content. Here, we introduce Stash, a novel hash-based data structure based on stochastic tile hashing (Stashing), which provides a lossy representation of nucleotide sequences, such as long reads.

Stash is implemented as a two-dimensional bit array and populated using sliding windows of spaced seed patterns to hash input sequences. The sequence hashes indicate the memory loci, and sequence ID hashes determine the stored value.

By measuring the number of tile matches for related Stash frames, one can detect whether two genomic regions are covered by the same set of sequencing reads. We report this score on a chromosome of the human genome reference after Stash is filled with experimental Oxford Nanopore Technology sequencing reads and show that as the distance between two loci of the reference contig increases, the metric decreases since a smaller number of common reads cover those regions.

We expect Stash to provide benefits to a variety of bioinformatics applications, including de novo genome assembly and misassembly detection.

Keywords: Hashing, Sequence Mapping, Genomics, Probabilistic Data Structures, High-throughput Algorithms

Johnathan Wong (BC Cancer, Genome Sciences Centre), Vladimir Nikolic (BC Cancer, Genome Sciences Centre), Lauren Coombe (BC Cancer, Genome Sciences Centre), Rene Warren (BC Cancer, Genome Sciences Centre) and Inanc Birol (BC Cancer, Genome Sciences Centre). *GoldRush-Path: A de novo assembler for long reads with linear time complexity.*

Abstract. De novo genome assembly is a cornerstone to a variety of genomic analyses. Long sequencing read technologies have enabled researchers to assemble draft genomes with high contiguity and few structural errors. Most long read assemblers adopt the overlap layout consensus paradigm, a quadratic run time algorithm in its naïve implementation, to address the high number of base errors present in long reads. Recently, ONT and PacBio have made tremendous strides in improving the quality of their long read sequencing technologies, and opportunities for new long read assembly algorithms have emerged. We present GoldRush-Path, a memory-efficient long read assembler algorithm that runs in linear time in the number of reads, as part of the GoldRush pipeline. GoldRush-Path iterates through the long reads and identifies a set of "golden path" sequences that cover ~1X of the target genome by querying each read against a multi-index Bloom filter and inserting it only if its associated sequence signatures are missing. GoldRush-Path, the costliest step in the GoldRush pipeline, consumes at most 73 GB of RAM when assembling human genomes. The selected golden path is then polished and scaffolded in the pipeline, yielding NGA50 lengths of 12 Mbp for human genome assemblies in our tests.

Keywords: de novo long read assembler, ONT, genomics, long reads, golden path

Shulan Tian (Division of Computational Biology, Department of Quantitative Health Sciences, Mayo Clinic, Rochester, MN 55905, USA), William Jenkinson (Division of Computational Biology, Department of Quantitative Health Sciences, Mayo Clinic, Rochester, MN 55905, USA), Alejandro Ferrer (Division of Hematology, Department of Internal Medicine, Mayo Clinic, Rochester, MN 55905, USA), Huihuang Yan (Division of Computational Biology, Department of Quantitative Health Sciences, Mayo Clinic, Rochester, MN 55905, USA), Saurabh Baheti (Division of Computational Biology, Department of Quantitative Health Sciences, Mayo Clinic, Rochester, MN 55905, USA), Terra Lasho (Division of Hematology, Department of Internal Medicine, Mayo Clinic, Rochester, MN 55905, USA), Joel Morales-Rosado (Division of Computational Biology, Department of Quantitative Health Sciences, Mayo Clinic, Rochester, MN 55905, USA), Mrinal Patnaik (Division of Hematology, Department of Internal Medicine, Mayo Clinic, Rochester, MN 55905, USA), Wei Ding (Division of Hematology, Department of Internal Medicine, Mayo Clinic, Rochester, MN 55905, USA), Konstantinos Lazaridis (Division of Gastroenterology & Hepatology, Department of Internal Medicine, Mayo Clinic, Rochester, MN 55905, USA) and Eric Klee (Division of Computational Biology, Department of Quantitative Health Sciences, Mayo Clinic, Rochester, MN 55905, USA). *A unified somatic calling of next-generation sequencing data enhances the detection of clonal hematopoiesis of indeterminate potential.*

Abstract. Clonal hematopoiesis (CH) of indeterminate potential (CHIP) is a premalignant state, in which leukemia-associated driver genes acquire somatic mutations in peripheral blood, at a variant allele frequency (VAF) of 2% or greater; yet the individual does not meet the World Health Organization diagnostic criteria for a hematologic neoplasm. CHIP represents a risk factor for various hematologic malignancies and cardiovascular diseases. The VAF cutoff of >=2% was set arbitrarily, which considers the limitation of standard next generation sequencing (NGS) platforms in detecting small clones due to the relatively high sequencing error and the rarity of clinical consequences associated with mutations at lower VAFs. However, individuals with CH at >=1% VAF in leukemia driver genes also had a significantly increased risk of developing AML as those with >=2% VAF. Popular variant calling algorithms for CHIP detection often lose power on variants with low VAFs. Currently, no analytical pipeline has been developed specifically for CHIP detection. this study presents UNIfied SOmatic calling of Next-generation sequencing data, or UNISON for short, which is a software toolkit designed for streamlined CHIP discovery from population studies, even with suboptimal sequencing coverage. UNISON should be broadly applicable to CHIP detection in large-scale WES and WGS projects.

Keywords: Next generation sequencing, Clonal hematopoiesis of indeterminate potential, Machine learning

Xiyu Peng (Iowa State University) and Karin Dorman (Iowa State University). *Accurate estimation of haplotypes and abundances from Illumina amplicon data by AmpliCI.*

Abstract. Amplicon sequencing is widely applied to explore heterogeneity and rare variants in genetic populations. Resolving true biological variants and accurately quantifying their abundance from noisy amplicon sequence data is crucial for downstream analyses, but measured abundances are distorted by stochasticity and bias in amplification, plus errors during Polymerase Chain Reaction (PCR) and sequencing. Previously we presented AmpliCI, a reference-free, model-based method for rapidly resolving the number, abundance and identity of error-free sequences in massive Illumina amplicon datasets. Here we present AmpliCI v2, that can take into account Unique Molecular Identifier (UMI) information to achieve higher resolution when denoising Illumina amplicon data. The v2 version includes a new module, DAUMI, a probabilistic framework to resolve haplotypes and deduplicated abundance from amplicon sequence data with UMIs. We demonstrate that AmpliCI v2 achieves better performance in haplotype identification and accurate abundance estimation compared to previous AmpliCI version and other UMI-aware clustering methods.

Keywords: Model-based Clustering, Amplicon Sequence, Unique Molecular Identifier

Fairlie Reese (University of California, Irvine), Elisabeth Rebboah (University of California, Irvine), Narges Rezaie (University of California, Irvine), Brian Williams (California Institute of Technology), Heidi Liang (University of California, Irvine), Magdalena Gantuz (University of California, Irvine), Barbara Wold (California Institute of Technology) and Ali Mortazavi (University of California, Irvine). *Characterizing alternative splicing in the ENCODE4 mouse postnatal time course using bulk and single-nucleus long-read RNA-seq*.

Abstract. Alternative isoforms that arise from internal splicing as well as transcription start site (TSS) or transcription end site (TES) choice are known to play key roles during postnatal development. Long-read RNA-seq (lrRNA-seq) sequences through the entire transcript, thus providing not only the ends but also the internal structure of each transcript, and can be applied to both bulk and single-cell samples.

As a part of the final phase of the ENCODE Consortium, we collected 5 tissues (adrenal glands, gastrocnemius muscle, heart, hippocampus, and cortex) from C57BL6J/Castaneus F1 hybrid mice at 7 postnatal timepoints (P4, P10, P14, P25, P36, P2mo, P18-20mo). We have sequenced all of these timepoints using bulk long-read RNA-seq in adrenal gland and gastrocnemius, and a subset of these at key developmental timepoints in the remaining tissues. We have also profiled adrenal gland and hippocampus using single-cell long-read RNA-seq (LR-Split-seq) with both PacBio and Oxford Nanopore (ONT) platforms. We call cell type and timepoint specific isoforms, TSSs, and TESs. We integrate the LR-Split-seq results with matching single-cell multiome data. This approach allows us to connect coaccessible regulatory DNA regions to alternative TSSs that we observe, giving us insight into the regulatory underpinnings guiding promoter choice.

Keywords: mouse, long-read RNA-seq, single-cell, isoform, transcriptomics, development

Rachael Aubin (University of Pennsylvania), Javier Montelongo (University of Pennsylvania) and Pablo Camara (University of Pennsylvania). *Clustering-independent estimation of dynamic cell abundance in bulk tissue using single-cell transcriptomic data*.

Abstract. Biological tissues are heterogeneous and comprise cells undergoing continuous biological processes like cell differentiation. Single-cell RNA-sequencing technologies enable the investigation of these processes. However, generating large cohorts of single-cell data is challenging compared to bulk transcriptomic data. Although many computational methods have been developed for inferring cell type abundance from bulk transcriptomic data, these approaches rely on cell type gene expression signatures and ignore intra-cluster variability. Continuous Deconvolution, ConDecon, is a clustering-independent deconvolution algorithm specifically developed to predict complex changes in single-cell abundance from bulk tissue. This approach estimates the probability that each cell in a reference single-cell data is present in a query bulk data. We compared ConDecon to 17 other methods and find that ConDecon performs comparably to state-of-the-art algorithms when inferring discrete cell type abundances. We then focus on ConDecon's novel ability to estimate dynamic cell abundances along continuous cellular processes. To that end, we applied ConDecon to well-characterized biological systems like B-cell maturation and immune activation. Finally, we use it to identify changes in the activation of tumor-infiltrating microglia during the mesenchymal transformation of pediatric ependymoma. We anticipate that ConDecon will extend the utility of current methods to characterize single-cell dynamics in bulk tissue.

Keywords: Cell type deconvolution, Single-cell RNA-seq., Bulk RNA-seq., Cell heterogeneity

Damla Senol Cali (Bionano Genomics), Gurpreet Singh Kalsi (Intel), Zülal Bingöl (Bilkent University), Can Firtina (ETH Zurich), Lavanya Subramanian (Facebook), Jeremie S. Kim (ETH Zurich), Rachata Ausavarungnirun (King Mongkut's University of Technology North Bangkok), Mohammed Alser (ETH Zurich), Juan Gómez Luna (ETH Zurich), Amiral Boroumand (Carnegie Mellon University), Anant Nori (Intel), Allison Scibisz (Carnegie Mellon University), Sreenivas Subramoney (Intel Labs), Can Alkan (Bilkent University, Department of Computer Engineering), Saugata Ghose (University of Illinois Urbana-Champaign) and Onur Mutlu (ETH Zurich). *GenASM: A High-Performance, Low-Power Approximate String Matching Acceleration Framework for Genome Sequence Analysis.*

Abstract. Rapid genome sequence analysis is currently bottlenecked by the computational power and memory bandwidth limitations of existing systems, as many of the steps in genome sequence analysis must process a large amount of data. A major contributor to this bottleneck is approximate string matching (ASM), which is used at multiple points during the mapping process.

We propose GenASM, the first ASM acceleration framework for genome sequence analysis. GenASM performs bitvector-based ASM, which can efficiently accelerate multiple steps of genome sequence analysis. We modify the underlying ASM algorithm (Bitap) to significantly increase its parallelism and reduce its memory footprint. Using this modified algorithm, we design the first hardware accelerator for Bitap.

We demonstrate that GenASM provides significant performance and power benefits for three different use cases of ASM in genome sequence analysis for both short and long reads: read alignment, pre-alignment filtering, and edit distance calculation. We show that GenASM is significantly faster and more power- and area-efficient than state-of-the-art software and hardware tools for each of these use cases.

Keywords: genome sequence analysis, approximate string matching, read mapping, read alignment, hardware acceleration, hardware/software co-design, sequencing, pre-alignment filtering, edit distance calculation

Vladimir Nikolic (BC Cancer Agency - Genome Sciences Centre), Lauren Coombe (BC Cancer Agency - Genome Sciences Centre), Johnathan Wong (Genome Sciences Centre), Janet Li (BC Cancer Agency - Genome Sciences Centre), Inanc Birol (BC Genome Sciences Centre) and Rene Warren (BC Cancer Agency). *GoldRush-Edit : A targeted, alignment-free polishing & finishing pipeline for long read assembly, using long read k-mers.*

Abstract. An increasing number of genome assembly projects are exclusively utilizing long sequencing reads despite their still appreciable error rates (87-98%), and polishing of the resulting assemblies would be highly desirable. Popular methods for polishing assemblies with long reads, e.g. Racon, rely on sequence alignments for nucleotide base error correction, a costly paradigm that, although robust, is not scalable for large (>3Gbp) genomes, requiring large memory servers and long run times. We present GoldRush-Edit, a RAM-efficient polishing pipeline to correct base errors in long read assemblies using a scalable and targeted k-mer-based method. GoldRush-Edit uses ntEdit for fast correction and low-quality sequence tagging, then Sealer for finishing of these tagged regions using an implicit de Bruijn graph. To achieve acceptable base accuracy, each genomic locus under scrutiny obtains its long read mappings from ntLink, a lightweight minimizer-based scaffolder and gap-filling application. Corresponding long read k-mers are extracted to build reusable targeted Bloom filters in-memory to be used by ntEdit and Sealer. Goldrush-Edit achieves more than 60% reduction in both indels and mismatches on an assembly of the genome of a human cell line, NA24385, using matched nanopore long read data exclusively, while using orders of magnitude less memory compared to Racon.

Keywords: long reads, polishing, alignment-free, efficient, scalable

Vladimir Nikolic (BC Cancer Agency - Genome Sciences Centre), Parham Kazemi (BC Cancer Agency - Genome Sciences Centre), Lauren Coombe (BC Cancer Agency - Genome Sciences Centre), Johnathan Wong (Genome Sciences Centre), Amirhossein Afshinfard (BC Cancer Genome Sciences Centre.), Rene Warren (BC Cancer Agency) and Inanc Birol (BC Genome Sciences Centre). *btllib: A C++ library with Python interface for efficient sequence processing*.

Abstract. Bioinformaticians often write one-off computer programs to perform a specific task instead of reusing existing code. This practice leads to lower software quality and non-reusable code. As bioinformatics analyses are becoming increasingly more complex and deal with ever more data, high quality code is needed for reliable and producible performance. The solution to this is well-designed and documented libraries, such as SeqAn – a C++ library that implements algorithms and data structures commonly used in bioinformatics. Here, we present the btllib library as an addition to this ecosystem with the goal of providing highly efficient, scalable, and ergonomic implementations of bioinformatics algorithms and data structures. The library is implemented in C++ with Python bindings available for a high-level interface. What sets it apart from other libraries is its focus on specialized algorithms with efficiency and scalability in mind as its aim is to enable sequence processing for large genomes. Parallelization, thread safety or race condition minimization when it helps performance (e.g. maximize throughput) are core fundamentals of btllib. The goal of btllib is not to compete, but to complement other available libraries with applications in bioinformatics and genomics research.

Keywords: c++, python, library, assembly, analysis, scalable, efficient

Naveen Duhan (Utah State University) and Rakesh Kaundal (Utah State University). *pySeqRNA: An automated Python package for Next-Generation Sequencing data analysis and report generation*.

Abstract. Every day, massive amounts of data are generated by Next-Generation Sequencing (NGS) technologies. However, streamlined analysis remains a major barrier to effectively utilizing the technology. In recent years, many algorithms, statistical methods, and software tools have been developed to perform the individual analysis steps of various NGS applications. We have developed a Python package (pySeqRNA), that allows fast, efficient, manageable, and reproducible RNA-Seq analysis with uniform workflow interface and support for running on the High-Performance Computing Cluster (HPCC) as well as on local computers. It is an extensible pipeline for performing end-to-end analysis with automated report generation. pySeqRNA workflow consists of quality check and pre-processing of raw sequence reads, accurate mapping of millions of sequencing reads to a reference genome including the identification of expression levels of genes in two ways: (i) Uniquely mapped reads, (ii) Multi-mapped groups, a novel feature added, and Differential analysis of gene expression among different biological conditions, functional enrichment analysis. By integrating several command-line tools and custom Python scripts, it allows effective use of existing software and tools with newly written modules without restricting users to a collection of pre-defined methods and environments. This package accelerates retrieval of reproducible results from NGS experiments. http://bioinfo.usu.edu/pySeqRNA/.

Keywords: RNA Sequencing, Transcriptomics, NGS data analysis, Multimapped Gene Groups

Kendell Clement (Massachusetts General Hospital / Harvard Medical School), Linda Lin (Boston Childrens Hospital / Harvard Medical School), Daniel Bauer (Boston Childrens Hospital / Harvard Medical School) and Luca Pinello (Massachusetts General Hospital / Harvard Medical School). *Quantification of complex genome editing events including large insertions and translocations using CRISPRlungo*.

Abstract. Genome editing technologies are rapidly evolving, and analysis of deep sequencing data from target and off-target regions is necessary for evaluating editing efficiency, precision and specificity. Our group has developed the widely-used tool, CRISPResso2, which standardized quantification of editing frequencies at predefined loci using amplicon sequencing. However, this and other methods are only able to detect small insertions and deletions. In order to quantify complex genome editing events including large insertions, inversions and translocations, assays have been proposed which enrich for DNA sequences using only one PCR origin as the anchor for amplification. We developed a novel analytic tool called CRISPRLungo to analyze sequencing data produced from single-anchor PCR which can quantify and visualize complex genome editing events without any a priori assumption of the expected outcomes. We generated single-anchor amplification data for a therapeutic genome editing experiment and show that our tool can take advantage of the richness of unidirectional sequencing data to both sensitively and specifically detect a variety of complex genome editing outcomes, including identifying rare chromosomal alterations not detectable using current analysis toolkits. CRISPRLungo is available as open-source software that enables researchers to comprehensively assess genome editing outcomes without the biases of amplicon sequencing.

Keywords: CRISPR, Genome Editing, Off-target, Translocation, amplicon sequencing

Ragnar Groot Koerkamp (ETH Zurich) and Pesho Ivanov (ETH Zurich). *Exact pairwise alignment using A* with chained-seeds heuristic and match pruning*.

Abstract. We present a fast exact algorithm for global pairwise alignment between related sequences with unit edit costs. It instantiates the A* shortest path algorithm with a novel chained-seeds heuristic that improves as the search progresses.

Chained-seeds heuristic. We partition the first sequence into short seeds, and find their matches in the second sequence. The heuristic is defined as the lowest cost of a chain of remaining matches. It combines the seed costs of skipped remaining seeds, and the gap costs between consecutive matches in a chain. To efficiently compute the heuristic, we introduce a contours data structure.

Match pruning. To further reduce the number of states that the A* explores, we ignore (prune) already expanded matches in the heuristic computation. We prove that pruning preserves a shortest path.

Results. Our aligner, A*PA, demonstrates near-linear runtime (best fit: n^1.09) on random related genetic sequences with e = 5% uniform error rate and length n ≤ 10^7 bp. Compared to the leading exact aligners Edlib and BiWFA, A*PA reaches > 400x speedup for n = 10^7 bp and e = 5%, and 2.5x speedup for simulated ONT reads from a human genome (n ≈ 10^6 bp, e ≈ 10% errors).

Keywords: pairwise alignment, exact algorithm, A* algorithm, seed heuristic, multiple-path pruning, dynamic heuristic, contours

Noa Oded Elkayam (Emendo Biotherapeutics Ltd), Michal Sharabi Schwager (Emendo Biotherapeutics Ltd), Malka Aker (Emendo Biotherapeutics Ltd), Ella Segal (Emendo Biotherapeutics Ltd) and Idit Buch (Emendo Biotherapeutics Ltd). *Comprehensive bioinformatics tools for quantitative analysis of gene editing*.

Abstract. The ability to generate and analyze massive data can accelerate our understanding of gene editing processes. However, the generation of such data imposes two major challenges. The first, is the experimental procedure which parallelizes many samples/conditions at once. The second is the computational analysis which aims to produce few metrics for fast meaningful comparison. Addressing these challenges must be scalable and reproducible, while limiting human intervention to reduce errors. At EmendoBio, we developed a procedure that takes thousands of samples and automatically forms a DNA library prep for next-generation sequencing (NGS), using a robotic Biomek i7 system. This step involves target specific amplification, different amplicon mixing and Illumina distinct indexing. Strict input validation steps are taken to meet pre-defined formats. Following sequencing procedures from various sources, the analysis of many samples is triggered at once (automatically or manually by user request) using Amazon serverless technology combined with parallel batch processing. The analysis space comprises many different bioinformatics tools such as CRISPR on/off target analysis, transcriptome characterization assays, mutations and SNPs phasing, RNA-Seq analysis and others. Each analysis is followed by specific post-processing calculations, visualization and summarized metrics. Our simplified and automated procedure enables efficient cross-experimental conclusions regarding gene editing processes.

Keywords: Gene editing, CRISPR, High-throughput sequencing, Automation

Gergely Csaba (LMU Munich), Evi Berchtold (LMU Munich), Armin Hadziahmetovic (LMU Munich), Markus Gruber (LMU Munich), Constantin Ammar (LMU Munich) and Ralf Zimmer (LMU Munich).
EmpiReS: Differential Analysis of Gene Expression and Alternative Splicing.

Abstract. While absolute quantification is challenging in high-throughput measurements, changes of features between conditions can often be determined with high precision.

Therefore, analysis of fold changes is the standard method sufficient for differential expression, but often, the analysis of “changes of changes” is required.

Differential alternative splicing is an application of such a doubly differential analysis.

EmpiReS is a quantitative approach for various kinds of omics data based on fold changes for appropriate features of biological objects.

Empirical error distributions for these fold changes are estimated from Replicate measurements and used to quantify feature fold changes and their directions.

We assess the performance of EmpiReS to detect differentially expressed genes applied to RNA-Seq using simulated data.

It achieved higher precision than established tools at nearly the same recall level.

Furthermore, we assess the detection of alternatively Spliced genes via changes of isoform fold changes on distribution free simulations and on experimentally validated splicing events.

EmpiReS achieves the best precision-recall values for simulations based on different biological datasets.

We propose EmpiReS as a general, quantitative and fast approach with high reliability and an excellent trade-off between sensitivity and precision for both differential expression and differential alternative splicing.

Keywords: differential expression, differential alternative splicing, alternative splicing, RNA-seq, Transcriptome

Janik Sielemann (Bielefeld University), Katharina Sielemann (Bielefeld University), Broňa Brejová (Comenius University in Bratislava), Tomas Vinar (Comenius University in Bratislava) and Cedric Chauve (Simon Fraser University). *plASgraph - using graph neural networks to detect plasmid contigs from an assembly graph*.

Abstract. Identification of plasmids from sequencing data is an important and challenging problem related to antimicrobial resistance spread. We provide a new architecture for identifying plasmid contigs in fragmented genome assemblies built from short-read data. Unlike previous machine-learning approaches for this problem, which classify individual contigs separately, we employ graph neural networks (GNNs) to include information from the assembly graph. Propagation of information from nearby nodes in the graph allows accurate classification of even short contigs that are difficult to classify based on sequence features or database searches alone.

Our new species-agnostic software tool plASgraph outperforms recently developed PlasForest, which uses database searches to supplement sequence-based features. Since our tool does not rely on existing plasmid databases, it is more suitable for classification of contigs in novel species. Our tool can also be trained on a specific species, and in that scenario it outperforms mlplasmids trained on the same species.

On one hand, our work provides a new, accurate, and easy to use tool for plasmid classification; on the other hand, it serves as a motivation for more widespread use of GNNs in bioinformatics, such as in pangenome sequence analysis, where sequence graphs serve as a fundamental data structure.

Availability: https://github.com/cchauve/plASgraph

Keywords: contig classification, graph neural network, machine learning, plasmids

Emma Jones (The University of Alabama at Birmingham), Avery Williams (The University of Alabama at Birmingham), Anisha Haldar (The University of Alabama at Birmingham), Vishal Oza (The University of Alabama at Birmingham), Timothy Howton (The University of Alabama at Birmingham) and Brittany Lasseigne (The University of Alabama at Birmingham). *Comparing transcriptional diversity metrics across brain regions and biological sex in long-read RNA sequencing data*.

Abstract. Third-generation (i.e., long-read) sequencing platforms like Oxford Nanopore and PacBio implement additional capabilities for collecting genomic information, including novel isoform detection, due to their ability to sequence the entire length of mRNA transcripts. While measuring which genes and/or transcripts are differentially expressed across conditions is common, it is only one way to compare gene expression and is susceptible to missing important biological information. As nothing in biology acts in isolation, there is a need to describe patterns present in entire gene expression profiles in addition to comparing individual differentially-expressed genes. Another way to measure transcriptional differences globally is transcriptional diversity. Transcriptional diversity can refer to the overall number of genes expressed, or it can refer to differential isoform usage. Transcriptional diversity has been previously described in many ways, but different measures of transcriptional diversity may distinctly capture biological and technical variation. Here, we compare transcriptional diversity metrics including coefficient of variation (CV), Shannon entropy, and the Gini index in publicly-available Genotype-Tissue Expression (GTEx) project long-read RNA sequencing data with respect to brain region and biological sex.

Keywords: long-read, RNA-seq, alternative splicing, transcriptional diversity, isoform usage, nanopore

Nathan Dwarshuis (NIST), Justin Wagner (NIST), Peter Tonner (NIST), Nathanael Olson (NIST), Jennifer McDaniel (NIST) and Justin Zook (NIST). *Using Machine Learning Models to Understand Errors in Human Genomic Variation*.

Abstract. The Genome in a Bottle consortium generates variant benchmarks for a set of human genomes to enable evaluation and comparison of sequencing technologies and variant detection methods. While these technologies can resolve most of the genome, correctly calling variants in complex or repetitive regions remains a challenge. We currently have general heuristics to predict incorrectly-called variants (more repetition is harder, etc); however, we lack a data-driven model to link variant caller performance to specific, quantifiable genomic contexts.

We aim to make such a model using explainable boosting machines (EBMs). EBMs are a linear combination of arbitrary univariate and bivariate functions (generalized additive models with interaction terms). Despite being flexible, the relative simplicity of EBMs will allow interpretation of the functional relationship and relative contribution of each feature. For example, the model revealed A/T homopolymers longer than ~15bp predict higher Illumina single nucleotide variant (SNV) and insertion/deletion (INDEL) error rates. For G/C homopolymers, any length above 0bp and increasing imperfect fraction predicted higher error rates for Illumina.

Ultimately, this will provide a data-driven foundation for comparing variant caller methods and/or sequencing technologies in difficult regions of the genome, and enable improved design of stratifications delineating difficult regions.

Keywords: benchmarks, variant calling, machine learning, glassbox

Timothy Howton (University of Alabama at Birmingham), Vishal Oza (University of Alabama at Birmingham), Elizabeth Wilk (University of Alabama at Birmingham), Michal Mrug (University of Alabama at Birmingham), Bradley Yoder (University of Alabama at Birmingham) and Brittany Lasseigne (University of Alabama at Birmingham). *Changes in Cellular Metabolism in Autosomal Dominant Polycycstic Kidney Disease*.

Abstract. Autosomal dominant polycystic kidney disease (ADPKD) is characterized by the development of cysts in the kidneys that increase in number and volume with age. The increase in the quantity and size of cysts eventually interferes with normal kidney function and ultimately leads to end-stage kidney disease. Roughly 1 in 1000 people are affected by ADPKD, and it is the fourth leading cause of end-stage kidney disease. ADPKD is a multisystem disease therefore patients can also suffer from hemorrhagic stroke, cardiac arrest, and/or complications from severe cystic liver disease. The disease is predominantly caused by mutations in the PKD1 and PKD2 genes which encode for polycystin 1 (PC1) and polycystin 2 (PC2), respectively. ADPKD displays metabolic changes including alternative glucose metabolism similar to the Warburg effect, oxidative phosphorylation, and fatty acid synthesis. Additionally, dietary modifications including caloric restriction have shown to improve symptoms. However, metabolic changes at a single-cell resolution have not been thoroughly examined. Here we use single-cell RNA-seq approaches to explore cell-specific metabolic pathway changes in publicly available human ADPKD and Pkd2 knock-out mice datasets.

Keywords: polycystic kidney disease, metabolism, scRNA-seq

Talha Murathan Goktas (Canada's Michael Smith Genome Science Centre), Vladimir Nikolic (Canada's Michael Smith Genome Science Centre), Ka Ming Nip (Canada's Michael Smith Genome Science Centre), Johnathan Wong (Canada's Michael Smith Genome Science Centre), Lauren Coombe (Canada's Michael Smith Genome Science Centre), Rene Warren (Canada's Michael Smith Genome Science Centre) and Inanc Birol (Canada's Michael Smith Genome Science Centre). *Mapping noisy long-reads with multi-indexed Bloom Filter: miBF-mapper*.

Abstract. Mapping genomic sequences to references is an essential step for genomic analysis. Since the early days of genomics research, genomic sequence mapping and alignment tools have placed great effort to improve accuracy and decrease resource usage. Throughout the years, the mapping software improved substantially fueled by the diversity of data structures and algorithms developed by the community. Here we present miBF-mapper a long-read mapping software where we indexed reference genome with our in-house data structure multi-indexed Bloom Filter(miBF). Considering >10% overlap with the true region as correct mapping, miBF-mapper had 99.9% mapping accuracy in mapping 45k simulated ONT reads to C.elegans reference in 5 minutes 32 seconds and required 1 GB of RAM, and 92.6% accuracy in mapping 50k simulated ONT reads to H.sapiens GRCh38 reference in 35 minutes 21 seconds requiring 148GB of RAM. Here we discuss miBF-mapper algorithm in detail which is a successful application of the novel miBF data structure that should be of interest to the community.

Keywords: mapping, long-read, nanopore

Felix Offensperger (LMU Munich), Evi Berchtold (LMU Munich) and Ralf Zimmer (LMU Munich).
EmpiReR: Model-free reliable analysis of higher order differentials in complex replicate count data.

Abstract. There is no end to innovation in modern biology. Experimental designs are becoming increasingly complex, encompassing perturbations of multiple dimensions and conditions. Despite the immense information gain, statistics, e.g. with DESeq2, often focus on a 1-vs-1 or 1-vs-all design. Moreover, these tests are all based on questionable null hypothesis testing and the resulting p-values. These are often misunderstood and misapplied.

Here we introduce the EmpiReR, which employs a fuzzy value based representation of the count data. By fuzzy binning it captures the empirical error distributions of the data and estimates whether the features are consistent compared to the rest of the data in the same condition and whether they show an significant change when comparing conditions.

EmpiReR allows a p-value free analysis of data and the analysis of higher order differentials. This is important not only for complex data but also for comparisons, like pattern extraction in iATAC, where combination of data types (RNAseq and ATACseq) is critical to the analysis. EmpiReR can be used not only to compute the 'best' higher order differential changes, but also to extract and visualize evidence for complex patterns, e.g. fuzzy differential flows of foldchanges for time series data along various conditions.

Keywords: High-dimensional data, Multiomics, Fuzzy logic, P-value free

Siyuan Cheng (Washington University in St. Louis), Benpeng Miao (Washington University in St. Louis) and Bo Zhang (Washington University in St. Louis). *Benchmark of analysis strategies for ATAC-seq and CUT&Tag-seq*.

Abstract. Tn5 was one of the first identified prokaryotic transposons, and Tn5 transposase is already widely adopted into different genomic protocols to explore the genome and epigenome in a high-throughput fashion. Specifically, ATAC-seq and CUT&Tag-seq are becoming the most widely used epigenomic experimental approaches to measure chromatin accessibility and detect the DNA-protein interactions. Along with large-scale data production, it is now the new bottleneck to process these epigenomic data correctly. Many bioinformatics tools were developed for processing ATAC-seq and CUT&Tag-seq data, however, a comprehensive comparison and benchmarking of these methods is still lacking. Here, we conducted a comprehensive benchmarking to evaluate the performance of eight popular software in processing ATAC-seq and CUT&Tag-seq data, including AIAP, MACS2, SEACR, HMMRATAC, CUT&RUNTools2.0, and ChromHMM. We further test the performance of differentially analysis strategies for ATAC-seq and CUT&Tag-seq data. In conclusion, our study supplied a comprehensive bioinformatics guidance of ATAC-seq and CUT&Tag-seq data processing and differential analysis. The recommended analysis strategy was complied into Docker/Singularity image, allowing biologists easily perform data analysis by executing one line of command.

Keywords: Benchmark, CUT&Tag-seq, ATAC-seq