
High Throughput Sequencing
Algorithms and Applications

*A special track of the ISMB-ECCB 2021 meeting*
Lyon, France, July 25-27, 2021
Virtual Meeting

# ISMB-ECCB 2021 HiTSeq Track Proceedings

Lyon, France (virtual)
July 25-27, 2021
http://www.hitseq.org

**Organizers:**

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

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

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

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

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

Kjong Lehmann
ETH-Zürich, Zürich, Switzerland
E-mail: kjong.lehmann@inf.ethz.ch

## *Estimated nucleotide reconstruction quality symbols of basecalling tools for Oxford Nanopore sequencing*

Wiktor Kuśmirek (Warsaw University of Technology).

Currently, one of the fastest growing DNA sequencing technologies is nanopore sequencing. One of the key stages of processing sequencer data is the basecalling process, which from the input sequence of currents measured on the pores of the sequencer reproduces the DNA sequences called DNA reads. Many of the applications dedicated to basecelling together with the DNA sequence provide the estimated quality of reconstruction of a given nucleotide.

Herein, we examinated the estimated quality of nucleotide reconstruction reported by another basecallers. The results showed that the estimated reconstruction quality reported by different basecallers may vary depending on the tool used. In particular, for some tools, along with successive symbols of the estimated reconstruction quality (which theoretically should mean more and more accurate reconstruction), the real quality of the nucleotide increases (the number of matched nucleotides increaces and the number of errors decreases). However, there are tools that report the estimated reconstruction quality in the basecalling results, but these values are in no way interpretable. What is more, the estimated reconstruction quality reported in basecalling process is not used in any investigated tool for processing nanopore DNA reads.

## *Automated Quality Control of NGS Data using Machine Learning*

Steffen Albrecht (Faculty of Biology, Johannes Gutenberg University Mainz, Mainz, Germany), Maximilian Sprang (Faculty of Biology, Johannes Gutenberg University Mainz, Mainz, Germany), Miguel Andrade-Navarro (Faculty of Biology, Johannes Gutenberg University Mainz, Mainz, Germany) and Jean-Fred Fontaine (Faculty of Biology, Johannes Gutenberg University Mainz, Mainz, Germany).

The versatility of next-generation sequencing (NGS) applications makes the sequencing technology a popular tool in biology and medicine. Yet, the complexity to evaluate raw data quality has a negative impact on downstream analyses. In a clinical context, patient data of unnoticed low-quality can lead to wrong diagnosis or ill-suited treatment.

To address this problem, we have characterized 47 quality features on 2642 human and mouse functional genomics NGS files from ENCODE (RNA-seq, ChIP-seq and DNase-seq). Over 1 Million predictive models were evaluated within a grid search for 10 classification algorithms. External validations were performed on 700 files from 38 datasets in GEO or Cistrome databases.

Tree-based or deep learning algorithms such as random forest or multilayer perceptron generated the most accurate models. A generic model trained on the all data performed similarly to specialized models (average auROC=0.925) and better generalized on external data, including ATAC-seq data not used for training.

Provided the limited usefulness of publicly available guidelines and the observed high dependency of quality features to experimental conditions, our predictive models represent a valuable resource for scientists to better understand quality issues and perform automatic quality control.

Availability: https://github.com/salbrec/seqQscorer

## *cdev: A ground-truth based measure to evaluate RNA-seq normalization performance*

Diem-Trang Tran (University of Utah) and Matthew Might (University of Alabama at Birmingham).

Normalization of RNA-seq data has been an active area of research since the problem was first recognized a decade ago. Despite the active development of new normalizers, their performance measures have been given little attention. To evaluate normalizers, researchers have been relying on ad hoc measures, most of which are either qualitative, potentially biased, or easily confounded by parametric choices of downstream analysis. We propose a metrics called condition-number based deviation, or cdev, to quantify normalization success. cdev measures how much an expression matrix differs from another. If a ground truth normalization is given, cdev can then be used to evaluate the performance of normalizers. To establish experimental ground truth, we compiled an extensive set of public RNA-seq assays with external spike-ins. This data collection, together with cdev, provides a valuable toolset for benchmarking new and existing normalization methods.

## *Strobemers: an alternative to k-mers for sequence comparison*

Kristoffer Sahlin (Stockholm University).

K-mer-based methods are widely used in bioinformatics for sequence comparison. However, a single mutation will mutate k consecutive k-mers and makes most k-mer based applications sensitive to variable mutation rates. Many techniques have been studied to overcome this sensitivity, e.g., spaced k-mers and k-mer permutation techniques, but these techniques do not handle indels well. For indels, pairs or groups of small k-mers are commonly used, but these methods first produce k-mer matches, and only in a second step, a pairing or grouping of k-mers is performed. Such techniques produce many redundant k-mer matches due to the size of k.

Here, we propose strobemers, a new data structure for sequence comparison. We use simulated data to show that, under several mutation rates, strobemers outperform k-mers and spaced k-mers for sequence similarity searches by producing more evenly spread sequence matches and higher match coverage. We further implement a proof-of-concept sequence matching tool StrobeMap. We use StrobeMap with synthetic and biological Oxford Nanopore sequencing data to show the utility of using strobemers for sequence comparison in different contexts such as sequence clustering and alignment scenarios. A reference implementation of our tool StrobeMap together with code for analyses is available at https://github.com/ksahlin/strobemers.

# *Impact of lossy compression of nanopore raw signal data on basecalling and consensus accuracy*

Shubham Chandak (Stanford University), Kedar Tatwawadi (Stanford University), Srivatsan Sridhar (Stanford University) and Tsachy Weissman (Stanford University).

Motivation: Nanopore sequencing provides a real-time and portable solution to genomic sequencing, enabling better assembly, structural variant discovery and modified base detection than second generation technologies. The sequencing process generates a huge amount of data in the form of raw signal contained in fast5 files, which must be compressed to enable efficient storage and transfer. Since the raw data is inherently noisy, lossy compression has potential to significantly reduce space requirements without adversely impacting performance of downstream applications.

Results: We explore the use of lossy compression for nanopore raw data using two state-of-the-art lossy time-series compressors, and evaluate the tradeoff between compressed size and basecalling/consensus accuracy. We test several basecallers and consensus tools on a variety of datasets at varying depths of coverage, and conclude that lossy compression can provide 35–50% further reduction in compressed size of raw data over the state-of-the-art lossless compressor with negligible impact on basecalling accuracy ($\lesssim$0.2% reduction) and consensus accuracy ($\lesssim$0.002% reduction). The results suggest the possibility of using lossy compression, potentially on the nanopore sequencing device itself, to achieve significant reductions in storage and transmission costs while preserving the accuracy of downstream applications.

Availability and implementation: The code is available at https://github.com/shubhamchandak94/lossy_compression_evaluation.

## *Comparative genome analysis using sample-specific string detection in accurate long reads*

Parsoa Khorsand (University of California, Davis), Luca Denti (Department of Computational Biology, C3BI USR 3756 CNRS, Institut Pasteur), Paola Bonizzoni (Department of Informatics, Systems and Communication, University of Milano-Bicocca, Milano), Rayan Chikhi (Department of Computational Biology, C3BI USR 3756 CNRS, Institut Pasteur) and Fereydoun Hormozdiari (University of California, Davis).

Motivation: Comparative genome analysis of two or more whole-genome sequenced (WGS) samples is at the core of most applications in genomics. These include discovery of genomic differences segregating in population, case-control analysis in common disease, and rare disorders. With the current progress of accurate long-read sequencing technologies (e.g., circular consensus sequencing from PacBio sequencers) we can dive into studying repeat regions of genome (e.g., segmental duplications) and hard-to-detect variants (e.g., complex structural variants).

Results: We propose a novel framework for addressing the comparative genome analysis by discovery of strings that are specific to one genome ("samples-specific" strings). We have developed an accurate and efficient novel method for discovery of samples-specific strings between two groups of WGS samples. The proposed approach will give us the ability to perform comparative genome analysis without the need to map the reads and is not hindered by shortcomings of the reference genome. We show that the proposed approach is capable of accurately finding samples-specific strings representing nearly all variation (>98%) reported across pairs or trios of WGS samples using accurate long reads (e.g., PacBio HiFi data).

Availability: The proposed tool is publicly available at https://github.com/Parsoa/PingPong.

## *Polishing Copy Number Variant Calls on Exome Sequencing Data via Deep Learning*

Furkan Ozden (Bilkent University), Can Alkan (Bilkent University) and A. Ercument Cicek (Bilkent University).

Accurate and efficient detection of copy number variants (CNVs) is of critical importance due to their significant association with complex genetic diseases. Although algorithms that use whole genome sequencing (WGS) data provide stable results with mostly-valid statistical assumptions, copy number detection on whole exome sequencing (WES) data shows comparatively lower accuracy. This is unfortunate as WES data is cost efficient, compact and is relatively ubiquitous. The bottleneck is primarily due to non-contiguous nature of the targeted capture: biases in targeted genomic hybridization,GC content, targeting probes, and sample batching during sequencing. Here, we present a novel deep learning model,DECoNT, which uses the matched WES and WGS data and learns to correct the copy number variations reported by any off-the-shelf WES-based germline CNV caller. We train DECoNT on the 1000 Genomes Project data, and we show that we can efficiently triple the duplication call precision and double the deletion call precision of the state-of-the-art algorithms. We also show that our model consistently improves the performance independent from (i) sequencing technology, (ii) exome capture kitand (iii) CNV caller. Using DECoNT as a universal exome CNV call polisher has the potential to improvethe reliability of germline CNV detection on WES data sets.

## *Alzheimer's Risk Factors Age, APOE Genotype, and Sex Drive Distinct Molecular Pathways*

Yingxue Ren (Mayo Clinic), Na Zhao (Mayo Clinic), Yu Yamazaki (Mayo Clinic), Wenhui Qiao (Mayo Clinic), Fuyao Li (Mayo Clinic), Lindsey Felton (Mayo Clinic), Siamak Mahmoudiandehkordi (Duke University), Alexandra Kueider-Paisley (Duke University), Berkiye Sonoustoun (Mayo Clinic), Matthias Arnold (German Research Center for Environmental Health), Francis Shue (Mayo Clinic), Jiaying Zheng (Mayo Clinic), Olivia Attrebi (Mayo Clinic), Yuka Martens (Mayo Clinic), Zonghua Li (Mayo Clinic), Ligia Bastea (Mayo Clinic), Axel Meneses (Mayo Clinic), Kai Chen (Mayo Clinic), J Will Thompson (Duke University), Lisa St John-Williams (Duke University), Masaya Tachibana (Mayo Clinic), Tomonori Aikawa (Mayo Clinic), Hiroshi Oue (Mayo Clinic), Lucy Job (Mayo Clinic), Akari Yamazaki (Mayo Clinic), Chia-Chen Liu (Mayo Clinic), Peter Storz (Mayo Clinic), Yan Asmann (Mayo Clinic), Nilufer Ertekin-Taner (Mayo Clinic), Takahisa Kanekiyo (Mayo Clinic), Rima Kaddurah-Daouk (Duke University) and Guojun Bu (Mayo Clinic).

Evidence suggests interplay among the three major risk factors for Alzheimer's disease (AD): age, APOE genotype, and sex. Here, we present comprehensive datasets and analyses of brain transcriptomes and blood metabolomes from human apoE2-, apoE3-, and apoE4-targeted replacement mice across young, middle, and old ages with both sexes. We found that age had the greatest impact on brain transcriptomes highlighted by an immune module led by Trem2 and Tyrobp, whereas APOE4 was associated with upregulation of multiple Serpina3 genes. Importantly, these networks and gene expression changes were mostly conserved in human brains. Finally, we observed a significant interaction between age, APOE genotype, and sex on unfolded protein response pathway. In the periphery, APOE2 drove distinct blood metabolome profile highlighted by the upregulation of lipid metabolites. Our work identifies unique and interactive molecular pathways underlying AD risk factors providing valuable resources for discovery and validation research in model systems and humans.

## *phasebook: haplotype-aware de novo assembly of diploid genomes from long reads*

Xiao Luo (Bielefeld University), Xiongbin Kang (Bielefeld University) and Alexander Schönhuth (Bielefeld University).

Haplotype-aware diploid genome assembly is crucial in genomics, precision medicine, and many other disciplines. Long-read sequencing technologies have greatly improved genome assembly thanks to advantages of read length. However, current long-read assemblers usually introduce disturbing biases or fail to capture the haplotype diversity of the diploid genome. Here, we present phasebook, a novel approach for reconstructing the haplotypes of diploid genomes from long reads de novo. Benchmarking experiments demonstrate that our method outperforms other approaches in terms of haplotype coverage by large margins, while preserving competitive performance or even achieving advantages in terms of all other aspects relevant for genome assembly.

## *ACE: Explaining single-cell cluster from an adversarial perspective*

Yang Lu (University of Washington), Timothy Yu (University of Washington), Giancarlo Bonora (gbonora@uw.edu) and William Noble (University of Washington).

A common workflow in single-cell RNA-seq analysis is to project the data to a latent space, cluster the cells in that space, and identify sets of marker genes that explain the differences among the discovered clusters. A primary drawback to this three-step procedure is that each step is carried out independently, thereby neglecting the effects of the nonlinear embedding and inter-gene dependencies on the selection of marker genes. Here we propose an integrated deep learning framework, Adversarial Clustering Explanation (ACE), that bundles all three steps into a single workflow. The method thus moves away from the notion of "marker genes" to instead identify a panel of explanatory genes. This panel may include genes that are not only enriched but also depleted relative to other cell types, as well as genes that exhibit differences between closely related cell types. Empirically, we demonstrate that ACE is able to identify gene panels that are both highly discriminative and nonredundant.

## *Victor: full-length de novo viral haplotype reconstruction from noisy long reads*

Xiao Luo (Bielefeld University), Xiongbin Kang (Bielefeld University) and Alexander Schönhuth (Bielefeld University).

Haplotype-resolved assembly of highly diverse virus genomes is critical in prevention, control and treatment of viral diseases. Current methods either only handle accurate short reads, or collapse haplotype-specific variations. Here, we present Victor, a novel approach to reconstruct viral haplotypes from noisy long reads. As a crucial novelty, Victor is the first approach that reconstructs viral haplotypes from error-prone long read data referring to RNA virus quasispecies both accurately and at full length. Benchmarking experiments on both simulated and real datasets of varying complexity and diversity confirm this, by demonstrating the superiority of Victor in terms of relevant criteria in comparison with the state of the art.

## *High-Throughput Chemical Safety Screening Using Targeted RNA-seq*

Logan J. Everett (U.S. Environmental Protection Agency), Joshua A. Harrill (U.S. Environmental Protection Agency), Derik Haggard (U.S. Environmental Protection Agency), Joseph Bundy (U.S. Environmental Protection Agency), Beena Vallanat (U.S. Environmental Protection Agency), Imran Shah (U.S. Environmental Protection Agency) and Richard Judson (U.S. Environmental Protection Agency).

Traditional toxicological testing is a costly and slow process and as a result thousands of chemicals lack sufficient safety data to protect human health and the environment. Transcriptomics has emerged as a cost-effective method for broadly assessing chemical toxicity across many target pathways and mechanisms of action in a single assay. US EPA has designed a rapid and automated in vitro screening platform using the TempO-seq targeted RNA-seq assay to profile chemical bioactivity across a range of concentrations and cell types. To date, this approach has been used to screen over 1,000 chemicals in three biologically distinct cell lines, resulting in over 100,000 targeted RNA-seq profiles covering ~20,000 human protein-coding genes. The scale and complexity of this data has necessitated extensive development of novel bioinformatic methods, including: 1) an open-source pipeline optimized for rapid and robust processing of TempO-seq data; 2) novel QC procedures tailored to data generated in large-scale automated experiments; and 3) signature-level dose-response models to summarize results for each chemical and link the observed in vitro bioactivity to known targets, pathways, and hazards. This abstract does not necessarily reflect US EPA policy. Use of product or company names do not constitute endorsement by US EPA.

## *Efficient targeted error resolution and automated finishing of long read genome sequence assemblies*

Janet Xin Li (British Columbia Genome Sciences Centre), Rene Warren (BC Genome Sciences Centre), Lauren Coombe (BC Genome Sciences Centre), Johnathan Wong (BC Genome Sciences Centre) and Inanc Birol (BC Genome Sciences Centre).

Nanopore long-read whole-genome sequencing is rapidly taking a foothold in research settings, enabling chromosome-scale genome assemblies across the tree of life. However, resulting long-read assemblies still contain appreciable base errors. Existing genome polishing solutions mostly rely on sequence alignments to provide fairly robust error correction, but suffer scalability issues. Here we present a protocol, which employs memory-efficient Bloom filter (BF) data structures to address this problem. Alignment-free sequence polisher ntEdit makes use of these BFs, iterating from long-to-short kmers to verify each genomic base, fixing mismatches and indels whenever possible, and labeling problematic regions for further targeting by soft-masking the corresponding loci. These labeled regions, along with unresolved (gap) regions of the genome, are then targeted using Sealer, an alignment-free gap-filler that uses an implicit de Bruijn graph stored in BFs to further resolve problem sequences. In our tests on human NA12878 Redbean and Shasta nanopore long read genome assemblies our pipeline, which needed no human intervention, ran in <6 h requiring at most 84.3 GB RAM and recovered 88.7% and 90.5% complete conserved BUSCO genes. The outlined operations provide a scalable and efficient automated genome finishing solution for targeted error resolution in long-read genome assembly using short reads.

## *Dynamic Adaptive Sampling During Nanopore Sequencing and Assembly using Bayesian Experimental Design*

Lukas Weilguny (European Molecular Biology Laboratory, European Bioinformatics Institute), Nicola De Maio (European Molecular Biology Laboratory, European Bioinformatics Institute) and Nick Goldman (European Molecular Biology Laboratory, European Bioinformatics Institute).

One particularly promising feature of nanopore sequencing is the ability to reject reads, enabling real-time selection of molecules without complex sample preparation. Previously, such decisions were based on a priori choice. Instead, they could also incorporate already-observed data in order to maximise information gain. For example, during resequencing the genotype of sites without variation is confirmed by few reads; whereas more data would be desirable at variable sites.

We present BOSS-RUNS, a mathematical model to calculate the expected benefit of new reads and an algorithm to generate dynamically updated decision strategies. During sequencing, we quantify the uncertainty at each site and for each novel read decide whether the potential decrease in uncertainty at the sites it will most likely cover warrants complete sequencing.

In simulations of a microbial community we show that this can mitigate coverage bias or lead to higher minimum coverage in regions of interest compared to sequencing without, or with a priori adaptive sampling. Further, we consider the problem of de novo assembly by adapting our framework to genome graphs, which allows for the rejection of fragments from well-assembled regions to focus on reads that extend contigs or resolve repeats instead.

## *pyTCR: a comprehensive cloud-based platform for TCR-Seq data analysis using interactive notebooks to facilitate reproducibility and rigor of immunogenomics research*

Kerui Peng (University of Southern California), Jaqueline Brito (University of Southern California), Guoyun Kao (University of Southern California) and Serghei Mangul (University of Southern California).

pyTCR is a comprehensive platform with a rich set of functionalities of TCR repertoire analysis for biomedical researchers. Our cloud-based easy to use platform is based on the interactive notebook with the enhancement of reproducibility and transparency, by providing comprehensive and integrative functions, and customizable manipulations. The platform that pyTCR utilizes is interactive notebooks which code and results are all available to the users. pyTCR provides six types of analysis, including basic statistical analysis, clonality analysis, overlap analysis, segment usage analysis, diversity analysis, motif analysis. In each analysis type, metrics, visualization, and statistical analysis are provided, which offers a comprehensive solution to TCR analysis.

## *Overcoming uncollapsed haplotypes in long-read assemblies of non-model organisms*

Nadège Guiglielmoni (Université libre de Bruxelles), Antoine Houtain (Université de Namur), Alessandro Derzelle (Université de Namur), Karine Van Doninck (Université libre de Bruxelles) and Jean-François Flot (Université libre de Bruxelles).

Long reads have revolutionized the field of genome assembly and have made highly contiguous assemblies accessible for all genomes. Most long-read assemblers aim to produce a haploid assembly, regardless of the actual ploidy of the genome being assembled. For diploid and polyploid genomes, haplotypes are collapsed into a single sequence to represent every region exactly once in the assembly. Haplotype collapsing is especially challenging for non-model diploid or polyploid genomes, as they often display variable levels of heterozygosity across their genomes, and haploid assemblies often contain artefactual duplications due to remaining haplotigs.

We designed a benchmark of haploid assembly strategies, combining read filtering, different long-read assemblers, and haplotig-purging tools. We tested these strategies on the genome of a non-model diploid organism, Adineta vaga, for which high-coverage Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (Nanopore) low-accuracy long reads were available. We defined four scores to identify the best haploid assemblies: assembly size, contiguity, completeness, and haploidy, the latter using a new metric implemented in the tool HapPy.

The end purpose of this benchmark is to provide users with a methodology to obtain haploid assemblies of non-model eukaryote organisms with high contiguity and completeness, and that suit their computational requirements.

## *Nested Stochastic Block Models Applied to the Analysis of Single Cell Data*

Leonardo Morelli (Center for Omics Sciences at the IRCCS Ospedale San Raffaele (COSR)), Valentina Giansanti (Center for Omics Sciences at the IRCCS Ospedale San Raffaele (COSR)) and Davide Cittaro (Center for Omics Sciences at the IRCCS Ospedale San Raffaele (COSR)).

Single cell profiling represents a powerful and well established tool to understand complex behaviour of heterogeneous biological systems. One of the key step in single cell analysis is identification of cell groups to describe functional properties of the cell mixture under investigation. To this end, several approaches have been implemented, nowadays many converge on community detection in neighbourhood graphs by optimization of modularity. We propose an alternative and principled solution to this problem, based on Nested Stochastic Block Models, which identifies cell groups in a probabilistic way, returning a hierarchical description cell partitions. As baseline results, we found that our approach correctly identifies cell populations in several datasets; in addition we are able to find that the hierarchic description is more conservative in terms of cell identity compared to arbitrary choice of a resolution parameter. Lastly, we exploit the properties of the underlying generative model to perform robust label transfer across single cell datasets. To facilitate the adoption of Nested Stochastic Block Models, we developed a python library, schist, that is compatible with the popular scanpy framework.

## *RENANO: a REference-based compressor for NANOpore FASTQ files*

Guillermo Dufort Y Álvarez (Universidad de la República), Gadiel Seroussi (Universidad de la República, Uruguay and XPERI Corp., CA, USA), Pablo Smircich (Universidad de la República), José Sotelo (Universidad de la República), Idoia Ochoa (Department of Electrical Engineering, TECNUN, University of Navarra) and Álvaro Martín (Universidad de la República).

Nanopore sequencing technologies are rapidly gaining popularity, in part, due to the massive amounts of genomic data they produce in short periods of time (up to 8.5 TB of data in < 72 hours). To reduce the costs of transmission and storage, efficient compression methods for this type of data are needed. In this note, we introduce RENANO, a reference-based lossless data compressor specifically tailored to FASTQ files generated with nanopore sequencing technologies. RENANO improves on its predecessor ENANO, currently the state of the art, by providing a more efficient base call sequence compression component. Two compression algorithms are introduced, corresponding to the following scenarios: (1) a reference genome is available without cost to both the compressor and the decompressor; (2) the reference genome is available only on the compressor side, and a compacted version of the reference is included in the compressed file. We compare the compression performance of RENANO against ENANO on several publicly available nanopore datasets. RENANO improves the base call sequences compression of ENANO by 40.8% in scenario (1), and by 33.4% in scenario (2), on average, over all the datasets. As for total file compression, the average improvements are 13.1% and 10.7%, respectively.

## *NanoSpring: reference-free lossless compression of nanopore sequencing reads using an approximate assembly approach*

Qingxi Meng (Stanford University), Shubham Chandak (Stanford University), Yifan Zhu (Stanford University) and Tsachy Weissman (Stanford University).

Motivation: The amount of data produced by genome sequencing experiments has been growing rapidly over the past several years, making compression important for efficient storage, transfer and analysis of the data. In recent years, nanopore sequencing technologies have seen increasing adoption since they are portable, real-time and provide long reads. However, there has been limited progress on compression of nanopore sequencing reads obtained in FASTQ files. Previous work ENANO focuses mostly on quality score compression and does not achieve significant gains for the compression of read sequences over general-purpose compressors like Gzip. RENANO achieves significantly better compression for read sequences but is limited to aligned data with a reference available.

Results: We present NanoSpring, a reference-free compressor for nanopore sequencing reads, relying on an approximate assembly approach. NanoSpring achieves close to 2.5-3x improvement in compression over state-of-the-art reference-free compressors. The computational requirements of NanoSpring are practical, although it uses more time and memory than previous compressors to achieve the compression gains. NanoSpring is available on GitHub at https://github.com/qm2/NanoSpring.

## *LongStitch: High-quality genome assembly correction and scaffolding using long reads*

Lauren Coombe (BC Cancer Genome Sciences Centre), Janet Li (BC Cancer Genome Sciences Centre), Theodora Lo (BC Cancer Genome Sciences Centre), Johnathan Wong (BC Cancer Genome Sciences Centre), Vladimir Nikolic (BC Cancer Genome Sciences Centre), Rene Warren (BC Cancer Genome Sciences Centre) and Inanc Birol (BC Cancer Genome Sciences Centre).

Generating high-quality de novo genome assemblies remains an essential step in many analysis pipelines for gaining new insights into both model and non-model organisms. Long-read sequencing has demonstrated great benefit to genome assembly scaffolding through providing long-range evidence to span problematic repetitive genomic regions. Here, we present LongStitch, an efficient pipeline that corrects and scaffolds draft genome assemblies using long reads. LongStitch incorporates multiple tools developed by our group: Tigmint-long, then ntLink, and optionally ARKS-long. Tigmint-long and ARKS-long are correction and scaffolding utilities, respectively, previously developed for linked reads, and are now adapted to use long reads. Within LongStitch, we introduce our new long-read scaffolder, ntLink, which utilizes lightweight minimizer mappings to join contigs. LongStitch was tested using short and long-read assemblies of three human individuals, and improves the contiguity of each assembly from 2.0-fold to 304.6-fold (measured by NGA50 length). Furthermore, LongStitch generates more contiguous and correct assemblies than a state-of-the-art long-read scaffolder, LRScaf, for most tests, and consistently runs in under 5 hours using less than 23GB of RAM. Due to its efficiency and flexibility in improving draft assemblies using long reads, we expect LongStitch to benefit a wide variety of de novo genome assembly projects.

## *Variability and transcriptional properties of human rDNA repeats using long-read sequencing technologies.*

Emiliana Weiss (ANU), Lex Maxim van Loon (ANU), Nadine Hein (ANU), Nikolay Shirokikh (ANU), Austen Ganley (The University of Auckland), Ross Hannan (ANU) and Eduardo Eyras (ANU).

Ribosomal RNA genes (rRNAs) are encoded in the genome in hundreds of copies (rDNA) to satisfy the high demand for ribosomes in a cell. The repetitive nature of the rDNA has hindered its study, and currently all rDNA repeat copies are generally assumed to be identical to each other. Using Nanopore sequencing data from the lymphoblastoid cell line GM24385, we identified 918 reads of length >100kb containing a total of 3300 candidate rDNA repeat units. The rDNA units had highly conserved sizes, suggesting that they maintain full coding potential. We further predicted the patterns of CpG methylation on these reads and found two starkly contrasting methylation patterns with similar proportions (~50% each). One with the rRNA genes and promoter unmethylated and another pattern with the rRNA genes and promoter methylated. Interestingly, most (~90%) of the 918 reads analyzed had the same methylation pattern in all units, rather than alternating between methylated and unmethylated. Moreover, we found reads with inversions resulting in units with diverging and converging transcriptional orientations. This variability allows us to describe the sequence determinants of transcriptional activity and how this is organised in the rDNA repeat arrays.

## *CNVpytor: a tool for CNV/CNA detection and analysis from read depth and allele imbalance in whole genome sequencing*

Milovan Suvakov (Mayo Clinic), Arijit Panda (Mayo Clinic), Colin Diesh (University of California, Berkeley), Ian Holmes (University of California, Berkeley) and Alexej Abyzov (Mayo Clinic).

Detecting copy number variations (CNVs) and copy number alterations (CNAs) based on whole genome sequencing data is important for personalized genomics and treatment. CNVnator is one of the most popular tools for CNV/CNA discovery and analysis based on read depth (RD). Herein, we present an extension of CNVnator developed in Python -- CNVpytor. CNVpytor inherits the reimplemented core engine of its predecessor and extends visualization, modularization, performance, and functionality. Additionally, CNVpytor uses B-allele frequency (BAF) likelihood information from single nucleotide polymorphism and small indels data as additional evidence for CNVs/CNAs and as primary information for copy number neutral losses of heterozygosity. CNVpytor is significantly faster than CNVnator—particularly for parsing alignment files (2 to 20 times faster)—and has (20-50 times) smaller intermediate files. CNV calls can be filtered using several criteria, annotated, and merged over multiple samples. Modular architecture allows it to be used in shared and cloud environments such as Google Colab and Jupyter notebook. Data can be exported into JBrowse, while a lightweight plugin version of CNVpytor for JBrowse enables nearly instant and GUI-assisted analysis of CNVs by any user. CNVpytor release and the source code are available on GitHub at https://github.com/abyzovlab/CNVpytor under the MIT license.

## *Distribution-free differential expression analysis for scRNA-seq data across patient groups*

Erika Dudkin (University of Bonn), Kevin Bassler (Universaty of Bonn), Joachim Schultze (University of Bonn) and Jan Hasenauer (University of Bonn).

Single cell RNA-sequencing (scRNA-seq) data provide insights into gene expression profiles of individual cells on a large scale. This contributed in recent years substantially to the understanding and identification of cell types and differences between them. To unravel differences between cell populations, a multitude of differential expression (DE) methods has been introduced to compare clusters of cells. However, these methods are not suited for the identification of differences between patient groups for which scRNA-seq data are available. The emergence of scRNA-seq datasets with replicated multi-conditions demands the development of new particular methods.

In this work, we present a method for the statistical comparison of replicated multi-conditions. The method uses Wilcoxon rank sum test for the pairwise comparison of samples. Differences between patient combinations are evaluated while taking all single cell read counts into account. After calculating the test statistic, its significance is determined with a permutation test.

The proposed method was tested with a simulation study. This study showed that the proposed method is able to detect differences in distributions across patient groups with a similar mean, while maintaining a low False Positive Rate.

## *Nucleosome positioning based identification of tissue contributions in cell-free DNA*

Sebastian Röner (Berlin Institute of Health (BIH)) and Martin Kircher (Berlin Institute of Health).

Cell-free DNA (cfDNA) is found in many bodily fluids and is believed to derive primarily from apoptosis of hematopoietic cells. In the context of certain physiological conditions or disease processes, the proportion of tissues contributing to cfDNA changes. These observations led to an increased research interest in cfDNA for so-called liquid biopsies.

Besides tracing genetic alleles and methylation states, past studies showed that cfDNA fragmentation is associated with nucleosome footprints and DNA binding (Snyder et al.,2016). We previously prototyped a pipeline based on Windowed Protection Scores and quantification of nucleosome distances from Fast Fourier Transformation. We showed that cfDNA from healthy individuals most strongly correlates with expression of hematopoietic cell-types. In contrast, in samples from late-stage cancer patients the major contributions align with the cancer's tissue-of-origin.

Here, we describe an easy-to-use computational pipeline implemented to identify these major contributions to cfDNA samples (https://github.com/kircherlab/cfDNA). Based on read alignments to GRCh37 or GRCh38, nucleosome-positioning signals around transcribed genes are automatically quantified and correlated with gene expression values of the Human Protein Atlas (Uhlén et al., 2015). The most correlated expression profiles are highlighted for each sample, with the option to contrast them to another sample (e.g. disease vs. control, time points).

## *clusterExplorer: an R/Shiny app for single-cell RNA-seq cluster visualizations*

Carolin Walter (Westfälische Wilhelms-Universität Münster) and Martin Dugas (Westfälische Wilhelms-Universität Münster).

Single-cell RNA sequencing (scRNA-seq) is a powerful biological technique that offers valuable insight into cellular processes and related structures, however the comparison between different datasets is no trivial task. We present the R/Shiny app clusterExplorer, which offers visualization routines for scRNA-seq clusters in datasets with two predefined subsets, e.g. integrated datasets containing original and published scRNA-seq data, or datasets with different biological conditions. Data preprocessing is conducted with the Seurat pipeline separately for each dataset or subset. clusterExplorer subsequently allows the user to select one of Seurat's UMAP clusters from the separate clusterings for parallel visualization in all cluster sets that contain subsets of the chosen cells. For the resulting cell cluster projections, all cluster identities are compared between the datasets, and information regarding the three best-matching clusters in the other dataset's clustering are provided. In addition, clusterExplorer offers a convex hull representation of each cluster, in which transparent overlays for chosen quantiles of the cells allow to assess the general cluster structure. Furthermore, the convex hull approach allows to classify the relative cluster structure of the projection as “compact” or “scattered”, and thus to identify conserved or split cluster structures between datasets.

## *Studying the reproducibility of RNA-seq experiments by generation of artificial replicates: a comparison of methods*

Babak Saremi (University of Veterinary Medicine Hannover), Frederic Gusmag (University of Veterinary Medicine Hannover), Ottmar Distl (University of Veterinary Medicine Hannover), Julia Metzger (Max-Planck Institute for Molecular Genetics Berlin), Stefanie Becker (University of Veterinary Medicine Hannover) and Klaus Jung (University of Veterinary Medicine Hannover).

Although RNA-seq experiments have been found to be highly reproducible, cases have been reported were technical replicates can improve the power to detect differentially expressed genes or to detect potential lane effects of the flow cell. Using technical replicates is, however, usually too expensive.

We evaluate the use of three different approaches for generating artificial replicates in RNA-seq experiments: 1) bootstrapping reads from FASTQ-files (FB), 2) mixing observations (MO), and 3) bootstrapping from the columns of the count data matrix (BC). We used a data set generated in our own lab with one group of virus infected samples versus one control group, and with two technical replicates per sample.

The three methods were run to generate 10 artificial replicates per sample, resulting in 10 additional lists of p-values and log fold changes (logFC). We found that logFCs and p-values from the artificial replicates generated with FB are closer to those from R1 and R2 than logFCs obtained from replicates generated by MO or BC. The preliminary results suggest that bootstrapping from FASTQ files produces artificial replicates that are close to true technical replicates.

## *A comparison of methods for copy number variation analysis from single-cell RNA-seq data*

Rongting Huang (The University of Hong Kong), Julia Lam (The University of Hong Kong) and Yuanhua Huang (The University of Hong Kong).

Somatic copy number variation (CNVs) are major mutations in various cancers for their development and clonal evolution. Analysing CNV in single-cell RNA-seq data is of critical importance for both detecting the CNV states in tumour cells and revealing its impact on transcriptional phenotypes. A few computational methods have been recently proposed to analyse CNV from scRNA-seq data. However, their accuracy in identifying various CNV types and computational efficiency have not been well benchmarked, partly due to the lack of gold standard data sets. Here, we compared four commonly used methods (inferCNV, CopyKAT, HoneyBADGER, CaSpER) on a well characterised and verified gastric cancer sample GX109 for their accuracy in detecting copy loss, gain and loss of heterozygosity. Their robustness was also assessed by varying a few key parameters. Additionally, we characterised two public datasets (TNBC1 and BCH869) that have clear subclonal structures to evaluate the capability of dissecting clonal structure within a tumor tissue. Taken together, this evaluation provides a reference for method choice when analysing CNV in scRNA-seq data and highlights the existing challenges, and the high-quality annotation may further accelerate the development of more tailored and sophisticated methods.

## *Efficient linked-read barcode mapping without read alignment*

Richard Lüpken (Berlin Institute of Health) and Birte Kehr (Regensburg Center for Interventional Immunology (RCI)).

When sequencing whole genomes, one is facing a tremendous amount of mostly unstructured data. Obtaining all reads corresponding to a specific genomic location currently requires the computationally expensive alignment of all reads. Linked-read sequencing technologies provide an additional level of structure in their reads through the use of barcodes. Reads with the same barcode originate from a small set of large DNA molecules. This provides opportunities that have not yet been used to their full potential. Here we introduce an efficient approach for determining barcode intervals in a reference genome without performing a costly read alignment. Simultaneously we construct an index to quickly retrieve all reads of a given barcode from the input read files. Our barcode mapping approach queries minimizers from an open addressing k-mer index, which are then clustered into barcode intervals using a sliding window approach based on a scoring function. Mapping barcodes of a full set of reads took us 6.5 CPU hours whereas aligning the same read set with BWA mem took 244 CPU hours. When faced with WGS data but interested in a specific genomic location, our approach can quickly return all barcodes and reads belonging to the locus of interest.

## *Improving long-read consensus sequencing accuracy with deep learning*

Avantika Lal (NVIDIA), Michael Brown (Pacific Biosciences), Rahul Mohan (NVIDIA), Joyjit Daw (NVIDIA), James Drake (Pacific Biosciences) and Johnny Israeli (NVIDIA).

The PacBio HiFi sequencing technology is based on single-molecule, real-time (SMRT) sequencing of a circularized DNA molecule in repeated passes, producing multiple subreads which are individually approximately 90% accurate. Combining these subreads into a consensus sequence yields HiFi sequence reads that are both long (10-25 kb) and highly accurate (>99.9%). Here, we explored the utility of deep learning models for improving HiFi read consensus accuracy. We trained deep learning models with a variety of architectures and encoding types on a human HiFi dataset (HG002). Data was encoded as frequency counts of nucleotides at each position in a pileup of subreads. These counts along with the HiFi read sequence itself and its estimated read quality served as input to the model. Based on cross-validation results, an architecture with multiple convolutional layers followed by a recurrent layer to integrate long-range information achieved the best performance, reducing errors by 20-40% depending on the dataset. Additionally, we took the HG002-trained model and tested it on a different species (E. coli), and reduced errors by 21%. Our work demonstrates the feasibility of deep learning for reducing error rates for PacBio's circular consensus sequencing data type.

## *Improving SV calling in FFPE samples with FilterFFPE*

Lanying Wei (Institute of Medical Informatics, University of Münster, Germany), Martin Dugas (Institute of Medical Informatics, University of Münster, Germany) and Sarah Sandmann (Institute of Medical Informatics, University of Münster, Germany).

Next-generation sequencing data from formalin-fixed paraffin-embedded (FFPE) samples is enriched with artifact chimeric reads. These reads are generated during the sequencing process, rather than formed due to real structural variants (SV). However, existing SV detection tools cannot distinguish between these two kinds of chimeric reads, thus resulting in a large number of false positive SV calls. To specifically remove artifact chimeric reads, we developed FilterFFPE, a two-step filtering algorithm. While the first step identifies all possible artifact chimeric reads, the optional second step is specifically designed for samples with low coverage and/or low SV frequency. To evaluate the benefit of the second step, we considered two common tools Delly and Lumpy for SV calling. For simulated data with low coverage or low SV frequency, both tools showed clearly superior performance for the 2-step procedure (F1_noFiltration=0.62, F1_1-step=0.67, F1_2-step=0.72). Evaluating simulated samples with high coverage and high SV frequency, only marginal difference between the 1- and 2-step procedure can be observed (F1_noFiltration=0.54, F1_1-step=0.75, F1_2-step=0.74). Therefore, we propose to add FilterFFPE to every SV calling pipeline in FFPE samples. The decision to use the second filtering step should be based on sample coverage and heterogeneity.

## *Modeling Coverage in Whole Exome Sequencing Data Using Machine Learning Techniques*

Marius Wöste (Institute of Medical Informatics, University of Münster), Frank Tüttelmann (Institute of Reproductive Genetics, University of Münster) and Martin Dugas (Institute of Medical Informatics, University of Münster).

Sequencing coverage is a metric commonly used in whole exome sequencing (WES) experiments for quality control purposes and detection of copy-number variants (CNVs). However, coverage in WES data is often biased (e.g. due to GC-content in target regions) and shows high variance even in samples from the same sequencing run. Precise WES coverage models could thus potentially improve CNV detection and may, thereby, contribute to identifying genetic variants associated with disease.

We used a data set of 370 exomes to model WES coverage using machine learning algorithms. We first trained models for each sample individually using features related only to target properties (e.g. target length and GC-content). Our trained models only marginally reduced error compared to naively guessing coverage as the median of the sample. Subsequently, we included coverage information of samples from the same sequencing run for model training, resulting in ~80% reduction of root-mean-square errors compared to models based only on coverage from a single sample.

Our results indicate that WES coverage models trained on a single sample using simple target features are of limited use. We thus recommend training WES coverage models on multiple samples, e.g. by utilizing samples from the same sequencing run.

## *Consensus-based identification and comparative analysis of structural variants by long and short-read sequencing technologies in selected human families*

Mateusz Chiliński (Centre of New Technologies, University of Warsaw, S. Banacha 2c, 02-097 Warsaw, Poland), Sachin Gadakh (Centre of New Technologies, University of Warsaw, S. Banacha 2c, 02-097 Warsaw, Poland) and Dariusz Plewczynski (Centre of New Technologies, University of Warsaw, S. Banacha 2c, 02-097 Warsaw, Poland).

We present a comprehensive analysis of Oxford Nanopore (ONT) sequencing technology compared with short-read techniques, such as Illumina. In our study, we focus on the structural variants, at least 50 bp segments of DNA in length that are unique for personal genomes, as identified by the 1000 Genomes project. We improve the quality of the Structural Variants identification from the whole genome sequencing (WGS) experiments by using the consensus approach. Fifteen gold-standard tools were used for obtaining the polished list of Structural Variants (SV) for daughters of families from the 1000 Genomes project using publicly available datasets from next-generation sequencing experiments performed by both short-read (Illumina) and long-read (ONT) technologies. The results of the SV callers were merged using the novel ConsensuSV algorithm, which integrates the SV sets using machine learning by combining decision trees and neural networks trained and benchmarked on the high-quality SVs from the 1000 Genomes Project. Finally, upon comparing the SV sets obtained from ConsensuSV algorithm between long and short read, our findings demonstrate the superiority of ONT across all SV sizes, long-read-based SV inference detected more SVs than the short-read one.

## *The statistics of kmers from a sequence undergoing a simple mutation process without spurious matches*

Antonio Blanca (Penn State), Robert S. Harris (The Pennsylvania State University), David Koslicki (Penn State University) and Paul Medvedev (The Pennsylvania State University).

K-mer-based methods are widely used in bioinformatics, but there are many gaps in our understanding of their statistical properties. Here, we consider the simple model where a sequence S (e.g. a genome or a read) undergoes a simple mutation process whereby each nucleotide is mutated independently with some probability r, under the assumption that there are no spurious k-mer matches. How does this process affect the k-mers of S? We derive the expectation and variance of the number of mutated k-mers and of the number of islands (a maximal interval of mutated k-mers) and oceans (a maximal interval of non-mutated k-mers). We then derive hypothesis tests and confidence intervals for r given an observed number of mutated k-mers, or, alternatively, given the Jaccard similarity (with or without minhash). We demonstrate the usefulness of our results using a few select applications: obtaining a confidence interval to supplement the Mash distance point estimate, filtering out reads during alignment by Minimap2, and rating long read alignments to a de Bruijn graph by Jabba.

## *Matrix prior for data transfer between single cell data types in latent Dirichlet allocation*

Alan Min (University of Washington), Timothy Durham (University of Washington, Broad Institute), Louis Gevirtzman (University of Washington) and William Noble (University of Washington).

Single cell ATAC-seq (scATAC-seq) enables the mapping of regulatory elements in fine-grained cell types, but analysis of the resulting data is challenging, and large scale scATAC-seq data are difficult to obtain and expensive to generate. This motivates a method to leverage information from previously generated large scale scATAC-seq or scRNA-seq data to guide our analysis of scATAC-seq data sets. We analyze scATAC-seq data using latent Dirichlet allocation (LDA), a Bayesian algorithm that was developed to model text corpora, condensing documents into mixtures of topics. Recently, this approach successfully identified topics that distinguished between cell types, but has focused on using symmetric priors in LDA, meaning that the prior puts equal weights on peaks or genes for every topic. We hypothesized that nonsymmetric priors constructed using auxiliary data, which give peaks or genes unequal weights, may enable more accurate resolution of cell types. We verified our method in simulated data, and then analyzed data from whole C. elegans nematodes, where we used large sets of scATAC-seq and scRNA-seq data to construct nonsymmetric priors for analysis of a target scATAC-seq data set. We show that these priors improved our ability to capture cell type information and form improved cell clusters.

## *BinSPreader: refine binning results for fuller MAG reconstruction*

Yury Kamenev (ITMO University), Roman Kruglikov (Lomonosov Moscow State University), Ivan Tolstoganov (Saint Petersburg State University) and Anton Korobeynikov (Saint Petersburg State University).

Despite the recent advances in high-throughput sequencing, analysis of the metagenome of the whole microbial population still remains a challenge. In particular, the metagenome-assembled genomes (MAGs) are often fragmented due to interspecies repeats, uneven coverage and vastly different strain abundance. MAGs are usually constructed via a dedicated binning process that uses different features of input data in order to cluster contigs that might belong to the same species. This process has some limitations and therefore binners usually discard contigs that are shorter than several kilobases. Therefore, binning of even simple metagenome assemblies can miss a decent fraction of contigs and resulting MAGs oftentimes do not contain important conservative sequences. In this work we present BinSPreader – a novel binning refiner tool that exploits the assembly graph topology and other connectivity information to refine the existing binning, correct binning errors, propagate binning from longer contigs to shorter contigs and infer contigs belonging to multiple bins. Furthermore, BinSPreader can split input reads in accordance with the resulting binning predicting reads potentially belonging to multiple MAGs. We show that BinSPreader could effectively complete the binning increasing the completeness of the bins without sacrificing the purity and could predict contigs belonging to several MAGs.

# *DNA methylation calling tools for Oxford Nanopore sequencing: a survey and human epigenome-wide evaluation*

Yang Liu (The Jackson Laboratory), Wojciech Rosikiewicz (The Jackson Laboratory), Ziwei Pan (The Jackson The Jackson Laboratory, UConn Health Center), Nathaniel Jillette (The Jackson Laboratory), Aziz Taghbalout (The Jackson Laboratory), Jonathan Foox (Weill Cornell Medicine), Christopher Mason (Weill Cornell Medicine), Martin Carroll (University of Pennsylvania), Albert Cheng (The Jackson Laboratory) and Sheng Li (The Jackson Laboratory, UConn Health Center, University of Connecticut).

DNA methylation is a fundamental epigenetic modification process in gene transcription. Nanopore sequencing enables long-range base modification detection, e.g., DNA 5-methylcytosine (5mC) at single-molecule, single-base resolution. DNA methylation calling tools for Nanopore sequencing is emerging. However, their robustness for human natural DNA at the epigenome scale remains unclear, especially at the single-molecule resolution, which is critical for long-range epigenetic allele detection. Thus, we benchmarked DNA methylation calling tools using multiple human Nanopore sequencing datasets. We compared Nanopolish, Megalodon, DeepSignal, Tombo, and DeepMod at single-base, single-molecule resolution for a systematic evaluation on various genomic regions, e.g., singleton and non-singleton sites, CpG island, generic and intergenic regions, transcription start sites (TSS), transcriptional factors CCCTC-binding factor (CTCF) binding peaks, running speed and computing resource usage. We revealed a new bottleneck that specific genomic locations such as discordant regions have an effect on prediction accuracy, and there is a trade-off between detection accuracy, CpG site coverage, and running time for methylations calling. Meanwhile, we offered a customized recommendation for both practitioners to maximize the DNA modification detection capabilities. In summary, our work is the first to benchmark state-of-the-art DNA methylation calling tools for Nanopore sequencing for 5mC detection for epigenome-wide native human genome study.

## *Robust and accurate estimation of paralog-specific copy number for duplicated genes using whole-genome sequencing*

Timofey Prodanov (University of California San Diego) and Vikas Bansal (University of California San Diego).

Copy number and sequence variation in more than 150 genes that overlap low-copy repeats (LCRs) is associated with risk for rare and complex human diseases. Such duplicated genes are problematic for standard NGS analysis pipelines since a large fraction of reads derived from these regions cannot be mapped unambiguously to the genome. We have developed a computational framework, Parascopy, to estimate the total and paralog-specific copy number of genes that overlap LCRs using whole-genome sequencing (WGS) data. Parascopy jointly analyzes reads aligned to a genomic region and its paralogous sequences without relying on read mapping quality, uses a multi-sample Hidden Markov Model (HMM) to infer aggregate copy number, and leverages an EM algorithm to jointly estimate paralog-specific copy number and identify invariant paralogous sequence variants or PSVs. Analysis of WGS data for 2504 samples from the 1000 Genomes project and validation using experimental data shows that Parascopy outperforms existing methods for several disease-relevant genes such as SMN1/2, RHCE and SRGAP2, can automatically identify invariant PSVs and can estimate copy number for more than 165 duplicated gene loci for a single human genome in less than 20 minutes.

## *Co-linear chaining with overlaps and gap costs*

Chirag Jain (Indian Institute of Science), Daniel Gibney (University of Central Florida) and Sharma V. Thankachan (University of Central Florida).

Co-linear chaining has proven to be a powerful technique for finding approximately optimal alignments and approximating edit distance. It is used as an intermediate step in numerous mapping tools that follow seed-and-extend strategy. Despite this popularity, subquadratic time algorithms for the case where chains support anchor overlaps and gap costs are not currently known. Moreover, a theoretical connection between co-linear chaining cost and edit distance remains unknown. We present algorithms to solve the co-linear chaining problem with anchor overlaps and gap costs in $\tilde O(n)$ time, where $n$ denotes the count of anchors. We establish the first theoretical connection between co-linear chaining cost and edit distance. Specifically, we prove that for a fixed set of anchors under a carefully designed chaining cost function, the optimal `anchored' edit distance equals the optimal co-linear chaining cost. Finally, we demonstrate experimentally that optimal co-linear chaining cost under the proposed cost function can be computed significantly faster than edit distance, and achieves high correlation with edit distance for closely as well as distantly related sequences.

## *Novel alternative splicing events detection in mouse genome with Spladder*

Agata Muszyńska (Malpolska Centre of Biotechnology), Ryszard Przewłocki (Department of Molecular Neuropharmacology, Institute of Pharmacology Polish Academy of Sciences, Kraków, Poland) and Paweł P. Łabaj (Małopolska Centre of Biotechnology of Jagiellonian University).

The mouse is a widely studied animal, but the complexity of its transcriptome is still not fully understood. One of the mechanisms that stands for this is alternative splicing. Currently, we are not fully aware of all the alternative splicing events (ASE) that can occur in a given transcriptome. One of the tools that allows us to investigate this is Spladder. It builds a splicing graph based on the current annotation and then expands it with new events. We applied Spladder to neuropathic pain data. There were 88 samples in total, but we initially focused on analyzing 24 samples for wildtype and neuropathic pain animals. Despite the fact that in previous analysis of differential gene expression the reproducibility was very low, great proportion of the novel ASE detected by Spladder were common between groups.The next step was to examine how the results overlapped for all 88 samples. Still, the overlap was surprisingly high, reaching up to 40% of the common events. This result might indicate that the mouse reference model lacks information for brain tissue. It could also reflect neuroplasticity - the fact that new connections are constantly being made in the brain and different tissues evolve over time.

## *New algorithms for accurate and efficient de-novo genome assembly from long DNA sequencing reads*

David Guevara-Barrientos (Universidad de los Andes), Laura Gonzalez (Universidad de los Andes), Daniela Lozano (Universidad de los Andes), Juanita Gil (University of Arkansas), German Andrade (Universidad de los Andes), Maria Camila Hoyos (Universidad de los Andes), Christian Chavarro (Universidad de los Andes), Natalia Guayazan (Universidad de los Andes), Luis Alberto Chica (Universidad de los Andes), Maria Camila Buitrago (Universidad de los Andes), Edwin Bautista (Universidad de los Andes), Juan Camilo Bojacá (Universidad de los Andes), Miller Trujillo (Universidad de los Andes), Maria del Rosario Leon (Universidad de los Andes), Fernando Reyes (Universidad de los Andes) and Jorge Duitama (Universidad de los Andes).

DNA sequencing using long-read technologies is becoming a common task in different research projects, producing high-quality de-novo haploid and diploid genome assemblies. Although current solutions achieve contiguous assemblies of complex genomes, new algorithmic techniques have the potential to further improve the accuracy and computational efficiency to build both haploid and diploid genome assemblies. We present here the design and implementation of new algorithmic approaches for assembly of large DNA sequencing reads, following the overlap-layout-consensus (OLC) process. We build a two-vertex-per-read undirected graph from minimizers with hash codes based on rankings of k-mers according to their distance from the k-mer count mode. Different statistics are collected from overlaps and used as features to identify edges for layout paths as a machine learning binary classification problem. Experiments with Pacific Biosciences HiFi data from three different species shows that our algorithms are efficient to generate accurate assemblies for all cases. Furthermore, we integrated previous works on single individual haplotyping into the layout construction to build phased assemblies covering important regions in the human genome such as the major histocompatibility complex. We expect that this work contributes to the development of algorithms to achieve chromosome-level assemblies of complex genomes.

# *ATAC-DoubletDetector: Multiplet detection in single nucleus ATAC-seq by exploiting its unique data features*

Asa Thibodeau (The Jackson Laboratory for Genomic Medicine), Alper Eroglu (The Jackson Laboratory for Genomic Medicine), Nathan Lawlor (The Jackson Laboratory for Genomic Medicine), Djamel Nehar-Belaid (The Jackson Laboratory for Genomic Medicine), Romy Kursawe (The Jackson Laboratory for Genomic Medicine), Radu Marches (The Jackson Laboratory for Genomic Medicine), George A. Kuchel (University of Connecticut Center on Aging, UConn Health Center), Jacques Banchereau (The Jackson Laboratory for Genomic Medicine), Michael L. Stitzel (The Jackson Laboratory for Genomic Medicine), A. Ercument Cicek (Computer Engineering Department, Bilkent University) and Duygu Ucar (The Jackson Laboratory for Genomic Medicine).

Similar to other droplet-based single cell assays, single nucleus ATAC-seq (snATAC-seq) data harbor multiplets that confound downstream analyses. Detecting multiplets in snATAC-seq data is particularly challenging due to data sparsity and limited dynamic range (0 reads: closed chromatin, 1: open on one parental chromosome, 2: open on both chromosomes). Yet, these unique data features offer an opportunity to identify multiplets. ATAC-DoubletDetector detects multiplets by studying the number of regions with >2 uniquely aligned reads across the genome, an effective alternative to methods based on artificially-generated multiplets. For benchmarking we generated data from two primary human tissues: peripheral blood mononuclear cells (PBMCs) and pancreatic islets. When a certain read depth per nucleus is achieved (>20K in PBMCs), ATAC-DoubletDetector captured 85% of simulated doublets. Moreover, ATAC-DoubletDetector was equally effective in identifying homotypic multiplets (i.e., multiplets from the same cell type), which are missed by simulation-based methods. Cell-specific marker peaks enabled accurate (85%) tracing of cellular origins of snATAC-seq multiplets. Accordingly, more abundant cells within a tissue are more likely to form multiplets and the majority of multiplets are homotypic. ATAC-DoubletDetector is a fast and effective multiplet detection/annotation tool for improved single cell epigenomic data analyses across diverse biological systems and conditions.

## *Single-cell transcriptional landscape of the human fallopian tube microenvironment and its implications for high grade serous 'ovarian' cancer*

Joshua Brand (McCardle Laboratory for Cancer Research, Department of Oncology, University of Wisconsin), Marcela Haro (Woman's Cancer Research Program at the Samuel Oschin Comprehensive Cancer Center, Cedars Sinai Medical Center), Kate Lawrenson (Woman's Cancer Research Program at the Samuel Oschin Comprehensive Cancer Center, Cedars Sinai Medical Center) and Huy Dinh (McCardle Laboratory for Cancer Research, Department of Oncology, University of Wisconsin).

High grade serous 'ovarian' cancer (HGSOC) is the most lethal form of ovarian cancer with evidence suggesting origins in the fallopian tube (FT). Characterizing FT heterogeneity is therefore vital to understanding HGSOC progression and for prognostic biomarker discovery. However, most of this work has been focused on fallopian epithelia, leaving the cellular composition of the microenvironment poorly understood.

Here, we defined cellular heterogeneity of non-epithelial compartments and sought for cell-cell interactions that may potentially be perturbed during HGSOC development. To this end, we have utilized a comprehensive pipeline to identify and annotate cellular subsets for further integrated analysis with HGSOC subtypes from publicly available bulk RNA-Seq datasets using computational deconvolution framework. We identified diverse fibroblast subsets with distinct functional enrichments including complement activation, collagen deposition, and antigen presentation. Immune transcriptional signatures showed high myeloid diversification, but less heterogeneity within the T cell compartment despite their higher abundance. Deconvolution analyses highlighted specific fibroblasts and immune cell-types that may disproportionately contribute to the tumor microenvironment establishment.

In summary, we provided the most comprehensive of cellular atlas of human FT to date. We will present our on-going effort in addressing how cell-cell communication influences epithelial cellular differentiation and their implications for tumorigenesis.

## *Partially reconstructing miRNA complements from sncRNA-seq without a reference genome*

Ernesto Aparicio-Puerta (University of Granada), Cristina Gómez-Martín (Amsterdam UMC) and Michael Hackenberg (University of Granada).

MicroRNAs have important roles in many biological processes and their expression can be routinely carried out using reference mature sequences. The prediction of novel miRNA genes however, generally requires the availability of genome sequences in order to assess important properties such as the characteristic hairpin-shaped secondary structure. However, although sequencing costs have decreased over the last years, many important species still lack a high quality genome assembly. We implemented an algorithm which exploits characteristic biogenesis features that can be assessed without genomic sequences such as the 5' processing homogeneity. We assessed its performance using sequencing datasets from several Homo sapiens and Mus musculus tissues and reference mature miRNA sequences from miRGeneDB. miRNAgFree was able to correctly predict XX/YY and ZZ/AA from human and mouse respectively with a precision of X% and Y%. Overall, 90-100% of the most expressed predictions for each tissue corresponded to bona fide miRNAs. Furthermore we found that XX and YY tissues were needed to recover Z% of the miRNA compliment, which suggests that miRNA-seq data alone can be sufficient to reconstruct a relevant portion of a species miRNAome.

## *MMseqs2 profile/profile: fast and ultra sensitive searches beyond the twilight zone*

Hyunjoo Ji (Seoul National University), Milot Mirdita (Max Planck Institute for Biophysical Chemistry), Hans-Georg Sommer (Max Planck Institute for Biophysical Chemistry), Clovis Galiez (Université Grenoble Alpes), Johannes Söding (Max Planck Institute for Biophysical Chemistry) and Martin Steinegger (Seoul National University).

Analyses of high-throughput sequencing studies are producing billions of novel, uncharacterized protein sequences of previously unstudied organisms. State-of-the-art sequence-to-sequence alignment methods are well-tuned to cope with the avalanche of data. However, they are limited by their sensitivity as they rely on existing homologous sequences in reference databases within the daylight-zone of detectability. The most sensitive alignment methods, HHblits and HHsearch, are based on HMM to HMM (profile-profile) alignments and are able to detect remote homologies across vast evolutionary distances, well into the midnight-zone. However, they are limited in their applicability since they take minutes to process a single query, despite large optimization efforts. Here, we propose MMseqs2 Profile/Profile, the first profile-profile alignment method to match the sensitivity of HHblits at a much higher runtime speed. We extend the fast SIMD-accelerated implementation of the Striped Smith-Waterman-Gotoh algorithm in MMseqs2 to support profile-profile alignments and introduce efficient workflows for reverse- and iterative-searches for the construction of extremely diverse multiple sequence alignments. At nearly 4100x the speed of HMMER3 while being 20% more sensitivity. At over 60x the speed of HHblits, we match its sensitivity. Furthermore, we expect MMseqs2 Profile/Profile to scale well onto large query datasets, due to its efficient parallelization.

## *A hybrid model of genome sequence and chromatin structure for noncoding variants effect prediction*

Wuwei Tan (Electrical & Computer Engineering, Texas A&M University) and Yang Shen (Electrical & Computer Engineering, Texas A&M University).

Most disease related single nucleotide polymorphisms are noncoding genomic variants. It is a critical task to quantify the functional effects of noncoding variants. The most powerful predictive model focuses on the genome sequences and uses advanced deep learning architectures. We have found that the for some genome sequences, which are closer in 3D space, they have very different sequences but similar epigenetic event patterns. So, we believe the 3D chromatin structure can help the model to correct this inconsistency pattern and have better predictive power. We proposed a hybrid framework that uses both genome sequences and chromatin structures. The genome sequences are embedded, using the SOTA genome sequences only model's sequence embedding section, as node features. And the whole genome chromatin structures(DNA-DNA interaction frequency from Hi-C) are used as the edge features. The graph model makes it possible to use both intra and inter-chromatin genome sequences to predict the noncoding variants effects. Our results indicate that our hybrid framework over-perform the SOTA sequence only models. And our framework makes it possible to identify and analyze the effects of co-exist long distance single necleotide polymorphisms(SNPs) and expression quantitative trait loci(eQTLs).

## *Biological discovery and consumer genomics databases activate latent privacy risk in functional genomics data*

Zhiqiang Hu (University of California, Berkeley) and Steven Brenner (University of California, Berkeley).

The privacy risks from individuals' genomes have garnered increasing attention. Recent research studies and forensics have underscored the ability to re-identify a person using genomic-identified relatives and quasi-identifiers, such as sex, birthdate and zip code. However, summary omics data, such as gene expression values and DNA methylation sites, are generally treated as safe to share, with low privacy risks – though research studies have indicated they could be linked to existing genomes. We have demonstrated that some types of summary omics data can be accurately linked to a unique genome. We developed methods to match against genotypes in consumer genealogy databases with their restricted tools. Thus, the theoretical privacy concerns regarding summary omics data are now practically relevant. The ability to link sets of quasi-identifiers can reveal a research participant's identity and protected health information. Most important, such risks increase over time, activated by new techniques, new knowledge, and new databases. Thus public omics data may become privacy time bombs: safe at the time of distribution, but increasingly likely to compromise personal information. The need to preserve individuals' genomic privacy for their lifetime and beyond (for descendants and relatives) poses unique challenges to the effective sharing of high-throughput molecular data.

## *GeneMark-ETP: automatic integration of genomic, transcriptomic and protein data for gene prediction in eukaryotic genomes.*

Tomas Bruna (Georgia Tech), Alexandre Lomsadze (Georgia Tech) and Mark Borodovsky (Georgia Tech).

Bioinformatic pipelines are conventionally used as primary means of annotation of protein coding genes in massively sequenced novel eukaryotic genomes. Still, the performance of the current methods integrating the main streams of input data – genomic, transcriptomic as well as cross-species proteins is far from satisfactory. We present a new computational method, GeneMark-ETP, the most comprehensive in a series of the GeneMark gene finders with unsupervised parameter training. Earlier, we have developed GeneMark-ET to integrate the ab initio derived genomic patterns with information on intron positions revealed by mapping RNA reads. Subsequently, we introduced GeneMark-EP, a tool that leveraged a protein database to extract hints to border sites of exon-intron structures and improve estimation of model parameters. Another tool, ProtHint was proposed to infer and score the hints to guide the gene finding algorithm. The new GeneMark-ETP integrates the noisy albeit redundant streams of the all three types of information to generate reliable evidence for exon-intron structures in each genomic locus. Special attention is devoted to scoring and weighting schemes combining the evidence of different nature. The focus of this development is on improving annotation of large eukaryotic genomes with low gene density and abundance of repetitive sequences.

## *Single-cell landscape of nuclear configuration and gene expression during stem cell differentiation and X inactivation*

Giancarlo Bonora (Genome Sciences, University of Washington), Vijay Ramani (Genome Sciences, University of Washington), Ritambhara Singh (Genome Sciences, University of Washington), He Fang (Genome Sciences, University of Washington), Dana Jackson (Genome Sciences, University of Washington), Sanjay Srivatsan (Genome Sciences, University of Washington), Ruolan Qiu (Genome Sciences, University of Washington), Choli Lee (Genome Sciences, University of Washington), Cole Trapnell (Genome Sciences, University of Washington), Jay Shendure (Genome Sciences, University of Washington), Zhijun Duan (Department of Medicine, University of Washington), Xinxian Deng (Department of Laboratory Medicine and Pathology, University of Washington), William Noble (Genome Sciences, University of Washington) and Christine Disteche (Department of Laboratory Medicine and Pathology, University of Washington).

Mammalian development is associated with extensive changes in gene expression, chromatin accessibility, and nuclear structure. We follow these changes during mouse embryonic stem cell (mESC) differentiation and X chromosome inactivation (XCI) by integrating allele-specific data from these three modalities obtained by high-throughput single-cell RNA-seq, ATAC-seq, and Hi-C. Allele-specific analysis and integration of our single cell data led to the following key findings: (1) The inactive X chromosome (Xi) in differentiated cells has a unique contact decay profile. (2) This Xi-specific structure is lost at mitosis, followed by its reappearance during interphase. (3) Differentiation of ESCs is associated with changes in genome structure that occur in parallel on both the X chromosomes and autosomes. (4) Trajectory analyses of single-cell Hi-C data reveals three distinct nuclear structure states. (5) Single-cell RNA-seq and ATAC-seq show evidence of a delay in female versus male cells, due to the presence of two active X chromosomes at early stages of differentiation. (6) The onset of the Xi-specific structure in single cells occurs later than gene silencing, consistent with chromatin compaction being a late event of XCI. (7) Novel computational approaches allow for the effective alignment of single-cell gene expression, chromatin accessibility, and 3D chromosome structure.

## *distinct: a novel approach to differential distribution analyses*

Simone Tiberi (University of Zurich), Helena L Crowell (University of Zurich), Lukas M Weber (Johns Hopkins Bloomberg School of Public Health), Pantelis Samartsidis (MRC Biostatistics Unit, University of Cambridge) and Mark D Robinson (University of Zurich).

Motivation: High-throughput single-cell data reveal an unprecedented view of cell identity and allow complex variations between conditions to be discovered; nonetheless, most methods for differential expression target differences in the mean and struggle to identify changes where the mean is only marginally affected. Results: Here, we present distinct, a general method for differential analysis of full distributions that is well suited to applications on single-cell data, such as single-cell RNA sequencing (scRNA-seq) and high- dimensional flow or mass cytometry (HDCyto) data. distinct is based on a hierarchical non-parametric permutation approach and, by comparing empirical cumulative distribution functions (ECDFs), identifies both differential patterns involving changes in the mean, as well as more subtle variations that do not involve the mean. We performed extensive benchmarks across both simulated and experimental data, where distinct shows very favourable performance, identifies more differential patterns than competitors, and displays good control of false positive and false discovery rates. Availability: distinct is available as a Bioconductor R package.

## *BANDITS: Bayesian differential splicing accounting for sample-to-sample variability and mapping uncertainty*

Simone Tiberi (University of Zurich) and Mark Robinson (University of Zurich).

Motivation. Alternative splicing plays a fundamental role in the biodiversity of proteins as it allows a single gene to generate several transcripts and, hence, to code for multiple proteins. Differential splicing (DS) analyses aim at identifying changes in splicing patterns between conditions (e.g., healthy vs. disease or across treatments). A significant challenge in DS analyses is that, unlike gene level studies, the RNA-seq counts at the transcript level, which are of primary interest, are not observed because most reads map to multiple transcripts (i.e., multi-mapping reads).

Results. We developed BANDITS, a statistical method to perform DS from RNA-seq data. BANDITS uses a Bayesian hierarchical structure to explicitly model the variability between samples. Our tool also models mapping uncertainty: it inputs the equivalence class of each read and treats the allocations of reads to the transcripts as latent variables. BANDITS tests for DS at both transcript and gene level, allowing scientists to investigate what specific transcripts are differentially used in selected genes. Furthermore, our tool also allows testing between more than two groups.

Ww show how, both in simulation studies and experimental data analyses, the proposed methodology has very favourable performance.

Availability. BANDITS is available as a Bioconductor R package.

## *Cell trajectory inference for mouse stem cell differentiation using cell type information from single-cell RNA-seq data*

Yutaka Izeki (Graduate School of Information Science and Technology, Osaka University), Mona Mizutome (School of Engineering Science, 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).

Cells in organisms have a wide variety of functions due to their different expression profiles of genes. Therefore, it is possible to classify cells by analyzing their gene expression, which is a fundamental research basis in cell biology. Recent technology has made it possible to analyze gene expression at a single-cell level. Using the data obtained from this technology, several trajectory analysis methods have been proposed to infer cell lineages that represent the process of cell differentiation. However, conventional methods have a problem that the accuracy of the estimated cell lineages may become low depending on the results of the dimensionality reduction since only limited information in the expression data is used. For example, the cell lineages are estimated as graph structures, such as minimum spanning trees, in the two-dimensional space. To cope with this problem, we propose a new method to estimate cell lineages using not only the information projected to 2D space but also the information of cell types based on a cell-type estimation method. The effectiveness of the proposed method will be demonstrated by the experiments using single-cell RNA-seq data of mouse stem cells compared with the conventional methods.

## *Parameter exploration improves the accuracy of long-read genome assembly*

Anurag Priyam (Queen Mary University of London), Alicja Witwicka (Queen Mary University of London), Anindita Brahma (Queen Mary University of London), Eckart Stolle (Zoological Research Museum Alexander Koenig) and Yannick Wurm (Queen Mary University of London).

Long-molecule sequencing is now routinely applied to generate high-quality reference genome assemblies. However, whole-genome sequence datasets differ in terms of repeat composition, heterozygosity, read lengths and error profiles. The assembly parameters that provide the best results could thus differ across datasets, and from the default settings of the assembly software. To determine the potential benefits of optimizing assembly parameters, we generated thirty-six assemblies by systematically varying three key parameters of the Canu genome assembler. To compare the assemblies, we devised novel metrics of assembly completeness and accuracy, and integrated them with the classical N50 and BUSCO metrics in a framework that weighs the metrics by their relative independence. We show that simple fine-tuning of assembly parameters can substantially improve the quality of long-read genome assemblies. In particular, modifying estimates of sequencing error rates improved some metrics more than two-fold. We present our metrics and our approach of combining assembly quality metrics as a flexible software — CompareGenomeQualities. Our software automates comparisons of assembly qualities for researchers wanting a straightforward mechanism for choosing among multiple assemblies.

## *Sabreur: fast, reliable, and handy barcode demultiplexing of fasta and fastq files*

Anicet Ebou (Bioinformatic team, Institut National Polytechnique Félix Houphouët-Boigny), Dominique Koua (Bioinformatic team, Institut National Polytechnique Félix Houphouët-Boigny) and Adolphe Zeze (Laboratoire de Biotechnologies végétales et microbiennes, Institut National Polytechnique Félix Houphouët-Boigny).

Biotechnology tooling advances in recent years have been observed by an increase in the throughput rate of sequencers. Indeed, next-generation sequencing tools are able to generate millions to billions of reads in a single run. To reach such a high rate in a cost-efficient manner, next-generation sequencers often take advantage of the barcoding of multiple samples or species. Therefore, obtained raw sequences from multiplexed sequencing need to be rapidly demultiplexed. Here we present sabreur, a fast, reliable, and handy barcode demultiplexing tool for fasta and fastq files. Sabreur easily manipulates different compression formats for better data volume management. Test conducted shows that sabreur is overall faster than existing demultiplexing tools both in single-end mode and in paired-end mode while allowing different format compression for output files. sabreur is implemented in Rust and available at https://github.com/Ebedthan/sabreur.

## *Investigating tumor genome instability with Ploidetect*

Luka Culibrk (Canada's Michael Smith Genome Sciences Centre), Erin Pleasance (Canada's Michael Smith Genome Sciences Centre), Karen Mungall (Canada's Michael Smith Genome Sciences Centre), Janessa Laskin (British Columbia Cancer Agency), Marco Marra (Canada's Michael Smith Genome Sciences Centre) and Steven Jones (Canada's Michael Smith Genome Sciences Centre).

Whole-Genome Sequencing (WGS) of tumors is being increasingly adopted to inform clinical decision-making for cancer treatment. Copy number variations (CNVs) frequently alter the quantity of specific regions of tumour genomes. We present Ploidetect, an R package which estimates tumor purity and ploidy, and calls CNVs from WGS data. Purity and ploidy estimation is conducted by fitting gaussian mixture models to read depth and allele frequency data. CNV segmentation uses a novel coarse-to-fine approach. Ploidetect was applied to a cohort of previously treated metastatic tumor WGS data (n = 735). Ploidetect demonstrated good concordance with manual estimation of tumour purity (r =0.807), and reduced oversegmentation of CNVs while maintaining greater sensitivity in identifying CNVs affecting known oncogenes and tumor suppressors when compared with other CNV software. Applying Ploidetect to examine genome instability in metastatic tumors revealed homozygous deletions recurrently targeting whole genes or specific exonic regions, and identification of genes associated with genomic stability related to DNA repair and cell cycle pathways. We demonstrate Ploidetect's effectiveness as a CNV caller, facilitating the use of WGS for analysis of complex, varied quality real-world clinical tumour samples, and showcase its utility in uncovering novel aspects of cancer CNV biology.

## *ABySS 2.5: Efficient repeat resolution with short reads*

Vladimir Nikolic (Canada's Michael Smith Genome Sciences Centre), Justin Chu (Canada's Michael Smith Genome Sciences Centre), Johnathan Wong (Canada's Michael Smith Genome Sciences Centre), Lauren Coombe (Canada's Michael Smith Genome Sciences Centre), Amirhossein Afshinfard (Canada's Michael Smith Genome Sciences Centre), Ka Ming Nip (Canada's Michael Smith Genome Sciences Centre), Rene Warren (Canada's Michael Smith Genome Sciences Centre) and Inanc Birol (Canada's Michael Smith Genome Sciences Centre).

It has been twelve years since the publication of the ABySS short-read de novo genome assembler, and four since its successor, ABySS 2. The first ABySS release in 2009 utilized a de Bruijn Graph distributed across multiple machines, making it the first assembler capable of assembling a human genome with short-read sequencing. ABySS 2, released in 2017, decreased the memory usage tenfold by employing a Bloom filter, enabling the assembler to run on a single machine. Here, we present new additions to the assembler with the ABySS 2.5 release. In addition to incremental improvements to the assembly algorithm, a major change is the inclusion of the RResolver algorithm for repeat resolution. RResolver follows the steps of ABySS 2 by using a Bloom filter to store short-read information which is used to evaluate graph path support. The improvements introduced since ABySS 2 increase a 2x150bp human assembly NGA50 contiguity by 10.6%. Additionally, the number of recovered complete BUSCO genes is increased from 76.2% to 77.8%. From the initial release to ABySS 2, and now ABySS 2.5, ABySS has come a long way in delivering high quality de novo genome assemblies with low resource usage.

## *Large-scale assessment of ChIP-seq quality metrics toward peak call-free quality control*

Hayato Anzawa (Tohoku University) and Kengo Kinoshita (Tohoku University).

Quality control (QC) is one of the inevitable subjects of ChIP-seq experiments. While FRiP (fraction of reads in peaks) is commonly used as a QC metrics to indicate ChIP-seq noise level, FRiP calculation requires peak calling analysis, and it depends on the selection of peak calling method. Recently, we introduced VSN (Virtual Signal-to-Noise ratio), a QC metric to inspect ChIP-seq signal-to-noise ratio without peak calling. Despite its advantages and potential, verification of VSN's performance or its statistical characterization in a large dataset has not been performed yet. In this study, we applied the VSN approach to a large-scale ENCODE human ChIP-seq dataset to clarify its performance on the variety of ChIP-seq experiments; different cell lines, ChIP targets, sample treatments, etc. Here, we report that the VSN approach can be widely applied for human ChIP-seq data and the distributions of logarithmic VSN values can be modelled as a normal distribution. This result implies the VSN would have a statistically advantageous property for establishing a threshold. Additionally, we introduce a web application based platform that enables researchers to obtain VSN analysis results of public data and to utilize the VSN approach for their ChIP-seq data.

## *Changes in the alternative splicing landscape of breast cancer cells caused by the treatment with the therapeutic monoclonal antibody trastuzumab*

Karsten Rinas (University of Waterloo), Brendan J McConkey (University of Waterloo) and Karsten Rinas (University of Waterloo).

Changes in alternative splicing patterns for different breast cancer types have been described multiple times. However, the effect of breast cancer therapeutics on the splicing landscape have not been studied to the same degree. In particular, research on the well-studied therapeutic monoclonal antibody trastuzumab for the treatment of HER2+ breast cancer shows a blatant gap in the study of the induced changes in alternative splicing. This is surprising considering that the complex expressional changes caused by trastuzumab are still not completely resolved. So far, aberrations in splicing associated with trastuzumab treatment have been primarily focused on the splicing of the HER2 gene. Here, we provide an untargeted alternative splicing analysis by using short read RNA-seq data of the treated breast cancer cell line. We detected complex changes in the expression of RNA binding proteins (RBPs), directed changes in splicing patterns within pathways, and change in known key genes such as HRAS or GRB7. Since the end of the patent for trastuzumab, biosimilars are entering the market. More research in alternative splicing events caused by trastuzumab treatment may provide an additional basis for determining biosimilarity.