1584S

Detection of Copy Number Variations in Cancer Genomes from High Throughput Sequencing Data. *G. Klambauer, S. Hochreiter.* Insitute of Bioinformatics, Johannes Kepler University Linz, Linz, Upper Austria, Austria.

"Copy Number estimation by a Mixture Of PoissonS" (cn.MOPS), is a well established and widely used method for detection of germline copy number variations (CNVs) in high-throughput sequencing data. cn.MOPS showed excellent performance at the detection of CNVs in HapMap samples, as well as in genomes of bacteria, fungi and plants. Since cn.MOPS constructs a model across samples for each genomic position, it is not affected by read count variations along chromosomes, and, therefore, geared to targeted sequencing. In a comparative study, cn.MOPS was the best performing method at the detection of CNVs in targeted sequencing data. However, the detection of somatic CNVs in cancer genomes is still challenging due to admixture of normal and tumor tissue, nondiploidy and very large copy number variations that affect normalization. Therefore, preprocessing, normalization, and the core algorithm of cn.MOPS have been optimized for CNV detection in cancer genomes. We demonstrate the improved performance of the enhanced cn.MOPS algorithm for cancer genomes on whole genome sequencing data from the International Cancer Genome Consortium (ICGC). cn.MOPS has been optimized for computation time and parallelized, which makes the method perfectly suited to analyze data sets of hundreds of cancer samples within a few hours.

1585M

Efficient variant pipeline for diagnosis of inherited cardiomyopathies associated genes using Ion Torrent PGM™ platform. *L. Cerdeira[1], T.G M. Oliveira[2], A. Pereira[2], M. Mitne-Neto[1].* 1) Research and Development, Fleury Group, São Paulo, SP, Brazil; 2) University of São Paulo - Heart Institute, São Paulo, SP, Brazil.

Hypertrophic cardiomyopathy (HC) is a primary cardiac disease characterized by hypertrophy of the left ventricle (LV) without dilation , usually asymmetrical and predominantly septal , in the absence of any other cardiac or systemic disease that can cause myocardial hypertrophy . Typically, CH is caused by mutations in genes encoding sarcomeric elements. Currently 19 genes have been discovered and linked to the CH spectrum, besides the filaments of the sarcomere, additional subgroups can be classified as related CH, as Z disc genes and calcium transport. Diagnosis is mainly clinical and usually only identified after the symptoms beginning. For that reason molecular genetic tests came up as a differential tool for the discovery of the mutations causing the phenotype. This study developed a bioinformatics pipeline for accurate molecular diagnosis of CH using Ion PGM data. The pipeline was developed using CLCBio Genomic Workbench 6.5 workflow and had as a first step a mapping assessment, with the 5 nucleotides at the 3'end trimmed and a Phred ≥ 20 used for quality control. Alignment against the human genome HG19 version was done using standard thresholds, followed by identification of variants by coverage and quality positioning. The identification of known variants was validated against the databases: dbSNP and clinicalvar and for further evaluation a prediction of splice site effect and amino acid change. The result were submitted to SIFT and Polyphen programs to obtain the values for protein damaging. To validate the pipeline we selected 15 DNA samples from previously analyzed patients, which had clinical and molecular diagnoses of HC from the Heart Institute (InCor - University of São Paulo, Brazil). The previous molecular diagnosis was performed by Sanger sequencing for the three most HC-associated genes: MYH7, MYBPC3 and TNNT2. All variants found were properly annotated for the three genes and were further used in the evaluation of NGS accuracy. The NGS pipeline presented here could identify > 97% of the Sanger sequencing identified mutations, showing its robustness and viability for HC and for other diseases with Mendelian heritability standard.

1586T

Assessing novel centromeric repeat sequence variation within individuals by long read sequencing. *K.H. Miga[1], J. Chin[2], A. Bashir[3,4].* 1) Center for Biomolecular Science and Engineering, University of California, Santa Cruz, CA 95064, USA; 2) Pacific Biosciences, Inc, 1380 Willow Rd, Menlo Park, CA 94025, USA; 3) Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, 1425 Madison Avenue, New York, NY 10029, USA; 4) Icahn Institute for Genomics and Multiscale Biology, Icahn School of Medicine at Mount Sinai, 1425 Madison Avenue, New York, NY 10029, USA.

Centromeres and other heterochromatic regions are commonly enriched with long arrays of near-identical tandem repeats, known as satellite DNAs, that offer a limited number of variant sites to differentiate individual repeat copies across millions of bases. This substantial sequence homogeneity challenges available assembly strategies, and as a result, centromeric regions are vastly underrepresented in genomic studies. Further, as these sites are known to be variable among individuals in the population, it is necessary to not only characterize the sequence organization of these regions in a single genome, but to develop high-throughput methods to study this new source of human sequence variation among individual genomes. To advance characterization in these regions we have designed alphaCENTAURI (centromeric automated repeat identification for alpha satellite DNA) that takes advantage of Pacific Biosciences' long reads from whole-genome sequencing. Long reads allow direct determination of satellite higher-order repeat structure as opposed to using indirect inference methods, like assembly, with reads shorter than the underlying lengths of the high order repeat unit. Here we demonstrate a comprehensive assessment of higher-order repeat patterns for two human cell lines: NA12878 (diploid) and the hydatidiform mole (CHM1, haploid-like) genomes. First, we show the reliability of the method by validating consistency with existing centromere repeat references. Additionally, we are able to identify changes in repeat unit directionality that exist within arrays and between individuals, representing polymorphisms in the population or errors within existent assemblies. The analysis also represents a robust and straightforward methodology for characterization of higher-order repeat variants within the array that differ between individuals. Based on this analysis, resolution of higher-order repeats could be readily performed at low depth and reasonable cost across a population, or in genomes without high-quality references. This study demonstrates the methods to generate a sequence survey for regions enriched in satellite DNA that are typically omitted from genomic studies. We believe it establishes a foundation to extend and improve genomic characterization of any higher-order repeat structure using long reads.

1587S

Anchored Assembly: An algorithm for large structural variant detection using NGS data. *J. Bruestle, B. Drees.* Spiral Genetics, Seattle, WA.

Statement of purpose Characterizing large indels, inversions, and multi-nucleotide variants is important for understanding cancer, bacterial pathogens, and neurological disorders. Standard pipelines often miss these variants. Spiral Genetics has developed Anchored Assembly, a novel method using direct, *de novo* read overlap assembly to accurately detect variants from next-generation sequence reads. Anchored Assembly's range of detection and low false discovery rates may be useful for characterizing structural differences between tumor and normal samples.

Methods used Anchored Assembly was evaluated against Pindel and BWA + GATK using simulated read data. Datasets were generated by populating chromosome 22 of the human genome reference sequence with a set of SNPs, insertions, deletions, inversions, and tandem repeats.

Summary of results On human chromosome 22 data, Anchored Assembly detected over 90% of indels and structural variants up to 50 kbp and SNPs with false discovery rates well below 1%. In comparison, Pindel and BWA + GATK had overall false discovery rates of 10% and 9%, respectively. We detect, on average, over 90% of indels and structural variants up to 30 kbp in non-repetitive regions. The ability to detect deletions and structural variants is undiminished by variant size, and the ability to accurately detect and assemble insertions continues well into the 30 kbp range.