De Novo Assembly and Structural Variation Discovery in in Human Disease and Non-Disease State Genomes Using Extremely Long Single-Molecule Imaging

A. Hastie, E. Lam, M. Imielinski1, C.-Z. Zhang1, J. Wala1, A. Pang, S. Chan, W. Andrews, H. Dai, Ž. Džakula, H. Cao
BioNano Genomics, San Diego, CA
1 Broad Institute, Boston, MA

Abstract
Structural variation analysis (SVA) of human genomes is usually a reference based process and therefore biased and incomplete. In order to have a comprehensive analysis of structural variation, a de novo approach is needed. De novo genome assemblies using only short read data are generally incomplete and highly fragmented due to the intractable complexity found in the human genome. This complexity, consisting mainly of large duplications and repetitive regions, hinders sequence assembly and subsequent comparative analyses. As a result of the remaining limitations of DNA sequencing and analysis technologies, it is not feasible to create high quality assemblies of individuals to detect and interpret the many types of structural variation that are refractory to high throughput or short-read technologies.

We present a single molecule genome analysis system (Irys®) based on NanoChannel Array technology that linearizes extremely long DNA molecules for direct observation. This high-throughput platform automates the imaging of single molecules of genomic DNA hundreds of kilobases in size to measure sufficient sequence uniqueness for unambiguous assembly of complex genomes. High resolution genome maps assembled de novo preserve long-range structural information necessary for structural variation detection and assembly applications. Dozens of human genomes have been de novo assembled by Irys to date, including cancer genomes. Structural variation analysis reveals insertions, deletions, inversions and translocations. Each genome shows dramatic structural variation even when considering only normal (non-disease state) individuals, including many megabases of variation within genomic regions not included in the public reference genome assembly (GRCh38), underscoring the need for more de novo approaches to genome analysis. In some cases, genome maps can identify translocations partners whose path pass through hundreds of kilobases of the genome which is absent from the reference. Extremely long single molecules can also be used to phase rearrangement breakpoints on the same derivative chromosome; something that can only be inferred by short read and copy number technologies.

Background
Generating high-quality finished genomes replete with accurate identification of structural variation and high completion (minimal gaps) remains challenging using short read sequencing technologies alone. The Irys platform provides direct visualization of long DNA molecules in their native state, bypassing the statistical inference needed to align paired-end reads with an uncertain insert size distribution. These long labeled molecules are de novo assembled into physical maps spanning the whole genome. The resulting order and orientation of sequence elements in the map can be used for anchoring NGS contigs and structural variation detection.

Methods
(1) Extremely long DNA is extracted from the source sample and (2) labeled with IrysPrep™ reagents by incorporation of fluorophore-labeled nucleotides at a specific sequence motif throughout the genome. (3) The labeled genomic DNA is then linearized in the IrysChip™ nanochannels and single molecules are imaged by Irys. (4) Irys performs automated data collection and image processing. Molecules are labeled with a unique signature pattern that is uniquely identifiable. (5) Molecules are assembled into genome maps and downstream analysis of maps is performed with the IrysView™ software suite.

Conclusions
BioNano Genomics genome mapping is a powerful tool for detection of structural variation in human individuals. Many SVs are common in our dataset suggesting that the hg19 reference contains mistakes or uncommon variants. Genome mapping and single molecule reads are useful for identifying translocation partners where repeat sequences break sequence assemblies and even through unreferenced DNA. Single molecule reads are able to phase multiple fragments of a rearranged derivative chromosome in a breast cancer sample.

References