Structural Variation Discovery by De Novo Genome Mapping of the Human Genome at the Single Molecule Level Using NanoChannel Linearization

Abstract
As a result of the remaining limitations of DNA sequencing and analysis technologies—even ten years after the completion of the human genome project—there remain about 400 gaps in the human reference sequence assembly, hundreds of millions of unassembled bases in those regions, and no effective tools to comprehensively characterize the structural variation in an individual’s genome. Despite the unappreciated reference sequence being of extremely high quality, it is not feasible to create similarly 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. We have used Irys genome mapping for the assembly and characterization of two human genomes. From these assemblies, we have detected many of the remaining gaps, identified known and novel structural variants and phase some haplotype blocks, including in the MHC region. We also resolve and measure long tandem repeat regions that are likely impossible to assemble by other methods.

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) IrysPrep™ reagents by incorporation of fluorophore-labeled nucleotides at a specific sequence motif throughout the genome. (2) The labeled genomic DNA is then linearized in the IrysChip™ nanochannels and single molecules are imaged by Irys. (3) Irys performs automated data collection and image processing. (4) 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 Irys enables visualization of extremely long, single DNA molecules for the direct characterization of complex structural events in the genome. This system permits rapid accurate genome-wide de novo assembly and detection of structural variants that typically confound short-read genome assembly and comparative genomic analysis. Here we demonstrate de novo human Genome Map assembly capabilities of the IrysChip nanochannel array and the Irys imaging system to characterize genome-wide structural variation in the human genome. By comparing de novo assemblies of a father-daughter pair we show that genome mapping is able to detect large structural variants with very good cross-validation. We are also able to map regions of the genome that are refractory to assembly by other methods (remaining gaps in the human reference genome).

References