**Abstract**

De novo genome assemblies using only short read data are generally incomplete and highly fragmented due to the intractable complexity found in most genomes. 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 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.

In this work, we describe the development and application of 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 several genomes, including human, plant, fungi, and bacteria.

In addition to describing the technology and analysis approaches useful for dissecting complex genomes, we demonstrate results from several genomes, where genome maps span remaining reference gaps, identify known and novel structural variants (including balanced rearrangements) and phase variation within haplotype blocks. 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 coverage (minimal gap) 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. (5) Molecules are labeled with a unique signature pattern that is uniquely identifiable. (6) Molecules are assembled into genome maps and downstream analysis of maps is performed with the IrysView™ software suite.

**Results**

- **Structural Variation Detection in Map**
  - Following de novo assembly of genome maps, structural variation is identified using a broad range of visual features, refractory to many high throughput and short-read assembly methods. Deletions are evident by the presence of novel label sites and expansion of adjacent labels. Exons are excluded by the absence of label sites or naming of inter-exon segments. Maps also identify variation in difficult-to-sequence highly repetitive regions, such as those involved in immune function (such as MHC) and near the telomeres and centromeres. A hypervariable locus known to be related to neurodevelopmental conditions, shows large deletions and inversions in two haplotypes.

- **Haplotype Assembly in MHC**
  - Individual haplotypes can be assembled in complex and variable regions such as the Human MHC due to Irys’ long single-molecule detection. De novo assembly of individual haplotypes in the resulting variation detected from different individuals. Further development work is ongoing to provide even longer phased/boxed de novo assembled maps.

- **Future Alternative Labeling**
  - Other labeling approaches are under development that can contribute additional information to Genome Maps. For example, dual-labeling of single molecule DNA in complex repetitive regions may contribute additional information to Genome Maps.

**References**

3. Das, S., et al. Single molecule linear analysis of DNA in nano-channel labeled with sequence specific fluorescent probes, Nucleic Acids Research (2010); 38: 8