Mapping the “Dark Matter” of Genome – Long Repeats and Complex SVs Leading To True Contiguity of De Novo Assembly with NanoChannel Technology

Abstract

In spite of advancement in high-throughput NGS in the past decade, a large portion of the human genome remains unresolved or ambiguously characterized. Especially, large genomic structural variations (SVs > 1 kb), known to be associated with complex traits diseases, are found more prevalent than we previously thought. During assembly, they leave gaps and unknown structural information as the “dark matter” of the genome, challenging for short read NGS and conventional low resolution cytogenetic techniques. Rapid comprehensive genome mapping in nanochannel arrays represents a new single-molecule platform independent of yet complementary to DNA sequencing for accurate genome assembly and structural variation analysis. Extremely long molecules of hundreds of kilobases fluorescently labeled at sequence motifs and elongated in nanofluidic channels enable direct image interrogation of comprehensive genome architecture at a high resolution. De novo assembly of these single molecules yields unprecedented long contiguous genome maps, advantageous in spanning over highly repetitive regions and complex structures in their native form.

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.

Conclusions

Irys enables visualization of extremely long, single DNA molecules for the direct characterization of complex structural events in the genome. Genome mapping in NanoChannel arrays is shown to be a rapid, accurate, powerful and robust method for detection of structural variation and the study of complex regions in the human genome.

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.

Hybrid Scaffolding Flow Chart

1. Align NGS derived map with BioNano derived maps
2. Identify potential chimerical contigs and flag them. (For now, we don’t use flagged contigs)
3. Run Marge Pipeline to generate “super-scaffolds”.
4. Identify PacBio – SuperContig Alignments and generate ASG output etc.

Copy Number Profiles in Cancer Samples

The motif pattern on long single molecules from BioNano’s Irys System were aligned to digitally nick labeled reference genome 19 to build a histogram map illustrating the depth profiles. Graph A depicts a “normal” healthy male Caucasian genome profile while graph B shows the cancer genome full of amplified regions (peaks), deletions and break points, polyploidy and aneuploidy. The green line is the normalized diploid level.

IrysSolve Computational Solutions

IrysSolve data analysis pipeline performs automated de novo assembly and structural variation detection in an integrated pipeline.

Options include:
- IrysSolve for Intel® Xeon Phi™, which features a specially optimized version of IrysView’s proven de novo assembly and structural variation detection pipeline, implemented on an open-source Linux platform, featuring full compatibility with IrysView.
- IrysSolve Cloud
- IrysSolve for your Existing Hardware, which allows you to integrate the IrysSolve data analysis pipeline on most pre-existing Linux compute hardware.

References

3) Das, S. K., et al. Single molecule linear analysis of DNA in nano-channel labeled with sequence specific fluorescent probes. Nucleic Acids Research (2010); 38:
5) http://dgv.tcag.ca/dgv/app/home