Abstract

Irys genome mapping technology provides direct analysis of extremely long genomic DNA (up to multi-megabases) without amplification. De novo assembly of these single molecules yields high-fidelity contiguous map information across long ranges. Its advantage over all other genome assembly methods is particularly dramatic in highly repetitive regions. Genome maps thus greatly complement assemblies using relatively short second- and third-generation sequencing reads. We have constructed genome maps of human NA12878 (cell line derived from the daughter in the CEU trio) which resulted in a consensus assembly measuring 2.9 Gb and with an N50 of 4.6 Mb. With Pacific Biosciences sequence from the same cell line, we also created a sequence-based assembly with N50 length of 930 kb in parallel. By combining data from these two technologies with a custom-designed hybrid scaffolding pipeline, we were able to generate an assembly having scaffold N50 length of greater than 30 Mb covering more than 2.7 Gb of the human genome. At the same time, we were able to identify potential misassemblies in the sequence assembly as well as in the genome maps by reviewing the inconsistencies between these two complimentary technologies. The hybrid scaffolds also discovered additional long range structural variations not identified in the sequence assembly.

Background

Generating high quality finished genomes repeals with accurate identification of structural variation and high completion (minimal gaps) remains challenging using short read sequencing technologies alone. Instead, Irys technology provides direct visualization of long DNA molecules in their native state, avoiding the statistical assumptions that are normally used to force sequence alignments of low uniqueness elements. The resulting order and orientation of sequence elements are demonstrated in anchoring NGS contigs and structural variation detection.

Methods

(1) Long molecules of DNA is labeled by IrysPrep™ reagents by (2) incorporation of fluorescent 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) Single-molecule data collected and detected automatically. (5) Molecules are labeled with a unique signature pattern that is uniquely identifiable and useful in assembly into genome maps. (6) Maps may be used in a variety of downstream analysis using IrysView™ software.

Hybrid Scaffolding Flow Chart

1. Flag Inconsistencies (OC and/or manual curate) 
2. Scaffolding Pipeline 
3. Export

NGS & Genome Map Hybrid Assembly

Step 1: Merge between NGS/BioNano contigs to generate hybrid contigs, keep merge cascade between hybrid contigs and NGS/BioNano contigs but not inside groups (NGS vs. NGS, BioNano vs BioNano and hybrid vs hybrid).

Step 2: Merge between hybrids to generate super scaffolds

Step 3: Extend hybrid with BioNano extend/merge pipeline

Step 4: Align NGS contigs to super scaffolds

Conclusions

Our data of this human sample NA12878 illustrates that pairing technologies can enable superior results than can be achieved with any single technology. The extremely long BioNano genome maps facilitate the correct ordering and orientation of Pacific Biosciences contigs. In fact, we achieved a hybrid assembly that has a larger N50 value than any short-gun sequencing project of the human genome to date. Furthermore, our pipeline can generate outputs in FASTA format, and so users can zoom in to base-pair resolution. Therefore, our hybrid data can provide insights from large structural level down to nucleotide level. We anticipate that future fully-integrated co-assembly of BioNano’s long molecules with sequence reads can further improve the contiguity and accuracy of the resulting assembly, thus enabling an even deeper understanding of the biology of genomes.

References

6) IrysView software assembles genome maps.