Abstract

De novo genome assemblies using purely short sequence reads are generally fragmented due to complexities such as repeats found in most genomes. These characteristics can hinder short-read assemblies and alignments, and that can limit our ability to study genomes.

The BioNano Genomics Irys System linearizes long DNA molecules, thus yielding single-molecules containing long-range information. These long-range hybrid scaffolds are able to identify novel chromosomal rearrangements undetectable by short-read alignment or reference-guided assembly approaches. We present a comprehensive analysis of a human genome by combining single molecule genome mapping with one of the most annotated sequence assemblies, the HuRef assembly. Overall, we found that the assemblies of sequencing and genome mapping technologies correspond well, and the resulting hybrid scaffolds are highly contiguous, with a N50 exceeding 35Mb, a value typically unachievable by short-read sequencing. In addition, we compared the structural variation with calls previously detected in the HuRef assembly, and found multiple novel variants spanning over hundreds of kilobases in size. Some of these variants reside in areas where the sequence assembly was poorly covered or was highly fragmented; yet these variants encompass numerous genes, and can be of functional importance. Finally, we identified genome maps that span over the remaining reference gaps, and maps that resolve and measure long tandem repeats.

Methods

1. IrysPrep kit extraction of long DNA molecules
2. IrysPrep reagents label DNA at specific sequence motifs
3. IrysChip linearizes DNA in NanoChannels
4. Irys automates imaging of single molecules in NanoChannels
5. Molecules and labels detected in images by instrument software
6. IrysView software assembles genome maps

(1) Long molecules of DNA are labeled with IrysPrep reagents by (2) 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) Single molecule data are 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.

Assembly Pipeline Overview

1. Generate in silico maps for sequence data
2. Align sequence and BioNano maps to flag potential chemical contigs
3. Run Merge to generate ‘Hybrid Scaffold’
4. Generate sequence – hybrid scaffold alignments (used generate ESPorted id)

HuRef Genome Map Assembly

Initial HuRef assembly was constructed using high-quality Sanger sequencing. Combining the sequence assembly with BioNano genome maps, we constructed hybrid scaffolds with unprecedented contiguity in length, with a N50 of 36.4 Mb (Left). In fact, the lengths of some larger scaffolds nearly reach entire chromosome arms, spanning over numerous assembly gaps (Right).

Hybrid Scaffold Pipeline Overview

1. Generate in silico maps for sequence data
2. Align sequence and BioNano maps to flag potential chemical contigs
3. Run Merge to generate ‘Hybrid Scaffold’
4. Generate sequence – hybrid scaffold alignments (used generate ESPorted id)

Example of a BioNano-Sequence Merge Event

Near chromosome-arm length hybrid scaffolds

Conclusions

We present here the results from the analysis of a diploid human genome. Using our de novo assembly pipeline, we constructed a highly contiguous genome map assembly, covering a large portion of the human reference and extending into previously unalignable, unannotated loci. We also detected large structural variants that are recallcitant to short-read, reference-based detection approaches, and we were able to elucidate the location, orientation and copy number of these events.

Moreover, our data shows that the correct pairing of technologies enables far richer results than can be achieved using each technology alone. With our hybrid scaffold pipeline, we integrated BioNano and sequence assembly, and achieved scaffolds with unprecedented lengths. Such scaffolding information can enable the correct anchoring of sequence assemblies, and the correct phasing of variation along chromosomes.

Reference