Abstract

In spite of advancements made in high-throughput next-generation sequencing 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 many human diseases, are found to be inconsistent across different individuals, which were previously thought. Often challenging to detect for short-read NGS and conventional low-resolution cytogenetic techniques, these large SVs—the "dark matter" of the genome—result in gaps and unknown structural information in assemblies. Rapid comprehensive genome mapping in nanochannel arrays provides a single-nucleotide imaging platform independent of, yet complementary to, DNA sequencing for accurate genome assembly and structural variation analysis. Extremely long molecules of hundreds to thousands 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. Here, we present results from analyses on normal human genomes, cancer genomes, and other complex genomes. We detected hundreds of large structural variants and haplotype differences in these genomes. In one human genome, we obtain hundreds of insertions/deletions and inversions larger than 1 kb. Without considering SVs that overlap with N-base gaps in hg19, 90% of these SVs are supported by orthogonal experimental methods or historical evidence in public databases. A larger portion of the complex genome is composed of previously unknown repeat material spanning tens of kilobases to multiple megabases, whose exact loci and copy numbers remain elusive to NGS. Without knowing the genomic context of these repeats or the amount of repeats, it is difficult to attach any biological significance to them. Using BioNano’s Irys® platform and novel algorithms designed specifically to investigate long repeat arrays, we were able to find and characterize clinically relevant repeat regions linked to human cardiovascular disease risk. For the first time, population-scale cross-sample genome comparison to identify comprehensive genomic structural variation is feasible on a single platform due to the quick turnaround time. Lastly, we present the first assembly of a diploid human genome that combines long-read NGS sequencing with Irys genome maps, resulting in a hybrid assembly with a map N50 exceeding 31 Mb, a dramatic improvement upon the assembly contiguity typically observed in shotgun sequencing approaches. Overall, genome mapping provides valuable structural information otherwise hard or impossible to decipher with short-read sequencing data alone, and paves the road for generating true contiguity in ultimate medical-grade and breeding-grade genome information.

Background

Generating high-quality finished genomes replete with accurate identification of structural variation and highly complete (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 detecting structural variation.

Methods

<table>
<thead>
<tr>
<th>Methods</th>
<th>IrysPrep kit extraction of long DNA molecules</th>
<th>IrysPrep reagents label DNA at specific sequence motifs</th>
<th>IrysChip linearizes DNA in NanoChannels</th>
<th>Irys automates imaging of single molecules in NanoChannels</th>
<th>Molecules and labels detected in images by instrument software</th>
<th>IrysView software assembles genome maps</th>
</tr>
</thead>
</table>

(1) Long molecules of DNA is 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.

200+ Unique Genomes Analyzed on Irys

Comparison of Repeats in Plant & Animal Genomes

Comparison of the frequency of labeled repeat motifs found in several datasets generated from Irys data and software. Only labeled repeats with 5 or more copies were considered. Notice the similarity between barley and wheat, both of which are cereal crops. The prominent peak at 10 kb for rockcress (Arabidopsis) was shown to correspond to rDNA. The 2.4 kb repeat in a human male is not as prominent in females and is believed to be primarily on the Y chromosome.

Toward True Contiguity – Examples of Chromosomal-Level Hybrid Scaffold

Conclusions

Here we have presented several examples in order to demonstrate the significant value added to genomic research which is only possible by directly interrogating extremely long reads. Long arrays of tandem repeats containing individual units, each of which are longer than nearly any NGS sequence read and full array lengths far beyond the longest sequence reads available today can be seen in single molecule Irys reads. Thousands of large SVs are identified in Irys data, some of which are invisible or unassembled by other methods. As demonstrated with the analysis on the Lp(a) gene, the Irys platform has the potential to provide quick, cost-effective diagnosis of genetic diseases linked to copy number and structural variation. These data underline the importance of complete de novo assembly by long read technology for personalized genomics.

Reference

4) 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: 8