De Novo Assembly and Structural Variation Discovery in Complex Genomes Using Extremely Long Single-Molecule Imaging

H VanSteenhouse[1], A Hastie[1], E Lam[1], H Dai[1], A Pang[1], T Chan[1], W Andrews[1], M Saghbini[1], C Lee[2], C Luo[3], J Ecker[3], P-Y Kwok[4], H Cao[1]
[1]BioNano Genomics, San Diego, California, USA
[2]The Jackson Laboratory for Genomic Medicine, Connecticut, [3]Salk Institute, San Diego, CA; [4]UCSF, San Francisco, CA

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 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.

We present 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. We demonstrate results from several of these 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.

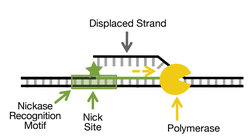
Background

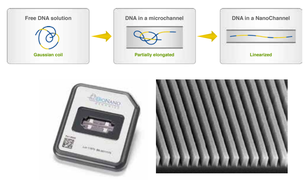
Generating high quality finished genomes replete with accurate identification of structural variation and high completion (minimal gaps) remains challenging using short read sequencing technologies alone. Instead, Irys technology provides direct visualization of long DNA molecules in their native state, 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.

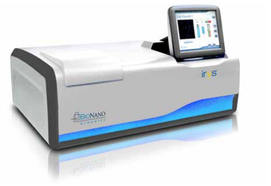
Methods

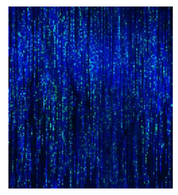
(1) DNA is labeled with IrysPrep™ reagents by incorporation of fluorophore labeled nucleotides at a specific sequence motif throughout the genome. (2) The labeled genomic DNA is then linearized in the IrysChip™ nanochannels and single molecules are imaged by Irys. (3) Single molecule data are collected and detected automatically. (4) Molecules are labeled with a unique signature pattern that is uniquely identifiable and useful in assembly into genome maps. (5) Maps may be used in a variety of downstream analysis using the IrysView™ software suite.

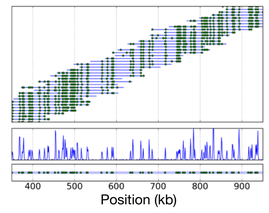
1) IrysPrep reagents label DNA at specific sequence motifs → 2) IrysChip linearizes DNA in NanoChannels → 3) Irys automates imaging of single molecules in NanoChannels → 4) Molecules and labels detected in images by instrument software → 5) IrysView software assembles genome maps





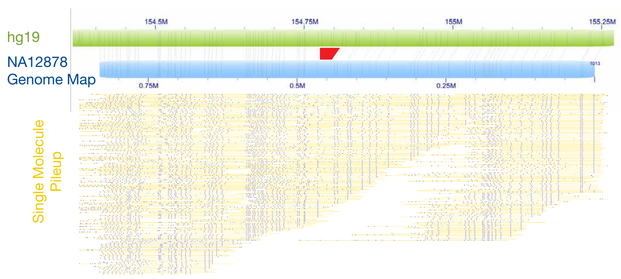




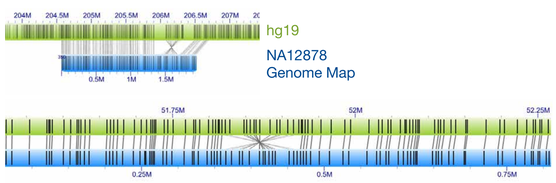


Structural Variation Discovery in Genome Maps

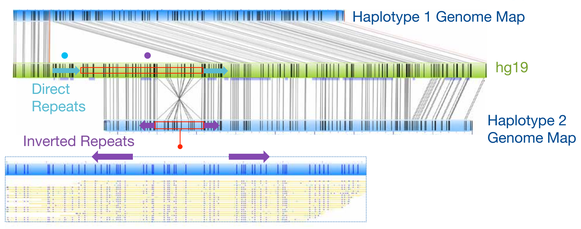
Deletion (ChrX: 154,773,676–154,814,880)



Inversions



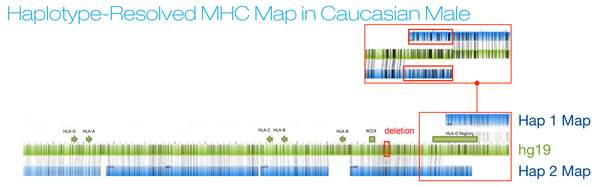
Inversion and Deletion at One Locus on Different Haplotypes



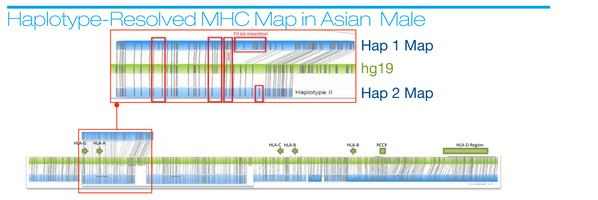
Genome-wide CNVs and structural variants are detected algorithmically as statistically significant deviations in label presence/absence and inter-label distances between a genome map (from NA12878 in these figures) and a reference genome (hg19 in this case). Inversions are a special case of label pattern deviation, where spans of several contiguous labels exhibit a reversed pattern in the genome map versus the reference. Long single molecule detection using Irys enables haplotypic analysis of such variants as well. The pattern of labels also suggest the source of variants: direct repeats (teal arrow) in the case of deletion, and inverted repeats (purple arrow) in the case of inversion.

Haplotype Assembly in MHC

Haplotype-Resolved MHC Map in Caucasian Male



Haplotype-Resolved MHC Map in Asian Male

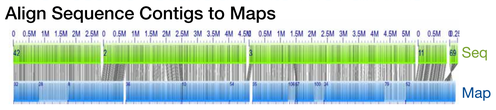


Again, by virtue of Irys long single molecule detection, individual haplotypes can be assembled in complex and variable regions such as the Human MHC. *De novo* assembly avoids reference bias in the resulting variation detected from different individuals. Further development work is ongoing to provide even longer phased blocks of *de novo* assembled maps.

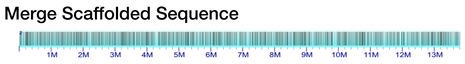
Scaffolding *De Novo* Sequence Assembly

Chromosome 1 Scaffolded into Two Full Arms

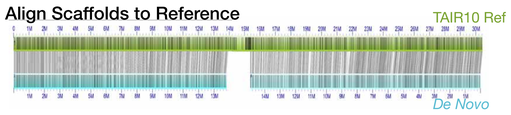
Align Sequence Contigs to Maps



Merge Scaffolded Sequence



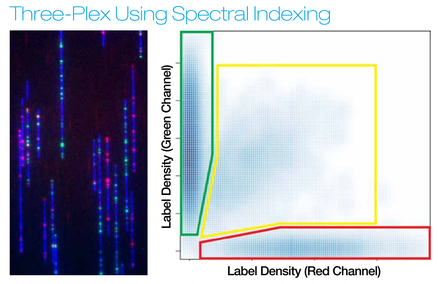
Align Scaffolds to Reference



De novo assembled genome maps provide long-range contiguity that spans repetitive elements that frequently break sequence assemblies. These maps can then be used to scaffold fragmented sequence assemblies. As an example, here, Arabidopsis sequence contigs (green) are scaffolded using genome maps (blue) to create two merged more contiguous scaffolds (teal) each of which constitute essentially the entire chromosome, broken only at the centromere. The whole-genome contig N50 was nearly doubled, from 6.2Mb to 10.8Mb.

Sample Multiplexing

Three-Plex Using Spectral Indexing



Other labeling approaches are under development to leverage the high throughput of Irys even for small genomes. For example, here three samples are labeled with three different spectrally resolved labels: Red, Green, and Red/Green Mixed. Each population is informatically separated by the colored zones in a scatter plot.

Live Demo in Suite 183

Human GM24631 is running right now, LIVE. Stop in the suite during a break, or make a private appointment: info@bionanogenomics.com

Streaming live at bionanogenomics.com/AGBT14

Conclusions

BioNano Genomics Irys enables visualization of single-molecules of extremely long DNA for the direct observation and measurement of genome complexities. This system permits accurate genome-wide assembly and detection of structural variants that typically confound short read genome assembly and comparative genomic analysis. Here we demonstrate the structural variation and genome assembly capabilities of the IrysChip nanochannel array and Irys imaging system to characterize genome-wide structural variation in a human genome, assemble separate haplotypes in the disease-relevant MHC region, and significantly improve the assembly of a plant genome to improve understanding of critical regions.

References

1) Lam, E.T., et al. Genome mapping on nanochannel arrays for structural variation analysis and sequence assembly. Nature Biotechnology (2012); 10: 2303
2) 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
3) Xiao, M et al. Rapid DNA mapping by fluorescent single molecule detection. Nucleic Acids Research (2007); 35:e16.
4) Hastie, A.R., et al. Rapid Genome Mapping in Nanochannel Arrays for Highly Complete and Accurate *De Novo* Sequence Assembly of the Complex Aegilops tauschii Genome. PLoS ONE (2013); 8(2): e55864.