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

T. Dickinson, A Hastie, E Lam, H Dai, A Pang, W Andrews, M Saghbini, X Yang, P-Y Kwok[1], M Rossi [3], H Cao
BioNano Genomics, San Diego, CA, USA
[1]UCSF, Cardiovascular Research Institute and Institute for Human Genetics, San Francisco, CA, USA
[3]Emory University, Atlanta, GA, USA

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 similarly 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 humans (cancer and phenotypically normal), plant, fungi, and bacteria.

In addition to describing the technology and analysis approaches useful for dissecting complex genomes, 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. Genome-wide structural variation detection includes specific positional information of CNVs, as well as identifying balanced variation such as translocations and inversions. We also resolve and measure long tandem repeat regions that are likely impossible to assemble by other methods.

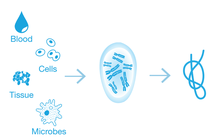
Background

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. 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.

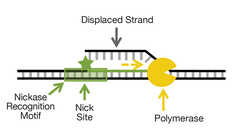
Methods

(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.

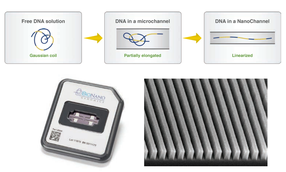
1) IrysPrep Kit extraction of long DNA molecules



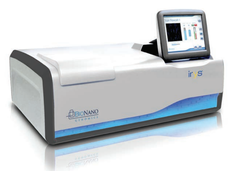
2) IrysPrep reagents label DNA at specific sequence motifs



3) IrysChip linearizes DNA in NanoChannels

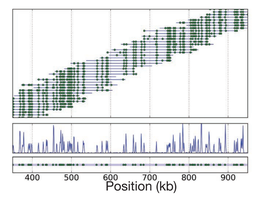


4) Irys automates imaging of single molecules in NanoChannels



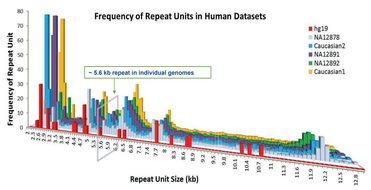
5) Molecules and labels detected in images by instrument software

6) IrysView software assembles genome maps



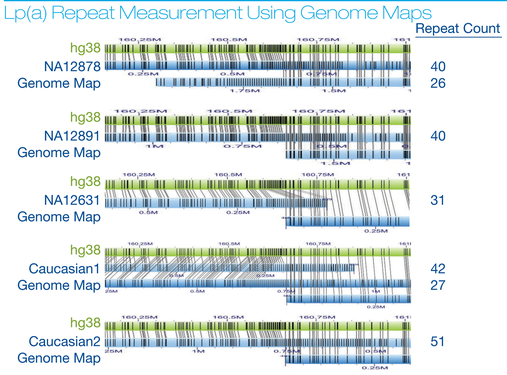
Lp(A) Repeat Copy Number Diversity in Humans

Repeat Distributions Vary Between Human Individuals



Irys single-molecule data retain repeat information, enabling genome-wide analysis of repeat content. Repetitive element contribution to the genome varies between individuals and is greatly underrepresented in the reference sequence.

Lp(a) Repeat Measurement Using Genome Maps

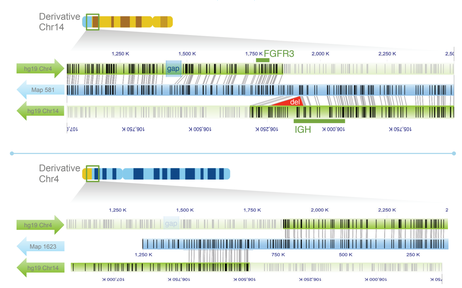


Despite being difficult to detect with traditional methods, repetitive content is known to have important biological functions. For example, Lp(a) protein levels are highly heritable and are chiefly determined by copy number variation at the LPA locus (10-50+ copies of kringle-IV type 2 (KIV2)-like domains) on chr 6q27. Elevated plasma lipoprotein(a) [Lp(a)] concentration has strong correlation to risk of cardiovascular disease (CVD), atherosclerosis, thrombosis, stroke and coronary heart disease (CHD).

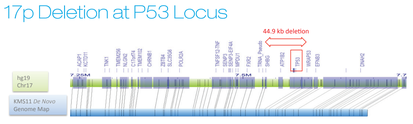
Genome maps assembled from long single-molecules that span the entire repeat region identify the haplotype-specific repeat copy number.

Multiple Myeloma Structural Variation

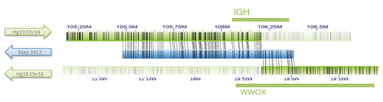
Balanced Translocation t(4;14)



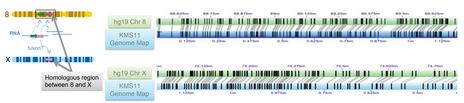
17p Deletion at P53 Locus



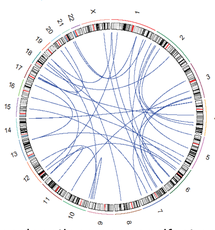
Fusion Gene Translocation t(14;16)(q32.3;q23)



Fusion Event Refuted



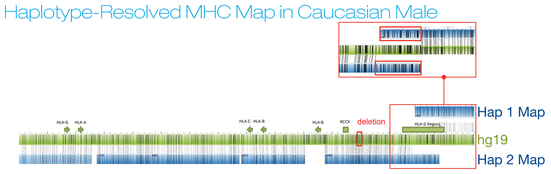
Widespread Translocations



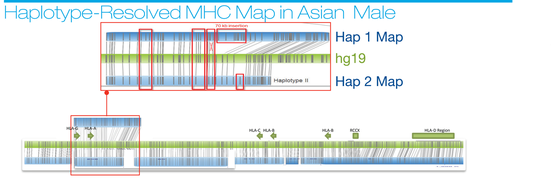
Translocations can manifest as aberrant expression of existing genes, nascent expression of resulting new fusion genes, and potential oncogenic activity, yet are difficult to detect with current methods. Irys long molecules retain positional information essential to discover translocations across the genome, as seen in the circos plot of the KMS11 multiple myeloma cell line. As examples, three archetypal variants are precisely detected: t(4;14)(p16.3;q32.3) reciprocal translocation results in FGFR3/IGH fusion and MMSET dysregulation; a deletion involving the P53 tumor suppressor locus; and another translocation involving the IGH locus t(14;16)(q32.3;q23). Errors made in short-read methods due to regions of homology can be overcome by genome map analysis.

Haplotype Assembly in MHC

Haplotype-Resolved MHC Map in Caucasian Male



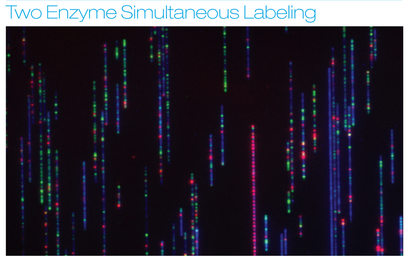
Haplotype-Resolved MHC Map in Asian Male



Individual haplotypes can be assembled in complex and variable regions such as the Human MHC due to Irys long single-molecule detection. *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.

Alternative Labeling

Two Enzyme Simultaneous Labeling



Other labeling approaches are under development that can contribute additional information to Genome Maps. For example, dual-labeling of human DNA shown here can be used for higher resolution analysis, and providing complementary labeling in complex repetitive regions.

Conclusions

BioNano Genomics Irys enables visualization of extremely long, single DNA molecules for the direct characterization of complex structural events in the genome. This system permits rapid accurate genome-wide *de novo* assembly and detection of structural variants that typically confound short-read genome assembly and comparative genomic analysis. Here we demonstrate *de novo* human Genome Map assembly capabilities of the IrysChip nanochannel arrays and the Irys imaging system to characterize copy number variation, complex rearrangements in cancer, and haplotype-phasing in the highly variable MHC region.

References

1) 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.
2) Lam, E.T., et al. Genome mapping on nanochannel arrays for structural variation analysis and sequence assembly. Nature Biotechnology (2012); 10: 2303
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: 8
4) Xiao, M., et. al. Rapid DNA mapping by fluorescent single molecule detection. Nucleic Acids Research (2007); 35:e16.