Rapid Structural Variation Detection and *De Novo* Assembly in Human and Complex Genomes Using Extremely Long Single-Molecule Imaging

A Hastie, E Lam, M Requa, H Dai, M Austin, F Trintchouk, M Saghbini, T Anantharaman, K Haden, S Rombauts[1], P-Y Kwok[2], K Robison[3], H VanSteenhouse, T Dickinson, X Yang, E Holmlin, H Cao
BioNano Genomics, San Diego, California, USA
[1]Ghent University, VIB, Gent, Belgium; [2]UCSF, San Francisco, CA, USA; [3]Warp Drive Bio, Carmbridge, MA, USA

Abstract

Despite continued cost reduction of raw base generation, improvement in base-calling accuracy, and recent advances in read length, complete *de novo* assembly and genome wide structural variant analysis of individual large complex genome remain expensive and challenging.

We present a rapid genome-wide analysis method based on the new NanoChannel Array technology (Irys) that dynamically streams and linearizes extremely long DNA molecules for direct image analysis of tens of gigabases per run. This high-throughput platform automates the imaging of genomic DNA hundreds to thousands of kilobases in length at the single-molecule level, for unambiguous assembly of complex genomes. High-resolution genome maps assembled *de novo* via unique sequence motif labeling, preserving long-range structural information that is intractable by current short read NGS platforms. This information is independent of current sequencing biochemistry and algorithms with built-in long range haplotyping and is critical to validate past and future genomic sequencing assembly data and discover new structural variants. It is simple and straightforward to set up and operate, amenable to whole genome structural variation comparative studies of large populations.

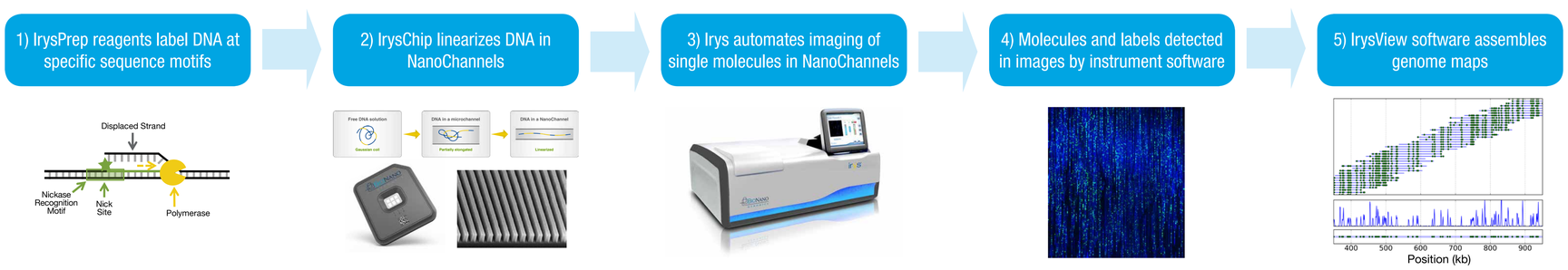
Here we report the first human *de novo* genome map assembly by the single molecule Irys system and analysis of complex regions of a variety of organisms using this approach. Unlike paired end sequencing approaches that are cumbersome and biased towards detecting more deletions than insertions, hundreds of large structural variants were uncovered with this direct view approach. We have corrected errors in previous assemblies, spanned many of the remaining gaps, identified known and novel structural variants and phased haplotype blocks—including in the highly variable regions involved with important immune system function. This technology and method will allow new discoveries and change our view toward understanding genome architecture and functions.

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.

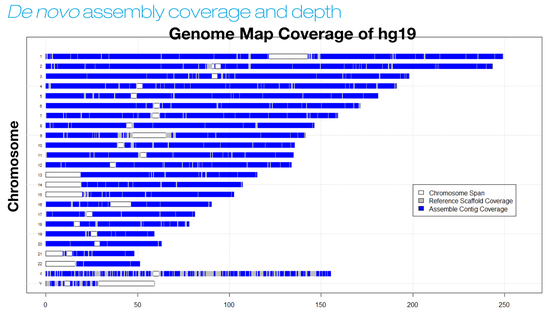
Methods

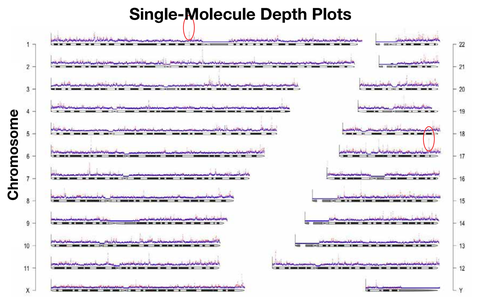
(1) DNA is labeled with IrysPrep™ reagents by incorporation of fluorophore-labeled nucleotides at a specific sequence motif throughout the genome. (2) The labeled genomic DNA is then linearized in the IrysChip™ nanochannels and single molecules are imaged by Irys. (3) Irys performs automated data collection and image processing. (4) Molecules are labeled with a unique signature pattern that is uniquely identifiable. (5) Molecules are assembled into genome maps and downstream analysis of maps is performed with the IrysView™ software suite.



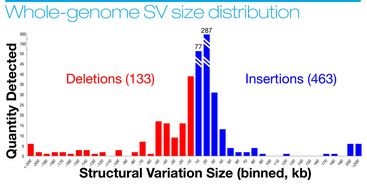
De Novo Human Genome Map Assembly and Genome-Wide Structural Variation Detection

De novo assembly coverage and depth

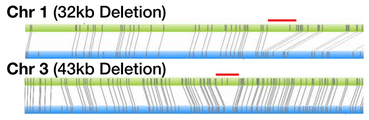




Whole-genome SV size distribution

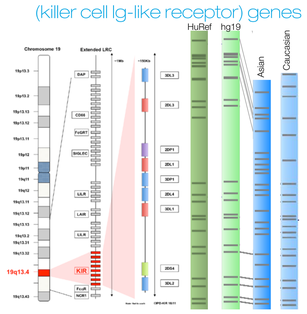


Deletions in genome maps

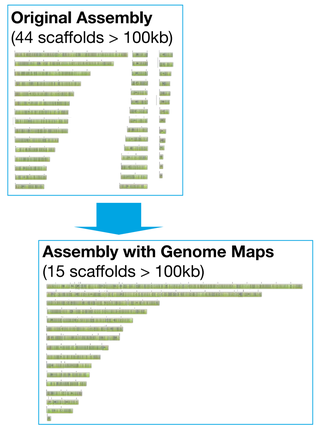


Single long motif maps (>100kb) from human cell lines were *de novo* assembled into genome maps at 55X depth. Molecules from two human samples were also aligned to hg19 as shown in the depth-plot, demonstrating broad genome-wide coverage with interspersed deviations indicating amplifications and deletions relative to each other and to the reference. Structural variation across a broad range of sizes refractory to many high throughput and short-read technologies was detected. Insertions are called by the presence of novel label sites and expansion of adjacent labels. Deletions are evident by the absence of label sites or narrowing of inter-label segments. Unlike indirect deduction methods such as arrays and mate-pair sequencing, there is no bias towards losses or only large variants. Maps also identify variation in difficult-to-sequence highly repetitive regions, such as those involved in immune function (such as LRC).

Human KIR (killer cell Ig-like receptor) genes



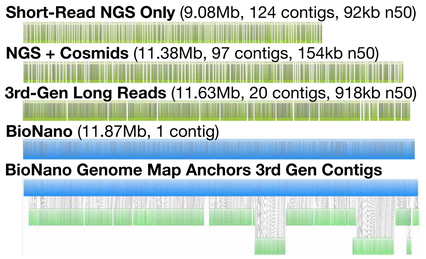
Scaffolding Gene Fragments in the Spider Mite Assembly



T. urticae DNA was used to create a *de novo* genome map and assemble sequence scaffolds and contigs. In addition to increasing whole-genome contiguity, the genome map was used to bridge important and repeat-rich silk genes as well as validate and correct the assembly.

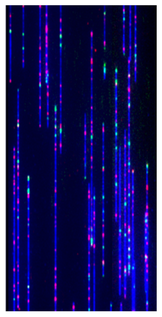
	Original Assembly	With Genome Maps
Size	90.8 Mb	90.8+ Mb
Scaffold N50	3.0 Mb	6.8 Mb
Largest Scaffold	7.8Mb	17.4 Mb
Scaffolds (>100kb)	44	15

Irys Genome Maps Are More Complete Than NGS or Third-Gen



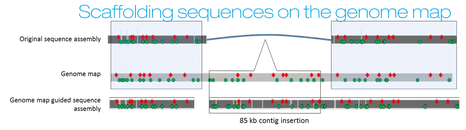
A *Streptomyces* genome was sequenced and assembled by a combination of different sequencing platforms (green maps). In contrast to these fragmented assemblies, one intact contiguous genome map (blue) was assembled *de novo* by the Irys system. The genome map anchored 19 of the 3rd-Gen Long Reads contigs.

Two-Color Genome Map of a Complex Region of Wheat Genome D



Dual-labeled single molecule images

Scaffolding sequences on the genome map



A BAC minimum tiling path for a 2.1 Mb region of *Ae. Tauschii*, the wheat D genome donor, was used to create a dual-motif genome map. Using this map, a physical map and the sequence assembly were corrected. The original sequence assembly (454 single read and paired-end reads) was 75% concordant with the genome map and was corrected to 95% accuracy by using the genome map.

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 genome-wide structural variation in human genome, the complete single-scaffold genome assembly of a model genome, two-color labeling of a complex region of wheat genome, and significant improvement of the assembly of an arthropod genome to enable better understanding of critical regions.

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.