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

A. Hastie[1], E. Lam[1], M. Requa[1], M. Austin[1], F Trintchouk[1], M. Saghbini[1], S. Rombauts[2], N. Rhind[3], Y. Gu[4], H. Cao[1]
[1]BioNano Genomics, San Diego, California, USA
[2]Ghent University, VIB, Gent, Belgium
[3]University of Massachusetts Medical School, Worcester, Massachusetts, USA
[4]Genomics and Gene Discovery Research Unit, US Department of Agriculture - Agricultural Research Service, Albany, California, 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. We present a single molecule genome analysis system (Irys™) based on NanoChannel Array technology that linearizes extremely long DNA molecules for 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* from the extremely long single molecules retain the original context and architecture of the genome, making them useful for structural variation and assembly applications.

Genome map-based scaffolding in shotgun sequencing experiments performed in parallel with second or third generation sequence production offers an integrated pipeline for whole genome *de novo* assembly solving many of the ambiguities inherent when using sequencing alone. Additionally, genome maps serve as a much-needed orthogonal validation method to NGS assemblies. As a result, genome maps improve contiguity and accuracy of whole genome assemblies, permitting a more comprehensive analysis of functional genome biology and structural variation.

In addition to providing an introduction to this newly available technology, we will demonstrate a number of examples of its utility in a variety of organisms, including an arthropod and crop plant.

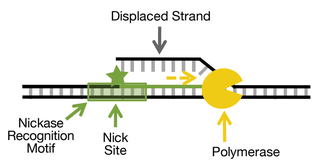
Background

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.

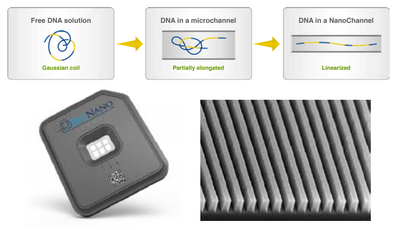
Methods

(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.

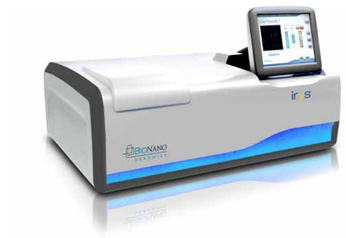
1) IrysPrep reagents label DNA at specific sequence motifs



2) IrysChip linearizes DNA in NanoChannels

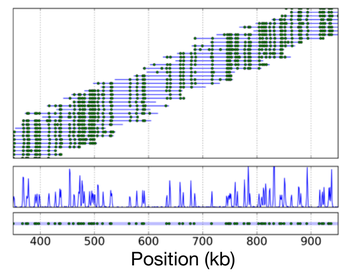


3) Irys automates imaging of single molecules in NanoChannels

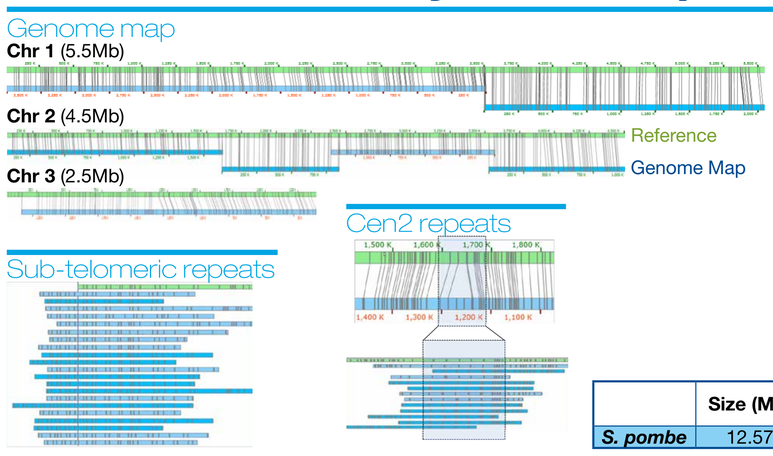


4) Molecules and labels detected in images by instrument software

5) IrysView software assembles genome maps



De Novo Assembly of the *S. pombe* Genome



Using the Irys platform, whole genome *de novo* assembly of the yeast, *S. pombe*, was performed. The genome maps cover 99% of the genome. All three centromeres are covered; previously unknown structural information for one of them (cen2) is shown. The sub-telemeric (rDNA) region was extended on the left end of chromosome 3.

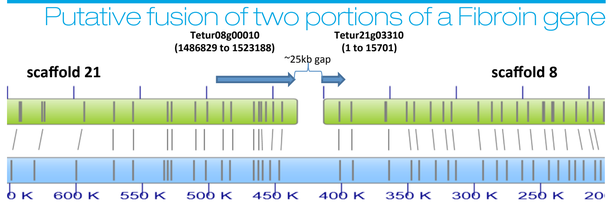
	Size (Mb)	Depth	Genome Maps	Genome Coverage
S. pombe	12.57	39x	7	99%

Spider Mite Genome Assembly

T. urticae DNA was used to create a *de novo* genome map and assemble sequence scaffolds and contigs. The complete *de novo* sequence assembly using genome maps for super-scaffolding is 90.8 Mb. The genome map was used to bridge important genes as well as validate and correct sequence assemblies. The scaffold N50 has so far been improved from 3Mb to 5.2 Mb (this assembly is still in progress).

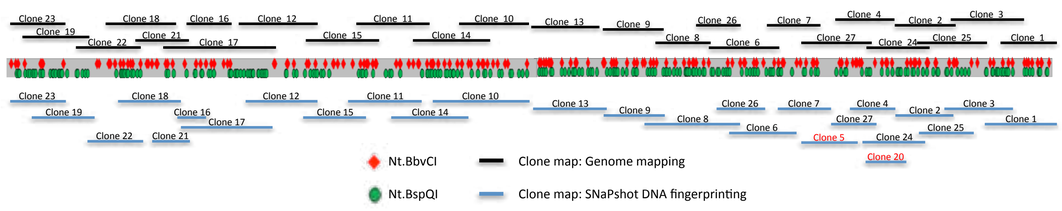
	Size (Mb)	N50 (Mb)	N90 (Mb)
Original assembly	90.8	3.0	0.9
Genome map assembly		5.2	1.8

Putative fusion of two portions of a Fibroin gene

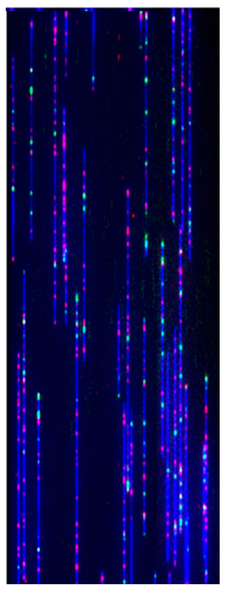


Assembly of a Region of the Wheat Genome

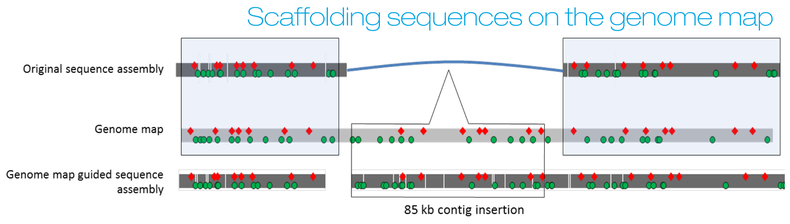
Genome map for 2.1 Mb of Ae. tauschii



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 dua-motif genome map. Using this map, a physical map and the sequence assembly were corrected. The genome map-independent 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 direct visualization of single-molecule, 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 the *S. pombe* centromeres and telomere, significantly improve the assembly of an arthropod genome, and assemble a difficult region of the wheat genome with multi-color mapping.

References

1) Hoskins, R.A. et al. Sequence finishing and mapping of Drosophila melanogaster heterochromatin. Science (2007); 15:1625-8
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.
5) Susan E Celniker et al, (2002) Finishing a whole-genome shotgun: Release 3 of the Drosophila melanogaster euchromatic genome sequence. Genome Biol. 2002; 3: research0079.1–79.14