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

BioNano GENOMICS

H. VanSteenhouse[1], A. Hastie[1], E. Lam[1], H. Dai[1], M. Requa[1], M. Austin[1], F. Trintchouk[1], M. Saghbini[1], S. Rombauts[2], N. Rhind[3], 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

Abstract

De novo genome assemblies using only short read data are generally incomplete due to intractable complexity found in most genomes. This complexity, consisting mainly of large duplications and repetitive regions, hinders sequence assembly and subsequent comparative analyses. Several "long-read" technologies have promised sufficient long-range information to span such repetitive elements and to measure adequate sequence uniqueness for unambiguous assembly of complex genomes. However, as yet, none have been reduced to practice with sufficiently long molecules, or on a platform available widely for routine use. 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. High-resolution genome maps assembled *de novo* from the extremely long single molecules retain the original context and architecture of the genome, making them extremely useful for structural variation and assembly applications.

Copy number and structural variation lack consistently robust methods of whole-genome detection. Comparison between individual genome maps generated by the Irys system permit accurate quantification of repeated elements, and precise positional localization of translocations or duplications.

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.

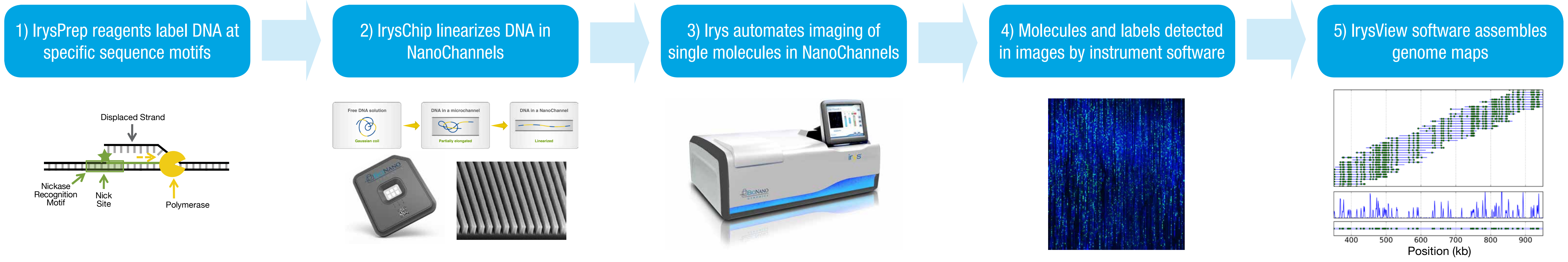
In addition to providing an introduction to this newly available technology, we will demonstrate a number of examples of its utility in a variety of organisms.

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

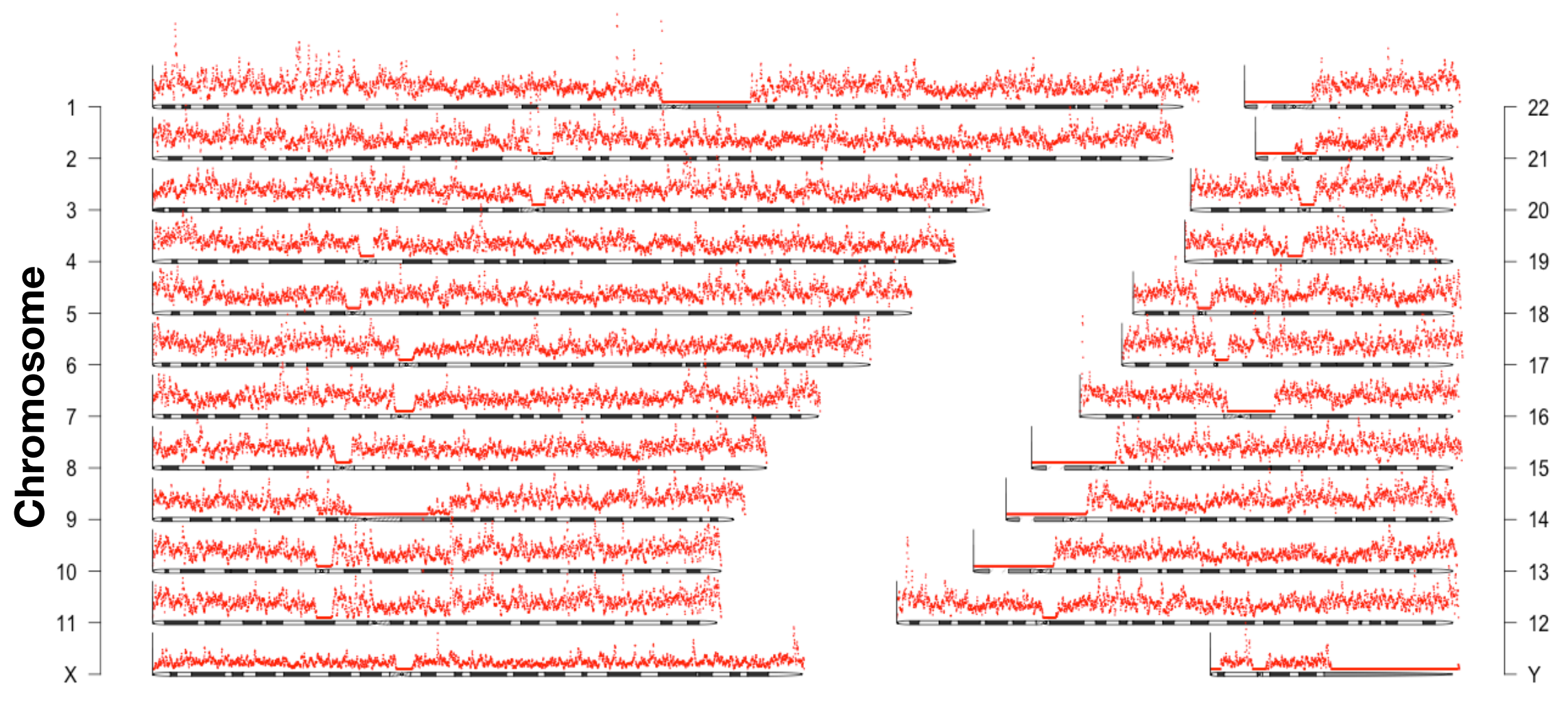
3) Irys automates imaging of single molecules in NanoChannels

4) Molecules and labels detected in images by instrument software

5) IrysView software assembles genome maps

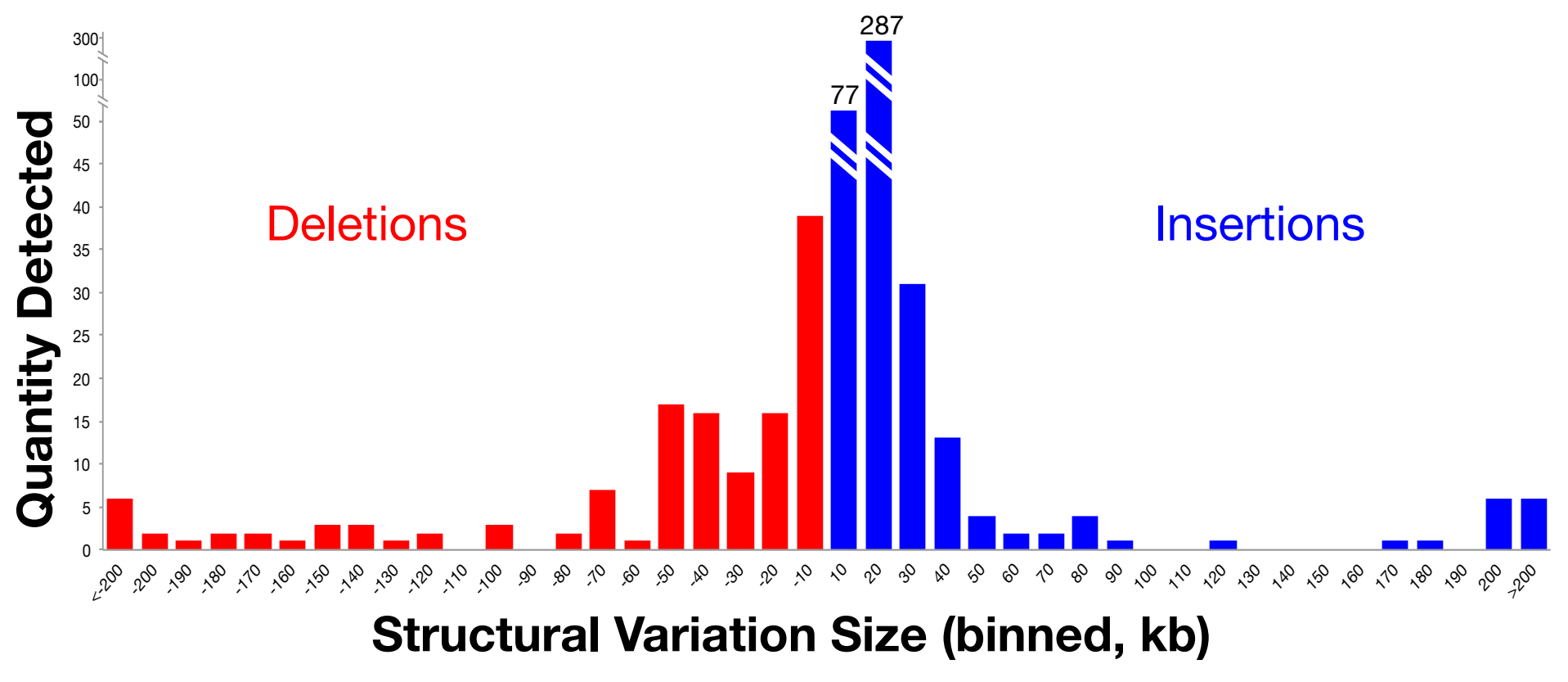
Human Genome-Wide Structural Variation Detection

Whole Genome Coverage of Long Molecules Mapped to hg19

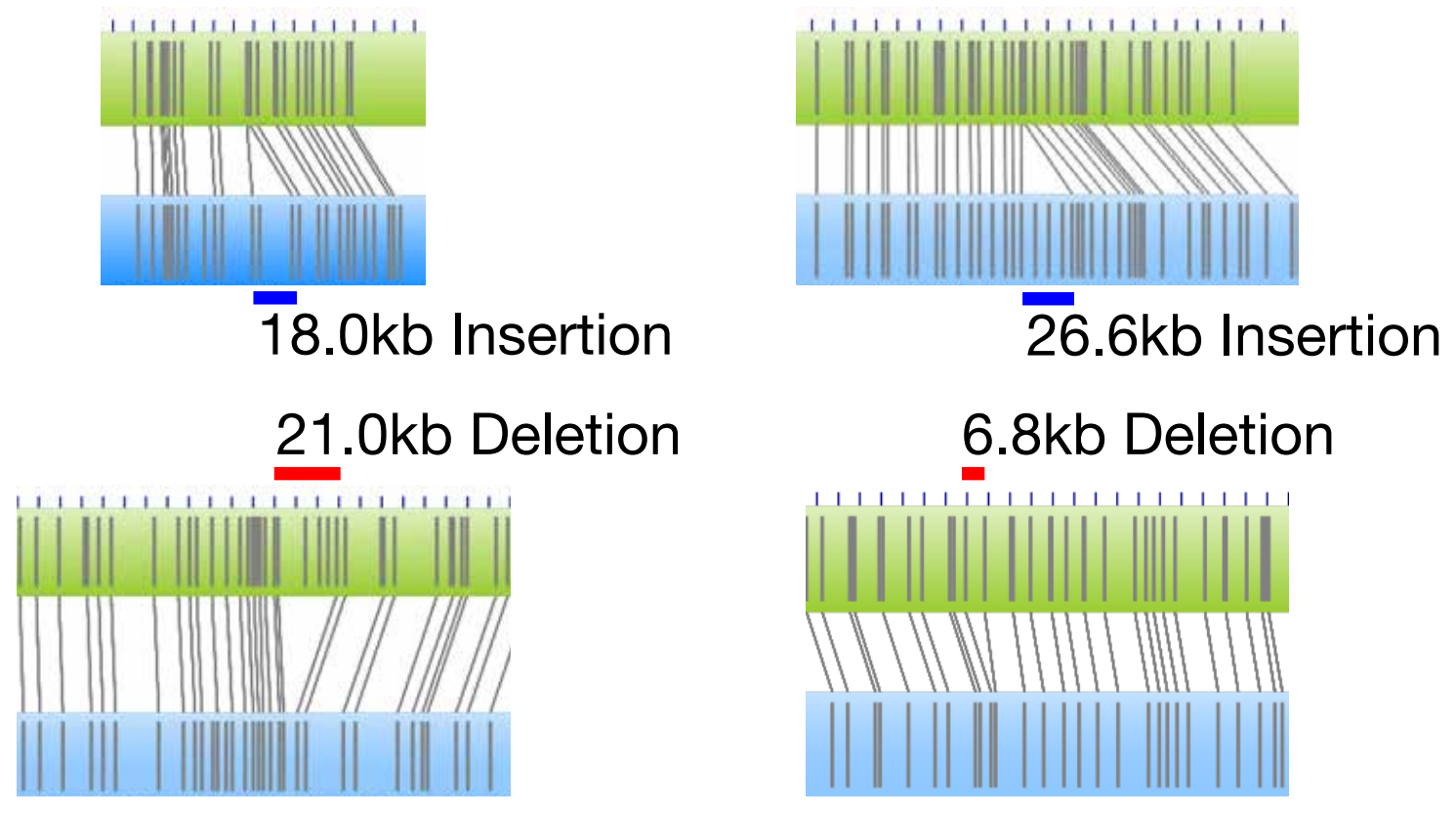


Single molecules from a human cell line mapped to hg19 provide a depth-plot (red plot) demonstrating uniform genome-wide coverage, with interspersed deviations indicating amplifications, deletions, and gaps in the reference. This dataset was then processed to comprehensively detect structural variation across sizes refractory to many high throughput and short-read technologies. Insertions (blue) are called by the presence of novel label sites and expansion of adjacent labels. Deletions (red) are evident by the absence of label sites. Unlike indirect deduction methods such as mate-pair sequencing, there is no bias towards loss or gain.

Whole-Genome SV Size Distribution

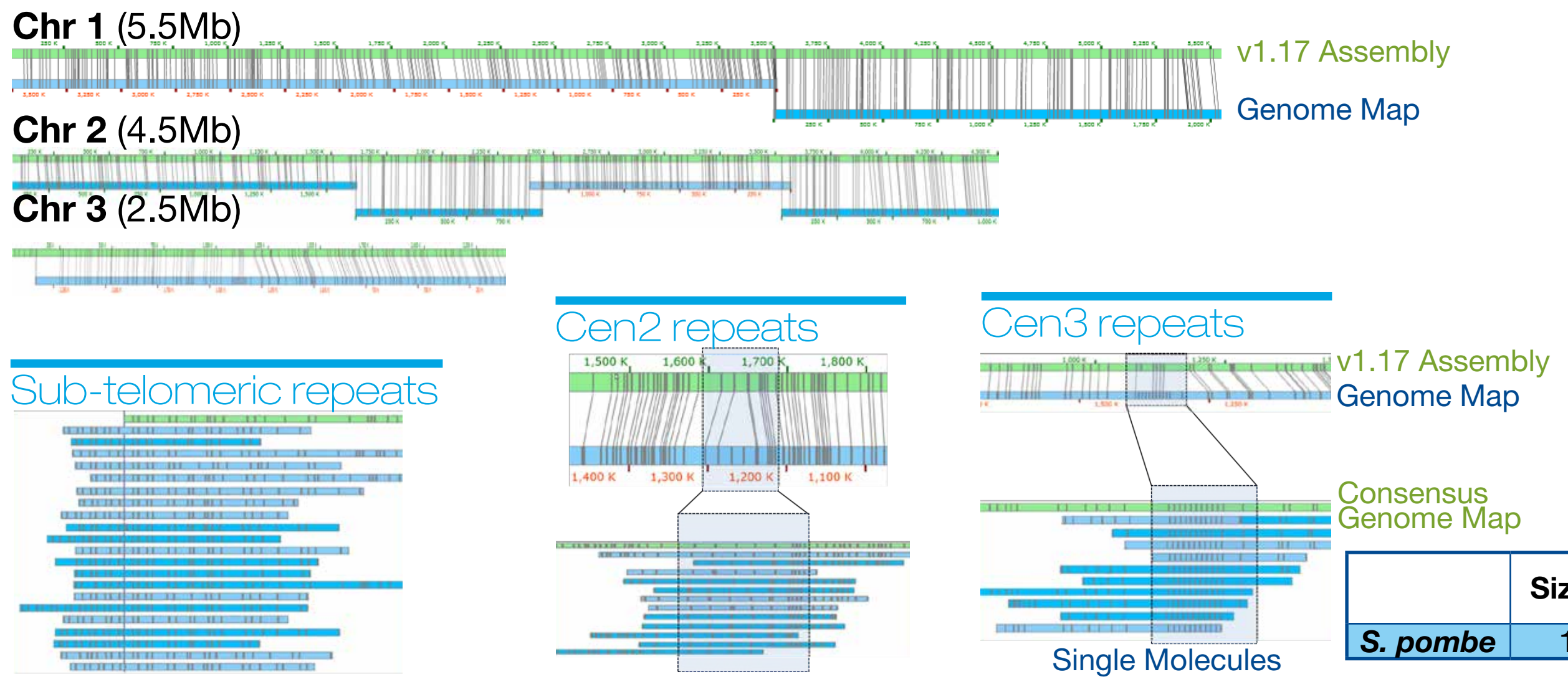


Kb-Scale Structural Variation Examples on Chr10



Repetitive Regions in *S. pombe* Genome

Genome map



Using the Irys platform, whole genome *de novo* assembly of the yeast, *S. pombe*, was performed. The genome maps cover 99% of the genome. The sub-telomeric (rDNA) region was extended on the left end of chromosome 3. All three centromeres are covered; previously unknown structural information for two of them (cen2 and cen3) are shown.

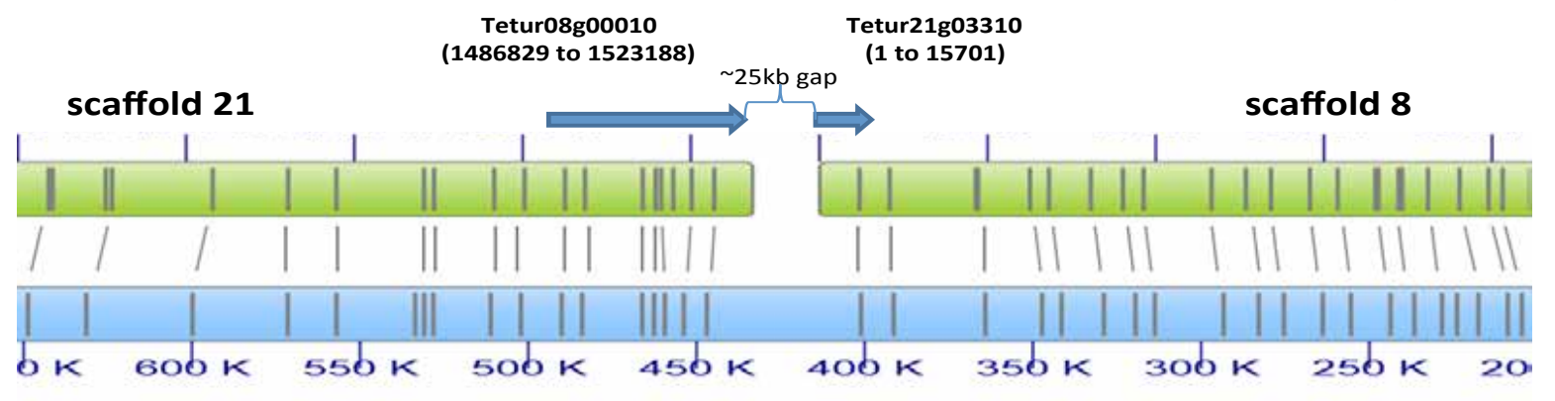
	Size (Mb)	Depth	Genome Maps	Genome Coverage
S. pombe	12.57	39x	7	99%

Merging Silk Gene Fragments in Spider Mite

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 and repeat-rich genes (such as those encoding silk proteins) 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



Conclusions

BioNano Genomics Irys enables 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 genome-wide structural variation in a human genome, the *S. pombe* centromeres and telomere, and significantly improve the assembly of an arthropod 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.