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

T Kelly, A Pang, H Dai, W Wang, A Hastie, E Lam, C Luo[1], J Ecker[1], H Stankova[2], J Dolezel[2], H Cao
BioNano Genomics, San Diego, California
[1]Salk Institute, San Diego, CA; [2]Institute of Experimental Botany, Czech Republic

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. The complexity present in large genomes consists mainly of large duplications and repetitive regions such as rDNA, centromeres, and telomeres. These features hinder sequence assembly and, in turn, narrow the scope of biological questions that can be addressed.

The BioNano Genomics Irys System linearizes extremely long DNA molecules and provides single-molecule reads containing this essential long-range information. These reads, which are hundreds of kilobases to megabases in length, retain far more structural information than is possible with sequencing platforms. Assembled genome maps are useful for scaffolding sequence contigs and validating sequence assemblies. Free from reference or amplification bias, de novo genome maps identify novel structural variations and repeats which are challenging to find with existing methods. Additionally, genome maps serve as a much-needed orthogonal validation method to NGS assemblies.

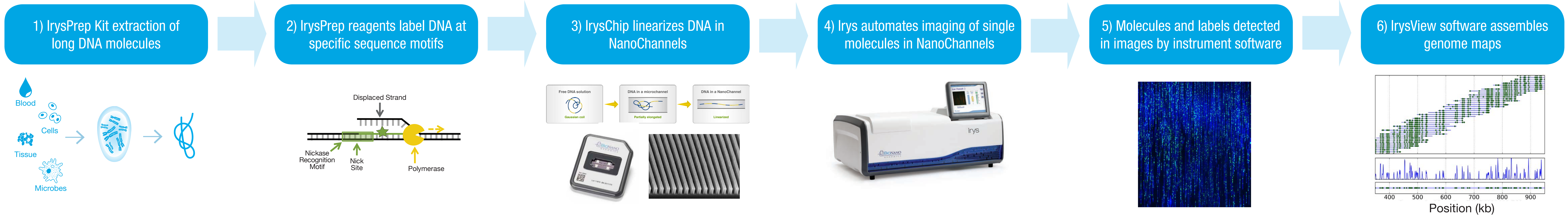
In addition to providing an introduction to this technology, we will demonstrate a number of examples of its utility in a variety of organisms, including examples from crop plants and model species. Genome maps are used to resolve repetitive functional elements and genes, validate or repair de novo sequence assemblies, and dissect cultivar differences and structural variants.

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) Long molecules of DNA is labeled with IrysPrep™ reagents by (2) 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) Single molecule data are collected and detected automatically. (5) Molecules are labeled with a unique signature pattern that is uniquely identifiable and useful in assembly into genome maps. (6) Maps may be used in a variety of downstream analysis using IrysView™ software.



1) IrysPrep Kit extraction of long DNA molecules

2) IrysPrep reagents label DNA at specific sequence motifs

3) IrysChip linearizes DNA in NanoChannels

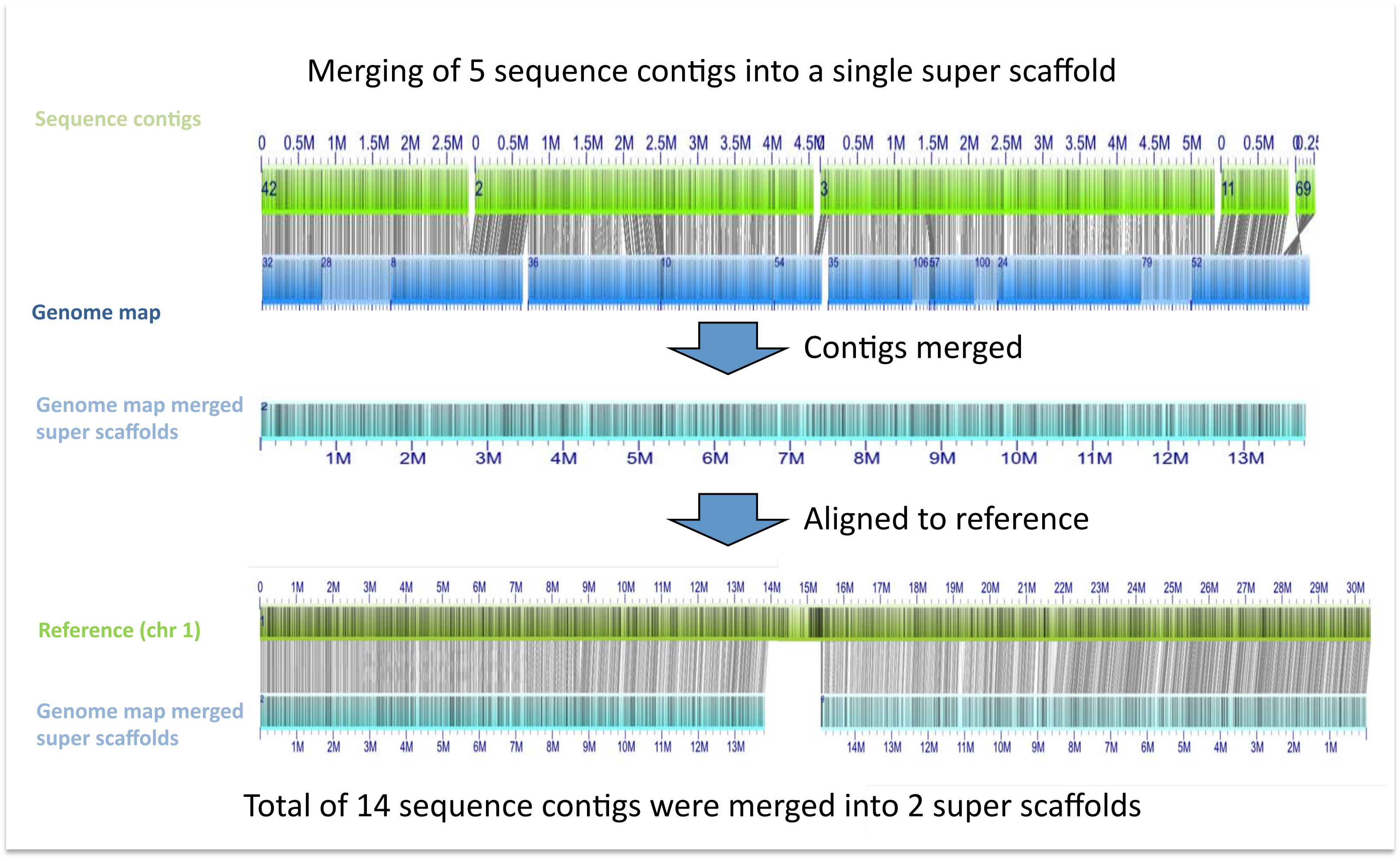
4) Irys automates imaging of single molecules in NanoChannels

5) Molecules and labels detected in images by instrument software

6) IrysView software assembles genome maps

Superscaffolding of *Arabidopsis*

In silico maps of sequence contigs have been merged by genome maps to produce superscaffolds, many of which represent whole chromosomal arms, as can be seen when aligned to the Col-0 reference.

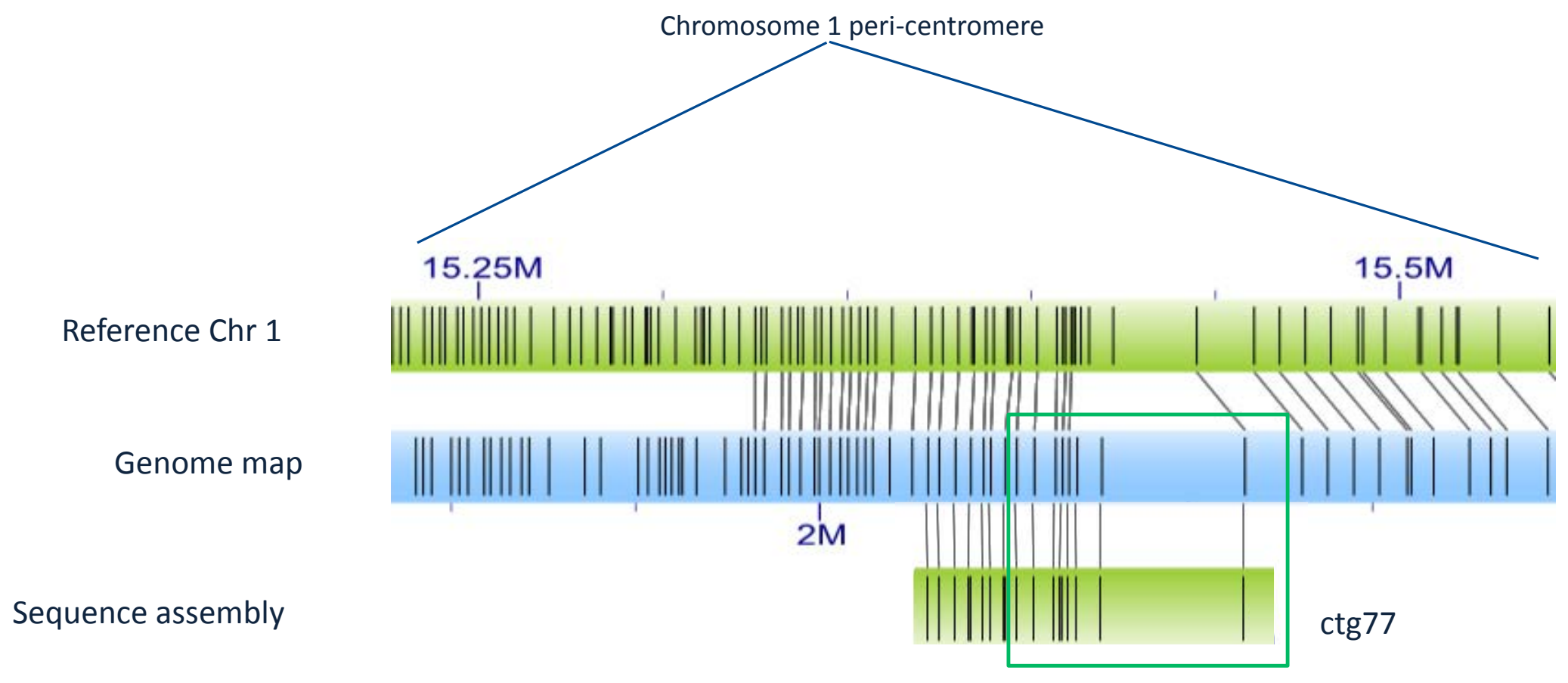


Merging of 5 sequence contigs into a single super scaffold

Contigs merged

Aligned to reference

Total of 14 sequence contigs were merged into 2 super scaffolds

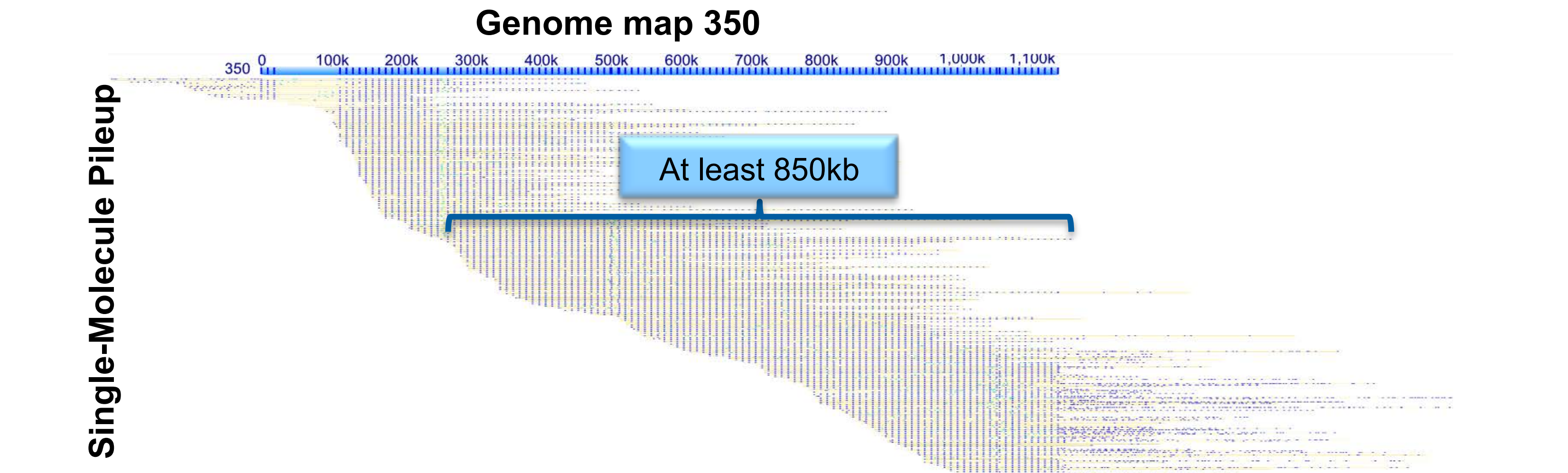


BioNano genome maps provide independent orthogonal evidence of structural variation, in this case, in a repetitive region near the centromere of chromosome 1.

De Novo Assembly of Wheat Flow-Sorted Chromosome 7DS

Extremely long single molecules from the Irys platform help to assemble extremely long tandem repeat unts. Here single molecules contain up to 850 kb of a 10 kb repeat unit in tandem.

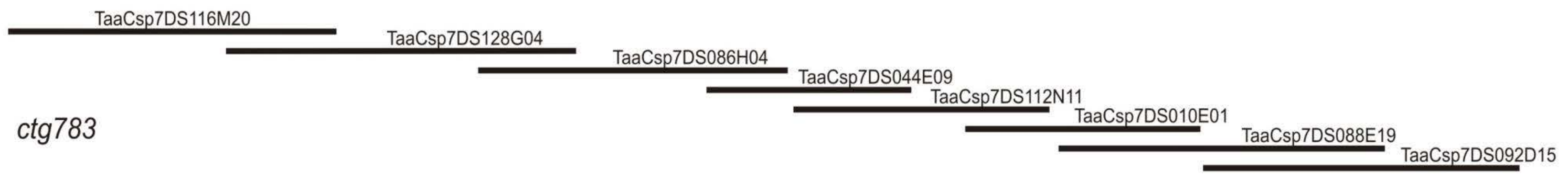
Chr7DS Assembly	
Total Map Length	350 Mb
Map N50	1.3 Mb
% Overlap Reference	92%



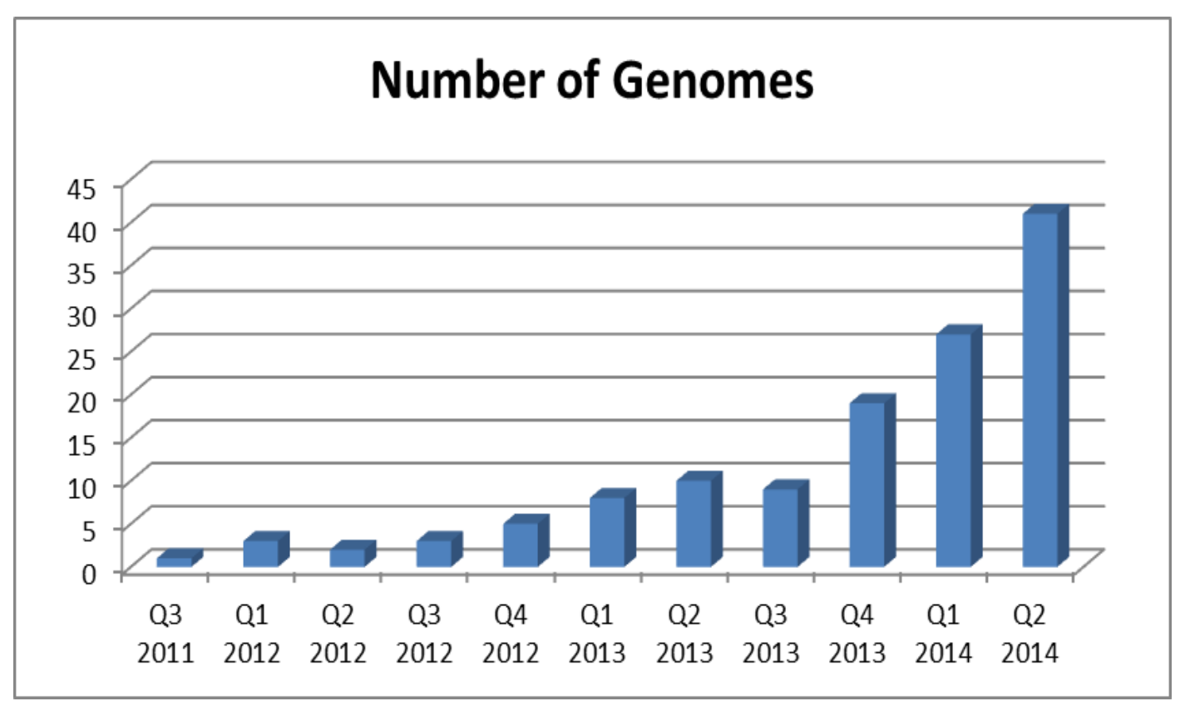
Genome map 350

At least 850kb

By aligning BAC clone sequences to the BioNano genome map, 8 BAC clones could be assembled as a single contig.



Genomes Mapped to Date



More than 80 unique organisms have already been mapped using the Irys genome mapping platform.

Conclusions

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. Together with sequencing information, we are able to construct superscaffolds which provide a more complete picture of the genome.

References

- Lam, E.T., et al. Genome mapping on nanochannel arrays for structural variation analysis and sequence assembly. Nature Biotechnology (2012); 10: 2303
- Frenkel et al. LTC: a novel algorithm to improve the efficiency of contig assembly for physical mapping in complex genomes (2010). BMC Bioinformatics 11: 584-601
- Hastie et al. Rapid Genome Mapping in Nanochannel Arrays for Highly Complete and Accurate De Novo Sequence Assembly of the Complex Aegilops tauschii Genome. (2013). PLoS ONE 8(2): e55864
- Kazakoff et al. Capturing the biofuel wellhead and powerhouse: the chloroplast and mitochondrial genomes of the leguminous feedstock tree Pongamia pinnata. (2012). PLoS ONE 7(12): e51687
- Kumar et al. Physical mapping resources for large plant genomes: radiation hybrids for wheat D-genome progenitor Aegilops tauschii. (2012). BMC Genomics 13: 597.
- Luo et al. A 4-gigabase physical map unlocks the structure and evolution of the complex genome of Aegilops tauschii, the wheat D-genome progenitor (2013). PNAS 110: 7940-7945