De Novo Assembly of Extremely Long Single-Molecule Genome Maps Imaged in Irys NanoChannel Arrays

M. Requa, M. Austin, H. Dai, H. Sadowski, M. Saghbini, H. Cao
BioNano Genomics, 9640 Towne Centre Dr, San Diego, CA 92121

Abstract

Despite significant advances in shotgun sequencing technology, *de novo* genome assemblies using only short read data are generally incomplete due to the complexity found in large genomes. This complexity, consisting mainly of large duplications and repetitive regions, hinders sequence assembly and subsequent comparative analysis. We present a single-molecule genome analysis system based on a NanoChannel Array™ technology that resolves these sequence ambiguities. This technology provides high throughput sequence motif maps of single molecule fragments hundreds of kilobases in size. Capitalizing on the information encoded in the extremely long single molecule maps, assembly algorithms unique to this format deliver high-resolution whole-genome sequence motif maps. Parallel assembly by shotgun sequencing and sequence motif mapping offers a multi-scale pathway for whole genome *de novo* assembly solving many of the ambiguities inherent in using short read assembly alone. Here we detail the BioNano genome assembly of *Drosophila melanogaster*, a eukaryotic model organism having a diploid genome estimated to be ~190 Mb haploid size. The genetic material was isolated from embryos. In this case, a high-quality genome draft exists, and the BioNano assembly is compared to the reference. Some structural variations are detected and previously unordered sequence contigs are anchored highlighting the utility of parallel multi-scale genome assembly for finishing applications. With this demonstration, *de novo* assembly of the human genome using BioNano technology is within reach.

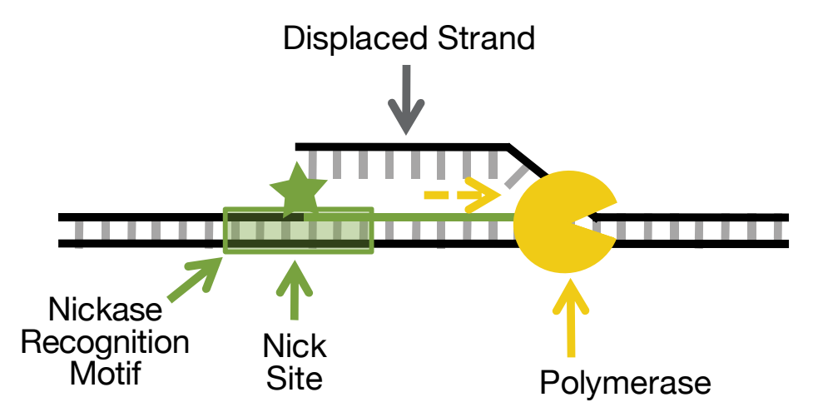
Background

Generating high quality finished genomes replete with accurate identification of structural variation and high completion (minimal gaps) remains extremely 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.

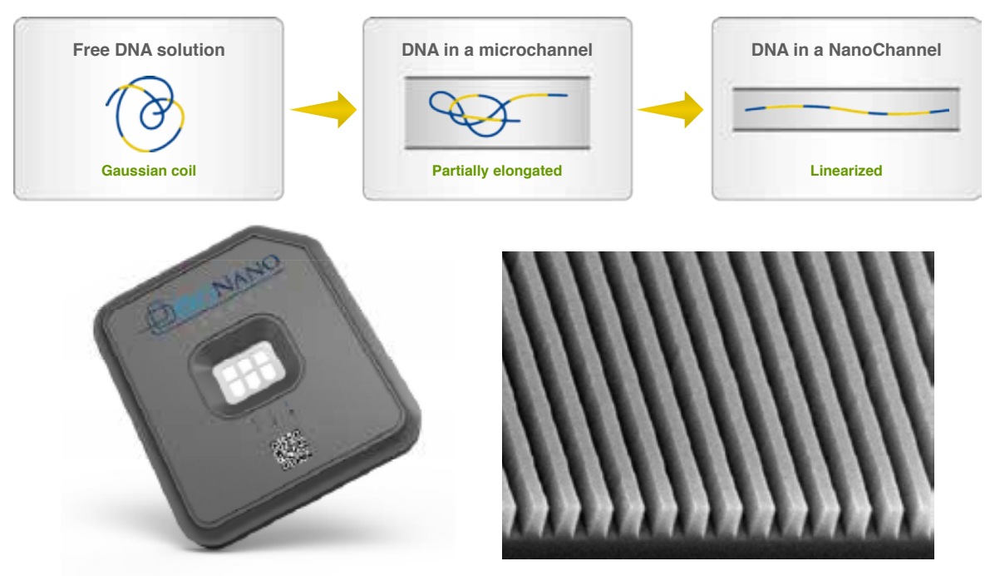
Methods

(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.

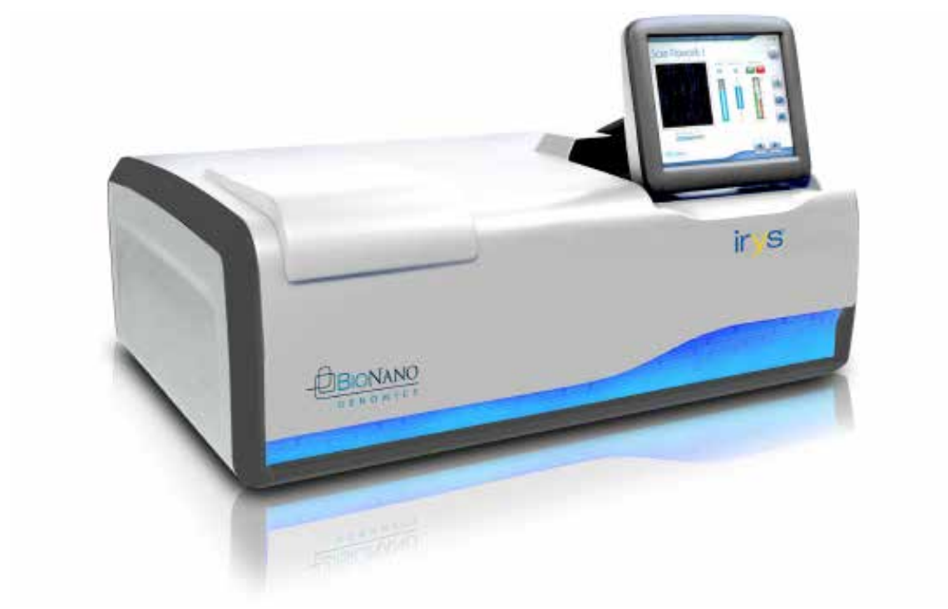
1) IrysPrep reagents label DNA at specific sequence motifs



2) IrysChip linearizes DNA in NanoChannels

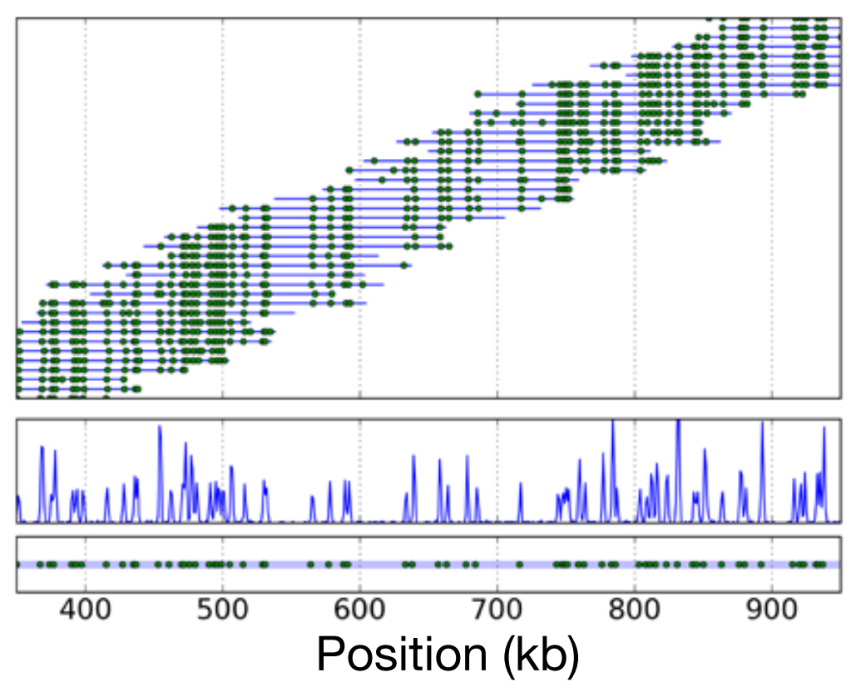


3) Irys automates imaging of single molecules in NanoChannels

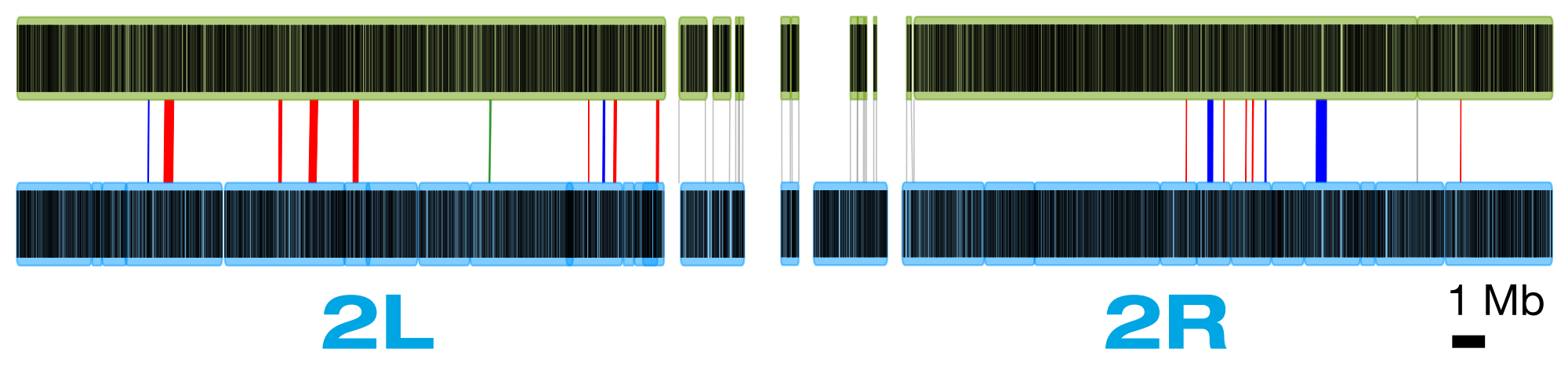


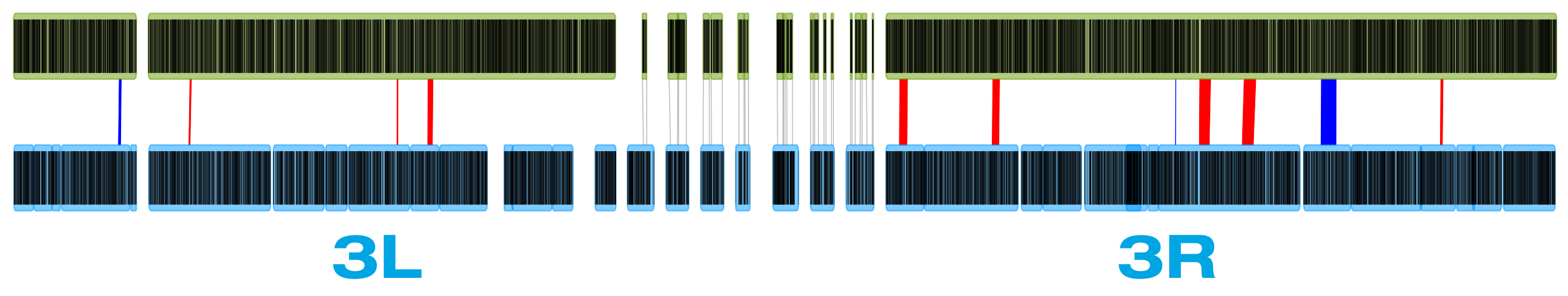
4) Molecules and labels detected in images by instrument software

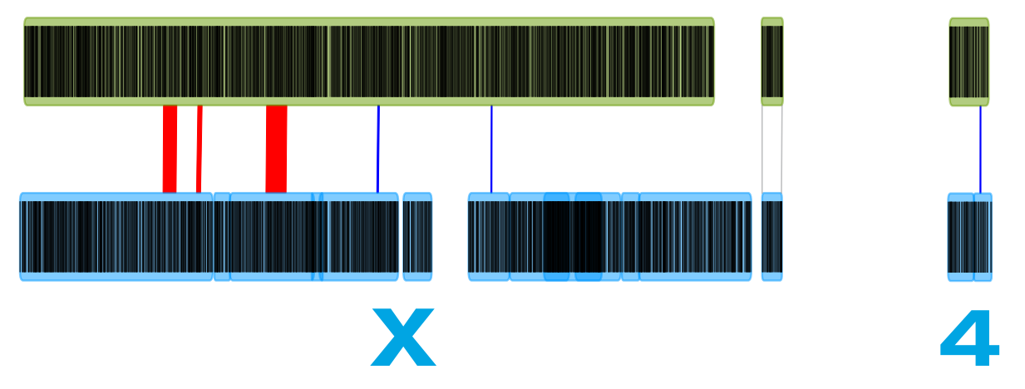
5) IrysView software assembles genome maps



De Novo Assembly and Anchoring







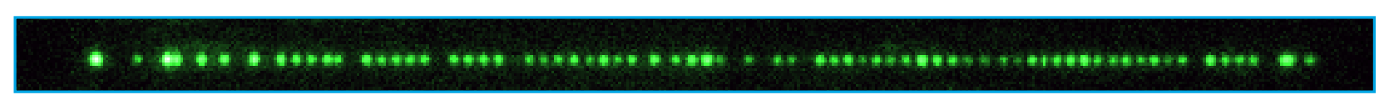
Whole Genome Map

Whole Genome *de novo* assembly of Drosophila melanogaster using the Irys platform was performed. Shown is the post-assembly comparison of consensus genome maps (blue) to the current reference sequence Dmel Rel 5.1[1] transformed into a corresponding sequence motif map (green). Consensus genome maps of low error and high contiguity allow for scaffolding (grey) previously unplaced reference contigs and discovery of structural variations, (red gains and blue losses).

Coverage Depth	75x
Genome Coverage	95%
Map N50	6.5Mb

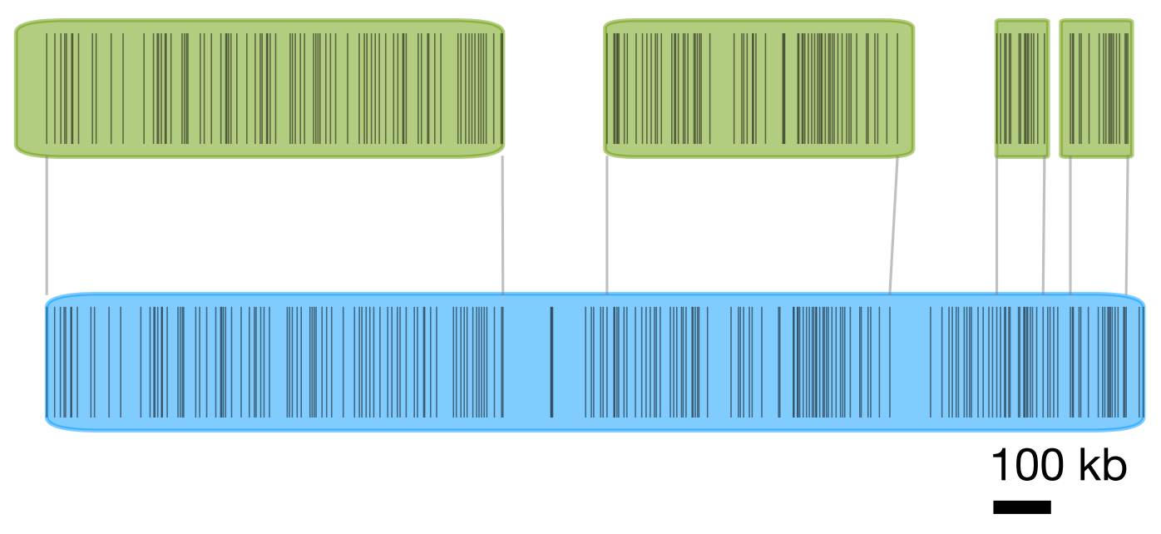
Visualize Highly Repetitive Regions

Irys provides long range information even in highly repetitive regions. A raw fluorescence micrograph depicts a molecule representing a ~250kb unfinished region containing a 2.5kb repeating element in the 2L arm of the reference genome. Repeat-masking algorithms are in development to automate the process of closing such expansive repetitive regions using Irys data.



Heterochromatin Contig Anchoring

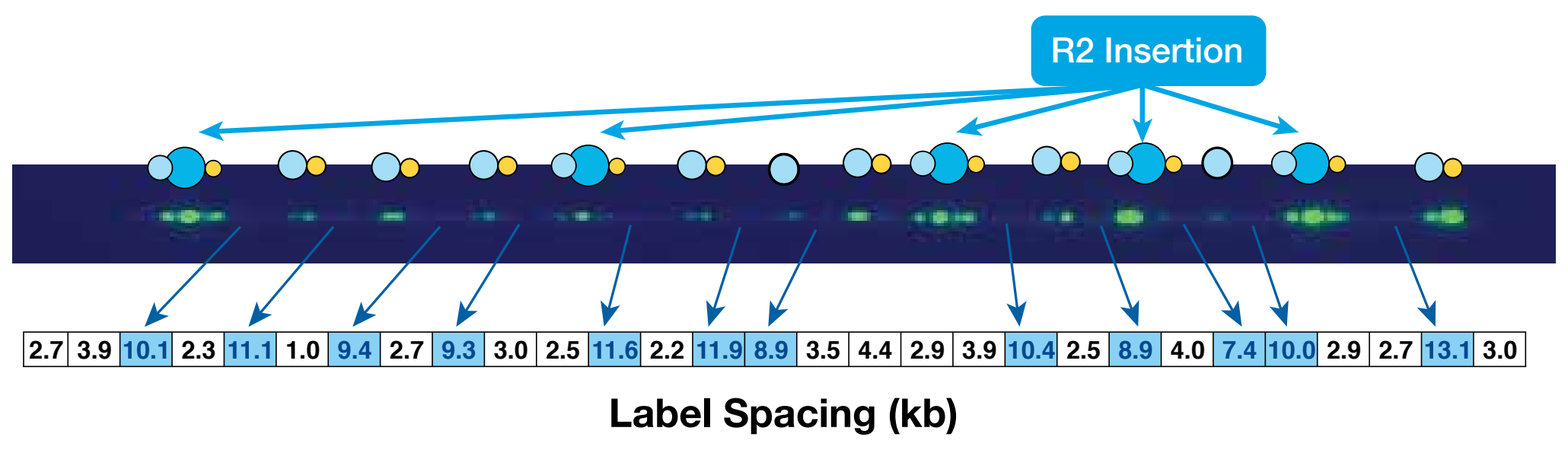
A *de novo* assembled consensus genome map provides contiguity information to anchor heterochromatin contigs with previously unknown gap sizes. Even in the case of Drosophila, a model organism with 20 years of significant research contributing to a numerous drafts of the reference sequence, there remain non-contiguities[1]. In this experiment we were able to assign measured distance of many euchromatin and heterochromatin contigs using *de novo* assembled consensus genome maps.

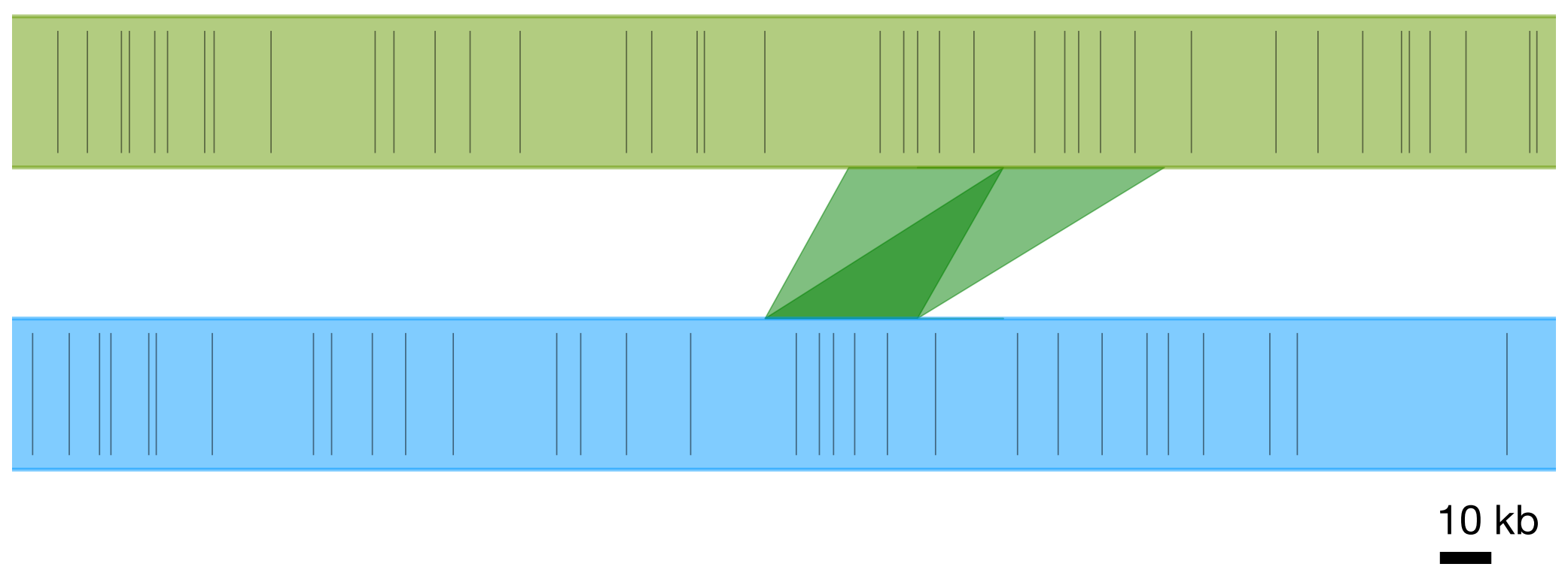


Structural Variation Analysis

rDNA Repeat Delineation

In drosophila, there are an estimated 100–200 copies of rDNA repeats on X and Y chromosomes. These tandem tracts each have two label sites 1.36kb apart within the 28S region, which results in a repeat pattern of twin labels separated by ~10kb spacing. Approximately half of rDNA carries site-specific insertion of non-LTR retrotransposons, which are 3–5kb, and produce altered label spacing of ~6kb.



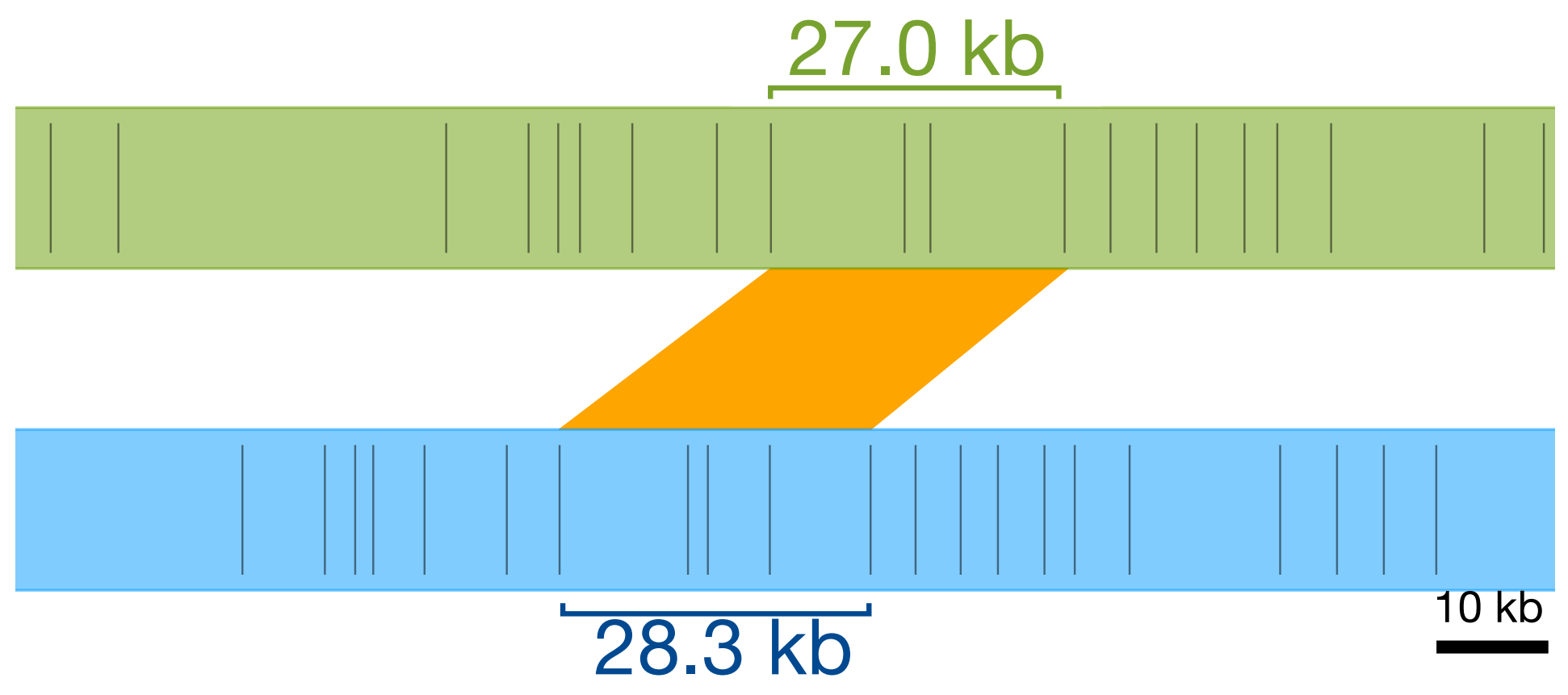


Copy Number Variation

A copy number polymorphism in the sample is detected relative to the reference sequence. We observe the expression of a single 25kb copy in the 2L chromosome arm whereas the reference contains two copies. The reference genome project described this polymorphism as heterozygous, yet the single-molecule data generated by Irys provides no evidence for hetrozygosity in the individual studied here.

Small Insertion

An insertion of 1.3 kb is detected in the 2L chromosome arm. High coverage depth allows for precise genome map feature localization. The inserted content contains an instance of the recognition sequence motif, adding an additional feature and increasing the segment size of adjacent labels. All 47 molecules that cover this region contain the insert, yielding high confidence and measurement precision.



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 a number of complex elements in a common model organism, *Drosophila melanogaster*. In addition to assembly of a high-coverage physical map, several examples detail the system's ability to resolve heterochromatin, repetitive regions, and copy number variation.

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
6) Polanco, C., Gonzalez, A.I., de la Fuente, A., Dover, G.A. (1998). Multigene family of ribosomal DNA in *Drosophila melanogaster* reveals contrasting patterns of homogenization for IGS and ITS spacer regions: A possible mechanism to resolve this paradox. Genetics 149: 243-256