Genome Map Assembly from NanoChannel Array Data for Structural Variation Detection in the Human Genome and Finishing in *Tribolium*

BioNano GENOMICS

W. Andrews[1], E. Lam[1], A. Hastie[1], H. Dai[1], M. Coleman[2], M. Austin[1], F. Trintchouk[1], M. Saghbini[1], T. Anantharaman[1], H. VanSteenhouse[1], K. Haden[1], T. Dickinson[1], S. Brown[2], X. Yang[1], E. Holmlin[1], H. Cao[1]
[1]BioNano Genomics, San Diego, California, USA
[2]Division of Biology, Kansas State University, Manhattan, KS

Abstract

We present the use of a newly available technology utilizing NanoChannel Arrays to analyze complex genomic architecture and functional regions by visualization of 100 kilobase and longer strands of intact genomic DNA. A successful de novo assembly of the human genome is presented and utilized in structural variation analysis. Structural variants are detected as low-scoring regions of the assembly flanked by high-scoring alignments. Examples of structural variants are presented from the KIR region of chromosome 19 and the IGH region of chromosome 14. These loci are important in the immune system function and both are known to be highly variable. The extremely long DNA molecules provide unique opportunities to study these complex structural variants which are difficult to analyze using sequencing alone.

The high assembly quality achievable with genome maps makes them useful for finishing where gaps between sequence contigs exist. We present an analysis of the Tribolium castaneum genome, and demonstrate the sizing of several gaps, as well as ordering and placing contigs with previously unknown locations. This organism's assembly has over 400 scaffolds and 7000 contigs and is complicated by a large fraction of repetitive heterochromatin sequence.

The Irys platform from BioNano Genomics overcomes the limitations of short fragment technologies to provide unprecedented insights into whole-genome biology. Irys is a single-molecule genome analysis system based on NanoChannel Array technology that linearizes extremely long DNA molecules for observation. This high-throughput platform automates massively parallel imaging of individual molecules of genomic DNA hundreds of kilobases in size to measure sufficient sequence uniqueness and long-range contiguity critical for unambiguous de novo assembly of complex genomes. High-resolution genome maps assembled de novo retain the original context and architecture of the genome, making them useful for sequence assembly scaffolding and structural variation detection applications.

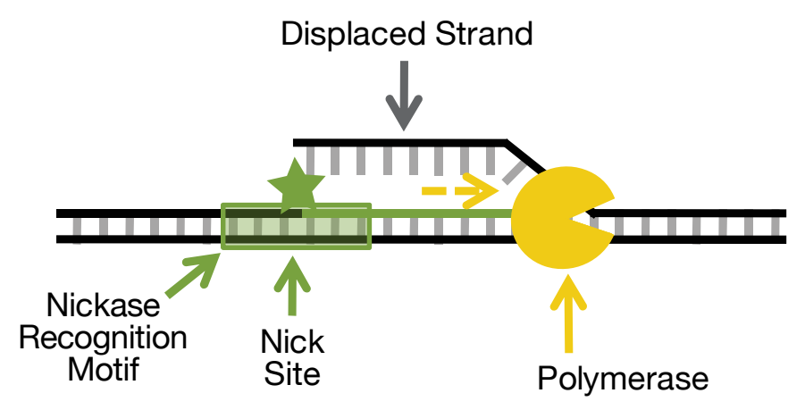
Background

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.

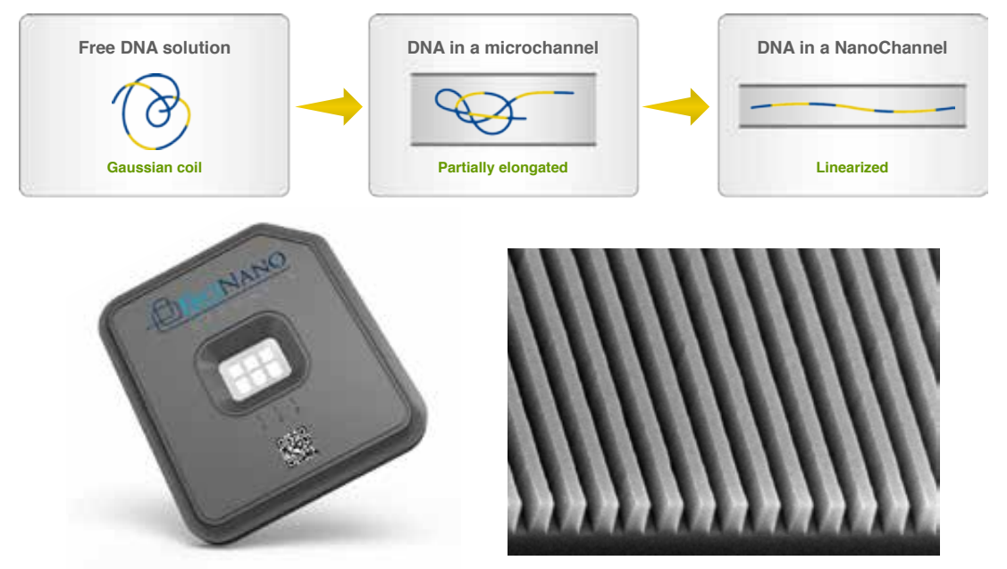
Methods

(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. (6) Downstream analyses include detection of structural variation.

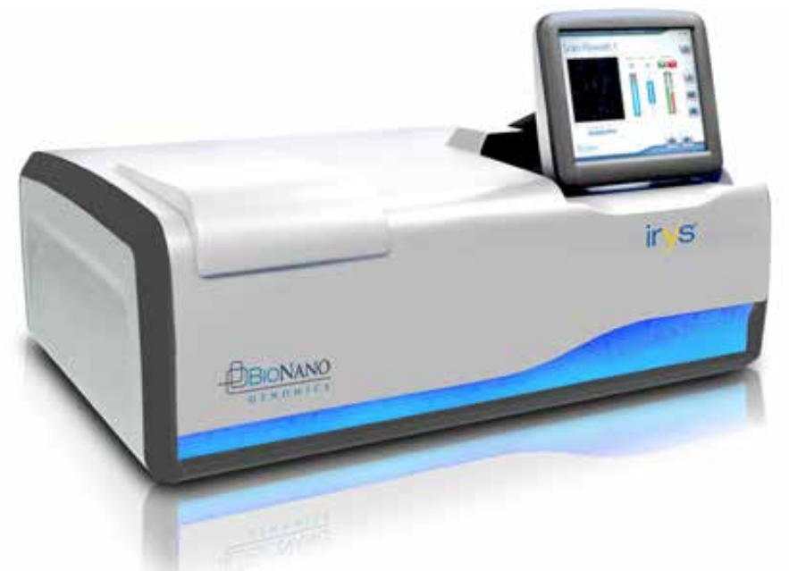
1) IrysPrep reagents label DNA at specific sequence motifs



2) IrysChip linearizes DNA in NanoChannels

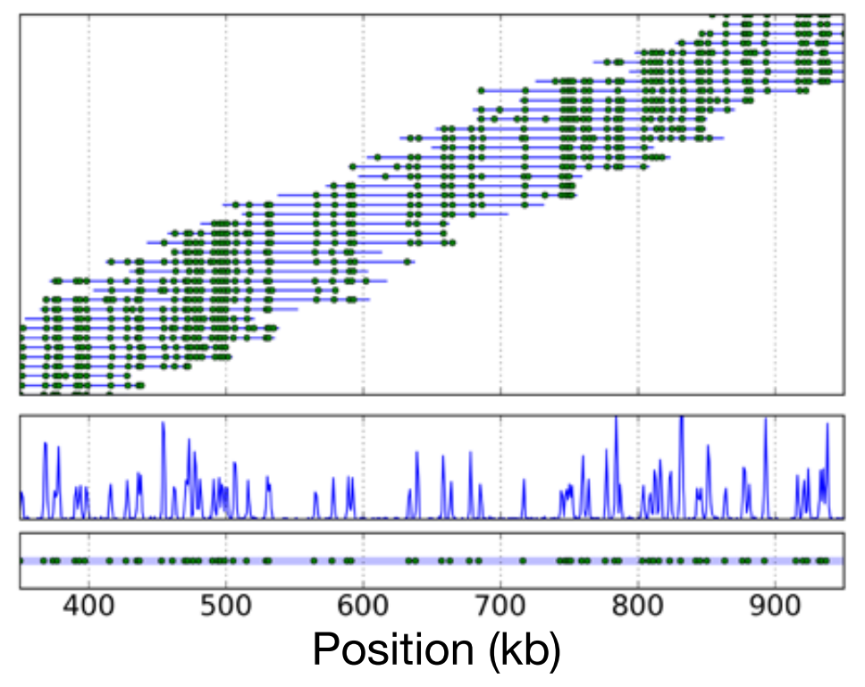


3) Irys automates imaging of single molecules in NanoChannels

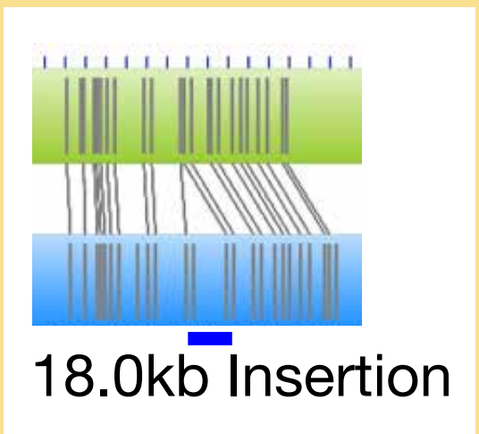


4) Molecules and labels detected in images by instrument software

5) IrysView software assembles genome maps

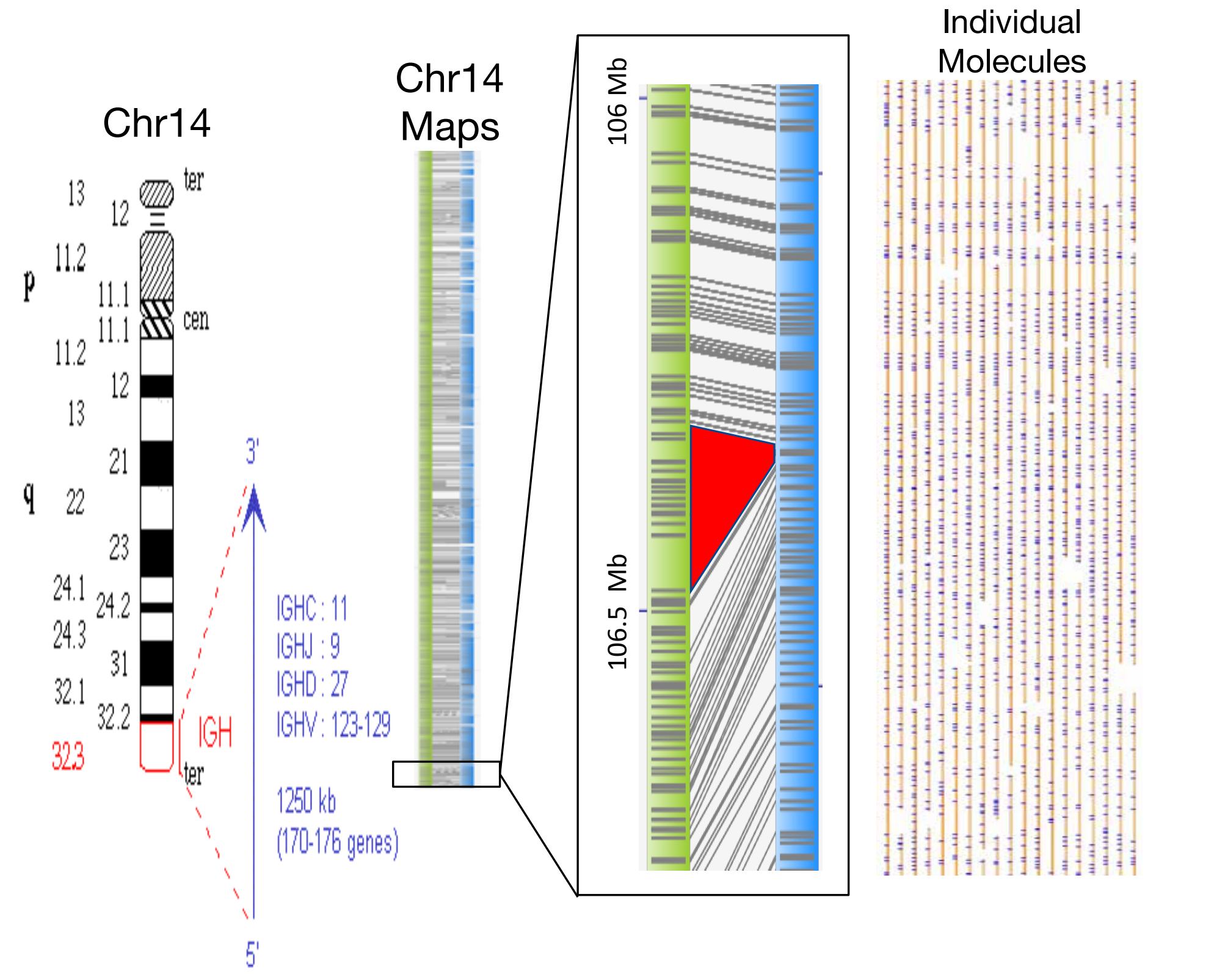


6) Automated Structural Variation Detection Algorithm



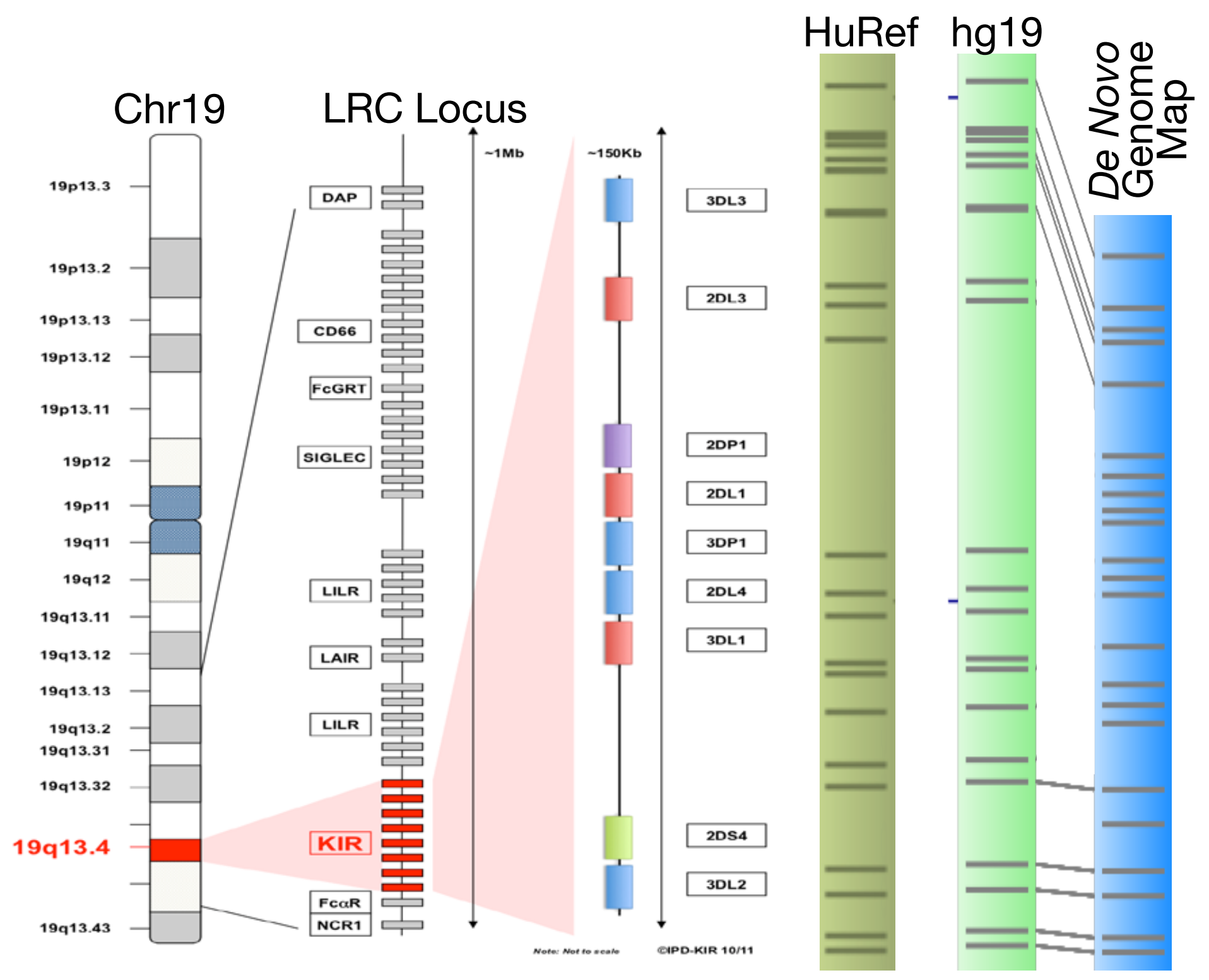
To align genome maps to each other or to a reference, a dynamic programming approach similar to classic Smith-Waterman is employed. However, instead of the units of comparison being single base-pairs, they are instead intervals between neighboring motif sites. Structural variations are called at regions that are discordant, having relatively low similarity. To gain confidence in such calls, it is required that the structural variations be flanked on both sides by high scoring regions.

Human Immunoglobulin Heavy (IGH) Locus Rearrangement



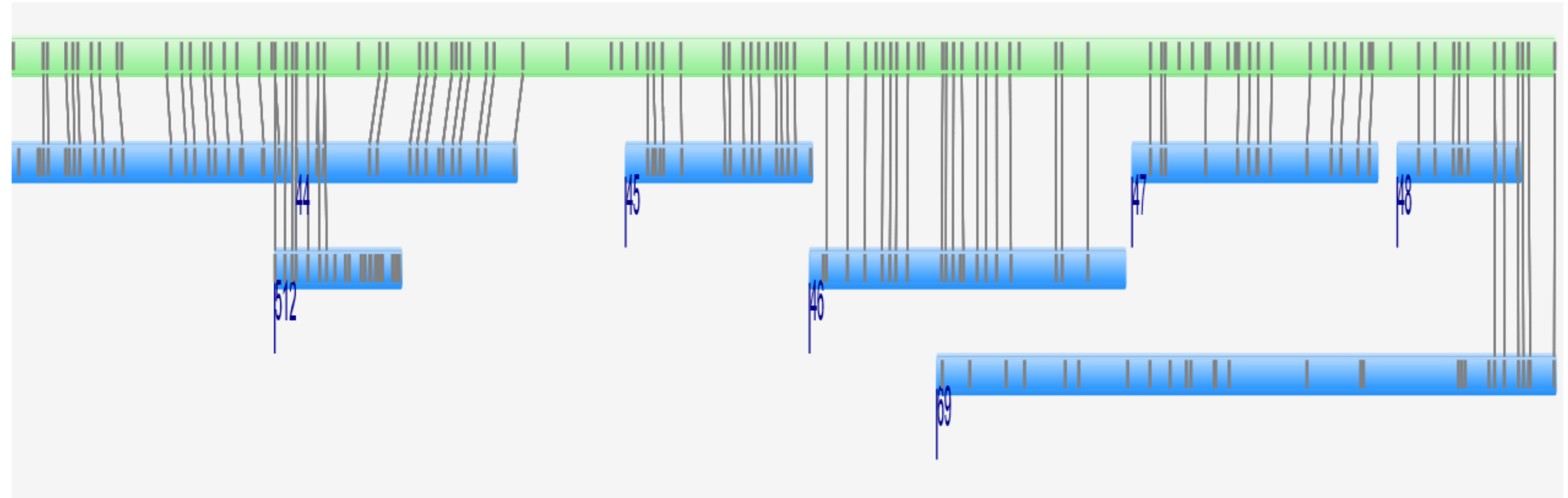
The Immunoglobulin Heavy Locus (IGH) region of chromosome 14 contains genes for immunoglobulins—heavy chains of human antibodies. The region codes for at least four different segments that are rearranged at various stages in B cell development. Due to the complexity of the region, rearrangements and other structural variations are common (1). In this case, we observe a large deletion event (~100kb) relative to hg19.

Human Killer Cell Immunoglobulin-like Receptors (KIR) Variability



Killer cell Immunoglobulin-like Receptors (KIR) are encoded by genes on chromosome 19. They interact with MHC molecules to detect pathogens or tumors, and therefore represent essential immune function. Due to the high diversity and combinatorial expression of these molecules, their coding regions are highly variable and contain many polymorphisms (2). Here we *de novo* assemble the rearranged region from a single individual, identifying several differences from HuRef or hg19.

Tribolium Genome Assembly and Scaffolding



The beetle *Tribolium castaneum* has a 200 Mb genome characterized by long stretches of repetitive heterochromatin sequence and low methylation. The current draft sequence has over eight thousand contigs and N50 of 43kb. BioNano genome maps provide great value in finishing such a genome, in ordering sequence scaffolds and in sizing gaps between sequence contigs. In this example, a single genome map (green, ~1 Mb) anchors five sequence contigs that are part of the same scaffold, thereby sizing four gaps. In addition, two contigs from different scaffolds also align to this 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 and variant detection capabilities and significant scaffolding improvement of the assembly of an arthropod genome.

References

1) JV Ravetch, et al. Structure of the human immunoglobulin mu locus: Characterization of embryonic and rearranged J and D genes Cell 27:583–591

2) C Vilches and P Parham. (2002) KIR: Diverse, Rapidly Evolving Receptors of Innate and Adaptive Immunity Annual Review of Immunology 20: 217–251