De Novo Assembly of Complex Genomes Using Extremely Long Single-Molecule Imaging Technology

W Wang[1], S Chan[1], A Hastie[1], H Wu[2], X Jiang[3], J Wu[4], H Cao[1]
[1]BioNano Genomics, San Diego, California, USA; [2]Genergy Biotechnology, Shanghai, China;
[3]The Institute of Medicinal Plant Development (IMPLAD), Beijing, China;
[4]The Institute of Apicultural Research, Chinese Academy of Agricultural Sciences, Beijing, China

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 most 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 and capture 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 also 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 honey bee (*Apis cerena*) and Zi Zhi (*Gandoerma Sinesis*). Genome maps are used to resolve repetitive functional elements, validate or scaffold de novo sequence assemblies, and discover differences in haplotypes.

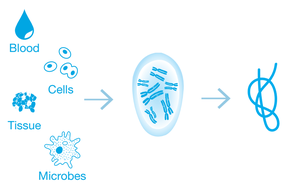
Background

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.

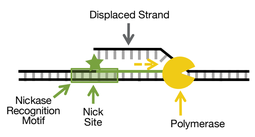
Methods

(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.

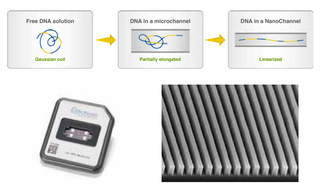
1) IrysPrep Kit extraction of long DNA molecules



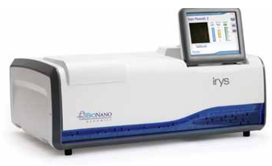
2) IrysPrep reagents label DNA at specific sequence motifs



3) IrysChip linearizes DNA in NanoChannels

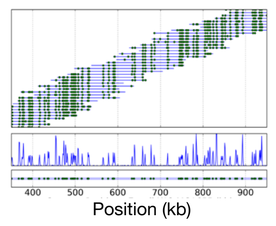


4) Irys automates imaging of single molecules in NanoChannels



5) Molecules and labels detected in images by instrument software

6) IrysView software assembles genome maps

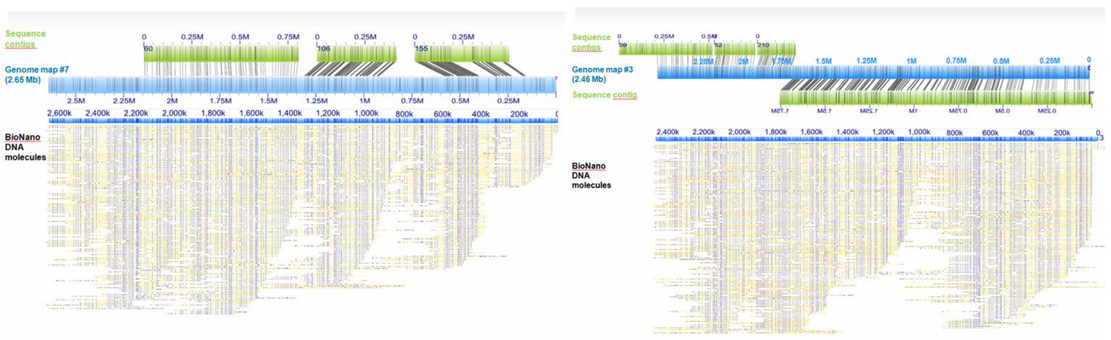


De Novo Assembly of Chinese Honey Bee

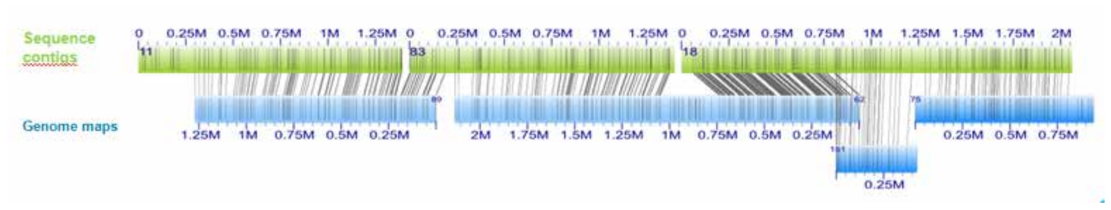
	Honey Bee *De Novo* Assembly
Assembly Size (Mb)	202.88
Map N50 (Mb)	1.38
#Maps	225
# of Aligned Genome Maps	179 (79.5%)

We generated high quality runs with Irys platform and assembled the Chinese Honey Bee (*Apis Cerena*) with BspQI nicking motif. The assembled genome maps had high confidence and total length was similar to expected genome size (210 Mb).

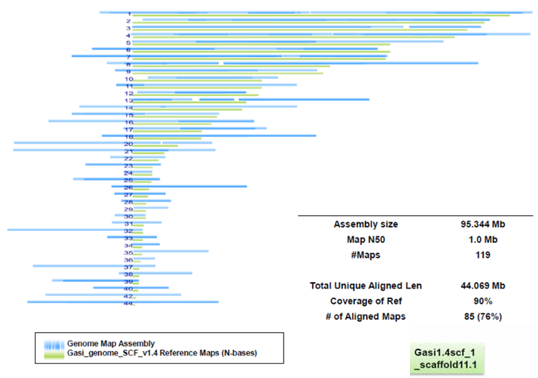
The *Apis Cerena* genome maps align well to the sequence assembly of the sample. In addition to validation of the sequence assembly, the genome maps scaffold sequence contigs.



We could improve assembly results and construct super scaffolds by merging the data from both sequencing contigs and BioNano genome maps.



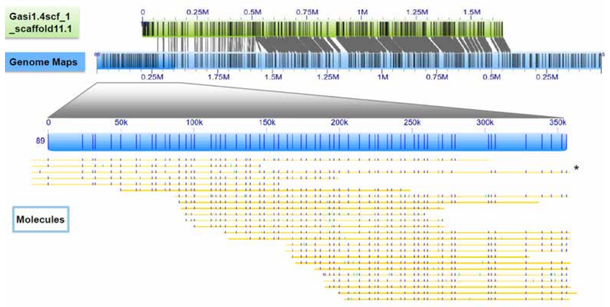
De Novo Assembly and Haplotype Discovery (*Gandoerma Sinesis*)

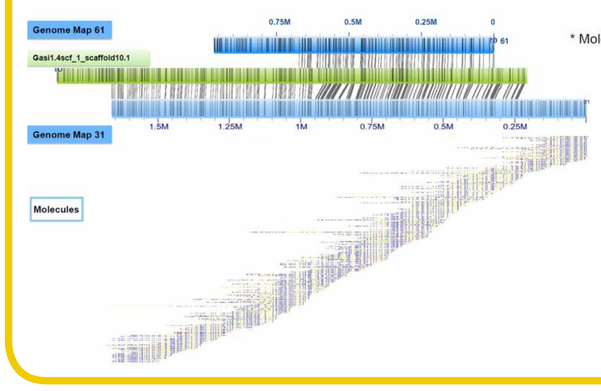


Assembly size	95.344 Mb
Map N50	1.0 Mb
#Maps	119
Total Unique Aligned Len	44.069 Mb
Coverage of Ref	90%
# of Aligned Maps	85 (76%)

We assembled the medicinal mushroom (*Gandoerma Sinesis*) with BbvCI nicking motifs and aligned the genome maps to the sequence scaffold assembly .

Long molecules from Irys platform provided extra information about the genome, such as repeat regions, which were not previously captured by sequence assembly.





* Molecule (ID#36890) spanning the whole 350 kb region

We identified haplotype differences using genome maps.

Conclusions

Irys enables understanding of complex genomes through visualizing extremely long single genomic DNA. With the system, we can *de novo* assemble genomes, validate sequencing assembly, detect structural variants, and identify genome features that typically confound short read genome assembly and comparative genomic analysis. Together with sequencing information, we're able to construct superscaffolds and identify haptotypes, 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
- 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
- Xiao, M et al. Rapid DNA mapping by fluorescent single molecule detection. Nucleic Acids Research (2007); 35:e16.