Integrated Genome Mapping in NanoChannel Arrays and Sequencing for Better Human Genome Assembly and Structural Variation Detection

Andy Wing Chun Pang[1], Alex Hastie[1], Palak Sheth[1], Thomas Anantharaman[1], Željko Džakula[1], Han Cao[1]

[1]BioNano Genomics, San Diego, California, United States of America

Abstract

De novo genome assemblies using purely short sequence reads are generally fragmented due to complexities such as repeats found in most genomes. These characteristics can hinder short-read assemblies and alignments, and that can limit our ability to study genomes.

The BioNano Genomics Irys® System linearizes long DNA molecules, yielding single molecules containing long-range genomic information. These molecules, hundreds of kilobases in length, provide the structural information often missed by sequencing technologies. Once assembled, the genome maps from these molecules scaffold sequencing assemblies to validate the accuracy of the sequences, and to anchor the adjacent sequences into the proper order and orientation. The resulting long-range hybrid scaffolds can then be used to identify novel chromosomal rearrangements undetectable by short-read alignments or reference-guided assembly approaches.

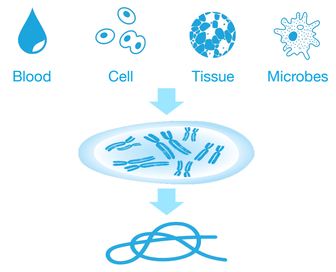
Here, we present a comprehensive analysis of a human genome by combining single molecule genome mapping with one of the most annotated sequence assemblies, the HuRef (Human Reference Genome – J. Craig Venter Institute) assembly. Overall, we found that the assemblies of the combined sequencing and genome mapping technologies correspond well, and the resulting hybrid scaffolds are highly contiguous, with a N50 exceeding 35 Mb, a value typically unachievable by short-read sequencing. In addition, we compared the structural variation with calls previously detected in the HuRef assembly, and found multiple novel variants spanning over hundreds of kilobases in size. Some of these variants reside in areas where the sequence assembly was poorly covered or was highly fragmented; yet these variants encompass numerous genes, and can be of functional importance. Finally, we identified genome maps that span over the remaining reference gaps, and maps that resolve and measure long tandem repeats.

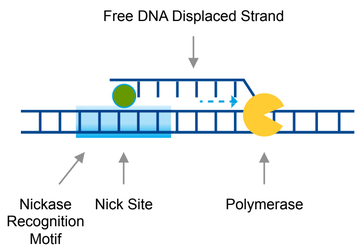
Background

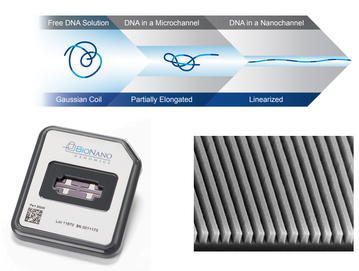
Generating high-quality finished genomes replete with accurate identification of structural variation and high completion (minimal gaps) remains challenging using short read-based sequencing technologies alone. The Irys System provides direct visualization of long DNA molecules in their native state, 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 assemblies and structural variation.

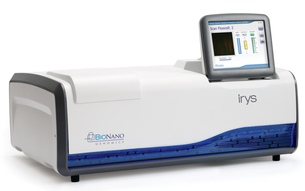
Methods

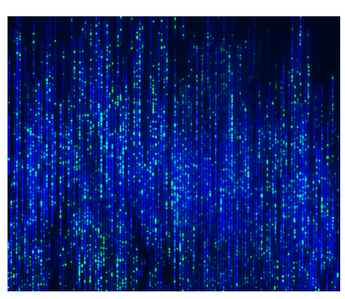
1. IrysPrep® kit extraction of long DNA molecules
2. IrysPrep® reagents label DNA at specific sequence motifs
3. IrysChip® linearizes DNA in NanoChannels
4. Irys® automates imaging of single molecules in NanoChannels
5. Molecules and labels detected in images by instrument software
6. IrysView® software assembles genome maps

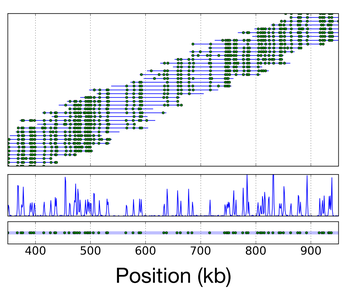






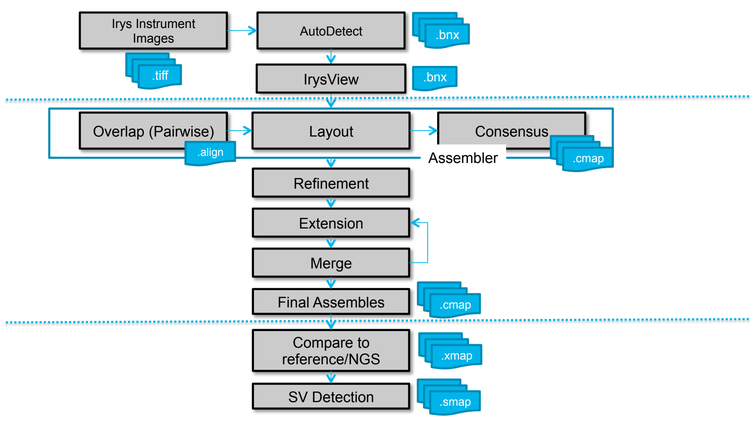




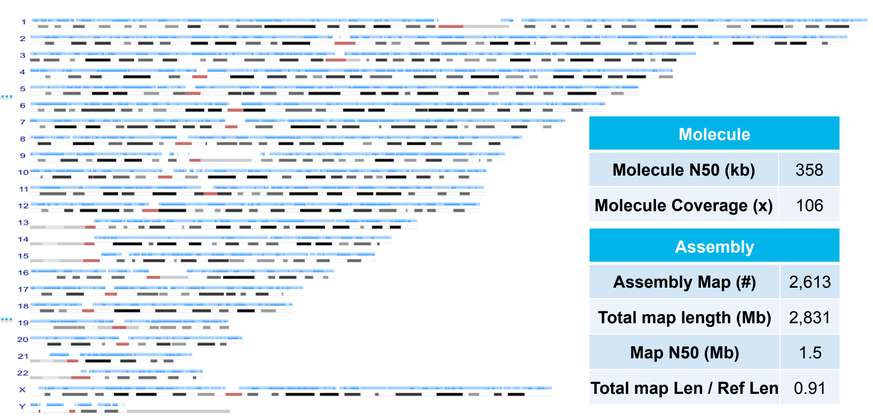


(1) Long molecules of DNA is labeled with IrysPrep® reagents by (2) incorporation of fluorophore labeled nucleotides at a specific sequence motif throughout the genome. (3) The labeled genomic DNA is then linearized in the IrysChip® NanoChannels and single molecules are imaged by Irys®. (4) Single molecule 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.

Assembly Pipeline Overview

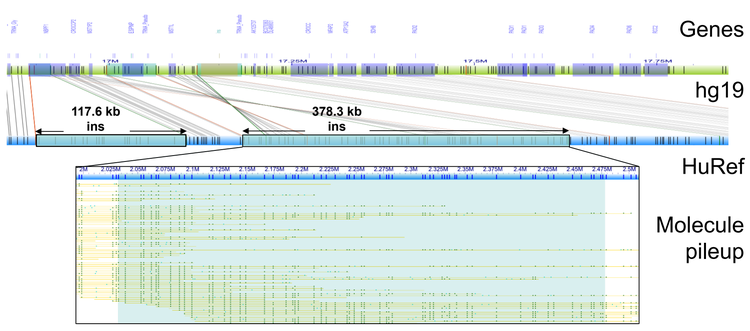


HuRef Genome Map Assembly



Molecule	
Molecule N50 (kb)	358
Molecule Coverage (x)	106
Assembly	
Assembly Map (#)	2,613
Total map length (Mb)	2,831
Map N50 (Mb)	1.5
Total map Len / Ref Len	0.91

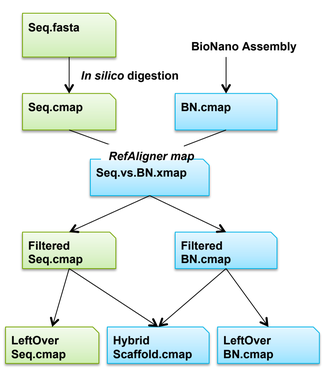
Complex Insertions on chr1 Uniquely Detected by BioNano



Our assembly data identified two novel neighboring insertions (117.6 kb and 378.3 kb) impacting multiple genes.

Hybrid Scaffold Pipeline Overview

1. Generate *in silico* maps for sequence data
2. Align sequence and BioNano maps to flag potential chimerical contigs
   - 2b. Alignment Visualization, QC, Manual Inspection
3. Run Merge to generate "Hybrid Scaffolds".
4. Generate sequence – hybrid scaffold alignments *(and generate AGP output etc.)*

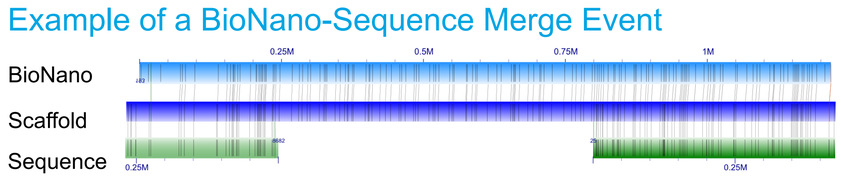


Ultra-long Hybrid Scaffolds

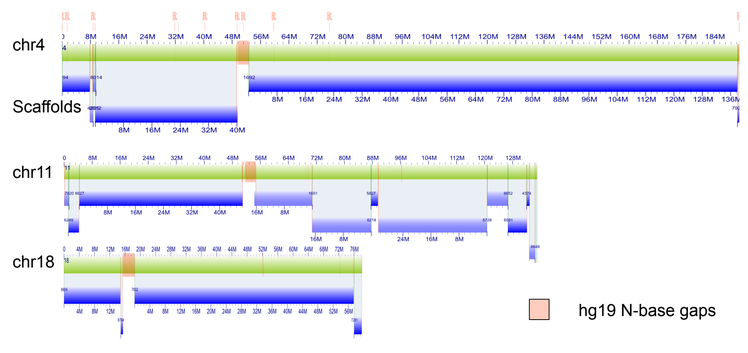
	Sequence	BioNano	Hybrid
#	4,899	2,613	228
Min size (kb)	0.84	98.28	228.53
N50 size (Mb)	**19.72**	**1.52**	**36.40**
Max size (Mb)	64.96	8.57	137.88
Total size (Gb)	2.83	2.83	2.84

Initial HuRef assembly was constructed using high-quality Sanger sequencing. Combining the sequence assembly with BioNano genome maps, we constructed hybrid scaffolds with unprecedented contiguity in length, with a N50 of 36.4 Mb **(Left)**. In fact, the lengths of some larger scaffolds nearly reach entire chromosome arms, spanning over numerous assembly gaps **(Right).**

Example of a BioNano-Sequence Merge Event



Near chromosome-arm length hybrid scaffolds



Conclusions

We present here the results from the analysis of a diploid human genome. Using our *de novo* assembly pipeline, we constructed a highly contiguous genome map assembly, covering a large portion of the human reference and extending into previously unalignable, unannotated loci. We also detected large structural variants that are invisible to short-read, reference-based detection approaches, and we were able to elucidate the location, orientation, and copy number of these events.

Moreover, our data shows that the correct pairing of technologies provides the potential for far richer results than can be achieved using each technology alone. With our Hybrid Scaffold pipeline, we integrated BioNano and sequence assembly data, and achieved scaffolds with unprecedented lengths. Such scaffolding information facilitates the correct anchoring of sequence assemblies, and the correct phasing of variation along chromosomes allowing researchers to gain a more complete view of their genomes. For more information about next-generation mapping, also see Posters #1832T, #3118T, #2721W and #2496F.

Reference

1 Cao, H., et al. Rapid Detection of Structural Variation in a Human Genome using NanoChannel based Genome Mapping Technology. Giga Science (2014); 3(December 2014): 34
2 Hastie, A.R., et al. Rapid Genome Mapping in NanoChannel Arrays for Highly Complete and Accurate *De Novo* Sequence Assembly of the Complex Aegilops tauschii Genome. PLoS ONE (2013); 8(2): e55864.
3 Lam, E.T., et al. Genome mapping on NanoChannel arrays for structural variation analysis and sequence assembly. Nature Biotechnology (2012); 10: 2303
4 Das, S. K., et al. Single molecule linear analysis of DNA in NanoChannel labeled with sequence specific fluorescent probes. Nucleic Acids Research (2010); 38: 8
5 Xiao, M., et. al. Rapid DNA mapping by fluorescent single molecule detection. Nucleic Acids Research (2007); 35:e16.
6 Levy, S., et al. The Diploid Genome Sequence of an Individual Human. PLOS Biology (2007); 5(10):e254.