Building high quality, chromosome-scale, de novo genome assemblies by scaffolding Next-Generation Sequencing assemblies with Bionano genome maps

A W C Pang, J Wang, E T Lam, B Clifford, S Bocklandt , S Oeser, T Anantharaman, A Hastie, H B Sadowski, M Oldakowski
Bionano Genomics, San Diego, California, United States of America

Abstract

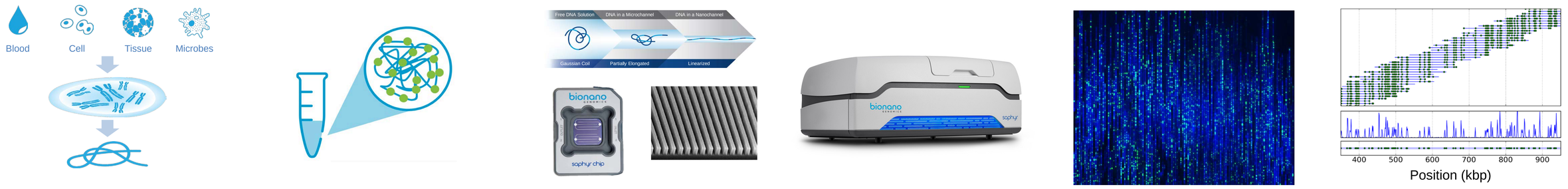
Except for a few model organisms, many biologically and economically important plants and animals still lack a reference-quality genome assembly that is crucial to the understanding of their biology. Their genomes are often complex and highly repetitive, making generation of high-quality assemblies almost impossible with next generation sequencing (NGS) alone and without access to long-range structural information. Bionano genome mapping provides a solution to reconstruct the full genomic architecture of large and complex genomes. Here, we present a direct enzymatic labeling approach which maintains the integrity of the DNA and enables us to create very contiguous Bionano maps which can then be used to scaffold NGS sequence assemblies to produce highly contiguous and structurally accurate hybrid assemblies that can span most repeat regions. This direct labeling method is compatible with a vast array of organisms. We validated our approach with the human NA12878 genome. Starting with NGS assemblies with N50 ranging from 0.18 to 0.9 Mbp, we produced hybrid assemblies with N50 from 70 to 80 Mbp. Chromosome-arm length scaffolds were assembled in 20 chromosomes, and alignments show that they were consistent with the hg19 reference. The hybrid assemblies incorporated 80-90% of total NGS sequences with over 99% scaffolding accuracy. We will also show equally impressive scaffolds for a variety of plants and animals. For a low cost and only several days from sample-to-scaffold, this new method promises to set a new standard for making high-quality genome assemblies.

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 Saphyr™ 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 entire diploid genome. The resulting provides the ability to correctly position and orient sequence contigs into chromosome-scale scaffolds and detect a large range of homozygous and heterozygous structural variation with very high efficiency.

Methods

1. Extraction of long DNA molecules
2. Label DNA at specific sequence motifs
3. Saphyr Chip linearizes DNA in NanoChannel arrays
4. Saphyr automates imaging of single molecules in NanoChannel arrays
5. Molecules and labels detected in images
6. Bionano Access software assembles optical maps

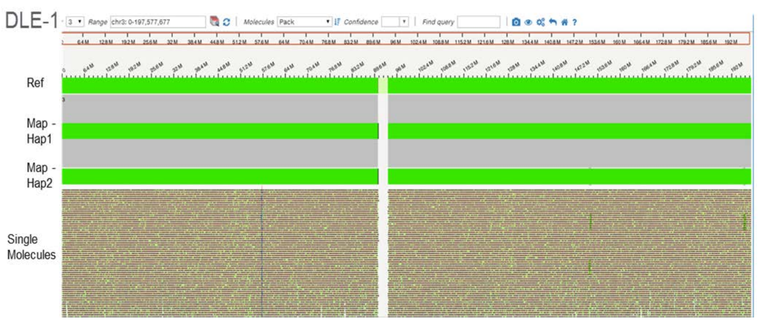


(1) Long molecules of DNA are labeled with Bionano reagents by (2) incorporation of fluorophores at a specific sequence motif throughout the genome. (3) The labeled genomic DNA is then linearized in the Saphyr Chip using NanoChannel arrays (4) Single molecules are imaged by Saphyr and then digitized. (5) Molecules are uniquely identifiable by distinct distribution of sequence motif labels (6) and then assembled by pairwise alignment into *de novo* genome maps.

DLS (Direct Label and Stain) Labeling Chemistry

- DLE-1 is one of Bionano DLS enzymes
- Highly specific, tag a fluorence label at particular recognition motifs
- Single enzymatic reaction; no nicking; no repair step
- Labeled molecule average length used for assembly is ~250 kbp in length
- Assembled genome maps can reach chromosome arm-length (human)

Full chromosome arms of human chr3 was assembled

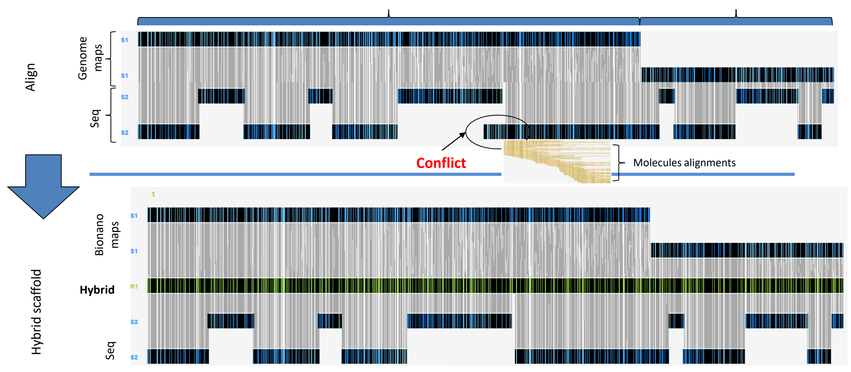


DLE-1 site density suitable for a wide range of genomes

Genome	Genome size (Mbp)	Site density (/100 kbp)
Frog	6016	11.9
Shark	4452	14
Plasmodium	23	14.9
Grass	243	15.1
Duckweed	143	15.2
Salamander	29038	16.5
Fish (fCotGob)	643	16.8
Fly	284	17.6
Chickpea	532	18.6
Bat	2126	18.6
Canola	850	19.5
Zebrafinch	1139	19.7
Goat	2924	19.8
Human	3137	20.1
Blackcap bird	1032	20.7
Brassica	579	20.9
Rabbit	2964	21
Orangutan	3043	21
Sorghum	727	21.5
Deer	2484	21.6
Cat	2670	21.7
Mouse	2804	22
Rat	2870	22.8
Mosquito	1268	22
Barley	4834	22.9
Clouded leopard	2437	22.9
Tobacco	4713	23
Pit viper	1418	24
Eucalyptus	691	24.3
Hummingbird	1057	24.3
Kakapo (falcon)	1060	25.6
Corn	2106	25.7
Sugarbeet	563	25.8
Soy	979	26.2

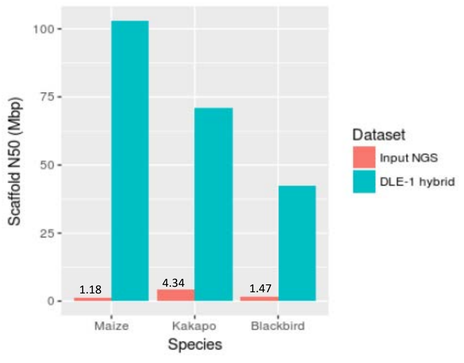
Hybrid-scaffolding NGS sequence assembly with Bionano maps:

1. Sequence contigs are converted into *in-silico* maps and aligned to Bionano maps
2. Assembly errors in sequence assembly were detected and corrected
3. NGS contigs are ordered and oriented into ultra long super-scaffolds

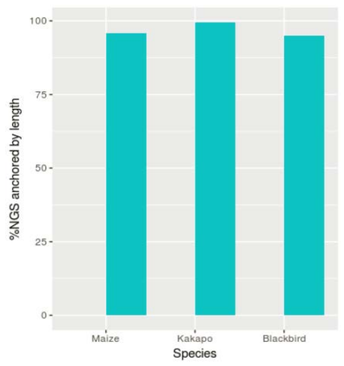


Hybrid-scaffold of plant and animal genomes

16-100x improvement in contiguity over input NGS

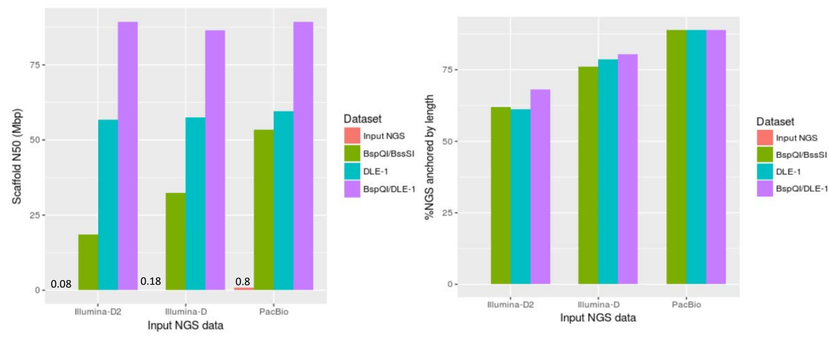


>95% of NGS contigs by total length incorporated in final scaffold

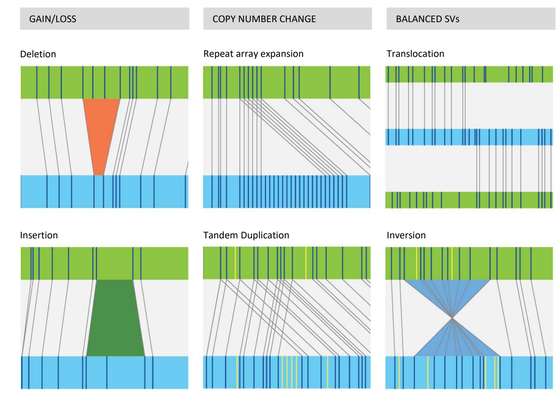


Key features of Bionano hybrid-scaffolds

- Contiguous: scaffold N50 over **80 Mbp**, up to **700x** improvement over input NGS
- Complete: up to **88-99%** of NGS contigs incorporated in hybrid scaffold
- Cost-effective: Sample to genome scaffolding in as little as 5 days **< 1000 dollars** in cost
- Compatible with many species: DLE-1 maps successfully generated for **> 15 different species**
- Compatible with different sequencing technologies



Bionano maps can detect structural variation across different genomes (See poster 534C by Jill Chiyu Lai)



Conclusions

Bionano Genomics's genome mapping solution provide an accurate and direct view of the global architecture of genome sequences. Integrating mapping data with NGS sequence data present both a global, top-down view along with single-nucleotide level details of the genome. The scaffolds generated with this data have set a new standard for genome assembly that can be accomplished in less than one week and for <1000 dollars.

Reference

1) Cao, H., et al., Rapid detection of structural variation in a human genome using NanoChannel-based genome mapping technology. Gigascience (2014); 3(1):34
2) Lam, E.T., et al. Genome mapping on NanoChannel arrays for structural variation analysis and sequence assembly. Nature Biotechnology (2012); 10: 2303
3) Pendleton, M., Sebra, R., et al. Assembly and diploid architecture of an individual human genome via single-molecule technologies. Nature Methods (2015); e3454
4) Huddleston J, C. M.-L. (2016, Nov 28). Discovery and genotyping of structural variation from long-read haploid genome sequence data. Genome Res, gr.214007.116.