Building High-quality, *de novo* Genome Assemblies by Scaffolding Next-Generation Sequencing with BioNano's Next-Generation Mapping

J Wang, A W Pang, E T Lam, W Andrews, T Anantharaman, A Hastie, M Stedman, H Sadowski, M Saghbini, Z Y Zhu, Ž. Džakula , M Austin , M Borodkin, H Cao

BioNano Genomics, San Diego, California, United States of America

Abstract

Combining next-generation sequencing (NGS) and next-generation mapping (NGM) information from BioNano Genomics' Irys System provides a solution that is being adopted to produce affordable, high-quality and chromosome-scale *de novo* genome assemblies.

We describe a novel workflow that utilizes two nicking endonucleases to increase the sequence motif barcoding pattern information density and improve contiguity through tiling of genome maps. We generated two sets of BioNano maps, each with a different nicking enzyme and developed novel algorithms that use the NGS sequences as a bridge to merge single-enzyme BioNano maps into combined maps that contain the sequence motif patterns from both nicking enzymes. Since the BioNano maps were generated independently they serve as orthogonal sources of evidences to detect and correct assembly errors in input data. The complementarity of different data also greatly improves the contiguity of the merged BioNano map while doubling the information density, which substantially increase the ability to anchor short NGS sequences in the final scaffolds.

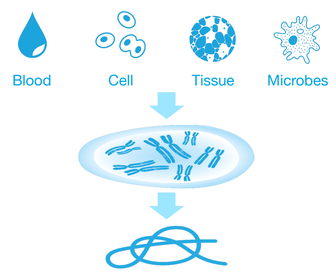
We first validated our approach in the well-studied human NA12878 genome. Compared to the published single-enzyme hybrid-scaffolds, the two-enzyme approach improved the scaffold contiguity by 300% and anchored 30% more sequence contigs in the final scaffolds and corrected 50% more assembly errors in NGS sequences. We then benchmarked our pipeline in both animal and plant genomes and showed that it performed robustly across those. This new approach can greatly expands the type of NGS data that can be integrated with BioNano maps to produce highly accurate and contiguous assemblies for complex genomes.

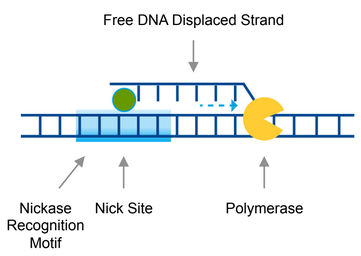
Background

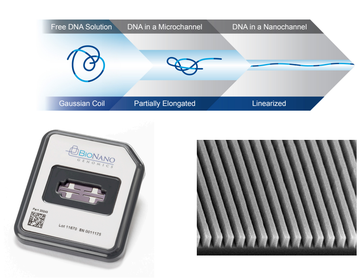
Generating high-quality finished genomes replete with accurate identification of structural variation and high completion (minimal gaps) remains challenging using short read sequencing technologies alone. The Irys system provides direct visualization of long DNA molecules in their native state, preserving the most trustworthy *de novo* long range genomic structural information instead of the statistical inference often needed in other indirect approaches. 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 detecting structural variation.

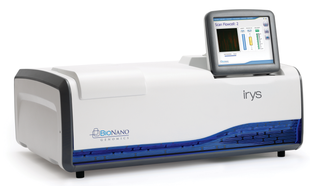
Methods

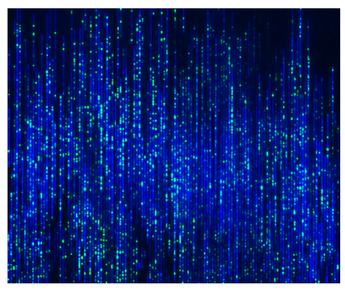
1. IrysPrep kit extraction of long DNA molecules
2. IrysPrep reagents label DNA at specific sequence motifs
3. IrysChip linearizes DNA in NanoChannel arrays
4. Irys automates imaging of single molecules in NanoChannel arrays
5. Molecules and labels detected in images by instrument software
6. IrysView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

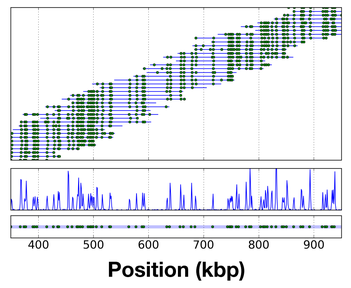








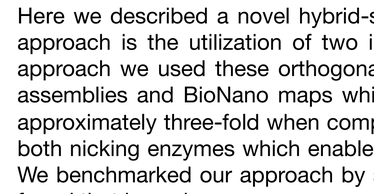




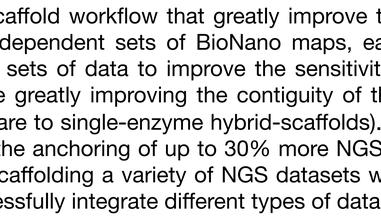
(1) Long molecules of DNA are labeled with IrysPrep® reagents by (2) incorporation of fluorophore labeled nucleotides at a specific sequence motif throughout the genome. (3) The labeled genomic DNA is then linearized in the IrysChip® using NanoChannel arrays and single molecules are imaged by Irys. (4)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

Overview of Two-Enzyme Hybrid-Scaffold

1. Combined conflict resolutions:



Conflicts not detected by BSPQI maps can be detected by BSSSI maps and vice, versa. The two independent sets of BNG maps also allow conflict resolution by consensus.



Inconsistency of NGS alignment between BSPQI and BSSSI maps allow the detection and correction of errors within BNG maps

2. Scaffolding BNG maps and NGS contigs

3. Anchoring NGS contigs

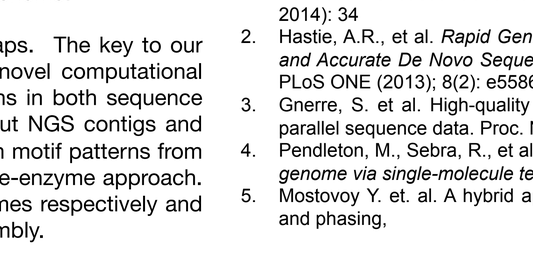
NA12878 benchmark datasets

We first benchmarked our pipeline by scaffolding a set of BSPQI and BSSI BNG maps with different NGS datasets from a well-studied human genome (NA12878).	**N50 (Mbp)**	**# of contigs**	**Total length (Mbp)**
BNG BSPQI genome maps	3.768	1306	2905.348
BNG BSSSI genome maps	1.865	2438	2896.648
Illumina-D ➢ Illumina short-read, 51.22x of 250bp pair-ends, ➢ assembled with DISCOVAR *de novo*	**0.179**	884719	3068.614
Illumina-S ➢ Illumina short-read, 39.8x of 101bp pair-end ➢ 24.7x of 2.5k-3.5k mate-pair sequencing ➢ assembled and scaffold with SOAP-*de novo*	**0.554**	1045521	2936.589
PacBio ➢ 46x of pac-bio long-read, mean read length 3.6kbp	**0.9**	24126	3065.158

1. *Improved sensitivity of conflict resolution*

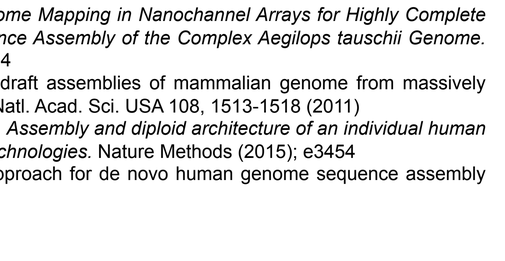
We evaluated sensitivity of conflict resolution by simulating chimeric joins (two distal genomic regions that are fused together) in NGS contigs or BNG maps and check if they can be detected in hybrid-scaffolds.

Conflict resolution in real-data: hybrid-scaffold substantially reduce the number of chimeric contigs in final scaffolds (as detected by local alignments to multiple distal regions in hg19)



2. *Improved scaffold contiguity*

3. *Higher information density improves the anchoring of short NGS contigs to scaffolds*



NA12878 Datasets (see table 1)	Assemblies/Scaffolds	# of NGS anchored	Total length of NGS in scaffolds (Mbp)
Illumina-D	Single-enzyme hybrid	8477	2081.313
	Sequential	9120	2165.69
	Two-enzyme hybrid	**12223**	**2340.351**
Illumina-S	Single-enzyme hybrid	5498	2557.519
	Sequential	5614	2584.65
	Two-enzyme hybrid	**6181**	**2619.42**
PacBio	Single-enzyme hybrid	3925	2665.313
	Sequential	3995	2682.4
	Two-enzyme hybrid	**4387**	**2703.96**

4. *Scaffolding plants and animal genomes*

Sample	NGS data	Scaffold pipeline	BNG Total Size (Mbp)	NGS Total Size (Mbp)	BNG N50 (Mbp)	NGS N50 (Mbp)	Total Scaffold size (Mbp)	Scaffold N50 (Mbp)	# of scaffolds	# NGS in scaffolds	Total Length (Mbp)
Sugar beet	30x PacBio	Single-enzyme	617 .0	350.440	1.95	0.135	501.787	2.376	275	744	136.776
		Two-enzyme	617.0, 588.246*	350.440	1.95, 1.21*	0.135	476.84 (245.71)**	3.74 (2.49)***	160 (100)**	1586	228.336
	40x PacBio	Single-enzyme	617.0	495.899	1.95	0.370	567.876	2.725	277	1209	398.207
		Two-enzyme	617.0, 588.246	495.899	1.95, 1.21	0.370	576.535 (40.602)	6.91 (5.79)	131 (63)	1688	452.762
	60x PacBio	Single-enzyme	617.0	538.580	1.95	0.742	570.097	3.970	195	904	491.012
		Two-enzyme	617.0, 588.246	538.580	1.95, 1.21	0.742	581.835 (15.725)	10.67 (10.03)	98 (24)	1071	511.324
	80x PacBio	Single-enzyme	617.0	562.760	1.95	1.400	576.11	7.295	134	606	526.456
		Two-enzyme	617.0, 588.246	562.760	1.95, 1.21	1.400	582.8534 (12.58)	14.56 (14.5)	63 (18)	697	538.753
Humming Bird	PacBio	Single-enzyme	1045.797	1105.676	0.723	4.073	1052.383	8.272	248	488	1023.477
		Two-enzyme	1045.797, 822.674	1105.676	0.723, 1.118	4.073	1038.564 (27.949)	15.12 (14.78)	143 (43)	514	1028.573

*two sets of BNG maps were used as input, thus two statistics for BNG maps
**number in parenthesis indicate statistics for unscaffolded single-enzyme scaffolds in two-enzyme scaffolds
*** number in parenthesis is the overall N50 that include both two-enzyme scaffolds and left-over single-enzyme scaffolds

Conclusions

BioNano Genomics' next-generation mapping (NGM) solution provides an accurate and direct view of the global architecture of genome sequences. Integrating NGM data with NGS sequence data present both a global, top-down view along with single-nucleotide level details of the genome.

Here we described a novel hybrid-scaffold workflow that greatly improve the integration of NGS sequence data with BioNano maps. The key to our approach is the utilization of two independent sets of BioNano maps, each from a different nicking endo-nuclease. With a novel computational approach we used these orthogonal sets of data to improve the sensitivity of automatic detection and resolution of chimeric joins in both sequence assemblies and BioNano maps while greatly improving the contiguity of the final scaffold (up to 100-fold when compared to input NGS contigs and approximately three-fold when compare to single-enzyme hybrid-scaffolds). The final scaffold generates genome maps that contain motif patterns from both nicking enzymes which enable the anchoring of up to 30% more NGS contigs into the final scaffolds when compared to single-enzyme approach. We benchmarked our approach by scaffolding a variety of NGS datasets with BioNano maps from human, animal and plant genomes respectively and found that in each case we can successfully integrate different types of data to produce a contiguous and high-quality genome assembly.

See also Other Bionano Posters: P0712, P0961, P0958 and P0033.

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. Gnerre, S. et al. High-quality draft assemblies of mammalian genome from massively parallel sequence data. Proc. Natl. Acad. Sci. USA 108, 1513-1518 (2011)
4. Pendleton, M., Sebra, R., et al. *Assembly and diploid architecture of an individual human genome via single-molecule technologies.* Nature Methods (2015); e3454
5. Mostovoy Y. et. al. A hybrid approach for de novo human genome sequence assembly and phasing,