Building Complete and Accurate Genome Assemblies using BioNano Next-Generation Mapping Technology

A W C Pang[1], T Anantharaman[1], P Sheth[1], X Zhou[1], J Wang[1], J Lee[1], A Hastie[1], Ž Džakula[1], H Cao[1]

[1]BioNano Genomics, San Diego, CA, USA

Abstract

High-quality assemblies are important when trying to understand the biology of genomes. Current short-read assemblers are memory intensive and have difficulties constructing contiguous assemblies; collecting deep coverage data by long-read technologies can be time-consuming and expensive. BioNano's next-generation mapping data complements short-read data and alleviates the need to collect high-coverage long-read data.

BioNano Genomics Irys® System utilizes 150 kilobase pair to megabase pair DNA to construct ultra-long genome maps. These assemblies reveal large structural variants and combined with sequencing assemblies produce hybrid scaffolds of unprecedented lengths, some spanning chromosomal arms. Chimeric joins are formed when short reads, molecules, or paired-end inserts are unable to span across long DNA repeats. These errors appear as conflicting junctions when one compares the sequence and the BioNano assemblies.

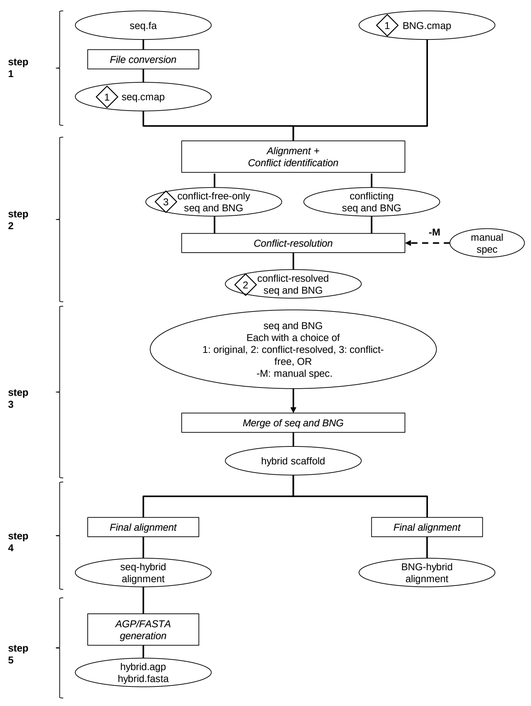
We developed a method to detect and resolve these conflicting junctions. At a conflicting region, where the BioNano genome map does not have long BioNano molecule support, the map is split into halves, thus resolving the conflicting junctions. If the BioNano genome map does have BioNano molecule support, the sequence fragment is split. Using this approach on a karyotypically normal human genome, we resolved 12 (85.7%) erroneous translocation calls from the sequence assembly and one (11.1%) erroneous translocation call from the BioNano assembly. Then, by combining the two refined assemblies, we built long hybrid scaffolds achieving an N50 of 21.3 Mbp (which is over 2.2 times the input sequence assembly N50).

Our plan is to use this novel hybrid scaffold functionality for constructing correct and contiguous reference assemblies for complex plants and animal genomes.

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, 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 genome maps spanning the whole genome. The resulting order and distance between the labels in the genome map can be used for anchoring sequencing assemblies and structural variation detection.

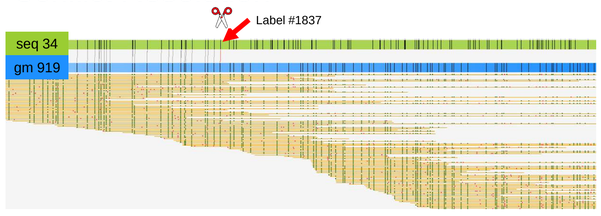
Methods

Hybrid Scaffold Pipeline Overview

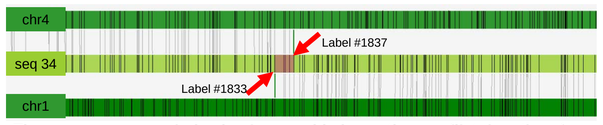


Workflow of the IrysSolve® Hybrid Scaffold Software. Oval boxes represent entities, such as assembly; rectangular boxes represent scripts or process; diamond boxes represent specific assemblies that are used as input to the merge step.

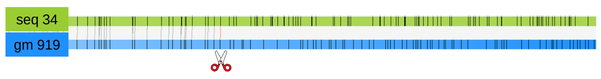
Conflict Resolution



An example from a karyotypically normal human sample. Conflict between an *in silico* digested sequence and BioNano assemblies at a location where their alignment terminated prematurely (arrow; the label number on the sequence contig is also shown). Note that the BioNano genome map configuration had support from numerous long molecules spanning both sides of the conflict junction, suggesting the sequence contig was misassembled. Here, the sequence contig was automatically cut by the IrysSolve Hybrid Scaffold Software at Label #1837.



The sequence was indeed misassembled, as the conflict junction from above was the result of a chimeric join between chr4 and chr1.



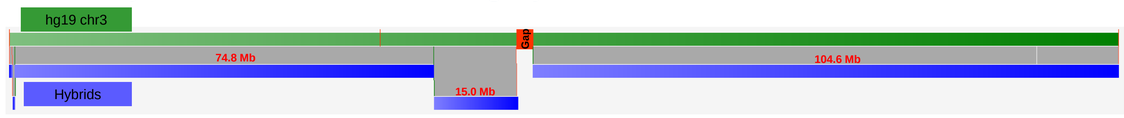
All automatic conflict resolution decisions can be manually altered. Continuing from the top example, users can alternately cut the genome map. To manually perform conflict resolution, users can edit the status of the conflict in the specification text file that has been generated by the IrysSolve Hybrid Scaffold Software. This manually edited specification file (shown below) can then be inputted into the IrysSolve Hybrid Scaffold Alignment Software again, and the conflicts handled accordingly. A new set of hybrid scaffolds would then be created.

Example of A Manually Edited Specification File of Conflicts

xMapId	refQry	Ref Id	Left Bkpt	Right Bkpt	Orientation	LeftBkpt_toCut	RightBkpt_toCut	ref_toDiscard
1309	ref	34	-1	15,496,008	+	okay	~~cut~~ okay	okay

refQry	qryId	Left Bkpt	Right Bkpt	Orientation	LeftBkpt_toCut	RightBkpt_toCut	qry_toDiscard
qry	919	-1	333,058	+	okay	okay ~~cut~~	okay

Assembly Correction and Ultra Long Hybrid Scaffolds



We applied this novel conflict-resolution pipeline, the IrysSolve Hybrid Scaffold Software, to scaffold a karyotypically normal human individual, and we identified and corrected 12 (85.7%) translocation calls in the sequence assembly and one (11.1%) translocation call in the BioNano assembly. The resulting hybrid scaffold reached an N50 length of 21.3 Mbp, which is 2.2 times the initial sequence assembly. Some long hybrid scaffolds spanning chromosome arms are shown.

IrysSolve Hybrid Scaffold's Assembly Statistics

Sample	Sequence method	Input BNG N50 (Mb)	Input Seq N50 (Mb)	Hybrid N50 (Mb)	Hybrid plus not-scaffolded Seq N50 (Mb)	% Seq in Hybrid	N50 fold increase
Human	Sanger	1.52	19.72	39.87	38.64	99%	2.02
Tomato	Illumina/Sanger/Clone	1.20	16.71	18.04	18.04	98%	1.08
Mammal	Illumina/Sanger/Clone/Optical map	1.59	14.44	42.85	42.16	97%	2.97
Human mole	PacBio	1.06	13.19	29.92	29.30	97%	2.27
Human (chromosome 20)	PacBio	1.13	10.98	29.25	29.25	99%	2.66
Human	PacBio	1.52	9.52	29.36	22.88	91%	3.08
Arabidopsis	PacBio	1.12	6.55	10.84	10.84	95%	1.65
Plant	Sanger	1.29	6.23	7.78	7.78	97%	1.25
Mammal	PacBio	1.59	4.68	22.33	21.98	97%	4.77
Human mole	PacBio	3.88	4.51	43.09	37.54	85%	9.55
Human mole	PacBio	1.33	4.51	23.08	19.36	85%	5.12
Human	PacBio	4.60	4.41	30.79	29.87	94%	6.98
Arabidopsis	PacBio	0.77	4.39	6.33	6.05	85%	1.44
Spider mite	Sanger	0.87	3.32	4.32	4.32	93%	1.30
Human	Illumina/Sanger/Clone	1.12	2.58	8.60	7.51	88%	3.33
Plant	Illumina/Sanger/Clone	1.12	2.02	4.20	3.48	86%	2.08
Plant	Illumina	1.12	1.17	2.42	2.19	86%	2.07
Plant	Illumina	0.90	1.09	2.92	1.41	50%	2.68
Human	PacBio	4.60	0.94	14.21	12.54	90%	15.12
Worm	Illumina	1.13	0.63	2.94	1.67	65%	4.67
Parasite	PacBio	1.05	0.63	1.61	1.61	95%	2.56
Parasite	PacBio	1.13	0.63	1.53	1.53	92%	2.43
Fish	Illumina	1.46	0.41	3.03	1.51	54%	7.39
Plant	Illumina	1.24	0.30	1.57	1.05	57%	5.23
Plant	Illumina	1.35	0.28	2.13	1.05	49%	7.61
Tomato	Illumina	0.89	0.22	1.25	0.61	41%	5.68
Plant	Illumina	1.42	0.19	1.76	0.41	21%	9.26
Mammal	Illumina	1.24	0.11	1.38	0.78	49%	12.55
Marine animal	Illumina	1.25	0.10	1.74	0.20	27%	17.40
Marine animal	Illumina	0.94	0.10	1.45	0.39	34%	14.50
Plant	Illumina	0.48	0.08	0.84	0.14	26%	10.50

The table shows IrysSolve Hybrid Scaffold's assembly statistics for various experiments where we integrated different sequencing platforms. The longer the input sequences (Input Seq), the better the integration (% Seq in Hybrid).

Conclusions

In addition to creating long hybrid scaffolds, our novel Hybrid Scaffold Software can detect and resolve chimeric joins in sequence and BioNano assemblies. It automatically detects these joins by examining alignments between the two assemblies, and it checks for chimeric scores, indicating the number of molecules supporting the genome map, on both sides of the junctions. In the case of a strong support, it would cut the sequence assembly, and alternately, cut the genome map. This auto-resolution approach has been applied to a human sample, and successfully removed erroneous chimeric joins. The resulting hybrid scaffolds reached an unprecedented contiguity. Moreover, all conflict-resolution decisions can be altered by users, enabling greater manual control in the scaffolding process.

We have applied the IrysSolve Hybrid Scaffold Software to a large number of genomes assembled by different platforms, and we observed that genome mapping and sequencing are highly compatible. In general, we observed that the scaffolding performance is better with longer sequence assemblies. Additionally, the final hybrid scaffolds always show improved contiguity. In conclusion, the IrysSolve Hybrid Scaffold Software is valuable in constructing accurate and long-range assemblies for complex plant and animal genomes. See also Posters: P0207, P0702, P0236, and P1272.

Reference

- Pendleton, M., Sebra, R., et al. Assembly and diploid architecture of an individual human genome via single-molecule technologies. Nature Methods (2015); e3454
- 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
- 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