Hybrid Scaffolding Improves Genome Assembly Accuracy and Contiguity

Bionano Genome Mapping Reveals True Long Range Structure of the Genome and Reduces Sequencing Costs

Generating high-quality finished genomes remains challenging. Accurate identification of structural variation with minimal gaps is difficult or impossible using short-read sequencing technologies alone.

The genomes of most higher organisms are highly repetitive. Two thirds of human and most mammalian genomes consist of repeats, and many plant genomes have even higher repeat content. Short reads usually fail to span long repeat arrays or disambiguate different copies of interspersed repeats that are not spanned. These failures can limit contig length and introduce chimeric joins and other assembly errors.

The widespread use of next-generation sequencing (NGS) has led to an accumulation of incomplete assemblies that contain large numbers of contigs and limited long-range information. NGS technology is based on fragmenting DNA molecules, reading just hundred(s) of basepairs, and using algorithms to reassemble these fragments.

The introduction of long-read sequencing has led to improved assembly contiguity and accuracy as well, but can be time-consuming and expensive, especially when deep coverage or the spanning of long tandem repeats is required. Read lengths are still limited to tens of kilobasepairs.

Recently, synthetic long-read technologies like that of 10x Genomics have gained momentum. Using a barcoding method to link short reads, some mid-range structural information is retained, thus improving the contiguity of NGS assemblies. Synthetic long reads are still plagued by some challenges inherent to NGS technology. These challenges include failure to disambiguate interspersed repeat units and correctly assemble and size long repeat arrays; assembly gaps due to incomplete coverage and GC bias in PCR amplification and sequencing; and lack of long range structural information caused by fragmenting of the DNA and reads that are too short to span and correctly resolve larger structural variation.

Only extremely long, megabase size molecules provide accurate structure of the genome. Bionano Genome Mapping visualizes long DNA molecules in their native state. Long range genomic structural information is preserved and directly observed instead of algorithmically inferred as in sequencing 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.

Megabase size molecules of genomic DNA are labeled, linearized and uniformly stretched in high density NanoChannel arrays, and imaged on Saphyr.™ A specific 6 or 7 basepair sequence is labeled. Traditionally, Bionano mapping used a nicking endonuclease to Nick the sequence motive, followed by Labeling, Repair, and Stain (NLRS). This process is highly robust and specific, but it introduced systematic double-stranded breaks that limited the contiguity of Bionano maps.

The recently introduced Direct Label and Stain method (DLS) does not nick the DNA, eliminating systematic molecule breaks. The DLS protocol consists of a single enzymatic labeling reaction, followed by cleaning and staining. There is no need to repair for a more streamlined protocol.

The label patterns generated by NLRS or DLS allow each long molecule to be uniquely identified and aligned. Using pairwise alignment of the single molecules, consensus genome maps are constructed, refined, extended and merged. DLS genome maps are 50-fold longer than NLRS maps on average, improving visualization of genome structure and creating the most contiguous and accurate assemblies. Chromosome arms and full chromosomes are often assembled in single maps. Genome maps can be created using different endonucleases to generate broader coverage and higher label density.

Sample	Molecule N50 > 150 Kbp (Kbp)	Bionano Map N50 (Mbp)
NA12878	293	55.9
Human Fresh Blood	307	56.9
Bionano Maize B73	260	100.0
Durum Wheat	364	13.0
Farro	300	32.7
Strawberry	241	13.3
Kakapo	247	69.3
Hummingbird	310	38.7
Blackbird	243	21.6
Fish	245	22.3
Ferret	262	66.1
Pig	335	65.2
Soybean	246	23.0
Brassica	270	12.4
Mouse	280	101.0

Table 1:
Organisms de novo assembled using Bionano direct labeling chemistry (DLS). De novo assemblies often cover whole chromosome arms, only broken at centromeres and other low complexity regions which are longer than molecule length.

Hybrid Scaffold Construction

The *de novo* Bionano genome maps can be integrated with a sequence assembly to order and orient sequence fragments, identify and correct potential chimeric joins in the sequence assembly, and estimate the gap size between adjacent sequences. In order to do so, the Bionano Solve software imports the assembly and identifies putative nick sites in the sequence based on the nicking endonuclease-specific recognition site.

These *in silico* maps for the sequence contigs are then aligned to the *de novo* Bionano genome maps. Conflicts between the two are identified and resolved, and hybrid scaffolds are generated in which sequence maps are used to bridge Bionano maps and vice versa. Finally, the sequence assembly corresponding to this hybrid scaffold is generated and exported as FASTA and AGP files.

The pipeline is fully integrated with Bionano Access™ which provides a convenient interface for running Hybrid Scaffold and viewing scaffolding results.

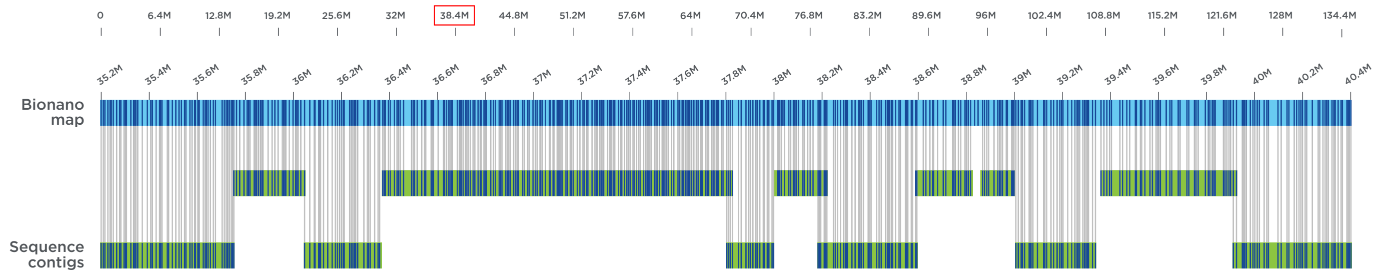

Figure 1:
Combining Bionano maps with sequence assemblies. Sequence contigs are anchored and oriented using the de novo generated Bionano maps.

Contiguity and Completeness

The hybrid scaffolding process reduces thousands of contigs found in the initial NGS assembly to a handful of scaffolds, improving assembly accuracy and quality while reducing the need for deep sequencing coverage.

The hybrid scaffolding approach can yield significant improvements in contiguity, as expressed by the assembly N50 values, across genomes as seen in Table 2. We created hybrid scaffolds for three genomes (maize, kakapo, and blackbird). This process improved contiguity by as much as 84-fold. The Bionano Solve pipeline makes near-complete use of the available input assemblies, taking into account 95-99.5% of the total length of the sequence (Table 2).

Bionano hybrid scaffolding is agnostic to the sequence technology used. Recent publications featured scaffolded assemblies using Illumina sequencing alone,[1] PacBio alone,[2] 10x Genomics assemblies,[3] nanopore sequencing[4] and combinations of those.[5]

NGS Dataset	NGS N50 (Mbp)	Total Scaffold Size (Mbp)	Scaffold N50 (Mbp)	# of NGS Anchored	Total NGS Anchored/Total NGS Length in Mbp (%)
Bionano Maize	1.185	2,120	100	2,809	99.5%
Kakapo	4.34	1,176	71	1,898	95.9%
Blackbird	1.47	1,018	42	977	95.0%

Table 2:
Contiguity of sequence assemblies of 3 organism is shown before (NGS N50) and after scaffolding (Scaffold N50) with Bionano maps built using DLS. Total assembly size and percentage of NGS sequence incorporated in the assembly is shown as well.

Assembly Conflicts and Resolution

The Bionano hybrid scaffold pipeline detects and resolves chimeric joins. Chimeric joins are typically formed when short reads, molecules, or paired-end inserts are unable to span across long DNA repeats. The errors appear as conflicting junctions in the alignment between the Bionano map and NGS assemblies.

When the hybrid scaffold pipeline detects a conflict, it analyzes the single-molecule data that underlies a Bionano map and assesses which assembly was incorrectly formed. If the Bionano map has long molecule support at the conflict junction, the sequence contig is automatically cut, removing the putative chimeric join (Figure 2). If it does not have strong molecule support, then the Bionano map is automatically cut. Both assemblies must have coverage spanning both sides of a chimeric join to detect and resolve these conflicts.

Users can manually inspect all conflict resolution results. Bionano Solve notes the IDs and coordinates of the sequences and maps where conflicts have been detected and the corresponding resolution approaches taken. The scaffold can be edited in Bionano Access™ and modified, and then run again in the hybrid scaffold pipeline to produce a new set of scaffolds based on the manual conflict resolution. This manual enhancement process can be performed multiple times, giving users fine control in generating high-quality, complete hybrid scaffolds.

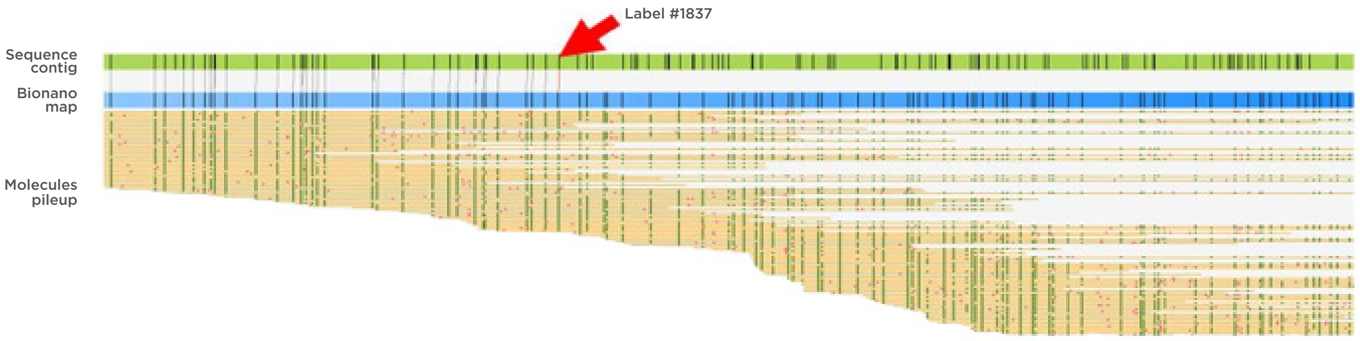

Figure 2:
Example of a conflict between a sequence contig and a Bionano map. (Top) The conflict junction as shown by the red arrow in the alignment between the sequence contig and the Bionano map. There is strong molecule support spanning the junction region on the genome map, so the sequence is cut at the label indicated.

Accuracy

A more accurate assembly doesn't just have better contiguity and fewer errors, but is more functional as well. Genes and their regulatory sequences need to be assembled, ordered and oriented correctly to allow for a meaningful functional analysis. Figure 3 illustrates a region containing a muscle skeletal receptor kinase (MuSK) gene in hummingbird—which may be of biological significance for the extreme flight skills of this bird. The *de novo* PacBio assembly failed at the dense repeats in the gene, leading to it being split between two sequence contigs and failing to measure repeat array copy number. Bionano hybrid scaffolding correctly brings the two pieces together to create one functional gene.

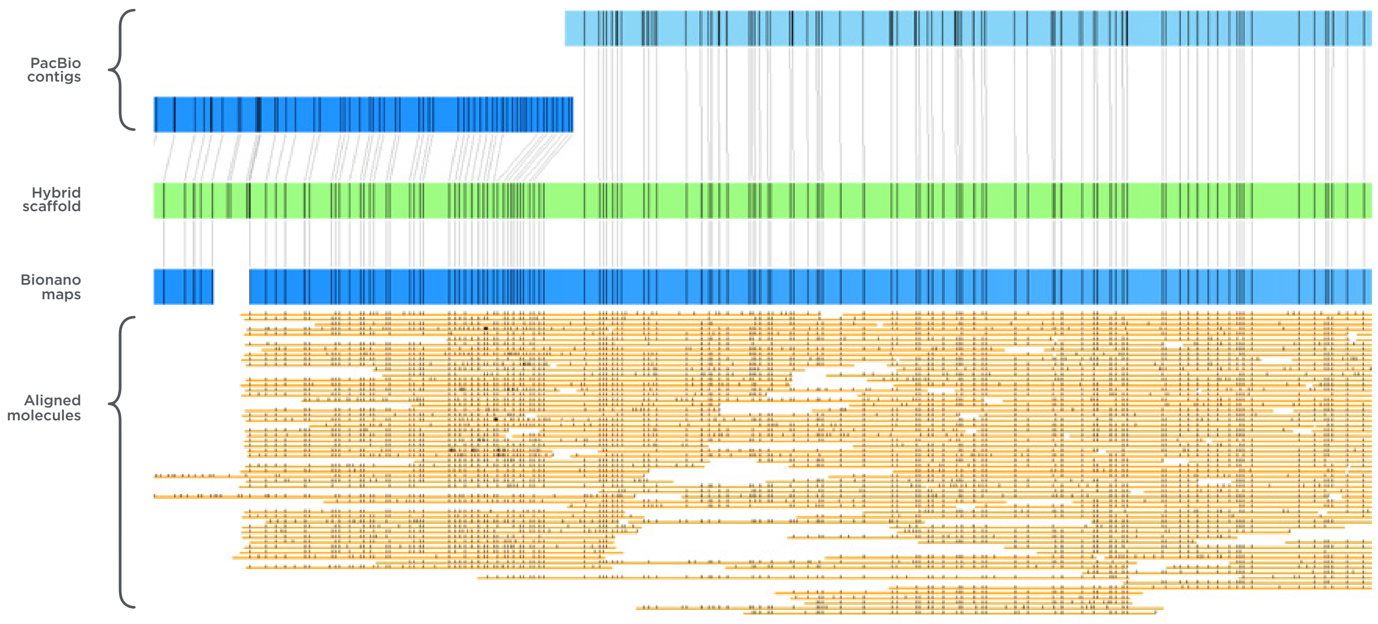

Figure 3:
The hummingbird Muscle Skeletal Receptor Kinase is split between two PacBio sequence contigs. Bionano hybrid scaffolding uses the genome map to anchor and orient the contigs, leading to a functional gene. Vertical lines represent Nt.BspQI recognition sites.

Higher Levels of Contiguity Using Two-Enzyme Hybrid Scaffolding

Assembly contiguity can be further increased by performing hybrid scaffolding with maps using two separate labeling enzymes. Two sets of Bionano maps, each generated with a different labeling enzyme, can be integrated with NGS sequences together. The two maps can both be generated using NLRS, or could be one DLS map combined with one NLRS map. This integration enables the NGS sequences to function as a bridge to merge single-enzyme Bionano maps into two-enzyme maps that contain the sequence motif patterns from both labeling enzymes. Since the Bionano maps are generated independently they serve as orthogonal sources of evidence 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 increases the ability to anchor short NGS sequences in the final scaffolds.

The two-enzyme approach was validated on the human NA12878 genome, a model data set for which sequence data is publicly available. Three different assemblies were tested: two Discovar assemblies of Illumina 250 bp pair-end sequence with different quality (Illumina-D2, N50: 0.08 Mbp, Illumina-D, N50: 0.18 Mbp); and PacBio, 46x with mean read length of 3.6 kbp. In each case, the assembly contiguity is greater with DLS alone than when combining two nicking endonuclease, but the two-enzyme approach combining DLS with NLRS maps improves the scaffold contiguity up to 1000-fold when compared to input NGS, anchors more sequence contigs in the final scaffolds and corrects more assembly errors in NGS sequences (Figure 4). The pipeline performs robustly in both animal and plant genomes as well. This approach greatly expands the type of NGS data that can be integrated with Bionano maps to produce highly accurate and contiguous assemblies for complex genomes.

The two-enzyme scaffolding method improves the error correction even further. Since the Bionano maps were generated independently they serve as orthogonal sources of evidence to detect and correct assembly errors in input data.

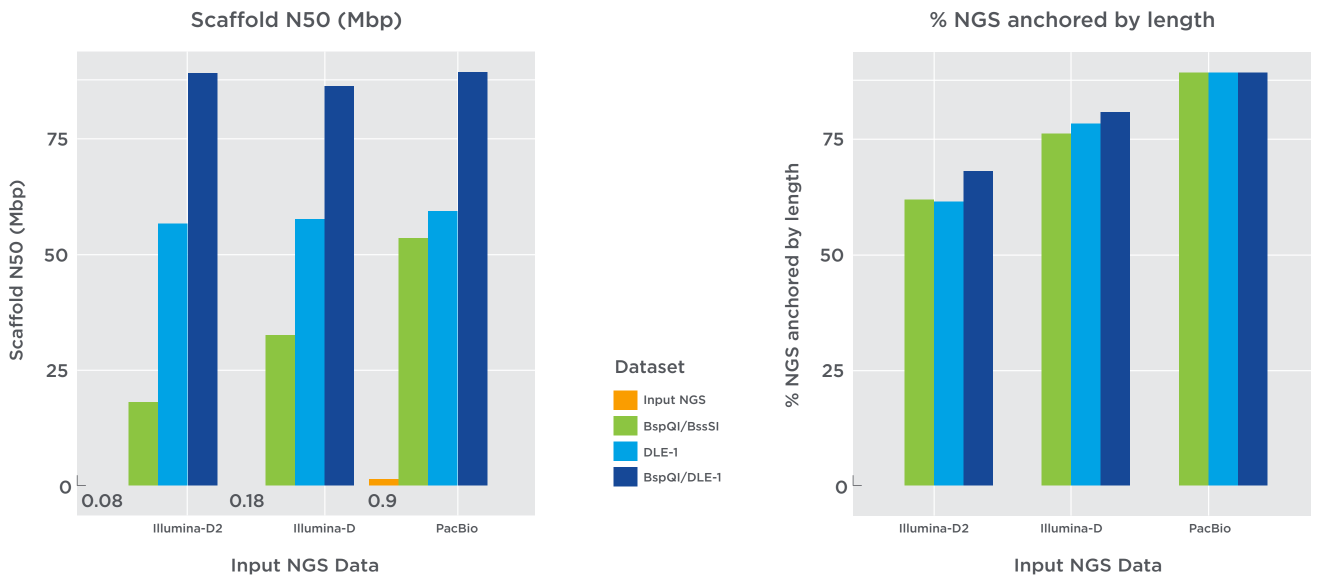

Figure 4:
Three sequence assemblies from the human NA12878 reference genome with various contiguity (Illumina-D2, Illumina- D, PacBio) were scaffolded with Bionano maps generated with nicking endonucleases (Nt.BspQI, Nb.BssSI) or with DLS (DLE-1).

Cost Considerations

Bionano hybrid scaffolding makes an assembly better for a low cost. Adding a Bionano genome map to your assembly costs as little as $500 in materials for up to human-size genomes, and remains affordable for larger genomes as well. This compares extremely favorable to PacBio sequencing, Dovetail or NRGene assemblies. No matter what your sequencing strategy is, adding Bionano to your assembly is a good decision. The significant improvements in contiguity and accuracy produce a better assembly and thus a superior publication at a reasonable cost.

Alternatively, the improvements in contiguity using Bionano hybrid scaffolding allow you to reduce the sequencing coverage necessary to produce an assembly of a certain quality. A 30x PacBio assembly scaffolded with Bionano maps typically produces an assembly of superior contiguity than 80x PacBio alone. Depending on the organism's genome size, a significant reduction in PacBio sequencing can reduce the cost by tens of thousands of dollars – far more than the cost to generate the Bionano data.

Discussion

Combining NGS and Bionano Genome Mapping data produces assemblies of the highest quality. This approach offers an affordable solution to improve fragmented draft assemblies and build the highest-quality assemblies containing accurate long-range information.

Bionano hybrid scaffolding is agnostic to the sequence technology used. Recent publications have scaffolded assemblies based on Illumina sequencing alone, PacBio alone, 10x Genomics assemblies, NRGene assemblies, nanopore sequencing, and combinations of those.

Bionano maps can error correct input sequence assemblies. Any of the scaffolding technologies using synthetic long reads or DNA cross-linking provide some sort of error correction compared to short-read assemblies alone. However, since they are NGS based, they suffer from most of the same problems plaguing short-read only assemblies. Only Bionano Genome Mapping provides non-sequencing based, orthogonal genome structure data in a high throughput way, allowing for a completely independent error correction.

Recent publications on reference genomes for wheat, banana, bed bug and maize[5,6,7,8] all included Bionano data to create higher contiguity and/or correct assembly errors. All major human reference genome publications used Bionano Genome Mapping data as well, including the NA12878 genome[2], the Chinese reference genome[9] and the Korean reference genome[10]. The contiguity of these recent genome publications combining *de novo* sequence assemblies with Bionano maps approach or surpass that of the hGCR38 reference genome (Table 3). **Including Bionano mapping data into *de novo* genome assemblies has become a *de facto* standard.**

Table 3: Assembly statistics for a number of human reference genomes (data from Genome in a Bottle Consortium[2,3,9,10]).

	AK1	HX1	NA12878	NA12878	NA24385	GRCh38
Sequencing	PacBio	PacBio	Illumina + 10x Genomics	PacBio	PacBio	Sanger
Scaffolding	Bionano	Bionano	Bionano	Bionano	Bionano two-enzyme	multiple
Input N50 (Mbp)	17.92	8.325	7.03	1.56	4.7	56.41
Hybrid Scaffold N50 (Mbp) scaffold	44.85	21.979	33.5	26.83	80.46	67.79
Fold Improvement after Bionano hybrid scaffold	2.5x	2.6x	4.8x	17.2x	17.1x	

bionano®
GENOMICS

References: 1.J. W. Clouse et al The Amaranth Genome: Genome, Transcriptome, and Physical Map Assembly The Plant Genome (2016) 2.Pendleton, M., Sebra, R., et al. Assembly and diploid architecture of an individual human genome via single-molecule technologies. Nature Methods (2015); e3454 3.Mostovoy J et al. A hybrid approach for de novo human genome sequence assembly and phasing Nature Methods (2016) 4.Deschamps S et al. A chromosome-scale assembly of the sorghum genome using nanopore sequencing and optical mapping. Biorxiv 2018 5.Zimin et al Hybrid assembly of the large and highly repetitive genome of Aegilops tauschii, a progenitor of bread wheat, with the mega-reads algorithm bioRxiv 2016 6.Martin et al. Improvement of the banana "Musa acuminata" reference sequence using NGS data and semi-automated bioinformatics methods BMC Genomics (2016) 7.Rosenfeld et al. Genome assembly and geospatial phylogenomics of the bed bug Cimex lectularius Nature Communications (2016) 8.Dong et al Analysis of tandem gene copies in maize chromosomal regions reconstructed from long sequence reads PNAS (2016) 9.Shi et al Long-read sequencing and de novo assembly of a Chinese genome Nature Communications (2016) 10.Seo JS et al de novo assembly and phasing of a Korean human genome Nature 2016

For general information about the Saphyr™ System, please contact info@bionanogenomics.com or visit bionanogenomics.com