Bionano Access®: Hybrid Scaffold Report Guidelines

Document Number: 30324

Document Revision: A

Table of Contents

Legal Notice

For Research Use Only. Not for use in diagnostic procedures.

Revision History

Revision	Notes
A	Initial Release

Introduction

The purpose of this document is to provide guidelines to customers who want to evaluate the quality of the hybrid scaffold using both Saphyr System and Bionano Solve pipeline (v3.3 or above). For comprehensive discussion of Hybrid Scaffolding principle, please refer to the Hybrid Scaffold Theory of Operations document.

Initially, *de novo* assembly of the molecules file (.bnx) into consensus genome maps (.cmap) is performed by the Bionano Solve pipeline (see Assembly Report Guidelines document for further information). The consensus Bionano genome maps (.cmap) are then compared to imported NGS data (.fasta, .fna) that have been converted into a reference file (.cmap). These two .cmap files are aligned. Bionano maps can be used to orient and scaffold NGS contigs and size gaps by bridging across complex elements.

Input NGS assemblies with contig/scaffold N50 of at least 150 kbp (with at least 9-10 labels in the contig/scaffold) are sufficient to produce high-quality hybrid scaffolds. It is possible to scaffold less contiguous NGS assemblies; however, the success depends on other factors such as label density across the genome and assembly accuracy. As to Bionano data, we recommend using as input a minimum of 60X effective molecule coverage in order to build an accurate and contiguous consensus genome map assembly. When using nickases, using more coverage does not significantly improve map contiguity. When using a DLS enzyme like DLE-1, effective coverage up to and beyond 100X has shown improved map contiguities for some plants and animals.

When there are discrepancies between the NGS data and the Bionano maps, the pipeline would either automatically cut the NGS contigs or Bionano cmaps for proper orientation, or flag them for manual review depending on selections that the user has made when launching the Hybrid Scaffold pipeline. We recommend choosing “Resolve Conflicts” for both Bionano and NGS assemblies.

When running the *de novo* assembly pipeline for hybrid scaffolding applications, we recommend using assembly parameters for non-haplotype-aware assembly. The current Hybrid Scaffold pipeline does not explicitly handle haplotype information and assumes there is only one genome map or NGS sequence contig covering a given genomic region. If multiple haplotypes are present, the pipeline may make false positive conflict cuts and incorrectly mix haplotypes in the final scaffolds.

Hybrid Scaffold Report

The results presented in the Hybrid Scaffold Report are divided into different sections corresponding to the scaffolding steps that occur in the pipeline process. Shown below is the Hybrid Scaffold report of a good sample. The values were generated using Bionano Access version 1.3, Solve version 3.3, (or Tools version 1.3). The sample was labeled with DLE-1. Please note the Hybrid Scaffold Report obtained with two-enzyme scaffolding will have a similar layout.

Original Bionano genome map statistics:

Count	: 29
Min length (Mbp)	: 0.652
Median length (Mbp)	: 78.975
Mean length (Mbp)	: 89.127
N50 length (Mbp)	: 166.519
Max length (Mbp)	: 323.539
Total length (Mbp)	: 2584.679

Original NGS sequence statistics:

Count	: 11453
Min length (Mbp)	: 0.000
Median length (Mbp)	: 0.002
Mean length (Mbp)	: 0.208
N50 length (Mbp)	: 1.769
Max length (Mbp)	: 12.266
Total length (Mbp)	: 2384.189

Conflict resolution from BNG-NGS alignment:

Number of conflict cuts made to Bionano maps	: 2
Number of conflict cuts made to NGS sequences	: 2603
Number of Bionano maps to be cut	: 2
Number of NGS sequences to be cut	: 1330

NGS FASTA sequence in hybrid scaffold statistics:

Count	: 3528
Min length (Mbp)	: 0.032
Median length (Mbp)	: 0.243
Mean length (Mbp)	: 0.590
N50 length (Mbp)	: 1.393
Max length (Mbp)	: 12.266
Total length (Mbp)	: 2082.351

Hybrid Scaffold FASTA statistics:

Count	: 23
Min length (Mbp)	: 0.081
Median length (Mbp)	: 104.001
Mean length (Mbp)	: 108.960
N50 length (Mbp)	: 162.374
Max length (Mbp)	: 317.250
Total length (Mbp)	: 2506.090

Hybrid Scaffold FASTA plus not scaffolded NGS FASTA statistics:

Count	: 10544
Min length (Mbp)	: 0.000
Median length (Mbp)	: 0.002
Mean length (Mbp)	: 0.266
N50 length (Mbp)	: 160.379
Max length (Mbp)	: 317.250
Total length (Mbp)	: 2807.755

1. Input Data Statistics

This section describes the input statistics of the hybrid scaffold. The original Bionano Genome Map Statistics are the summary statistics of the genome maps produced during the *de novo* assembly. NGS Sequences Statistics are the summary statistics of the input NGS data.

Original Statistics Definitions:

Count – The number of genome maps/contigs in your final assembly.
Min Length (Mbp) – The length of the shortest genome map or contig.
Median Length (Mbp) – The length of the central most/midpoint genome map or contig.
Mean Length (Mbp) – The average length of the genome map or contig.
N50 Length (Mbp) – The length of the genome map or contig at the point of half mass of distribution.
Max Length (Mbp) – The length of the longest genome map or contig.
Total Length (Mbp) – The sum of the lengths of all genome maps or contigs.

2. Results of Conflict Resolution

This section includes important statistics related to conflicts between the NGS data and Bionano data. When sequence and Bionano mapping data diverge, one entity is cut to align to the other. If “Resolve Conflicts” is chosen, this section will show in the Hybrid Scaffold Report. If either “Keep Conflicts” or “Remove Conflicts” is selected, this section would not be shown.

Conflict Resolution from BNG-NGS Definitions:

Number of Conflict Cuts Made to Bionano Maps – When there are discrepancies between NGS and Bionano data, the number of individual cuts that have been made to Bionano maps.
Number of Conflict Cuts Made to NGS Sequences – When there are discrepancies between NGS and Bionano data, the number of individual cuts that have been made to the NGS data.
Number of Bionano Maps to be Cut – How many Bionano maps will be cut.
Number of NGS Sequences to be Cut – How many NGS contigs will be cut.

3. NGS Data Incorporation Statistics

The amount and metrics of incorporated NGS sequence data in the scaffold.

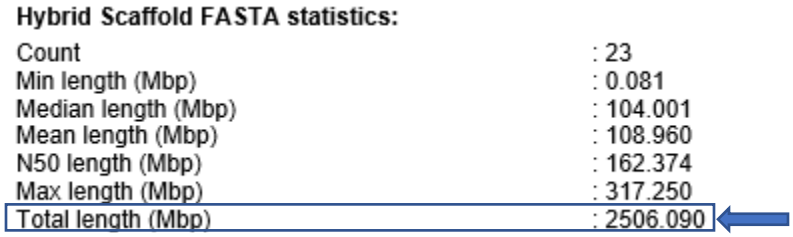
4. Hybrid Scaffold Statistics

This section summarizes the final hybrid scaffold results and is split into two sections. The first section corresponds to just the consensus hybrid scaffold of all NGS data that can be anchored or scaffolded by the Bionano maps. Unanchored NGS data is available in the second section and contains all of the scaffolding data, but additionally it includes the unanchored NGS sequences that were either too short to scaffold, or did not align to the Bionano mapping data.

Hybrid Scaffold Report Evaluation

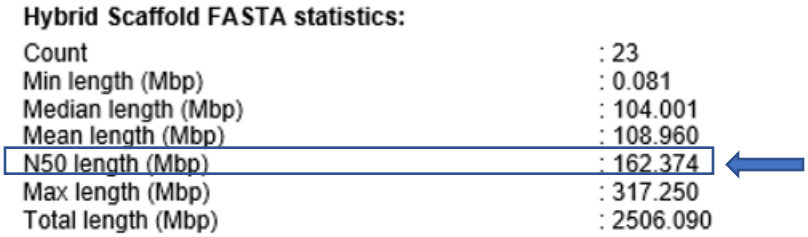
The quality of the hybrid scaffold can be evaluated in its completeness (total assembly length), contiguity (number of hybrid scaffolds and the N50 of the hybrid scaffold), and amount of sequence incorporation.

Completeness. Compare the total length of the hybrid scaffold to the expected genome size. In the following example, the total length of the hybrid scaffold is 2506.090 Mbp. The expected genome size of this particular organism is ~2500 Mbp. Therefore, the hybrid scaffold is considered complete in this example.



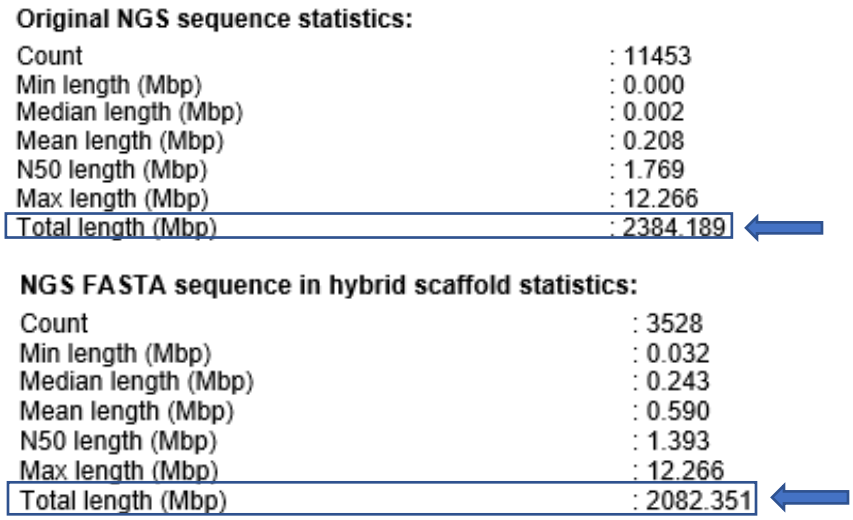
Hybrid Scaffold FASTA statistics:	
Count	: 23
Min length (Mbp)	: 0.081
Median length (Mbp)	: 104.001
Mean length (Mbp)	: 108.960
N50 length (Mbp)	: 162.374
Max length (Mbp)	: 317.250
Total length (Mbp)	: 2506.090

Contiguity. Examine the N50 of the hybrid scaffold. In the example shown below, the N50 length of the hybrid scaffold is 162.374 Mbp. The longer the hybrid scaffold N50, the more contiguous is the scaffold. Please keep in mind that ideally the scaffold counts should be close to the number of chromosome arms. Furthermore, the total length in the Hybrid Scaffold FASTA statistics should be close to the genome size of the organism. In this example, the number of scaffold counts is 23, which is close to the number of chromosome arms of this organism (18). The total length of the hybrid scaffold is 2506.090 Mbp, which is close to the genome size of this organism, 2500 Mbp.



Hybrid Scaffold FASTA statistics:	
Count	: 23
Min length (Mbp)	: 0.081
Median length (Mbp)	: 104.001
Mean length (Mbp)	: 108.960
N50 length (Mbp)	: 162.374
Max length (Mbp)	: 317.250
Total length (Mbp)	: 2506.090

Amount of sequence incorporation. Determine how much of the original NGS sequence was incorporated in the hybrid scaffold. Ideally, the hybrid scaffold incorporates the majority of input sequences, which is the total length value in the Original NGS sequence section in the Hybrid Scaffold statistics report. In the example shown below, out of a total 2384.189 Mbp original NGS sequence, 2082.351 Mbp was incorporated in the hybrid scaffold, which represents 87.3% or a majority of input sequences.



Original NGS sequence statistics:

Count	: 11453
Min length (Mbp)	: 0.000
Median length (Mbp)	: 0.002
Mean length (Mbp)	: 0.208
N50 length (Mbp)	: 1.769
Max length (Mbp)	: 12.266
Total length (Mbp)	: 2384.189

NGS FASTA sequence in hybrid scaffold statistics:

Count	: 3528
Min length (Mbp)	: 0.032
Median length (Mbp)	: 0.243
Mean length (Mbp)	: 0.590
N50 length (Mbp)	: 1.393
Max length (Mbp)	: 12.266
Total length (Mbp)	: 2082.351

Important Notes

When running the Hybrid Scaffold pipeline, we suggest that the input data (both NGS and Bionano) come from the same sample. If this is not possible, the input data should come from the same species. The quality of the hybrid scaffold, as evaluated in its completeness, contiguity, and amount of sequence incorporation, depends on the quality of the input data, which includes both Bionano data as well as NGS data. In turn, the quality of input data can be evaluated based on their respective completeness, contiguity, and accuracy. There are only a few cases when data from a hybrid scaffold are not optimal:

1. Poor Bionano assembly data (i.e., lower fraction of input molecules aligned to the assembly as evaluated in the assembly report document)
2. Poor NGS data (i.e., the contigs have low N50 or the NGS total length is not similar to the expected genome size)
3. High heterozygosity of the organism
4. High discordance between the individual being mapped and the individual that was used to obtain the NGS data.

Appendix

In this section, we show how results from a Hybrid Scaffold Report may be summarized and presented in a table format. Further, we show two examples of suboptimal hybrid scaffold results. In the first example, we see that out of a total of 2181.231 Mbp original NGS sequence, only 1170.938 Mbp (or 53.7%) was incorporated into the hybrid scaffold. We see that the total length, 3374.518 Mbp, is larger than the lengths of the input assemblies. As a result, a large portion of the sequences and maps are incompatible.

	Original BNG	Original NGS	NGS used in hybrid	Hybrid	Hybrid + not scaffolded NGS
Number of maps	225	4471	2457	140	4613
N50 (Mbp)	55.341	1.091	0.712	56.866	35.454
Total Length (Mbp)	2663.205	2181.231	1170.938 (53.7%)	2364.225	3374.518

To further troubleshoot this particular hybrid scaffold, we recommend the following:

1. Evaluate the Bionano data by referring to the Bionano Access Assembly Report.
2. Determine whether there is discordance between the organism being mapped and the organism used to obtain the NGS data.

In the second example below, we see that out of a total of 5067.549 Mbp original NGS sequence, only 1724.610 Mbp (or 34.0%) was incorporated into the hybrid scaffold. Of the original 151816 NGS contigs, only 15283 contigs were used in the hybrid scaffold (141335 NGS contigs were not scaffolded at all). Furthermore, we observe that the Original NGS contig has an N50 length of only 0.059 Mbp.

	Original BNG	Original NGS	NGS used in hybrid	Hybrid	Hybrid + not scaffolded NGS
Number of maps	7834	151816	15283	4075	141335
N50 (Mbp)	1.908	0.059	0.121	2.000	0.741
Total Length (Mbp)	11646.940	5067.549	1724.610 (34.0%)	4660.640	8003.524

To further troubleshoot this hybrid scaffold, we recommend the following:

1. Determine the completeness of the sequence assembly. Determine whether the NGS total length is similar to the expected genome size.
2. Evaluate NGS data. A contig N50 of at least 150 kbp is required in order to be sufficiently scaffolded with Bionano data.
3. Determine the heterozygosity of the organism

Technical Assistance

For technical assistance, contact Bionano Genomics Technical Support.

You can retrieve documentation on Bionano products, SDS's, certificates of analysis, frequently asked questions, and other related documents from the Support website or by request through e-mail and telephone.

Type	Contact
Email	**support@bionanogenomics.com**
Phone	**Hours of Operation:** **Monday through Friday, 9:00 a.m. to 5:00 p.m., PST** **US: +1 (858) 888-7663**
Website	**www.bionanogenomics.com/support**

Bionano Genomics, Inc.
9540 Towne Centre Drive, Suite 100
San Diego, CA 92121