Bionano Access®: Assembly Report Guidelines

Document Number: 30255

Document Revision: C

Table of Contents

Legal Notice

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

Revision History

Revision	Notes
A	Initial Release
B	Adjust coverage amounts to reflect guidance for haplotype-aware assemblies.
C	Adjust wording for precision. Adjust metrics to reflect most up-to-date recommendations.

Interpreting the Bionano Access Assembly Report

Introduction

This document provides guidelines for evaluating the quality of *de novo* assemblies generated using data from the Saphyr System. The guidelines described herein are based on internal experiences at Bionano Genomics® and are provided as-is. The assembly report, referred in this guidenline, is generated using Bionano Solve 3.5.1 or above and displayed in Bionano Access 1.5.1 or above.

The *De Novo* Assembly Pipeline

The Bionano Solve software suite is a collection of scripts and binaries. It contains a pipeline for *de novo* assembly of the Bionano molecules (.bnx) into consensus genome maps (.cmap).

The minimum input files for a *de novo* assembly include the molecules file (.bnx), the assembly arguments file (.xml) for specifying assembly parameters, and the cluster arguments file (.xml) for specifiying compute resources. To launch a *de novo* assembly in Bionano Access, please refer to the document Bionano Access Software User Guide. To start *de novo* assembly using command line instead, please refer to the document Running Bionano Solve Pipeline on Command Line.

We strongly recommend providing a good-quality reference for the assembly if one is available. When a reference is provided, the assembly pipeline generally proceeds through the following stages:

1. *Sort/filter the molecules
2. *Autonoise
3. Pairwise alignment
4. Assembly
5. Extension and pairmerge (5 rounds, by default)
6. *Non-haplotype/haplotype refinement to final consensus maps
7. *Structural variation (SV) calling
8. Copy number (CN) analysis

*key stages with metrics displayed in the Bionano Access Assembly Report.

After filtering, the pipeline first aligns molecules to the supplied reference to estimate error parameters for the molecules. These “autonoise” parameters are used in subsequent molecule alignment steps for evaluation. The molecules within each subset of every scan are stretched, such that the average sizing between the labels of the molecules would match that of the reference. This rescaling step helps alleviate systematic stretch differences between scans and subsets of scans. The reference is not utilized when constructing the consensus maps during the assembly, extension and pairmerge, or non-haplotype/haplotype refinement stages of the *de novo* assembly pipeline.

Assembly Report

The Assembly Report displayed in Bionano Access® provides summary statistics of key stages (with * mark above) of a *de novo* genome map assembly. It is a simplified version of the full assembly report file (exp_informaticsReport.txt). The “exp_informaticsReport.txt” file is saved in the assembly job folder on the web server and/or the computation server where the assembly is performed. The full report is generated based on results from all the stages in the assembly, as mentioned above. If a reference is provided in the assembly, the metrics of molecule-to-reference alignment and the genome map to reference alignment including structural variation calling will be included as well in the report. Additionally, the format of the Assembly Report may differ, depending on whether haplotype refinement is enabled. Please see below for more details.

The Assembly Report

The results that are presented in the Assembly Report are divided into different sections corresponding to the filtering and assembly steps that occur in the pipeline process.

1. Versions
2. Input Molecule Statistics
    a. Unfiltered
    b. Filtered
3. Molecules Aligned to the:
    a. Reference (if reference is provided)
    b. *De novo* assembly
4. *De Novo* Assembly
5. Structural Variation (SV) Summary (if reference is provided)

1. Versions

This section describes the version for all components of the bioinformatics assembly pipeline.

Tools Version – Version of "Bionano Tools." Bionano Tools includes Bionano Solve and additional scripts designed for running Bionano Solve on Bionano Compute Servers. The version information may not be available if an imported assembly was generated using command line, because the version information is provided by Bionano Access.

Solve Version – Version of "Bionano Solve." Bionano Solve includes the analysis pipeline for Bionano data processing in different applications, such as *de novo* assembly, Rare Variant Analysis, and hybrid scaffolding. The version information may not be available if an imported assembly was generated using command line.

Pipeline Version – Version of "Pipeline." Pipeline includes the binaries and scripts for assembly and assembly-related scripts.

RefAligner Version – Version of "RefAligner." RefAligner is the major binary that filters, aligns molecules, and generates .cmap files.

2. Input Molecule Statistics (Unfiltered and Filtered)

Information in this section are important statistics of the molecules file (.bnx). The statistics are similar to those found in the Molecule Quality Report (MQR); the numbers may differ from MQR because of differences in alignment parameters/stringency. The definitions below apply to both the unfiltered and filtered molecule statistics. For the unfiltered statistics, by default, these are calculated for all molecules ≥ 20 kbp by default. For the filtered statistics, by default, these are calculated for all molecules ≥ 150 kbp with ≥ 9 labels per molecule. If the molecules file (.bnx) has been filtered manually by the user after data collection, and a new molecules file (.bnx) was generated and used for a new assembly, the same thresholds would be applied by default.

Input Molecule Statistics Definitions:

Total Number of Molecules – The number of molecules present in the .bnx file.

Total Length (Mbp) – The summed length of the total number of molecules.

Average Length (kbp) – The average (mean) length of the molecules.

Molecule N50 (kbp) – The N50 (the point of half mass of the distribution) of the molecules.

Label Density (/100 kbp) – The average number of labels detected per 100 kbp of molecule length.

Coverage of the Reference (X) – The average depth of coverage of the genome as calculated by dividing the total length of molecules by the length of the reference. Reference must be provided.

2a. Input Molecule Stats (Unfiltered), Guidelines:

The guidelines below apply to assembly with input molecule file, which has no additional filters applied. The default molecule filter setting in the system is to keep all molecules with length longer than 20 kbp. If filtering has been performed to the molecules file (.bnx) prior to import the BNX file and start the assembly, then the values displayed will reflect that filtering accordingly.

Metric	Guidelines
Total number of molecules	Dependent on the size of the genome and quality of data.
Total length (Mbp)	Dependent on the size of the genome and quality of data.
Average length (Mbp)	Generally ≥ 100 kbp is acceptable. 150 kbp or higher is considered good for human samples.
Molecule N50 (kbp)	Generally ≥ 150 kbp is acceptable. 230 kbp or higher is considered good for human samples.
Label density (/100 kbp)	Dependent on labeling chemistry used and the genome content (i.e. this is sequence dependent). Should correlate to *In Silico* Digestion statistics as calculated by Bionano Access. Minimum should be above 7. When seeing +/- 2 differences from the expected labeled density, one may want to double check either reference quality or Bionano data quality.

2b. Input Molecule Stats (Filtered) Guidelines:

The descriptions below apply only to filtered data with default setting (molecule length ≥ 150 kbp and with ≥ 9 labels per molecule). If filtering has been performed to the molecules file (.bnx) prior to import the BNX file and start the assembly, then the value displayed in this section will reflect that accordingly.

Metric	Guidelines
Total number of molecules	Dependent on the size of the genome and quality of data.
Total length (Mbp)	Dependent on the size of the genome and quality of data.
Average length (Mbp)	Generally, 240 – 300 kbp for human samples is good (higher is better).
Molecule N50 (kbp)	Generally, 240 – 400 kbp for human samples is good (higher is better).
Label density (/100 kbp)	Dependent on labeling chemistry used and the genome content. Should correlate roughly to calculated *In Silico* Digestion statistics.
Coverage of the reference (X)	Only displayed when a reference is provided. For human SV analysis, 100X (minimum) Saphyr data are acceptable. Please refer to Data Collections Guidelines for details.

3. Molcules Aligned to the Reference/Assembly

The definitions below apply to molecules aligned to the supplied reference before assembly takes place (Molecules Aligned to the Reference) and also to molecules aligned to consensus genome maps after *de novo* assembly (Molecules Aligned to the Assembly), which is the very last step of the assembly process. If no reference is provided, the molecule aligned to the reference section would be absent in the report.

Molecule Alignment Definitions:

Total Number of Molecules Aligned – The number of molecules after filtering (≥ 150 kbp) that align to the *in silico* digested reference file (.cmap). This is the human reference or user-supplied sequence assembly (Reference) or Bionano consensus genome maps (Assembly).

Fraction of Molecules Aligned - The proportion of filtered molecules that align to the consensus genome maps (Assembly only).

Effective Coverage of the Reference/Assembly (X) - The total length of filtered (≥ 150 kbp) and aligned molecules divided by the length of the reference or consensus assembled maps after *de novo* assembly.

Average Confidence – The average alignment score for all the molecules that align to the reference or consensus assembled maps. Scores are estimates of the probability that the labels on a map match the labels on the reference purely by chance and that the motifs are unrelated. The scores are calculated as -log10 of that probability. The higher the score, the better.

3a. Molecules Aligned to the Reference Guidelines:

The guidelines below apply only to filtered molecule data, which is then compared to the user-specified reference. These values are also dependent on the quality of the reference supplied by the user. It could be difficult to interpret them with a poor quality reference. Below are examples assuming a high quality labeled sample and a high quality reference. Please see the Molecules Quality Report Guidelines document for more details.

Metric	Guidelines
Total number of molecules aligned	This depends on the quality of the reference and Bionano data. For a human sample with DLS labeling, 75-95% of the original filtered molecules should align.
Effective coverage of the reference (X)	The desired effective coverage of the genome will depend on the application. Refer to Data Collections Guidelines and Theory of Operations documents for SV Calling and Hybrid Scaffolding for further explanation, but ≥ 70X is preferred for most applications.*
Average Confidence	The average confidence is typically above 20 (higher is better).

**Coverage in this assembly report is calculated differently as is in molecule quality report (MQR). 70X here roughly corresponds to 80X in MQR.*

3b. Molecules Aligned to the Assembly Guidelines:

The definitions below apply only to filtered data which is then assembled in a *de novo* fashion into consensus genome maps (the assembly). The input molecules from that assembly are then compared back to the consensus genome maps to generate these metrics. This section is listed in the Assembly Report after the *de novo* assembly results, but is listed here for better continuity.

Metric	Guidelines
Total number of molecules aligned	Ideally, the difference should be within 10 – 15% with the total number of filtered molecules which would indicate a well-labeled, good quality sample.
Fraction of molecules aligned	Ideally, this should be 0.85 – 0.9, though the higher the better as it indicates better data quality; for example, how well the molecules are labeled. Values ≥ 0.6 are typically acceptable.
Effective coverage of the assembly (X)	The effective coverage of the assembly depends on the assembly size, and application-specific *effective coverage of reference (X)* target. ≥40X is typical for SV Calling with human haplotype-aware assemblies.
Average Confidence	The average confidence is typically ≥ 20.

4. *De Novo* Assembly

This section summarizes the final assembly results. The definitions below apply to data that is generated when molecules are assembled into consensus genome maps in a *de novo* fashion. Some of these metrics require the input of a reference. If one is not provided, those values will be empty. The haploid values are provided by default, but the diploid numbers require haplotype-aware assembly to be peformed which will attempt to separate maps based on alleles. Additionally, the number of maps and N50 will vary depending how complex multi-path regions (CMPRs) are treated. CMPRs are ambiguous regions of the genome, such as large segmental duplications (please see Structural Variant Calling Theory of Operation document for additional information about CMPRs). If CMPRs are selected to be cut, this will likely increase the number of maps and decrease genome map N50.

***De Novo* Assembly Definitions:**

Diploid/Haploid Genome Map Number – The total number of assembled genome maps (.cmaps) created after the assembly process. In NGS terms, this would be referred to as contig number. The lower the number of genome maps, the higher the contiguity of your assembly.

Diploid/Haploid Genome Map Length (Mbp) – The summed length of the assembled genome maps.

Diploid/Haploid Genome Map N50 (Mbp) – The N50 (the point of half of the mass of the distribution) of the assembled genome maps in the assembly.

Total Reference Length (Mbp)* – The summed length of the maps in the specified reference.

Total Number of Genome Maps Aligned (fraction)* – Number (and fraction) of maps that align completely or partially to the reference.

Total Unique Aligned Length (Mbp)* – The summed aligned length of the reference. This can be thought of as the amount of the reference "covered."

Total Unique Aligned Length / Reference Length* – The fraction of the reference that is uniquely covered by the genome map *de novo* assembly.

***** Indicates a reference must be provided to generate these metrics.

De Novo Assembly Guidelines:

The definitions below apply only to filtered data which is then assembled in a *de novo* fashion. These values are used to evaluate assembly quality.

Metric	Guidelines
Diploid/Haploid genome map count	Lower values are better. With increasing molecule N50, this number typically decreases. Dependent on size of the genome and N50. DLS data generally has a much lower number than NLRS due to higher contiguity. It also depends on whether complex multi path regions (CMPRs) are cut or not; cutting CMPRs would increase this number. The degree of increase depends on how many such complex regions there are in a given genome. For a good quality human sample using DLS labeling and CMPR cutting, we typically see approximately 500 diploid and 300 haploid genome maps.
Diploid/Haploid genome map length (Mbp)	Depends on the genome, in theory this would be approaching twice the total haploid genome length for diploid assembly, or the same as the haploid genome length for a haploid assembly.
Diploid/Haploid Genome map N50 (Mbp)	Depends on chromosome/chromosome arm size. For human samples with DLS labeling, a usual value is between 50 - 100 Mbp; for NLRS a usual value is between 1 - 4 Mbp. The number of CMPRs are inversely correlated to genome map N50; therefore, cancer genomes can have much lower N50 values.
Total reference length (Mbp)	Genome size dependent, based on the user-specified reference.

Total number of genome maps aligned (fraction)	The fraction of aligned maps, this depends on the quality of the reference, generally expected ≥ 0.80 for human samples.
Total unique aligned length (Mbp)	With a good reference, this should approach the reference length. May be reduced considerably with low reference contiguity.
Total unique aligned length/reference length	Dependent on the quality of the reference. Generally expect ≥ 0.9 for human samples; may be reduced considerably with low reference contiguity.

5. Structural Variant (SV) Summary

The definitions below apply to data after consensus genome maps are generated in a *de novo* fashion and then compared to a reference. This section is only provided when a reference is supplied. Additional information can be found in the Structural Variant Theory of Operations document. No guidelines for values can be supplied as they will vary depending on the difference between the individual sample and the provided reference, the quality of the reference, as well as whether an SV Mask file (.bed) was selected. The alignment and SV calling rely on sufficient similarity between the genome maps and the reference.

Structural Variant Summary Definitions:

SV Type – This is the header under which the below SV's are listed.

Deletions – The number of segments in the genome map assembly which are shorter when compared to the reference.

Insertions – The number of segments in the genome map assembly which are longer when compared to the reference.

Duplications – The number of duplications in the genome map assembly compared to the reference.

Inversion Breakpoints – The number of inversion breakpoints in the genome map assembly compared to the reference.

Interchromosomal Translocation Breakpoints – The number of translocation breakpoints between chromosomes in the genome map assembly compared to the reference.

Intrachromosomal Translocation Breakpoints – The number of translocation breakpoints within a chromosome in the genome map assembly compared to the reference.

Example Assembly Report (human sample labeled with DLE-1)

	Target
Tools Version: 1.5.2	
Solve Version: Solve3.5.1_01142020	
Pipeline Version: 10322	
RefAligner Version: 10436	
Input molecule stats (unfiltered):	
Total number of molecules: 1342962	
Total length (Mbp) : 365017.777	
Average length (kbp) : 271.801	> 150
Molecule N50 (kbp) : 283.424	> 150
Label density (/100kb) : 14.863	14 - 17
Input molecule stats (filtered):	
Total number of molecules : 1108393	
Total length (Mbp) : 316793.483	
Average length (kbp) : 285.813	> 230
Molecule N50 (kbp) : 294.996	> 230
Label density (/100kb) : 16.021	14 - 17
Coverage of the reference (X): 102.334	> 100
Molecules aligned to the reference:	
Total number of molecules aligned : 1041816	
Fraction of molecules aligned: 0.940	
Effective coverage of reference (X): 79.432	> 70
Average confidence : 40.4	> 20
De novo assembly:	
Diploid genome map count : 482	
Diploid genome map length (Mbp) : 5773.880	
Diploid genome map N50 (Mbp) : 68.828	> 50*
Haploid genome map count : 311	
Haploid genome map length (Mbp) : 3039.368	
Haploid genome map N50 (Mbp) : 59.544	> 50*
Total reference length (Mbp) : 3095.677	
Total number of genome maps aligned (Fraction): 373 (0.77)	(0.70)*
Total unique aligned length (Mbp) : 2859.255	
Total unique aligned length / reference length: 0.924	> 0.85
Molecules aligned to the assembly:	
Total number of molecules aligned : 1044963	
Fraction of molecules aligned : 0.943	> 0.8
Effective coverage of assembly (X): 52.403	> 40

Average confidence	:	53.4	> 20

SV summary:

SV type	:	N
Deletions	:	1310
Insertions	:	2883
Duplications	:	48
Inversion breakpoints	:	93
Interchr. translocation breakpoints:		0
Intrachr. translocation breakpoints:		0

The values above list typical numbers we see for a good haplotype- aware human Assembly Report using a high quality reference. For other organisms or lower quality references, these numbers will vary.

* These are typical metrics for human DLS data; these do not apply to NLRS data.

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.bionanongenomics.com/support

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