Bionano Access®: *De Novo* Assembly Informatics Report Guidelines

DOCUMENT NUMBER:
CG-30255

DOCUMENT REVISION:
E

Effective Date:
08/10/2023

Table of Contents

Legal Notice

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

Patents

Products of Bionano Genomics® may be covered by one or more U.S. or foreign patents.

Trademarks

The Bionano logo and names of Bionano products or services are registered trademarks or trademarks owned by Bionano Genomics, Inc. (“Bionano”) in the United States and certain other countries.

Bionano™, Bionano Genomics®, Saphyr®, Saphyr Chip®, Bionano Access™, VIA™ software, and Bionano EnFocus™ are trademarks of Bionano Genomics, Inc. All other trademarks are the sole property of their respective owners.

No license to use any trademarks of Bionano is given or implied. Users are not permitted to use these trademarks without the prior written consent of Bionano. The use of these trademarks or any other materials, except as permitted herein, is expressly prohibited and may be in violation of federal or other applicable laws.

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.
D	Document updates for Solve 3.7 release • Rename to Bionano Access De Novo Assembly Informatics Report • Update pipeline stages to include VAF, AOH/LOH, Annotation • Update De Novo Informatics Report format
E	Document updates for Solve 3.8 release • Update with additions of raw and clustered variant counts • Update with addition of haplotype assembly stats • Combine example report with recommended values for human assemblies

Interpreting the Bionano Access *De Novo* Assembly Informatics 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 *De Novo Assembly Informatics Report*, referred in this guideline, is generated using Bionano Solve 3.7 or above and displayed in Bionano Access 1.7.

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). also included in the *de novo* assembly pipeline are tools for calling structural variation compared to a reference genome, determining relative copy number state across the genome after calibrating the median state to 2 copies for autosome chromosomes, detection of segments where heterozygosity is absent (AOH), and annotation of structural and copy number variants. Some analyses are supported for specific genomes only.

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 specifying compute resources. To launch a *de novo* assembly in Bionano Access, please refer to the *Bionano Access Software User Guide* (CG- 30142) To start *de novo* assembly using command line instead, please refer to Running *Bionano Solve Pipeline on Command Line* (CG-30205).

With human data, the user may select `hg19_DLE1_0kb_0labels.cmap or hg38_DLE1_0kb_0labels.cmap` as reference to utilize downstream controls, masks, and variant annotations. In plant and animal systems, we strongly recommend providing a good-quality reference for the assembly if one is available. When a reference is selected, 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. Confidence score calculation
9. Allele fraction determination
10. Copy number (CN) analysis
11. AOH/LOH Detection
12. Variant annotation

*key stages with metrics displayed in the *Bionano Access De Novo Assembly Informatics 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.

DE NOVO ASSEMBLY INFORMATICS REPORT

The *De Novo Assembly Informatics 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 *De Novo Assembly Informatics Report* may differ, depending on whether haplotype refinement is enabled. Please see below for more details.

The *De Novo* Assembly Informatics Report

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

1. Job Details
2. Molecule Stats
3. Input molecule stats (filtered)
4. Molecules aligned to the reference
    a. **NOTE**: only if a reference is selected
5. *De novo* assembly
6. Molecules aligned to the assembly
7. Structural Variation (SV) summary (raw counts)
    a. **NOTE**: only if a reference is selected
8. Structural Variation (SV) summary (clustered counts)
    b. **NOTE**: only if a reference is selected
9. AOH/LOH Detection Stats
    a. **NOTE**: only if `hg19_DLE1_0kb_0labels.cmap or hg38_DLE1_0kb_0labels.cmap` is selected as reference
10. CNV Statistics
    a. **NOTE**: only if a reference is selected

JOB DETAILS

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

- Access Version – Version of Bionano Access. This software is the user's data management and visualization hub.
- Solve Version – Version of Bionano Solve. This software 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.
- Compute On Demand Version – Version of Bionano Solve as run using Bionano Compute On Demand. The version information may not be available if an imported assembly was generated using command line.

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.

MOLECULE ALIGNMENT 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 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.

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 *Molecules Quality Report Guidelines* (CG-30223) for more details.

Table 1. Molecules Aligned to the Reference Guidelines

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* (CG-30173), *Bionano Solve Theory of Operation: Structural Variant Calling* (CG-30110), and *Bionano Solve Theory of Operation: Hybrid Scaffold* (CG-30073) 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 De Novo Assembly Informatics Report is calculated differently as is in molecule quality report (MQR). 70X here roughly corresponds to 80X in MQR.

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 *De Novo Assembly Informatics Report* (CG-30255) after the *de novo* assembly results but is listed here for better continuity.

Table 2. Molecules Aligned to the Assembly Guidelines

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.

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 performed 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* (CG-30255) 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 the 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."
- Assessment of stable regions of the genome* - Quality control assessment of regions of the genome observed to be stable across assemblies. Reported as PASS/FAIL.

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

D*E 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.

Table 3. Filtered Data definitions

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.

STRUCTURAL VARIANT (SV), AOH/LOH, AND CNV STATS

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 *Structural Variant Theory of Operations* (CG-30110). 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 appropriate Mask file (.bed) was selected. The alignment and SV calling rely on sufficient similarity between the genome maps and the reference. Structural variant counts are presented for 'raw' counts as well as 'clustered'. Raw counts are counts of SV calls made locally to individual consensus maps (which may be redundant) while clustered counts are counts of distinct genomic variants.

Example *De Novo* Assembly Informatics Report (human sample labeled with DLE-1)

Below is a sample *de novo* assembly informatics report along with target values, where applicable, for a good haplotype-aware human analysis using a high-quality reference. For other organisms or lower quality references, this numbers will vary. These are typical metrics for human DLS data; these do not apply to NLRS data. **NOTE:** Many metrics, such as the number of structural variants or number of input molecules, are dataset dependent and do not have recommended values.

Table 4. Job Details section

Label	Value	Description
Job ID	456	Job Identifier
Server name	192.168.48.138	Name or IP of the server that ran job
Created at	2021-08-25T21:15:41.102Z	Date job was created
User Name	Bionano User	Full name of user who launched job
Job type	Annotated *de Novo* Assembly	The type of operation performed
Access Version	1.8	Bionano Access Version
Solve Version	Solve3.8	Bionano Solve Version
Compute On Demand Version	Solve3.8	Version of pipeline for Compute On Demand
Job Name	My_sample_001 – *De novo*	Alias for Job
Project Name	Interesting_cases	Name of the project
Sample Name	My_sample_001	Name of the sample
Sample UID	c6fb6e5c-efb8-11eb-a7dd-3cfdfe7f3f60	System generated global unique identifier
Reference	hg38_DLE1_0kb_0labels.cmap	Name of the reference genome this sample was aligned to

Label	Value	Description
Input job 1 - Job ID	345	Job Identifier
Input job 1 - Job Name	My_sample_001	Alias for Job
Input job 1 - Job type	Import Molecule	The type of operation performed
Enzyme	CTTAAG	Name of enzyme: DLE-1

Table 5. Molecule statistics section

Label	Sample Value	Target Value	Description
Total number of molecules	3,456,789	Dataset dependent	The number of molecules present in the .bnx file.
Total length	456,789.01 Mbp	Dataset dependent	The summed length of all molecules in the bnx files.
Average length	145.56 kbp	> 100 kbp	The average (mean) length of the molecules.
Molecule N50	256.78 kbp	> 150 kbp	N50 can be described as a weighted median statistic such that 50% of the summed molecule length is contained in molecules equal to or larger than this value
Label density	15.96 /100kb	14 -17 / 100 kbp	The average number of labels detected per 100 kbp of molecule length.

Table 6. Input molecule stats (filtered) section

Label	Sample Value	Target Value	Description
Total number of molecules	1,234,567	Dataset dependent	The number of molecules present in the .bnx file.
Total length	345,678.9 Mbp	Dataset dependent	The summed length of all molecules in the bnx files.
Average length	320.24 kbp	> 230 kbp	The average (mean) length of the molecules.
Molecule N50	333.33 kbp	> 230 kbp	N50 can be described as a weighted median statistic such that 50% of the summed molecule length is contained in molecules equal to or larger than this value
Label density	16.1 /100kb	14 – 17 / 100 kbp	The average number of labels detected per 100 kbp of molecule length.
Coverage of the reference	111.93 X	> 100 X	The average depth of coverage of the genome is calculated by dividing the total length of molecules by the length of the reference. Reference must be provided.

Table 7. Molecules aligned to the reference section

label	sample value	target value	description
Total number of molecules aligned	1,123,456	Dataset dependent	The number of molecules after filtering ($\geqslant$ 150 kbp) that align either to the reference file (.cmap), e.g., GRCh37 or GRCh38.
Fraction of molecules aligned	0.91	Dataset dependent	The proportion of filtered molecules that align to the reference or consensus genome maps (assembly only).
Effective coverage of reference	95.14 X	> 70 X	The total length of molecules divided by the length of the reference or consensus assembled maps after *de novo* assembly.
Average confidence	36.6	> 20	The average alignment score for all the molecules that align to the reference or assembly.

Table 8. *De novo* assembly section

Label	Sample Value	Target Value	Description
Diploid genome map count	699	Dataset dependent, usually 500-1500	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 the assembly.
Diploid genome map length	5,980.55 Mbp	Dataset dependent, usually 5,800-6,000 Mbp	The total length of the genome assembled
Diploid genome map N50	61.39 Mbp	> 50 Mbp	N50 can be described as a weighted median statistic such that 50% of the summed molecule length is contained in molecules equal to or larger than this value
Haploid genome map count	673	Dataset dependent, usually 250-750	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 number of genome maps, the higher the contiguity of the assembly
Haploid genome map length	3,176.54 Mbp	Dataset dependent, usually 2,900-3,200 Mbp	The summed length of the assembled genome maps
Haploid genome map N50	57.3 kbp	> 50 Mbp	The N50 (the point of half of the mass of the distribution) of the assembled genome maps in the assembly.
Total reference length	3,088.27 Mbp	Reference dependent	The summed length of the chromosomes (or other maps) in the specified reference.
Total number of genome maps aligned (Fraction)	619 (0.89)	(>0.70)	Number (and fraction) of maps that align completely or partially to the reference.
Total unique aligned length	2,876.9 Mbp	>2,750 Mbp	The summed aligned length of the reference. This can be thought of

Label	Sample Value	Target Value	Description
			as the amount of the reference "covered."
Assessment of stable regions of the genome	PASS	PASS	

Table 9. Molecules aligned to the assembly section

Label	Sample Value	Target Value	Description
Total number of molecules aligned	1,131,211	Dataset dependent	The number of molecules after filtering (⩾ 150 kbp) that align either to the reference file (.cmap), (e.g., GRCh37 or GRCh38) or the assembly.
Effective coverage of assembly (X)	59.86	> 40X	The total length of molecules divided by the length of the reference or consensus assembled maps after *de novo* assembly.
Average confidence	50.6	> 20	The average alignment score for all the molecules that align to the reference or assembly.

Table 10. SV summary (raw and clustered counts) section

Label	Sample Value	Target Value	Description
Deletions	1,614	Dataset dependent	The number of segments in the genome map assembly which are shorter when compared to the reference, indicating that a deletion occurred in the segment.
Insertions	1,533	Dataset dependent	The number of segments in the genome map assembly which are longer when compared to the reference, indicating that an insertion occurred in the segment.
Duplications	63	Dataset dependent	The number of nonredundant segmental tandem and tandem-inverted duplication patterns found in the genome map assembly compared to the reference.

Label	Sample Value	Target Value	Description
Inversion breakpoints	23	Dataset dependent	The number of inversion breakpoints found in the genome map assembly compared to the reference based on the observation that patterns are in opposite orientation compared to adjacent sequence patterns.
Interchr. translocation breakpoints	2	Dataset dependent	The number of nonredundant interchromosomal translocation breakpoints resulting in fusion between different chromosomes in the genome map assembly compared to the reference.
Intrachr. fusion breakpoints	0	Dataset dependent	The number of translocation breakpoints within a chromosome in the genome map assembly compared to the reference.

Table 11. AOH/LOH Detection Stats section

Label	Sample Value	Target Value	Description
Total number of calls in autosome	9	Dataset dependent	Total number of regions in the autosome identified as a having absence or loss or heterozygosity (AOH/LOH).
Total number of calls greater than 25 Mbp in autosome	1	Dataset dependent	Total number of regions greater than 25 Mbp in the autosome identified as a having absence or loss or heterozygosity (AOH/LOH).
Total number of calls greater than 40 Mbp in autosome	0	Dataset dependent	Total number of regions greater than 40 Mbp in the autosome identified as a having absence or loss or heterozygosity (AOH/LOH). Large calls are higher confidence calls.
Fraction of autosome called as AOH/LOH	0.021	Dataset dependent	Total fraction of the autosome identified as a having absence or loss or heterozygosity (AOH/LOH); high values may indicate consanguinity.

Table 12. CNV Statistics section

Label	Sample Value	Target Value	Description
Sex	male	Dataset dependent	Sex determined based on coverage of sex chromosomes
Median Coverage	99	Dataset dependent	Median number of molecules covering each position genome-wide
Global coefficient of variation	0.22	< 0.162	Coefficient of variation in coverage genome- wide
Median local coefficient of variation (2Mbp)	0.1	< 0.155	Median coefficient of variation observed within 2Mbp intervals
Percent above expected (2 Mbp window)	-6.82	(-20) – 20	Percent difference between the expected coefficient of variation in sample with low systematic biases (2 Mbp windows) and the observed coefficient of variation in the sample. If the observed percent different is greater than 20%, the sample should be considered to contain systematic bias.
Median local coefficient of variation (6Mbp)	0.11	< 0.145	Median coefficient of variation observed within 6Mbp intervals
Percent above expected (6 Mbp window)	-1	(-20) - 20	Percent difference between the expected coefficient of variation in sample with low systematic biases (6 Mbp windows) and the observed coefficient of variation in the sample. If the observed percent different is greater than 20%, the sample should be considered to contain systematic bias.
Correlation with label density	0.038	< 0.25	Correlation between coverage of genomic regions and the label density in the given regions. Value greater than 0.25 indicates high systematic biases and sample may have more false positive CNV calls.

Label	Sample Value	Target Value	Description
Wave template correlation	0	< 0.4	Correlation between the coverage of the query sample and the coverage profiles from known samples with large systematic bias. Value greater than 0.4 indicates high systematic biases and sample may have more false positive CNV calls.

Technical Assistance

For technical assistance, contact Bionano 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@bionano.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.bionano.com/support
Address	Bionano, Inc. 9540 Towne Centre Drive, Suite 100 San Diego, CA 92121