Saphyr Molecule Quality Report Guidelines

Document Number: 30223

Document Revision: C

Table of Contents

Legal Notice

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

Revision History

Revision	Notes
C	Document updates for Solve 3.7 release • Expand QC indicators • Update MQR format • Remove NLRS data example • Move manual molecule-to-reference alignment to Appendix

Saphyr Molecule Quality Report Guidelines

Important: The guidelines described herein are based on internal experiences at Bionano Genomics and are provided as-is. The purpose of this technical note is to provide guidelines to customers who want to evaluate the quality of data generated from the Saphyr System. For questions, please contact the Technical Support Team at support@bionanogenomics.com.

The Molecule Quality Report

The Molecule Quality Report (MQR) provides a summary report on the quantity and qualities of molecules in the BNX file. The report contains basic details about the molecule BNX file and associated job. For datasets for which a reference is selected, additional metrics are generated based on results from a molecule-to-reference alignment.

Bionano Solve aligns Bionano molecules to a given reference and identifies regions of similarity between Bionano molecules and reference CMAP. The MQR identifies and outputs the best alignment of each molecule to the reference, provided that the alignment meets the minimum alignment quality criteria.

To determine if the data quality is sufficient to proceed to secondary data analysis, the best indicators are the following:

1) Map rate*: What percentage of the Bionano molecules aligns to the reference (meeting minimum alignment quality criteria)?

2) Molecule N50*: A proxy for size, the N50 values indicate a weighted average length of DNA molecules in the dataset

3) Label Density, Negative Label Variance (NLV), and Positive Label Variance (PLV)*: What is the frequency of labeled sites in the Bionano molecules, and what percentage of labels are missing (NLV) or extraneous (PLV) in sample molecules relative to reference?

4) Effective coverage*: What is the pre-analysis estimate of molecule coverage of the reference, and is it adequate for the intended bioinformatic pipeline? See **30173 Data Collection Guidelines** for more details.

   If no reference is available, the user may evaluate Total DNA and estimated genome size to approximate raw coverage.

5) Noise parameters*: How different are the aligned Bionano molecules when compared to the reference?

***** The evaluation of the MQR results is highly dependent on the accuracy and completeness of the given reference and the identity of the sample with the reference. Many sequence assemblies, even at advanced stages, could have a high degree of structural inaccuracy that may compromise the use of the MQR. See *Appendix A: Manual Interpretation of Molecule-To-Reference Alignment* for details.

Molecule Quality Report Example

Beginning in Bionano Access 1.7 and Bionano Solve 3.7, the Molecule Quality Report consists of a Job Details section and an MQR Report Details section. Each entry contains a label, value, and brief description of its meaning. Using Bionano Access, the user may select to print the MQR and/or access a JSON of the contents.

Job Details

label	value	description
Job ID	569	Job Identifier
Server name	192.168.49.224	Name or IP of the server that ran job
Created at	2021-08-23T19:40:38.528Z	Date job was created
User Name	Bionano User	Full name of user who launched job
Job type	Import Molecule	The type of operation performed
Access Version	1.7	Bionano Access Version
Solve Version	Solve3.7	Bionano Solve Version
Compute On Demand Version	Solve3.7_	Version of pipeline for Compute On Demand
Job Name	My_sample_001 - Molecules	Alias for Job
Project Name	Interesting_cases	Name of the project
Sample Name	My_sample_001	Name of the sample
Sample UID	f231294a-0449-11ec-a21c-3cfdfe7f3f60	System generated global unique identifier
Reference	hg38_DLE1_0kb_0labels.cmap	Name of the reference genome this sample was aligned to

MQR Report Details

label	value	description
Reference	hg38_DLE1_0kb_0labels.cmap	Name of the reference genome this sample was aligned to.
Reference Length	3,088,269,832 bp	Total length of reference sequence
Enzyme	DLE-1	Name of the enzyme used in this sample.
Site	CTTAAG	Recognition sequence of the enzyme used.
Maximum molecule length	2.15 Mbp	The longest molecule detected during the chip run.
N50 (>= 20 kbp)	248.18 kbp	N50 of the molecules that are 20kbp or longer)
Total DNA (>= 20kbp)	516.18 Gbp	Total amount of DNA from molecules that are 20 kbp or longer

label	value	description
N50 (>= 150kbp)	311.07 kbp	N50 of DNA molecules that are 150kbp or longer
Total DNA (>= 150kbp)	407.75 Gbp	Total amount of DNA from molecules that are 150kbp or longer
N50 (>= 150kbp and min sites >=9)	312.2 kbp	Same as other N50 fields, but molecules must have at least 9 labels
Total DNA (>= 150kbp and min sites >= 9)	390.34 Gbp	Same as other Total DNA fields, but molecules must have at least 9 labels
Map rate	92.7 %	Percentage of molecules that are 150kbp or longer mapped to the reference
Effective coverage	107.56	Total amount of aligned DNA divided by the size of the reference genome times the map rate.
Average label density (>= 150kbp)	15.31 /100kbp	Average number of labels per 100 kbp for the molecules that are 150kbp or longer
Site SD	0.084	Constant term in sizing error relative to reference
Relative SD	0.013	Quadratic term in sizing error relative to reference
Scaling SD	0	Linear term in sizing error relative to reference
integrity_num	0.09	
Negative label variance (NLV)	8.53	Percentage of reference labels absent in molecules
Base pairs per pixel	492.6	Calculated base pairs per pixel in the alignment by comparing molecules to the reference
Label color	BNGFLGR001	Label color used for detection.
version	1	
Positive label variance (PLV)	3.08	Percentage of labels absent in reference

Interpret Molecule Quality Report Results

Below is a list of desirable metric ranges, based on human DNA labeled with DLE-1 and collected with a Saphyr instrument:

label	value	description
Reference	**application dependent**	Name of the reference genome this sample was aligned to.
Reference Length	**application dependent**	Total length of reference sequence
Enzyme	DLE-1	Name of the enzyme used in this sample.
Site	CTTAAG	Recognition sequence of the enzyme used.
Maximum molecule length	**no set recommendation**	The longest molecule detected during the chip run.
N50 (>= 20 kbp)	**> 150 kbp**	N50 of the molecules that are 20kbp or longer)
Total DNA (>= 20kbp)	**application dependent**	Total amount of DNA from molecules that are 20 kbp or longer
N50 (>= 150kbp)	**>230 kbp**	N50 of DNA molecules that are 150kbp or longer
Total DNA (>= 150kbp)	**application dependent**	Total amount of DNA from molecules that are 150kbp or longer
N50 (>= 150kbp and min sites >=9)	**>230 kbp**	Same as other N50 fields, but molecules must have at least 9 labels
Total DNA (>= 150kbp and min sites >= 9)	**application dependent**	Same as other Total DNA fields, but molecules must have at least 9 labels
Map rate	**>70%**	Percentage of molecules that are 150kbp or longer mapped to the reference
Effective coverage	**application dependent**	Total amount of aligned DNA divided by the size of the reference genome times the map rate.
Average label density (>= 150kbp)	**14 - 17 /100kbp**	Average number of labels per 100 kbp for the molecules that are 150kbp or longer
Site SD	**< 0.25**	Constant term in sizing error relative to reference
Relative SD	**< 0.04**	Quadratic term in sizing error relative to reference
Scaling SD	**-0.07 – 0.05**	Linear term in sizing error relative to reference
integrity_num	**< 20**	
Negative label variance (NLV)	**< 15.0**	Percentage of reference labels absent in molecules
Base pairs per pixel	**450 – 510**	Calculated base pairs per pixel in the alignment by comparing molecules to the reference
Label color	BNGFLGR001	Label color used for detection.
version	1	
Positive label variance (PLV)	**< 10.0**	Percentage of labels absent in reference

Appendix A: Manual Interpretation of Molecule-To-Reference Alignment

To interpret MQR results, check the molecule-to-reference map rate (%) first. The map rate is also closely tied to the completeness and accuracy of the reference. For example, the fraction of the genome that is assembled into large contigs or scaffolds in the reference, and how much error, such as repeat collapse, is in the reference assembly. Additionally, the map rate depends on the degree of identity of the Bionano sample with the reference sample (i.e. is the sample from the same individual as the reference?).

For example, the human reference is highly complete, so the map rate can be higher than 90% for a good molecule dataset. If the reference or sequence assembly is only 50% complete, then the expected map rate range may be halved to 30-40%, even if the Bionano molecules are of good quality.

If the obtained map rate is significantly lower than the minimum desired map rate (i.e., < 60% for high quality reference or 30% for half-complete reference), check the noise parameters. If the noise parameters are within the desired range (see *Interpret Molecule Quality Report Results* table), it could mean that the part of the Bionano data that does align to the reference is of good quality. In this case, these molecules can be used for *de novo* assembly; however, users may need to collect extra depth of the same data to compensate for low mapping rate.

When interpreting the results, it is important to consider the accuracy of the provided reference. However, evaluating the reference accuracy is often challenging. If the map rate is lower than expected based on the completeness of the reference, it is possible that the molecule quality is still good, but because of the inaccuracy of the reference, some molecules do not align. In this case, it is difficult to evaluate molecule quality using MQR.

Another way to evaluate alignment between the Bionano molecules and reference CMAP is to view alignments in Bionano Access. The aligned molecules should cover most of the reference genome (or reference contigs) relatively uniformly and without large errors (see Figure 1).

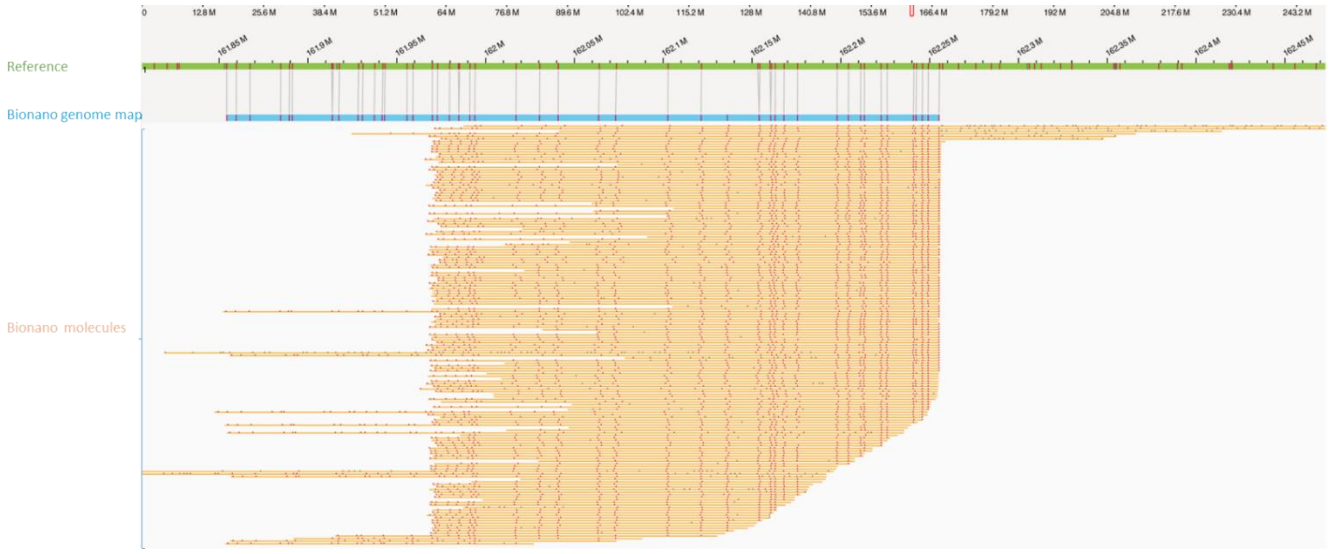

***Figure 1**: An example of a good alignment.*

In cases where it is difficult to evaluate the reference completeness or accuracy or when it is difficult to obtain reliable noise parameters from MQR, we recommend that users perform *de novo* assembly using the pre-assembly option and default noise parameters starting with at least 100X coverage data (refer to the tutorial video *Assembly Objects*). If most of the genome (> 80%) can be assembled with reasonable contiguity, the expected assembly length, and acceptable molecule alignments (i.e. a good alignment of the Bionano molecules to the assembled map as visualized; see Figure 1), then the data is likely to be of good quality.

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