Guidelines for Interpreting the Bionano Molecule Quality Report

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 Irys® System. For questions, please contact the Technical Support Team at support@bionanogenomics.com.

Molecule Quality Report (MQR)

The Molecule Quality Report provides a summary report on the molecule quality. The report is generated based on results from a molecule-to-reference alignment. The RefAligner tool aligns Bionano molecules to a given reference and identifies regions of similarity between Bionano molecules and reference. The input includes the molecule BNX file and the reference CMAP file. See the *Molecule Quality Report Output Files* section for more details.

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 continue data collection and to run *de novo* assembly, 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) 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 the *Interpret Molecule Quality Report Results* section for details.

Run the Molecule Quality Report

1. For instructions on performing a MQR in IrysView, see the IrysView® v2.5.1 Software Training Guide, section 6.18.
2. The recommended default MQR alignment parameters for human samples in IrysView v2.5.1 are the following:

-nosplit 2 -BestRef 1 -biaswt 0 -Mfast 0 -FP 1.5 -FN 0.15 -sf 0.2 -sd 0.0 -A 5 -outlier 1e-3 -outlierMax 40 -endoutlier 1e-4 -S -1000 -sr 0.03 -se 0.2 -MaxSF 0.25 -MaxSE 0.5 -resbias 4 64 -maxmem 64 -M 3 3 -minlen 150 -T 1e-11 -maxthreads 32 -hashgen 5 3 2.4 1.5 0.05 5.0 1 1 3 -hash -hashdelta 10 -hashoffset 1 -hashmaxmem 64 -insertThreads 4 -maptype 0 -PVres 2 -PVendoutlier -AlignRes 2.0 -rres 0.9 -resEstimate -ScanScaling 2 -RepeatMask 5 0.01 -RepeatRec 0.7 0.6 1.4 -maxEnd 50 –usecolor 1 -stdout –stderr –randomize –subset 1 5000

3. Modify the following alignment parameters as necessary:

M designates how many alignment iterations to perform. After each iteration of alignment, the noise parameters are estimated, and those noise parameters are used for the next iteration.

The number of iterations is chosen, considering a tradeoff of computational time and potentially more accurate noise estimates. Larger genomes or datasets could require a significant computation.

Using the default argument **-M 3 3**, the hash table is regenerated 3 times (second “3” in the argument) and perform 3 iterations for each hash table result (first “3” in the argument). This process gives more accurate error estimates than just **-M 3** or **-M 9**.

Without hashing arguments (-hashgen 5 3 2.4 1.5 0.05 5.0 1 1 3 -hash -hashdelta 10 -hashoffset 1 -hashmaxmem 64), **-M 3 3** is treated the same as **-M 9**, in which the alignment is repeated for 9 times.

T is the P-value cutoff. The cutoff should be set according to the genome complexity, which is scaled with the genome size and average label density. Therefore, the P-value can be adjusted based on the size of genome and average label density. We recommend 1e-11 for genomes larger than 1 Gbp in size with an average label density less than 15 labels per 100 kbp.

Genomes (larger than 1 Gbp) with higher label densities (> 15/100 kbp) require a lower P-value (more stringent). We suggest to lower the P-value by the factor of 100, per 1 label density increase from 15/100 kbp.

The table below lists the suggested P-value for example species with different genome size and average label density.

Species	Genome Size	Average Label Density (Nicking Enzyme)	Suggested P-value (-T)
E. coli	5 Mbp	9/100 kbp (Nt.BspQI)	1e-7
Drosophila	120 Mbp	11/100 kbp (Nt.BspQI)	1e-9
Human	3.3 Gbp	9/100 kbp (Nt.BspQI)	1e-11
Sample Species	2 Gbp	16/100 kbp	1e-13

outlierMax limits the maximum size of outlier* in kbp. It controls how tolerant the RefAligner is of the size of the outliers. If significant structural differences are expected between the sample and the reference, this argument (-outlierMax 40*)* may be modified or removed. See Figure 1.

*The size difference between aligned label intervals in reference and Bionano molecules (see Figure 1).

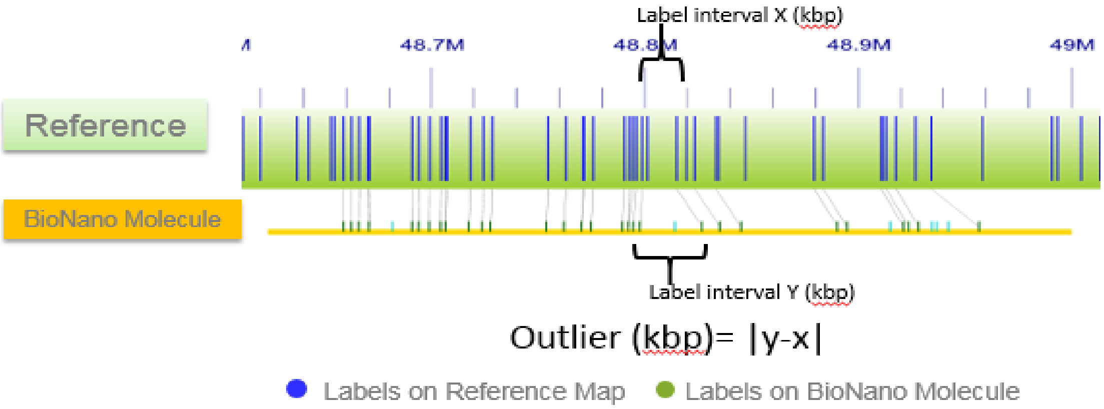

Figure 1*: Outlier is the size difference between aligned label intervals in reference and Bionano molecules*.

Molecule Quality Report Output Files

1. The key output files generated in MQR are the following:

File	Description
MoleculeQualityReportInput.tar.gz*	The ZIP file containing input BNX file and reference CMAP. This file is transferred from IrysView to the server when the MQR computation starts.
MQR_files.tar.gz*	The ZIP file containing multiple MQR result files. The file is transferred back to IrysView from the server to display the MQR results in IrysView, when the MQR computation finishes.
MoleculeQualityReport.stdout	The log file for the alignment.
MoleculeQualityReport_rescaled.bnx	When per-scan scaling of molecules is enabled (via default parameter "–ScanScaling 2"), all the scaled molecules are included in this file, not just subset of molecules used in alignment.
MoleculeQualityReport.maprate	The summary of map rate at increasingly stringent P-values.
MoleculeQualityReport.err	The summary of noise parameters for each alignment iteration (-M).
MoleculeQualityReport.errbin	The binary ERR file (same contents as *.err)
MoleculeQualityReport.xmap	The alignment result between Bionano molecules and reference CMAP. The alignment can be viewed in IrysView by opening the XMAP file.
MoleculeQualityReport_q.cmap	The alignment result of the aligned Bionano molecules for viewing.
MoleculeQualityReport_r.cmap	The alignment result of the aligned reference maps for viewing.
MoleculeQualityReport.scan	The alignment result of the Bionano molecules in each scan to the reference, using the same argument.
Molecules.bnx	The input BNX file for MQR alignment, which contains Bionano molecule and label information.
Reference.cmap	The input reference CMAP file for MQR alignment, which contains in silico digestion result of reference (contig length and label information)
MoleculeQualityReport.errbias	The information in these files are not used.
MoleculeQualityReport_intervals.txt	
MoleculeQualityReport.map	

*These 2 files are only generated when users run MQR on a remote server via IrysView.

2. The metric results of MQR in IrysView are the following:

Metrics	**Description**
Map Rate (%)	The percentage of molecules (N Molecules) aligned to reference.
N Molecules	The number of molecules used in the alignment to reference
FP (/100 kbp)	The density of unaligned labels in molecules (relative to reference length).
FP (%)	The percentage of unaligned labels in molecules (relative to number of labels in molecules).
FN (%)	The percentage of unaligned reference labels (relative to number of reference labels).
SiteSD (kbp) : sf	These parameters describe which inter-label distances in Bionano molecules match the reference. This is referred to as sizing error relative to reference. These parameters are components of the variance of the distance observed in the Bionano molecules for a given interval (distance between two labels) on the reference. $variance(x) = sf^2 + x\vert sd\vert sd + x^2*sr^2$ The **x** value is the interval length (kbp) and the other noise parameters are reported in the ERR file. SMin is not reported in the ERR file; it is either the minimum value of sqrt(variance) at x > 1 kbp OR the value of sqrt(variance) at x = 1 kbp, whichever is lower.
ScalingSD (kbp^1/2) : sd	
RelativeSD : sr	
SMin (kbp)	
Bpp	The calculated base pairs per pixel in the alignment by comparing molecule intervals to reference intervals.
Stretch (%)	The stretch factor of DNA molecules in the chip. This is computed from Bpp and average chip stretch factor.

Interpret Molecule Quality Report Results

1. List of metrics with ranges based on Bionano internal human data (source: Irys instrument).

Metrics	**Range**
Map Rate (%)	60%-80%
N Molecules	User-specified; recommended to be at least 5000
FP (/100kbp)	< 1.7
FP (%)	< 15%
FN (%)	< 21%
SiteSD (kbp): sf	< 0.25
ScalingSD (kbp^1/2): sd	(-0.07) ~ 0.05
RelativeSD: sr	< 0.04
SMin (kbp)	< 0.25
Bpp	450 ~ 510
Stretch (%)	83% ~ 94%

2. Interpret the MQR Results

To interpret MQR results, check the molecule-to-reference map rate (%) first. The map rate is also closely tied to the completeness and accurateness of the reference (i.e. how much non-sequenced part, gap and ambiguities in the sequence assembly?).Additionally, the map is tied to 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 as high as 90% for a good molecule dataset. If the reference or sequence assembly is only 50% complete, then the expected map rate range may be half or 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 recommended range, 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 accurateness 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 challenging to evaluate molecule quality using MQR.

Another way to evaluate alignment between the Bionano molecules and reference CMAP is to view alignments in IrysView. The aligned molecules should cover most of the reference genome (or reference contigs) relatively uniformly and without large errors (see Figure 2 and 3).

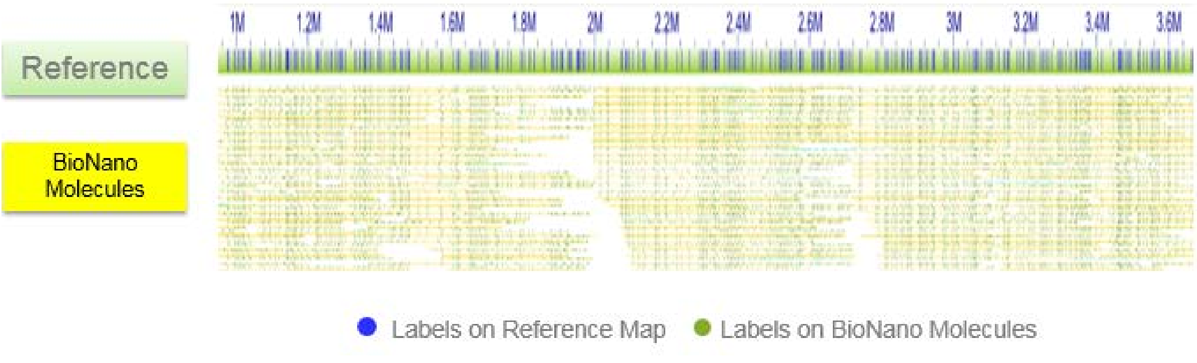

Figure 2*: An example of a good alignment.*

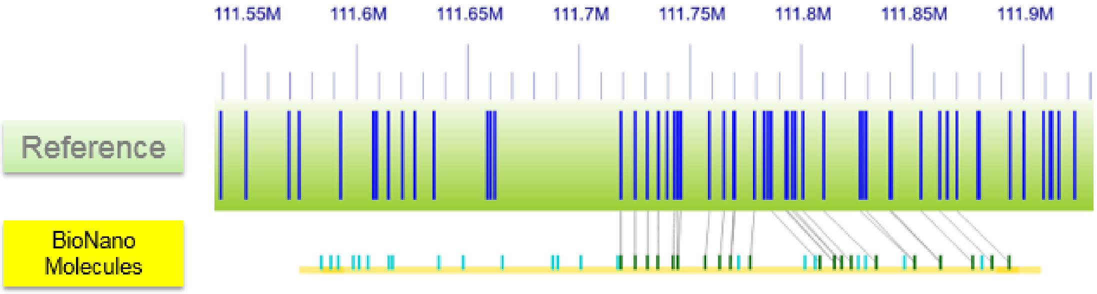

Figure 3*: An example of a chaotic alignment.*

In cases when it is difficult to evaluate the reference completeness or accurateness or when it is not sufficient to obtain reliable noise parameters from MQR, we recommend that users perform *de novo* assembly using default noise parameters with at least 100X coverage data. When most of the genomes (>50%) can be assembled with reasonable data quality (i.e. a good alignment of the Bionano molecules to the assembled map is visualized; see Figure 2), the data quality are more likely to be sufficient.