Data Collection Guidelines

Document Number: 30173

Document Revision: D

Table of Contents

Legal Notice

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

Revision History

Revision	Release Date	Notes
A	5/9/2017	
B	4/5/2018	
C	5/16/2019	
D	10/16/2020	Add guidance for RVA, FSHD. Add example scenarios

Data Collection Guidelines

Introduction

The Saphyr® System collects data by drawing ultra-high molecular weight DNA into nanochannel arrays, imaging the molecules in linear confinement, and detecting molecule and label positions from the images. Once a batch of molecules in the nanochannels have been imaged, the Saphyr clears the array and loads in new molecules. The system repeatedly cycles through this process (each cycle is one "scan"). The greater the number of scans that accumulate, the more raw data collects for a given dataset. Operational details are covered in the **Saphyr System User Guide - 30247**.

Per flowcell limits on maximum accumulated output and runtime vary by chip part number and Saphyr Instrument Control Software (ICS) version. Within those limits, it is up to the user to determine how much data to collect. This document will provide guidance, and discuss data collection strategies for various downstream applications.

The key factors for success in using the Saphyr System include:

- Design your experiments to meet research goals;
- Isolate high molecular weight gDNA from sample sources, then efficiently label at enzyme-specific recognition sequence;
- Collect sufficient amount of high quality data for downstream data analysis.

The Saphyr System collects whole genome imaging data. As such, throughput and coverage targets discussed here are genome-wide in scope, regardless of any more focused downstream application. Furthermore, the guidance here assumes that data is of reasonably good quality. For a discussion of raw data quality, please refer to **Bionano Access Dashboard Guidelines – 30304** and/or **Molecule Quality Report Guidelines – 30223**. For guidance on troubleshooting sample prep or quality from a particular sample type, please refer to the appropriate Bionano Prep™ documentation.

Throughput Targets and Coverage

Prior to collecting Saphyr data, the user enters chip and flowcell information in Bionano Access. A user is prompted to specify the Throughput Target (in Gbp) for each flowcell. This throughput target should reflect an appropriate fold coverage (X) of the haploid genome size investigated, specific to the intended application.

Conceptually, coverage refers to the average depth of molecules representing a given locus in the genome. It can be calculated from the predicted haploid genome size and the amount of input data.

$$Coverage\ (X) = \frac{Total\ DNA}{Genome\ size}$$

For example, 100 Gbp Total DNA Throughput of a sample with a 1.0 Gbp haploid genome size equals 100X Coverage (100 Gbp ÷ 1.0 Gbp = 100).

In practice, not every molecule in a raw dataset adequately represents a locus. Similarly, reference assemblies are not perfect approximations of the genome. It is useful to define *effective coverage of the reference*, which uses alignment (*i.e.* Map Rate) and genome size as in reference assembly as parameters.

$$Effective\ coverage\ of\ reference\ (X) = \frac{Total\ DNA \times [Map\ Rate]}{Reference\ size}$$

Throughput and Coverage are further refined into specific terms found throughout Bionano software, outlined in Table 1.

Table 1. Glossary of Terms relating Throughput and Coverage

Term	Meaning	Source	Comment
Pre-Analysis metrics: *prior to secondary bioinformatic analysis*			
Raw coverage (X)	[Total DNA (>=150 kbp)] ÷ [haploid genome size]		Useful baseline when reference and/or data quality unclear
Throughput Target	Minimum total flowcell throughput (≥150 kbp) that Saphyr targets to collect	Bionano Access; Saphyr Instrument Control Software (Saphyr ICS)	
Total DNA (>= 20 kbp)	Cumulative length of DNA molecules that are ≥ 20 kbp	Dashboard; Molecule Quality Report	Raw molecules detectable by Saphyr
Total DNA (>=150 kbp)	Cumulative length of DNA molecules that are ≥ 150 kbp	Dashboard; Molecule Quality Report	Counts toward Throughput Target
Est Effective coverage* (X)	[Total DNA (>=150 kbp)] × [Map Rate %]* ÷ [reference size]	Dashboard; Molecule Quality Report	Approximates effective coverage of reference (X) prior to analysis
Total DNA (>= 150 kbp & minSites >= 9)	Cumulative length of DNA molecules that are ≥ 150 kbp and have ≥ 9 labels	Molecule Quality Report	Pass filters for downstream analyses
Post-Analysis metrics: *output from secondary bioinformatic analysis pipelines*			
Effective coverage of reference (X)*	[Total DNA aligned to the reference in pipeline]* ÷ [reference size]	De Novo Assembly Report; Rare Variant Analysis Report	

Effective coverage of assembly (X)	[Total DNA aligned to the assembly in pipeline] ÷ [assembly size]	De Novo Assembly Report	

*****Calculation of *Map Rate* and *Effective Coverage of the Reference (X)* relies on quality and completeness of the provided reference, as well as sequence similarity between reference and sample. Human genome references are considered relatively complete, so these metrics are meaningful and can be clearly benchmarked. With non-human organisms or custom references, molecule-reference alignment can be hindered by a myriad of factors, which may underestimate coverage and data quality. Please see **Assembly Report Guidelines - 30255** for additional details.

In human samples, the Est Effective coverage (X) provided pre-analysis is generally higher than the effective coverage of reference (X) in the pipeline results. This stems from different stringencies in alignment parameters between pre-analysis mapping and eventually running the pipeline. Effective coverage recommendations outlined here refer to pre-analysis metrics.

Coverage Recommendations for Different Analyses

Coverage recommendations vary considerably by downstream application. As of Bionano Solve 3.5, there are pipelines for Structural Variant and Copy Number Analysis (De Novo Assembly and Rare Variant Analysis), facioscapulohumeral muscular dystrophy (FSHD) Analysis, and Hybrid Scaffolding. Tables 2-5 provide a summary of minimum coverage recommendation for Direct Label and Stain (DLS) data for each application. For a more detailed view of rationale (such as SV sensitivity or CN calling performance at different coverages) please refer to the Theory of Operations guide appropriate to the analysis at www.bionanogenomics.com/support/ .

Beginning in Solve 3.4, Bionano offers two distinct pipelines for whole-genome calling of structural variants, variant annotation, and copy number variants: the *de novo* assembly pipeline and rare variant analysis. The *de novo* assembly pipeline is intended for downstream analysis of constitutional, germline, and “typically” diploid SVs. It requires a modest amount of coverage (Table 2) to assemble large genome maps *de novo*, then aligns maps to a provided reference. Genome structure is elucidated, and SVs called are generally assumed to be homozygous or heterozygous.

Table 2. *De Novo* Assembly recommended minimum input, targeting human SV analysis

Pipeline	Minimum Pre-Analysis Est. Effective Coverage (X)	hg19 / hg38 length (Gbp)	Example Map Rate	Min. DNA (Gbp)	Post-Analysis Effective Coverage of Reference(X)
***De Novo* Assembly**	**80**	**× 3.1**	**÷ 78%**	**= 318**	**≈70**

Bionano Genomics has observed very high sensitivity and confidence (PPV) at 80X effective coverage pre-analysis for heterozygous and homozygous SVs. 80X is a recommended minimum. More coverage will yield small additional sensitivity for heterozygous and homozygous SVs with diminishing returns beyond 120X. If a different Est. effective

coverage is desired, simply substitute a new coverage target in the calculation table above.

In contrast to *de novo* assembly, Rare Variant Analysis is intended for analysis of cancer, somatic, or otherwise heterogenous samples. It leverages deeper input coverage (Table 3) to target SVs at low allelic fractions. The pipeline directly aligns molecules to hg19 or hg38 to capture SVs with a relatively low fraction of molecule support.

Table 3. Rare Variant Analysis recommended minimum input, for targeting ≥5% Variant Allele Fraction

Pipeline	Minimum Pre-Analysis Est. Effective Coverage (X)	hg19 / hg38 length (Gbp)	Example Map Rate	Min. DNA (Gbp)	Post-Analysis Effective Coverage of Reference(X)
Rare Variant Analysis	**340†**	**× 3.1**	**÷ 78%**	**= 1351**	**≈300†**

†Bionano Genomics has observed at least 90% sensitivity to structural variants at 5% variant allele fraction, with datasets at 300X effective coverage of reference post-analysis. This performance is shown in detail in the **Theory Of Operation – Structural Variant Calling - 30110**.

To ensure adequate effective coverage, we generally recommend a Throughput Target of 1500 Gbp for datasets intended for Rare Variant Analysis targeting ≥5% Variant Allele Fraction. The pipeline has been validated with datasets of up to 5000 Gbp Total DNA (>=150 kbp). This very high degree of coverage can improve sensitivity beyond typical low variant allele frequency applications. If a different Est. effective coverage is desired, simply substitute a new coverage target in the calculation table above.

The Bionano EnFocus™ FSHD Analysis uses a targeted assembly approach. The pipeline defines minimum acceptable criteria for data input and performs a PASS/FAIL assessment of molecule quality. The whole genome coverage recommendation provided is a safe minimum shown to confidently resolve targeted regions. Criteria are shown in Table 4.

Table 4. EnFocus™ FSHD Analysis minimum data criteria

Analysis	Pipeline	Minimum Pre-Analysis Effective Coverage (X)	Map Rate	N50 (>= 150 kbp)	Min. DNA (Gbp)
FSHD Analysis	EnFocus™ FSHD Analysis	≥75	≥70%	≥200 kbp	≥400

Finally, the hybrid scaffolding pipeline utilizes one or two *de novo* assembly output(s) to scaffold NGS contigs/scaffolds. It is normally used with plant and animal haploid NGS assemblies, frequently of genomes that are otherwise poorly characterized. The first step in a hybrid scaffolding workflow is collecting data to support *de novo* assembly to be used as input alongside NGS data. The assembly coverage recommendation is what is required to build contiguous and accurate consensus genome maps.

Table 5. *De Novo* Assembly recommended minimum input targeting downstream use with Hybrid Scaffolding, using Saphyr data and reference derived from sample-matched haploid NGS with reasonable completeness and contiguity

Pipeline	Minimum Pre-Analysis Est. Effective Coverage (X)*	Haploid genome size (Gbp)	Example Map Rate*	Min. DNA (Gbp)	Post-Analysis Effective Coverage of Reference (X)
***De Novo* Assembly**	**80**	**× 1.1**	**÷ 55%**	**= 160**	**≥70**

***Calculation of *Map Rate* and *Effective Coverage of the Reference (X)* relies on quality and completeness of the provided reference, as well as sequence similarity between reference and sample. For example, the *Map Rate* shown above (55%) would be considered low for human data against the highly complete hg19. It may be sufficient in a non-model system. Following calculation in Table 5, reduced map rates could cause an overestimation of Min DNA (Gbp) needed to meet target pre-analysis effective coverage. See example scenario 2 below.

If reference quality is poor or none is available, 100-120X raw haploid coverage can serve as minimum input target.

Genome researchers have enjoyed significant size and accuracy gains from scaffolding NGS contigs/scaffolds with *de novo* assembled Bionano maps. 80X is a recommended minimum for this application. When using a Direct Label and Stain (DLS) approach, effective coverage up to and beyond 100X has shown improved map contiguities for some plants and animals. Expect pipeline output to be highly genome-specific. If a different Est. effective coverage is desired, simply substitute a new coverage target in the calculation table above.

Resource Considerations

Collecting and/or analyzing high volumes of data incurs costs. These costs include *i.* time a chip spends occupying the Saphyr System, *ii.* data storage consumption, and *iii*. token (for Compute On Demand operations) and runtime costs of completing analysis pipelines. Runtime estimates provided by Bionano are based on our recommended coverage levels. We generally recommend not to collect more data than necessary for desired downstream application(s).

- Throughput of G2.3 chips in a Saphyr System can support collecting three datasets of ≥320 Gbp (fit for human *de novo* assembly; Table 2) every 8 hours. For targeting ≥1500 Gbp (Rare Variant Analysis; Table 3), data collection for three samples generally can take around 24 hours. Applications requiring higher data collection will take longer. For example, Bionano has observed collection of full 5000 Gbp datasets take an average of 96 hours. Any time dedicated to collecting a data volume beyond necessary coverage, poses some availability cost against collecting the next datasets.

- Larger volume datasets require larger files which consume storage space faster. Large file sizes may create difficulties with file transfer and data backup.

- Submitting higher volume datasets for bioinformatic analyses increases the computational burden. Higher

data inputs increases job runtime, imposing an availability cost on analysis goals. For analyses through Compute On Demand, higher coverage than necessary needlessly increases token cost.

If a dataset is larger than necessary for a desired application, we recommend to downsample it in Bionano Access (see **Bionano Access® Software User Guide - 30142**, and example scenario 1b).

Example Scenarios

Scenario 1a.

Research Goal: Rare Variant Analysis, targeting ≥5% variant allele fraction

In this scenario, a cancer researcher wishes to interrogate a sample for structural variants at ≥5% variant allele fraction. The researcher enters deidentified sample information into Bionano Access and sets the Throughput Target (Gbp) to 1500. The flowcell completes in 22 hours, and generates the following Molecule Quality Report (MQR):

```
-----------------------------------------------

Total DNA (>= 20 kbp):                  2213.9251 Gbp
N50 (>= 20 kbp):                        0.2250 Mbp
Total DNA (>=150 kbp):                  1527.1018 Gbp
N50 (>=150 kbp):                        0.3030 Mbp
Total DNA (>= 150 kbp & minSites >= 9): 1483.5802 Gbp
N50 (>=150 kbp & minSites >= 9):        0.3060 Mbp

-----------------------------------------------

Enzyme:                                 DLE-1
Enzyme recognition sequence:            CTTAAG
Label color:                            BNGFLGR001
Average label density (>= 150 kbp):     15.93/100 kbp
Reference:                              hg38_DLE1_0kb_0labels.cmap

-----------------------------------------------

Map rate (>= 150 kbp):          82.3%    Percentage of molecules aligned to reference
Effective coverage:             395.40x  Effective molecule coverage
```

```
Molecule integrity number:     N/A      Measure of molecule quality
Positive label variance (PLV): 6.8%     Percentage of labels absent in reference
Negative label variance (NLV): 10.7%    Percentage of reference labels absent in molecules
SiteSD:                        0.11     Constant term in sizing error relative to reference
ScalingSD:                     0.0000   Linear term in sizing error relative to reference
RelativeSD:                    0.0208   Quadratic term in sizing error relative to reference
SMin:                          0.11kbp  Minimum expected sizing error relative to reference
Base pairs per pixel:          487.66   Base pairs per pixel
...
```

Referring to Table 3, she notes that pre-analysis effective coverage has reached the minimum for her application. The research team proceeds with Rare Variant Analysis.

Scenario 1b.

Research Goal: *De Novo* Assembly with dataset from 1a

Another researcher from the lab accesses the data and would like to assemble germline structure of the genome. The researcher decides to separately analyze the raw data with the De Novo Assembly pipeline. Based on the original MQR data from Scenario 1a, the pre-analysis effective coverage is much higher than necessary for the application. To save on cost and runtime, the researcher generates a random downsample targeting 80X using Access (Table 2):

[80X pre-analysis coverage] × [3.1 hg38 length] ÷ [0.823 Map Rate] = [**301** Total DNA (>=150 kbp)]

The researcher confirms desired coverage in the downsampled dataset:

```
...
-------------------------------------------------


Map rate (>= 150 kbp):         82.3%    Percentage of molecules aligned to reference
Effective coverage:            80.32x   Effective molecule coverage
...
```

and proceeds with De Novo Assembly.

Scenario 2.

Research Goal: *De Novo* Assembly to support Hybrid Scaffolding a small plant genome

An agricultural center wants to improve the reference genome of a novel plant species. The NGS assembly that will

be scaffolded is not available at the time Bionano data is collected. However, reference sequences are available for related plants. The most similar one is chosen for *In Silico* Digestion (see **Bionano Access® Software User Guide – 30142**). The Files Summary produced is assessed for site density with DLE-1:

```
Channel 1 site density (sites/100kbp): 26.0
Channel 1 estimated label density (labels/100kbp) for Saphyr:    19.2
Channel 1 estimated label density (labels/100kbp) for Irys:      16.8
```

As site density falls within recommended 8 – 31 / 100kbp range for direct labeling - the cmap is added to references in Access, and a technician labels plant DNA with DLE-1.

The researchers expect a haploid genome size of 850 Mbp and decide to target 110X raw coverage. They enter sample information into Bionano Access and set the Throughput Target (Gbp) to 94. The flowcell completes in five hours, and generates the following Molecule Quality Report (MQR):

```
-------------------------------------------------


Total DNA (>= 20 kbp):                  225.3994 Gbp
N50 (>= 20 kbp):                        0.1160 Mbp
Total DNA (>=150 kbp):                  94.5186 Gbp
N50 (>=150 kbp):                        0.2016 Mbp
Total DNA (>= 150 kbp & minSites >= 9): 93.3931 Gbp
N50 (>=150 kbp & minSites >= 9):        0.2032 Mbp


-------------------------------------------------


Enzyme:                                 DLE-1
Enzyme recognition sequence:            CTTAAG
Label color:                            BNGFLGR001
Average label density (>= 150 kbp):     18.19/100 kbp
Reference:                              similar-plant_DLE1_0kb_0labels.cmap


-------------------------------------------------


Map rate (>= 150 kbp):         29.9%    Percentage of molecules aligned to reference
Effective coverage:            32.85x   Effective molecule coverage
Molecule integrity number:     N/A      Measure of molecule quality
Positive label variance (PLV): 18.3%    Percentage of labels absent in reference
```

```
Negative label variance (NLV): 23.4%    Percentage of reference labels absent in molecules
SiteSD:                        0.10     Constant term in sizing error relative to reference
ScalingSD:                     0.0000   Linear term in sizing error relative to reference
RelativeSD:                    0.0245   Quadratic term in sizing error relative to reference
SMin:                          0.18kbp  Minimum expected sizing error relative to reference
Base pairs per pixel:          500.39   Base pairs per pixel
```

The map rate of the sample to this reference is poor (<30%), with positive and negative label variance both rather high. The label density is within ~15% of expectation. These factors suggest labeled DNA quality may be relatively good, with sample-reference mismatch causing low map rate and coverage estimate problems. Sequence dissimilarity between sample and reference, and/or poor reference assembly contiguity, can cause these issues in non-model systems.

The operator may continue to collect data in the flowcell or begin analysis with what he has collected. To account for the uncertainty, the operator decides to do both. A De Novo Assembly is launched on the dataset; here is output from the assembly report:

```
...
Molecules aligned to the reference:
Total number of molecules aligned  :        310769
Effective coverage of reference (X):        32.924
Average confidence                 :        21.1

...

Molecules aligned to the assembly:
Total number of molecules aligned  :    857354
Fraction of molecules aligned      :         0.891
Effective coverage of  assembly (X):        82.489
Average confidence                 :        36.2
...
```

The researchers refer to **Assembly Report Guidelines – 30255** to evaluate their result, and observe that molecule alignment against the assembly is much improved over alignment to the reference. Also, the assembly size was consistent with their expectation. They note that this particular reference is a poor guide for the sample's map rate and pre-analysis effective coverage. The completed assembly looks adequate, and more data is incoming from the running flowcell if added coverage is desired. In the absence of a good reference, they may also consider running the de novo assembly with the pre-assembly option enabled.

Scenario 3.

Research Goal: *De Novo* Assembly in human datasets to support *Variant Annotation: trio*. One dataset map rate is lower than expected.

In this example, a researcher wants to analyze germline SVs in a child with apparent genetic disease. The researcher plans a trio analysis of the proband and both parents. The lab draws three blood samples, which are labeled with DLE-1. The researcher adds a G2.3 chip to the experiment and populates the three flowcells with deidentified mother/father/proband information. Throughput Target (Gbp) is set to 320 for each. The chip completes in six hours, and generates the following three map rates and pre-analysis coverages:

```
...
-------------------------------------------------

Map rate (>= 150 kbp):         85.8%    Percentage of molecules aligned to reference
Effective coverage:            92.52x   Effective molecule coverage
...
-------------------------------------------------

Map rate (>= 150 kbp):         83.3%    Percentage of molecules aligned to reference
Effective coverage:            89.90x   Effective molecule coverage
...
-------------------------------------------------

Map rate (>= 150 kbp):         71.9%    Percentage of molecules aligned to reference
Effective coverage:            75.21x   Effective molecule coverage
...
```

The Saphyr operator observes that one samples finished below the recommended 80X effective coverage minimum. Its map rate is somewhat lower than expectation, which reduces input coverage as per calculation in Table 2. During data collection, its map rate stabilized lower than the other two samples.

The Throughput Target was adjusted accordingly. The flowcell was rehydrated and run by itself to a new total Throughput Target (Gbp) of 400. The additional runtime completes in 90 minutes, and the run generates the following total dataset map rate and pre-analysis coverage:

```
...
-------------------------------------------------
```

```
Map rate (>= 150 kbp):          71.8%    Percentage of molecules aligned to reference
Effective coverage:             93.70x   Effective molecule coverage
...
```

With all three datasets now ≥80X effective pre-analysis coverage, the research team is satisfied all SVs annotated in the trio will have been generated with comparable sensitivity. They begin analyses with the *De Novo* Assembly and Variant Annotation Pipelines.

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