Table of Contents

The NGS Quality Workgroup developed these documents and tools as examples for use by nextgeneration sequencing laboratories. These documents and tools are not controlled files; format and content must be modified as needed to meet the document control, quality management system or regulatory requirements within your laboratory. It is the responsibility of the laboratory to take any necessary actions to ensure the information within these documents remains applicable.

1.0 Purpose

1.1 This document provides quality control (QC) guidance for nucleic acid sequencing using the Illumina technology. The guidance takes into account specific QC checkpoints between laboratory processes to ensure each step is completed correctly, with high confidence, and to generate quality data metrics that are informative for downstream bioinformatics processes.

1.2 The quality of nucleic acid extraction and manipulation, fractionation and size selection, and library preparation affects fragment size uniformity and library diversity, which is important for achieving complete and even coverage of the total nucleic acid to be sequenced. Gaps resulting from poor sample preparation cannot be corrected downstream by error correction methods employed by some sequencing technologies. In addition, quality scores do not reflect errors introduced during sample preparation, as the sequencing signal will appear clean and error-free. The maximal achievable accuracy of most sequencing platforms is limited by the sample accuracy.

1.3 NOTES:

a. The expected results included are based on standard NGS methods for Illumina sequencing at the time of document development. The advancement of new methods and technologies may allow for successful sequencing with QC results differing from those listed in this document.

Automated systems that perform multiple process steps during operation, or generate unique sample preparations that are difficult to analyze using conventional QC methods, will still require QC checkpoints. Custom procedures may need to be developed to satisfy the recommendations listed below

2.0 NGS QC Checkpoints

The following sections correspond to the process steps prior to sequencing, as outlined in **Figure 1** (see **Appendix A** for a detailed process map

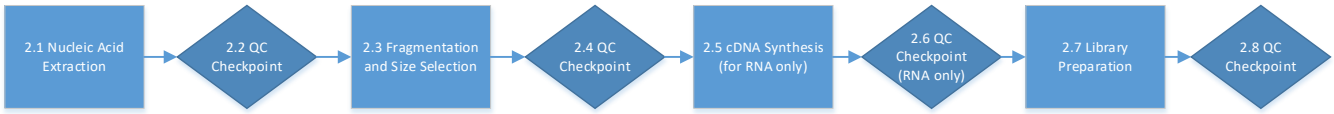


Figure 1. NGS QC Checkpoints for Illumina Workflows

1.1. Nucleic Acid Extraction

a. The Illumina sequencing workflows utilize either DNA or RNA as starting material. High quality nucleic acid purification is essential for obtaining accurate NGS data. The extraction method depends greatly on the sample type and matrices involved.

1.2. Post-Extraction Nucleic Acid QC Checkpoint

a. Nucleic acid purity and concentration should be quantitated after extraction to ensure the preparation is primarily the nucleic acid the user is analyzing (dsDNA, RNA, or cDNA) and mostly free of contamination. Quantitation of purity and concentration is highly recommended for all applications.

b. Purity is measured spectrophotometrically as the ratio of absorbance measurements at 280 nm and 260 nm. The NanoDrop instruments and procedures are commonly used to measure absorbance of nucleic acid samples and to quantitate purity. For accurate quantitation of nucleic acid, fluorescent dyes are used to complex with select forms of nucleic acid (dsDNA and RNA)

and their concentration is measured at a particular fluorometric emission and excitation wavelength.

c. Quantitation of dsDNA and RNA concentration are commonly completed with fluorometric instruments and methods such as Qubit or Quant-iT. In addition, real-time or quantitative PCR (qPCR) can also be used for very sensitive quantitation of dsDNA (e.g., KAPA qPCR). This is much more time consuming than fluorometric analysis described above and is usually not used until after library preparations are made, where very sensitive quantitation of pooled samples at low concentrations is crucial for successful sequencing.

d. In some cases, instruments that perform complete electrophoretic separation of nucleic acid such as the TapeStation, Bioanalyzer, or BluePippin can also measure absorbance and fluorescence of the electrophoretically separated nucleic acid. Hence, these instruments can measure nucleic acid purity and concentration, and can be used in lieu of the NanoDrop and Qubit or Quant-iT, but are more costly and time-consuming to perform. This is not recommended for QC checks where qPCR is recommended. In addition, these instruments provide assessments of nucleic acid integrity by visualizing the abundance and size range of fragments. This additional information can assist the user in understanding sample quality such as DNA sample integrity and provide insight into any issues identified during extraction. Conventional gel electrophoresis may also be used to assess nucleic acid integrity and sample quality, but does not quantitatively measure purity and concentration.

e. DNA Quality Check

i. Purity and concentration of the isolated dsDNA should be quantitated after extraction to ensure successful extraction, as well as prior to use if not used immediately.

ii. dsDNA Purity: Spectrophotometric measurement of absorbance ratio (A_{260}/A_{280}) should be used to quantitate purity of the DNA sample (e.g., NanoDrop). Although these methods also have the ability to quantitate DNA, the measurements are frequently affected by any present RNA or other biomolecules absorbing in the UV range and should not be used as quantitation for NGS library preparation. The Thermo Scientific T009-Technical Bulletin on 260/280 and 260/230 ratios contains additional information on purity analysis.

- **Expected Results:** A_{260}/A_{280} = 1.8-2.0; a ratio of ~1.8 is generally accepted as "pure" for DNA.
- If the ratio is appreciably lower, it may indicate the presence of protein, aromatic chemicals such as phenol, or other contaminants that absorb strongly at or near 280 nm.

iii. dsDNA Concentration: Fluorescent measurements using a particular fluorescent dye and fluorimeter should be used to measure dsDNA. These methods, such as Qubit and Quant-iT, are able to quantitate dsDNA exclusively, even in the presence of contaminating RNA and other single-stranded DNA.

- **Expected Results:** Concentration of preparation should be > 500 ng in a 20 -100 µL sample.
- Be aware that if any contaminating DNA or other reagents are present that also generates a signal at the same emission/excitation wavelengths of the fluorometric assay, this will result in a false positive reading.

iv. Alternative methods for quantitating purity, concentration, and assessing dsDNA quality: Electrophoretic instruments designed for NGS (e.g. TapeStation, Bioanalyzer, BluePippin) in

some cases can be used to quantitate purity and concentration of DNA (read vendor instructions) and can be used in lieu of methods described in **2.2.5.b** and **2.2.5.c**. In addition, this type of analysis can also provide qualitative analysis on DNA sample integrity (DNA quality). Conventional gel electrophoresis can only be used to visually assess DNA quality and does not quantitate purity and concentration.

f. RNA Quality Check

i. Purity and concentration of the isolated RNA should be quantitated after extraction to ensure successful extraction, as well as prior to use if not used immediately.

ii. RNA Purity: see **2.2.5.b**, which also applies to RNA, except a ratio of ~2.0 is generally accepted as "pure" for RNA.

iii. RNA Concentration: Fluorescent measurements using a particular fluorescent dye and fluorimeter should be used to measure RNA. Fluorescent dyes that are selective for RNA can be used, such as Ribogreen.

- **Expected Results:** Concentration of preparation should be > 500 ng in a 20-100 µL sample.
- Be aware that if any contaminating RNA or other reagents are present that also generates a signal at the same emission/excitation wavelengths of the fluorometric assay, this will result in a false positive reading.

iv. Alternative Methods for quantitating purity, concentration, and assessing RNA Quality: Electrophoretic instruments designed for NGS (e.g. TapeStation, Bioanalyzer, BluePippin) in some cases can be used to quantitate purity and concentration of RNA (read vendor instructions) can be used in lieu of methods described in **2.2.6.b** and **2.2.6.c**. Conventional gel electrophoresis can only be used to visually assess RNA quality and not quantitate purity and concentration.

1.3. Fragmentation and Size Selection

a. The fragmentation process shears the gDNA using one of three general techniques, depending on the downstream application: mechanical (e.g. Covaris), nebulization, or enzymatic (e.g. NEB Fragmentase, tagmentation).

1.4. Fragmentation and Size Selection QC Checkpoint

a. The success of fragmentation and size selection is best confirmed using an electrophoretic instrument designed for NGS. These instruments provide gel images and electropherograms, which are important qualitative and quantitative measurements of median fragment size and distributions of fragments sizes within the sample. The results provide important information on quality of fragmentation, especially during method development or processing novel and unknown sample types. Conventional gel electrophoresis can be used to qualitatively assess fragmentation and size selection, but does not provide comprehensive and quantitative analysis, such as electropherograms.

i. Expected Results: The electropherogram/gel band should reveal a single peak/band of desired size with no tailing and excessive broadening per laboratory specifications.

ii. Several documents from the electrophoretic instrument vendor manual give examples on good and poorly fragmented samples and provide insight into the root cause of inconsistencies and errors.

1.5. Synthesis of cDNA from mRNA in a total RNA preparation

a. After extraction, target RNA may be enriched from total RNA, fragmented (optional), and then converted to complementary DNA (cDNA) before library preparation. This is to increase stability of the sample as well as allow amplification. RNA to cDNA conversion is done using reverse transcriptase PCR.

1.6. cDNA Synthesis QC Checkpoint

a. Refer to **2.2** for quantitation of purity and concentration.

1.7. Library Preparation

a. There are many library preparation kits available specific to the kind of sequencing and downstream application. Generally, the fragments of nucleic acid become fused with adaptors recognized by the sequencing chip followed by PCR amplification.

1.8. Library Preparation QC Checkpoint

a. It is recommended that libraries are quantified prior to pooling and loading into the sequencer to ensure optimum cluster densities across every lane of every flow cell are achieved. Libraries can be quantified using a fluorometric measurement or by real-time PCR (e.g., KAPA qPCR). In addition, it is recommended to verify the size of your fragments and check for template size distribution through electrophoretic instrument analysis optimized for NGS. Conventional gel electrophoresis can also be used as a qualitative visualization of fragment size, distribution, and purity.

b. **Expected Results:** Library concentration > 1mM. The electropherogram/gel band analysis should reveal a single peak/band of desired size with no tailing and excessive broadening per laboratory specifications.

3.0 Related Documents

NOTE: Always check the Illumina website for updates and recent versions of guides.

Document Title	Document No.
Agilent 2200 TapeStation Assay	*Specify number*
Bioanalyzer DNA 7500 and DNA 12000 Assay	*Specify number*
BluePippin DNA Size Selection System	*Specify number*
Fragment Analyzer Assay	*Specify number*
KAPA Library Quantification for Illumina Platforms	*Specify number*
KAPA Library Quantification for Ion Torrent Platforms	*Specify number*
NanoDrop Nucleic Acid Quantitation Assay	*Specify number*
Quant-it Nucleic Acid Quantitation Assay	*Specify number*
Qubit dsDNA Quantitation Assay	*Specify number*
Qubit RNA Quantitation Assay	*Specify number*
ThermoScientific T009-Technical Bulletin	*Specify number*

4.0 Appendices

Appendix A – Illumina NGS QC Checkpoints Process Map

Appendix B – Illumina NGS QC Checklist

3.0 Revision History

Rev #	DCR #	Change Summary	Date

4.0 Approval

Approved By: _____ Date: _____

Author

_____

Print Name and Title

Approved By: _____ Date: _____

Supervisor

_____

Print Name and Title

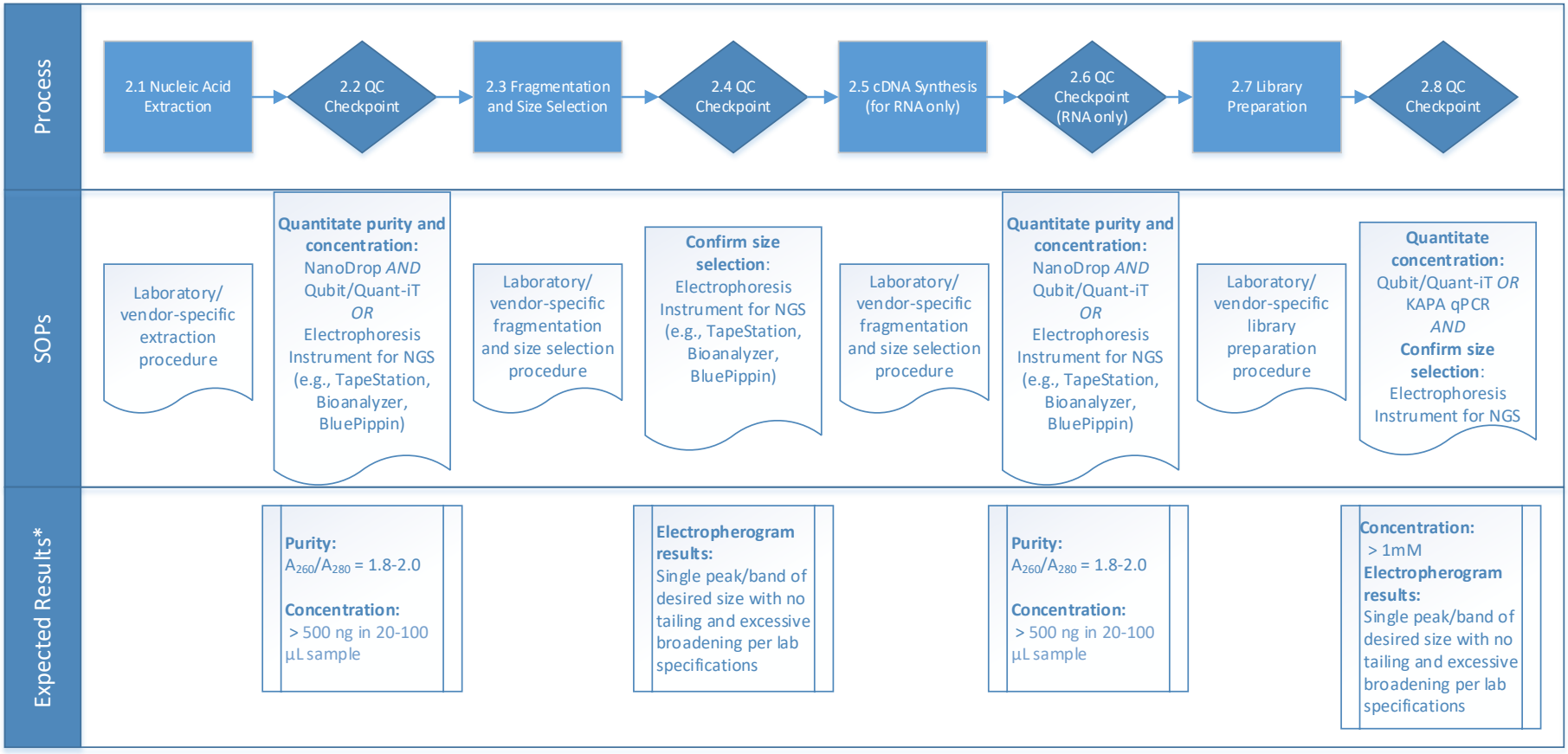
Approved By: _____ Date: _____

Quality Manager

_____

Print Name

Appendix A – Illumina NGS QC Checkpoints Process Map



Appendix B – Illumina NGS QC Checklist

QC Checkpoint (Process Step)	Method (SOPs)	Expected Results*
Post-Extraction Nucleic Acid (2.2) Quantitate purity and concentration	Purity (choose one): ☐ NanoDrop Nucleic Acid Quantitation Assay ☐ Other _____ *AND* Concentration (choose one): ☐ Qubit dsDNA or RNA Quantitation Assay ☐ Quant-iT Assay ☐ Other _____ *OR* Electrophoresis Instrument for NGS (choose one): ☐ TapeStation Assay ☐ Bioanalyzer Assay ☐ BluePippin Assay ☐ Other _____	Purity: A_{260}/A_{280} = 1.8-2.0 Concentration: > 500 ng in a 20-100 µL sample
Fragmentation and Size Selection (2.4) Confirm size selection	Electrophoresis Instrument for NGS (choose one): ☐ TapeStation Assay ☐ Bioanalyzer Assay ☐ BluePippin DNA Size Selection Assay ☐ Other _____	Electropherogram results: Single peak/band of desired size with no tailing and excessive broadening per lab specifications

QC Checkpoint (Process Step)	Method (SOPs)	Expected Results*
cDNA Synthesis* (2.6) Quantitate purity and concentration **for RNA sample only*	Purity (choose one): ☐ NanoDrop Nucleic Acid Quantitation Assay ☐ Other _____ *AND* Concentration (choose one): ☐ Qubit dsDNA or RNA Quantitation Assay ☐ Quant-iT Assay ☐ Other _____ *OR* Electrophoresis Instrument for NGS (choose one): ☐ TapeStation Assay ☐ Bioanalyzer Assay ☐ BluePippin DNA Size Selection Assay ☐ Other _____	Purity: A_{260}/A_{280} = 1.8-2.0 Concentration: > 500 ng in a 20-100 µL sample
Library Preparation (2.8) Quantitate concentration and confirm size selection	Concentration (choose one): ☐ Qubit dsDNA or RNA Quantitation Assay ☐ Quant-iT Assay ☐ KAPA qPCR ☐ Other _____ *AND* Electrophoresis Instrument for NGS (choose one): ☐ TapeStation Assay ☐ Bioanalyzer Assay ☐ BluePippin DNA Size Selection Assay ☐ Other _____	Concentration: > 1 mM Electropherogram results: Single peak of desired size with no tailing and excessive broadening per lab specifications

*The expected results included are based on standard NGS methods in use at the time of document development. The advancement of new methods and technologies may allow for successful sequencing with QC results differing from those listed in this document.

1.0 Purpose

This document provides quality control (QC) guidance for the analysis of nucleic acid next generation sequencing (NGS) data using bioinformatics. Following the generation of this NGS data, this guidance should be utilized with the analytical techniques used to process this data. The guidance takes into account specific QC checkpoints between computational processes to ensure each step is completed correctly, with high confidence, and to generate quality data metrics that yield an informative study.

QC checkpoints are necessary at several stages of bioinformatics analysis including evaluation of run metrics, filtering of raw sequences, alignment/assembly, and characterization stages. These steps ensure the sequence data meets standards for analysis, allows removal of low quality reads, and reduces false negatives and positives. This guidance also aims to promote standardized best practice measures in order to improve reproducibility of results. Due to the diverse and rapidly advancing number of pipelines used in bioinformatics, this guidance document will describe the general steps that should undergo QC.

2.0 Related Documents

NOTE: Always check for updates and recent versions of guides.

Title	Document Control Number
Sequencing QC SOP	
Pre-Analysis QC SOP	
Assembly QC SOP	

3.0 Bioinformatics QC Checkpoints

The following sections correspond to the process steps involved in bioinformatics for NGS, as outlined in the figure below (see Appendix A for a detailed process map).

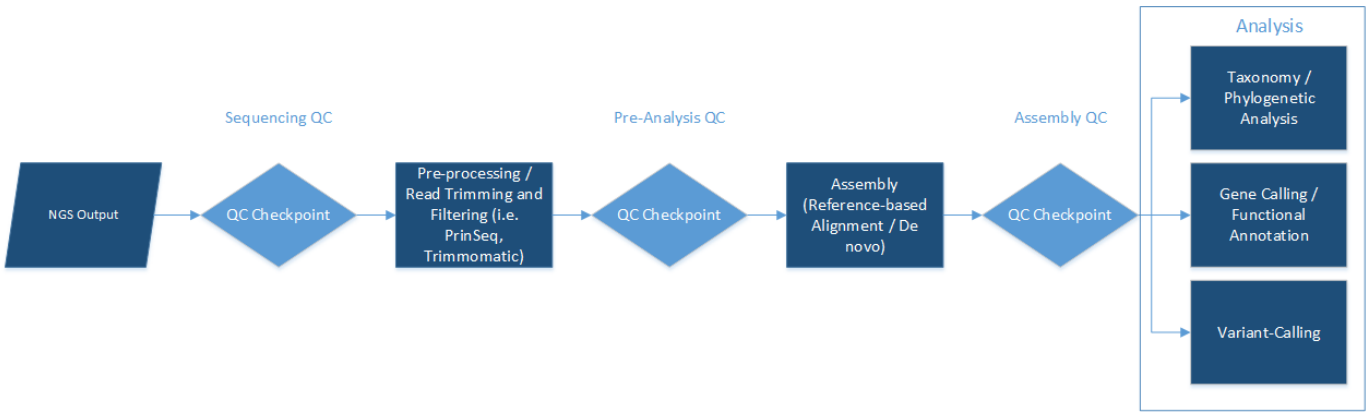


Figure 1. Bioinformatics Checkpoints

3.1. NGS Output QC- Initial Filter and Sequencing Run QC: Run metrics from the sequencer are evaluated using Illumina Sequencing Analysis Viewer (SAV). The key metrics involved in this step include the following:

a) **Cluster Density:** The density of clusters for each tile (in thousands per mm^2).

b) **% Clusters PF:** The percentage of clusters passing filter for each tile.
c) **Yield Total:** Check that the read length is what is expected for the NGS platform and chemistry used for the sequenced organism.
d) **% Aligned (PhiX):** The percentage of the passing filter clusters that aligned to the PhiX genome.
e) **Accuracy of base (%Q>/=30):** Base calling accuracy describes the probability that the sequencer incorrectly assigned a nucleotide base. This is most commonly given as Q score which is calculated as: $Q = - 10 \log_{10} P$, where P is the probability of error. Data with low Q scores may mean the data is unusable for further analysis. Q >30 is a standard threshold, which corresponds to 99.9% base calling accuracy.

3.2. Pre-Analysis QC – Trimming, Filtering and Quality Assessment using FASTQC: This stage of QC follows after the generation of a FASTQ file. This guidance should be utilized to assess the quality of sequence data prior to assembly and further analysis. This step includes the use of a quality based trimming and filtering tool such as PrinSeq and quality assessment using a tool such as FASTQC. The key metrics assessed at this stage include:

a) **Total Sequences -** A count of the total number of sequences processed. There are two values reported, actual and estimated. At the moment, these will always be the same. In the future, it may be possible to analyze just a subset of sequences and estimate the total number, to speed up the analysis, but since we have found that problematic sequences are not evenly distributed through a file we have disabled this for now.
b) **Filtered Sequences -** If running in Casava mode, sequences flagged to be filtered will be removed from all analyses. The number of such sequences removed will be reported here. The total sequences count above will not include these filtered sequences and will be the number of sequences actually used for the rest of the analysis.
c) **Sequence Length -** Provides the length of the shortest and longest sequence in the set. If all sequences are the same length only one value is reported.
d) **% GC -** The overall %GC of all bases in all sequences
e) **Per base sequence quality -** For each position a BoxWhisker type plot is drawn. The elements of the plot are as follows: the central red line is the median value, the yellow box represents the inter-quartile range (25-75%), the upper and lower whiskers represent the 10% and 90% points, and the blue line represents the mean quality.

3.3. Alignment and Assembly QC: At this stage, overlapping reads are aligned to create contigs and scaffolds for paired end reads. Homologous samples (i.e. bacterial isolate) are mapped to preexisting consensus genomes. Novel or heterogenous samples (i.e. isolates with no reference genome and/or metagenomics sample like stool) require *de novo* assembly. The quality of these assemblies is then evaluated using QUAST. The key metrics involved in the assessment stage for your assembly can be referenced below. Assembly joins reads that overlap into contigs (contiguous sequences). This is controlled by establishing minimum coverage, **N50**, **L50** and **minimum length of contigs** cutoff values that must be met. N50 describes a contig (contiguous sequence) length whereas L50 describes a number of contigs.

a) **Number of Contigs:** Total number of contigs of length
b) **Total Length:** Total number of bases in the assembly

c) **Minimum Coverage:** The minimum average depth of coverage and uniformity of coverage necessary for good assembly.
d) **N50 length:** A statistical measure of average length of a set of sequences. N50 is the length (in basepairs) of the smallest contig that takes the sum length of all contigs – when summing from longest to shortest – past 50% of the total size (in basepairs) of the assembly.
e) **L50 count:** The number of contigs evaluated at the point when the sum length exceeds 50% of the assembly size.
f) **Minimum length of contig:** For very large assemblies the number of contigs can be over a million and mapping reads back to contigs will take a long time. Set a minimum contig length to reduce the number of contigs that have to be incorporated into the data structure.

3.4. Reference-based Assembly QC: The metrics below are relevant to the evaluation of reference-based assembly quality only and should be used in tandem with the other metrics for assembly outlined above:

a) **Percentage of Genome Covered**: Assessed by calculating genome coverage or the average number of reads that align to the reference genome. How well the reads map to the reference genome indicates a certain level of confidence that any conclusions made downstream are reliable.
b) **Uniformity of coverage:** This refers to the distribution of coverage within specific targeted regions. Although the average coverage may meet the laboratory established threshold, the depth of coverage will vary across the genome resulting in variable accuracy across the genome. Check that there is uniformity of coverage across the regions that are sequenced. This is calculated by the variance in sequencing depth across the genome post mapping. Non-uniformity can increase rate of false positives.
c) **Choosing a reference genome** Curated reference genomes are available for some species and should be utilized when possible. These are high quality sequence data, often closed or finished genomes. Reference sequences also satisfy these requirements:
   - Genome sequences with less than 1 error per 100,000 base pairs
   - Each replicon is assembled into a single contiguous sequence with a minimal number of possible exceptions documented in the submission record
   - All sequences are complete and have been reviewed and edited
   - All known misassemblies have been resolved
   - Repetitive sequences have been ordered and correctly assembled

3.5. De Novo Contig Assembly QC: A *de novo* assembly joins reads that overlap into contigs (contiguous sequences). *De novo* assembly by definition lacks a reference sequence to use as a basis, therefore the quality of such an assembly should be evaluated using the aforementioned metrics in 2.3. which include **Minimum Coverage**, **N50**, **L50** and **Minimum length of contig**.

3.6. Analysis QC

3.6.1. Variant calling QC: The quality of variant calling is controlled by establishing the following parameters and thus, variant calling only occurs at positions that meet these requirements.

d) A non-reference base (a variant) is detected.

e) **Allele call score**: The allele call score refers to the probability of an incorrect base call (e.g. a score of 3 is equivalent to a Phred score of Q30 meaning the likelihood of an incorrect call for the base is 1 in 1000). The default Illumina setting for allele call score is >/=10.
f) **Minimum coverage**: Illumina recommends a mean coverage of 30x for DNA sequencing assuming a Phred score of Q30. Additionally, the depth at the SNPs position should be no greater than three times the chromosomal mean.
g) **Heterozygous calls:** Both alleles should have an allele call score >/= 10, and the ratio of their scores should be </=3.

4.0 Appendices

Appendix A – Bioinformatics QC Checkpoints Process Map

5.0 References

1) Andrews, S. (2010). FastQC: a quality control tool for high throughput sequence data.
2) Alexey Gurevich, Vladislav Saveliev, Nikolay Vyahhi and Glenn Tesler, QUAST: quality assessment tool for genome assemblies, *Bioinformatics* (2013) 29 (8): 1072-1075.
3) Schmieder, R., & Edwards, R. (2011). Quality control and preprocessing of metagenomic datasets. *Bioinformatics*, *27*(6), 863-864.
4) Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics, btu170. Chicago.
5) Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., ... & Pyshkin, A. V. (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of computational biology*, *19*(5), 455-477.
6) Simpson, J. T., Wong, K., Jackman, S. D., Schein, J. E., Jones, S. J., & Birol, I. (2009). ABySS: a parallel assembler for short read sequence data. *Genome research*, *19*(6), 1117-1123.
7) *De novo* bacterial genome sequencing: millions of very short reads assembled on a desktop computer. Hernandez D, François P, Farinelli L, Osterås M, Schrenzel J. *Genome Research.* 18:802-809, 2008.
8) Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature methods*, *9*(4), 357-359.
9) Ponstingl, H. (n.d.). SMALT. Retrieved May 15, 2017, from http://www.sanger.ac.uk/science/tools/smalt-0
10) MiSeq Specifications. (n.d.). Retrieved May 16, 2017, from https://www.illumina.com/systems/sequencing-platforms/miseq/specifications.html

6.0 Revision History

Rev #	DCR #	Change Summary	Date

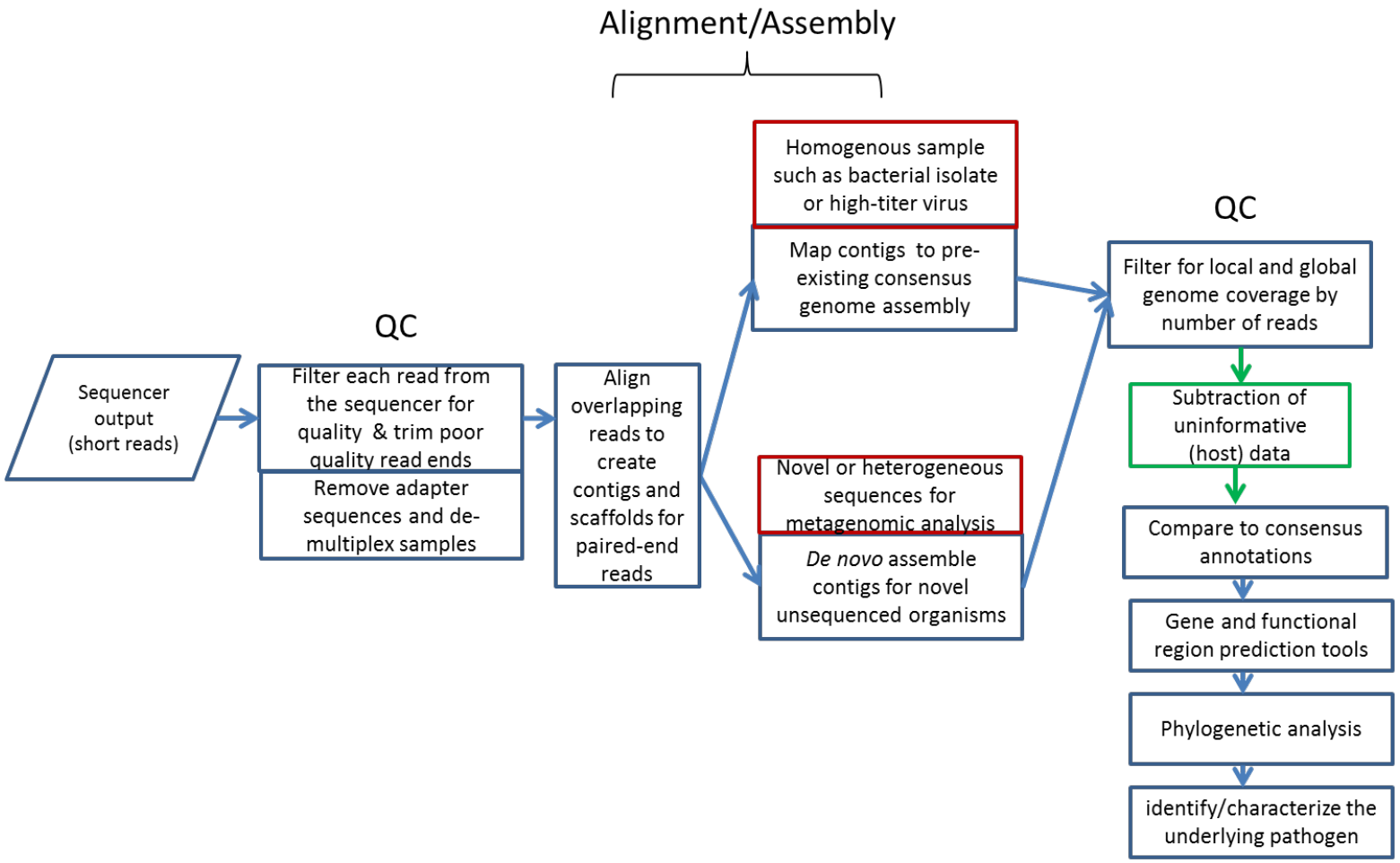
7.0 Approval

This document has been approved by the CDC CLIA Laboratory Director as the standard practice for CLIA-regulated CDC Infectious Diseases Laboratories under certificates 11D0668319 and 11D2030855.

Approved:

_____

Appendix A – Bioinformatics QC Checkpoints Process Map



1.0 Purpose

This document provides quality control (QC) guidance for the analysis of nucleic acid next generation sequencing (NGS) data. Following the generation of NGS data, this guidance should be utilized with the analytical techniques used to process sequence data. The guidance takes into account specific QC checkpoints between computational processes to ensure each step is completed correctly, with high confidence, and to generate quality data metrics that yield an informative study.

QC checkpoints are necessary at several stages of bioinformatics analysis including filtering raw read sequences, *de novo* or reference based alignment/assembly and characterization stages. These steps ensure that NGS data generated through the sequencing process meets standards for analysis through removal of low quality reads and reduction of false negatives and positives. This guidance also aims to promote standardized best practice measures in order to improve reproducibility of results.

2.0 Scope

This document provides information on sequencing QC: quality control steps to be performed on NGS data after it comes off the sequencing instrument and before Pre-Analysis QC.

3.0 Related Documents

Title	Document Control Number
Bioinformatics QC Workflows	

4.0 Responsibilities

Position	Responsibility
All Laboratory Staff	• Follow documented procedures
Team Lead	• Ensure documented procedures for data quality checks are established • Ensure documented procedures are followed
Quality Manager	• Ensure documented procedures are available to the end user • Review records of data quality checks as required

5.0 Definitions

Term	Definition
SAV	Sequence Analysis Viewer
Intensity (also referred to as P90)	The 90% percentile extracted intensity for a given image (lane/tile/cycle/channel combination). On platforms using four-channel sequencing, 4 channels (A, C, G, and T) are shown.
FWHM	The average full width of clusters at half maximum (representing their approximate size in pixels).
Corrected Intensity	The intensity corrected for cross talk between the color channels and phasing and prephasing.

Called Intensity	For a given base in a lane/tile/cycle, the average intensity for all clusters that were called as that base.
% No Calls	The percentage of clusters on a tile for which no base (N) has been called
% Base	The percentage of called (non-N) clusters for which the selected base has been called.
Signal to Noise	The signal to noise ratio is calculated as mean called intensity divided by standard deviation of noncalled intensities.
Error Rate	The calculated error rate, as determined by a spiked PhiX control sample. If a PhiX control sample is not run in the lane, this number is not available.
% Perfect Reads	The percentage of reads that align perfectly, as determined by a spiked in PhiX control sample. If a PhiX control sample is not run in the lane, this number is not available.
%Q >/= 20, %Q >/= 30	The percentage of bases with a phred or Q quality score of 20 or 30 or higher, respectively.
Median Q-Score	The median Q-Score for each tile over all bases for the current cycle. These charts are generated after the 25th cycle. This metric is best used to examine the Q-scores of your run as it progresses. The %Q30 plot can give an over simplified view due to its reliance on a single threshold.
Density	The density of clusters for each tile (in thousands per mm^2).
Desnity PF	The density of clusters passing filter for each tile (in thousands per mm^2).
Clusters	The number of clusters for each tile (in millions).
Clusters PF	The number of clusters passing filter for each tile (in millions).
% Pass Filter	The percentage of clusters passing filter.
% Phasing, % Prephasing	The average rate (percentage per cycle) at which molecules in a cluster fall behind (phasing) or jump ahead (prephasing) during the run.
% Aligned	The percentage of the passing filter clusters that aligned to the PhiX genome.
Time	The date and time the tile was processed for that cycle.
Minimum / Maximum Contrast	The 10th and 99.5th percentiles per channel of selected columns of the raw image, respectively.

6.0 Equipment

N/A

7.0 Reagents and Media

N/A

8.0 Supplies, Other Materials

N/A

9.0 Safety Precautions

N/A

10.0 Sample Information / Processing

Upon completion of the NGS run, transfer data to Isilon. *(Specify your laboratory data storage location here.)*

11.0 Quality Control

N/A

12.0 Workflow Chart

N/A

13.0 Process Overview

N/A

14.0 SAV Procedure

14.1 Once the sequencing run is complete load data into SAV

- **i.** Double click the Illumina Sequencing Analysis Viewer Software desktkop shortcut, or go to C:\Illumina\Illumina Sequencing Analysis Viewer Software and doublc click Sequencing Analysis Viewer Software.exe.
- **ii.** The Sequencing Analysis Viewer Software opens.
- **iii.** Click the tab containing the appropriate query information.
- **iv.** In the Run Folder field, copy the folder location or click Browse to select a run folder. Make sure to highlight the run folder and not the parent folder or any folder/file inside the run folder.
- **v.** Click Refresh. The SAV Software starts loading data showing quality metrics for that run.

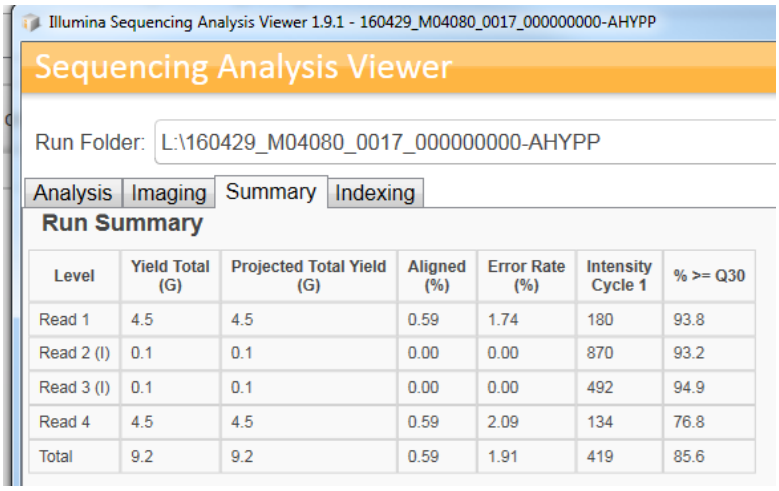
14.2 Under the Summary Tab (see Figure 1), review the following metrics:

a. In the top table review:

Metric	Expected Value
Level	The sequencing read level
Yield Total	MiSeq Reagent Kit V2: Output Max: 7.5 Gb 2.25 Gb at 2 x 300 bp 4.5 Gb at 2 x 150 bp 7.5 Gb at 2 x 250 bp MiSeq Reagent Kit V3: Up to 15 Gb at 2 x 300 bp Up to 3.75 Gb at 2 x 75 bp
Projected Total Yield	The projected number of bases expected to be sequenced at the end of the run, which is updated as the run progresses.
Aligned	The percentage that aligned to the PhiX genome i.e. if 1% PhiX was the initial input quantity, the aligned value should be equal to 1% or below.
Error Rate	The calculated error rate of the reads that aligned to PhiX.

Intensity Cycle 1	The average of the A channel intensity measured at the first cycle averaged over filtered clusters.
%Q >/= 30	MiSeq Reagent Kit V2: > 90% bases higher than Q30 at 1 × 36 bp > 90% bases higher than Q30 at 2 × 25 bp > 80% bases higher than Q30 at 2 × 150 bp > 75% bases higher than Q30 at 2 × 250 bp MiSeq Reagent Kit V3: > 85% bases higher than Q30 at 2 × 75 bp > 70% bases higher than Q30 at 2 × 300 bp

Figure 1. Sequence Analysis Viewer version 1.9.1 Summary Tab



Level	Yield Total (G)	Projected Total Yield (G)	Aligned (%)	Error Rate (%)	Intensity Cycle 1	% >= Q30
Read 1	4.5	4.5	0.59	1.74	180	93.8
Read 2 (I)	0.1	0.1	0.00	0.00	870	93.2
Read 3 (I)	0.1	0.1	0.00	0.00	492	94.9
Read 4	4.5	4.5	0.59	2.09	134	76.8
Total	9.2	9.2	0.59	1.91	419	85.6

a. In the Read Table (see Figure 2), review

Metric	Expected Value
Organism	*(Specify the organism here.)*
Range	*(Specify the range here.)*
Tiles	Standard Flow Cell in MiSeq Reagent Kit v3 (38 tiles) PGS Flow Cell in MiSeq Reagent Kit v3 (38 tiles) Standard Flow Cell in MiSeq Reagent Kit v2 (28 tiles) Micro Flow Cell in MiSeq Reagent Micro Kit v2 (8 tiles) Nano Flow Cell in MiSeq Reagent Nano Kit v2 (4 tiles)
Density	Kit V2: Loading Concentration: 10-15 pM Cluster Density:1000-1200 k/mm^2 Kit V3: Loading Concentration: 15 pM Cluster Density:1200-1400 k/mm^2
Clusters PF	80-95%
Phas./Prephas.	<0.25
Reads	Kit V3: 25M Kit V2 : 15M Micro Kit V2: 4M Nano Kit V2: 1M

Reads PF	Kit V2: Single Reads: 12-15M Paired End Reads: 24-30M Kit V3: Single Reads 22-25M, Paired End Reads: 44-50M
%Q >/= 30	Kit V2: > 90% bases higher than Q30 at 1 × 36 bp > 90% bases higher than Q30 at 2 × 25 bp > 80% bases higher than Q30 at 2 × 150 bp > 75% bases higher than Q30 at 2 × 250 bp Kit V3: > 85% bases higher than Q30 at 2 × 75 bp > 70% bases higher than Q30 at 2 × 300 bp
Yield	Kit V2: Output Max: 7.5 Gb 2.25 Gb at 2 x 300 bp 4.5 Gb at 2 x 150 bp 7.5 Gb at 2 x 250 bp Kit V3: Up to 15 Gb at 2 x 300 bp Up to 3.75 Gb at 2 x 75 bp
Cycles Err Rated	The number of cycles that have been error-rated using PhiX, starting at cycle 1.
Aligned	The percentage that aligned to the PhiX genome.
Error Rate	The calculated error rate, as determined by the PhiX alignment. Subsequent columns display the error rate for cycles 1–35, 1–75, and 1–100.
Intensity Cycle 1	The average of the A channel intensity measured at the first cycle averaged over filtered clusters.

Figure 2. Sequence Analysis Viewer version 1.9.1 Read Table

Read 1

Lane	Tiles	Density (K/mm2)	Cluster PF (%)	Phas/Prephas (%)	Reads (M)	Reads PF (M)	% >= Q30	Yield (G)	Cycles Err Rated	Aligned (%)	Error Rate (%)	Error Rate 35 cycle (%)	Error Rate 75 cycle (%)	Error Rate 100 cycle (%)	Intensity Cycle 1
1	28	1070 +/- 25	89.42 +/- 1.19	0.045 / 0.045	20.04	17.92	93.8	4.5	250	0.59 +/- 0.02	1.74 +/- 0.11	0.11 +/- 0.01	0.14 +/- 0.02	0.20 +/- 0.02	180 +/- 11

Read 2 (I)

Lane	Tiles	Density (K/mm2)	Cluster PF (%)	Phas/Prephas (%)	Reads (M)	Reads PF (M)	% >= Q30	Yield (G)	Cycles Err Rated	Aligned (%)	Error Rate (%)	Error Rate 35 cycle (%)	Error Rate 75 cycle (%)	Error Rate 100 cycle (%)	Intensity Cycle 1
1	28	1070 +/- 25	89.42 +/- 1.19	0.000 / 0.000	20.04	17.92	93.2	0.1	0	0.00 +/- 0.00	0.00 +/- 0.00	0.00 +/- 0.00	0.00 +/- 0.00	0.00 +/- 0.00	870 +/- 65

Read 3 (I)

Lane	Tiles	Density (K/mm2)	Cluster PF (%)	Phas/Prephas (%)	Reads (M)	Reads PF (M)	% >= Q30	Yield (G)	Cycles Err Rated	Aligned (%)	Error Rate (%)	Error Rate 35 cycle (%)	Error Rate 75 cycle (%)	Error Rate 100 cycle (%)	Intensity Cycle 1
1	28	1070 +/- 25	89.42 +/- 1.19	0.000 / 0.000	20.04	17.92	94.9	0.1	0	0.00 +/- 0.00	0.00 +/- 0.00	0.00 +/- 0.00	0.00 +/- 0.00	0.00 +/- 0.00	492 +/- 30

Read 4

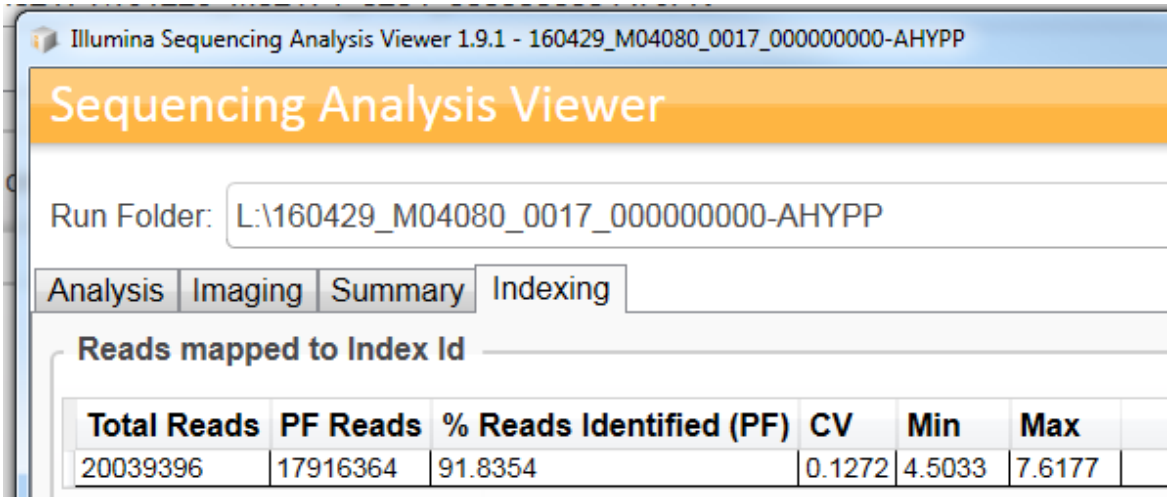
Lane	Tiles	Density (K/mm2)	Cluster PF (%)	Phas/Prephas (%)	Reads (M)	Reads PF (M)	% >= Q30	Yield (G)	Cycles Err Rated	Aligned (%)	Error Rate (%)	Error Rate 35 cycle (%)	Error Rate 75 cycle (%)	Error Rate 100 cycle (%)	Intensity Cycle 1

14.3 On the Indexing Tab (see Figure 3) select the displayed lane through the drop-down list.

a. Review the first table that contains a summary of the indexing performance for that lane.

Metric	Expected Value
Total Reads	Kit V2 : 15M Micro Kit V2: 4M Nano Kit V2: 1M
PF Reads	Kit V2: Single Reads: 12-15M Paired End Reads: 24-30M Kit V3: Single Reads 22-25M, Paired End Reads: 44-50M
% Reads Identified (PF)	The total fraction of passing filter reads assigned to an index.
CV	The coefficient of variation for the number of counts across all indexes.
Min	The lowest representation for any index.
Max	The highest representation for any index.

Figure 3. Sequence Analysis Viewerv version 1.9.1 Indexing Tab Overview



b. In the Indexing Tab (see Figure 4a and 4b) review the below:

Metric	Expected Value
Organism	*(Specify the organism here.)*
Range	*(Specify the range here.)*
Index Number	A unique number assigned to each index by SAV for display purposes.
Sample ID	The sample ID assigned to an index in the sample sheet.
Project	The project assigned to an index in the sample sheet.
Index 1 (I7)	The sequence for the first Index Read.

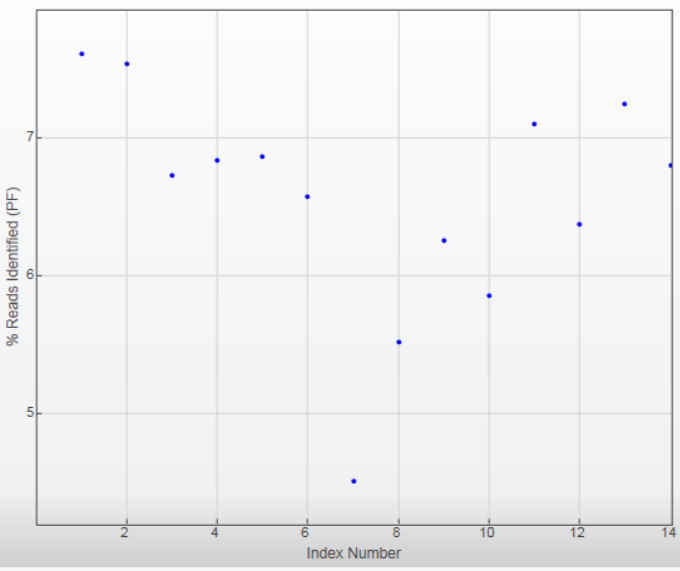
Index 2 (I5)	The sequence for the second Index Read.
% Reads Identified (PF)	The number of reads (only includes Passing Filter reads) mapped to this index.

Figure 4a. Sequence Analysis Viewer version 1.9.1 Indexing Tab Table

Index Number	Sample Id	Project	Index 1 (I7)	Index 2 (I5)	% Reads Identified (PF)
1	WHI5263A1	Elizabethkingia_outbreak_set_10	ATTCAGAA	TATAGCCT	7.6177
2	WHI5263A9	Elizabethkingia_outbreak_set_10	GAATTCGT	TATAGCCT	7.5386
3	WHI5263A2	Elizabethkingia_outbreak_set_10	ATTCAGAA	ATAGAGGC	6.7343
4	WHI5263A10	Elizabethkingia_outbreak_set_10	GAATTCGT	ATAGAGGC	6.8359
5	WHI5263A3	Elizabethkingia_outbreak_set_10	ATTCAGAA	CCTATCCT	6.8699
6	WHI5263A11	Elizabethkingia_outbreak_set_10	GAATTCGT	CCTATCCT	6.5767
7	WHI5263A4	Elizabethkingia_outbreak_set_10	ATTCAGAA	GGCTCTGA	4.5033
8	WHI5263A12	Elizabethkingia_outbreak_set_10	GAATTCGT	GGCTCTGA	5.517
9	WHI5263A5	Elizabethkingia_outbreak_set_10	ATTCAGAA	AGGCGAAG	6.2531
10	WHI5263A13	Elizabethkingia_outbreak_set_10	GAATTCGT	AGGCGAAG	5.8577
11	WHI5263A6	Elizabethkingia_outbreak_set_10	ATTCAGAA	TAATCTTA	7.1076
12	WHI5263A14	Elizabethkingia_outbreak_set_10	GAATTCGT	TAATCTTA	6.3766
13	WHI5263A7	Elizabethkingia_outbreak_set_10	ATTCAGAA	CAGGACGT	7.248
14	WHI5263A8	Elizabethkingia_outbreak_set_10	ATTCAGAA	GTACTGAC	6.7991

Figure 4b. Sequence Analysis Viewer 1.9.1 Indexing Tab Plot

% Reads Identified (PF)
7.6177
7.5386
6.7343
6.8359
6.8699
6.5767
4.5033
5.517
6.2531
5.8577
7.1076
6.3766
7.248
6.7991



14.4 Based on your MiSeq configuration, data is either stored locally or automatically transferred to the network storage.

a. If data is stored locally, after reviewing data in SAV, transfer data using an FTP-based program (e.g. WinSCP or FileZilla) to a specified directory *(include your storage location here)*.

15.0 Trending over Time

a. Several of the aforementioned values shown by SAV might indicate decreasing health of a sequencer. Ensure that you are not seeing a decrease in these values over time (provided other variables remain constant). Keep in mind that they could also be the result of poor library prep or faulty templates/kit(s). These metrics include: **Number of reads**, **Percentage >Q30**, **Error rate** and **Demultiplexing.**

16.0 Method Performance Specifications

N/A

17.0 Calculations

N/A

18.0 Reference Values, Alert Values

N/A

19.0 Interpretation of Results

a. Assess the eveness and consistency of yield across all samples.

i. Low yield in one sample and high/double yield in another with all others having consistent yield may indicate mixed tags. In this case, consult with a bioinformatician on the best way to proceed.

ii. Low yield overall may indicate an issue in library prep. Consult with a bioinformatician or prepare a new library.

b. Mean quality score should be above 30 for each sample. Run FastQC (a quality control tool for high throughput sequence data) on each sample with a mean quality score less than 30.

20.0 Results Review and Approval

N/A

21.0 Sample Retention and Storage

Store data in compliance with all applicable regulations, CDC records retention policy, and laboratory data storage procedures. *(Update to specify your laboratory's data retention and storage policy)*

22.0 References

22.1 Illumina Sequence Analysis Viewer v1.11 Part # 15066069 v02 February 2016

23.0 Appendices *(Include example screen shots of good and poor quality data applicable to your laboratory methods)*

SAV Sample Screenshots:

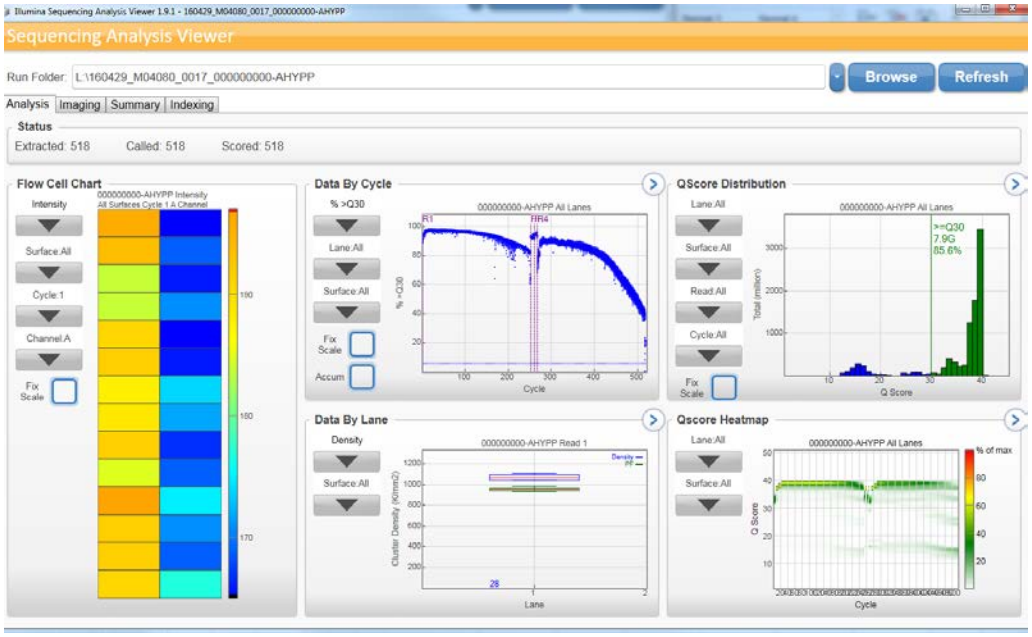


Figure A-1. Sequence Analysis Viewer 1.9.1 Analysis Tab

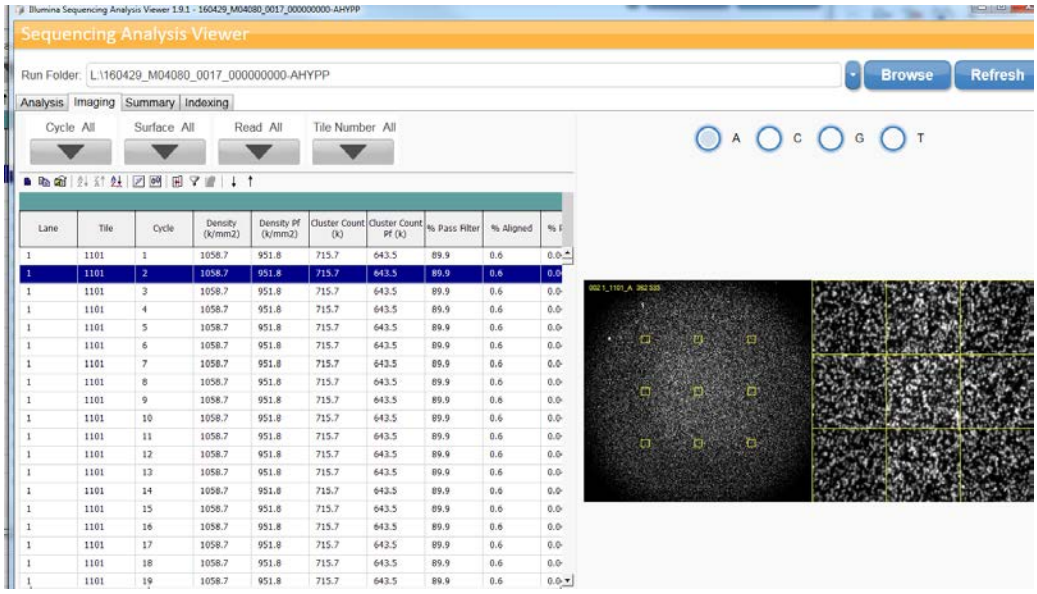


Figure A-2. Sequence Analysis Viewer 1.9.1 Imaging Tab

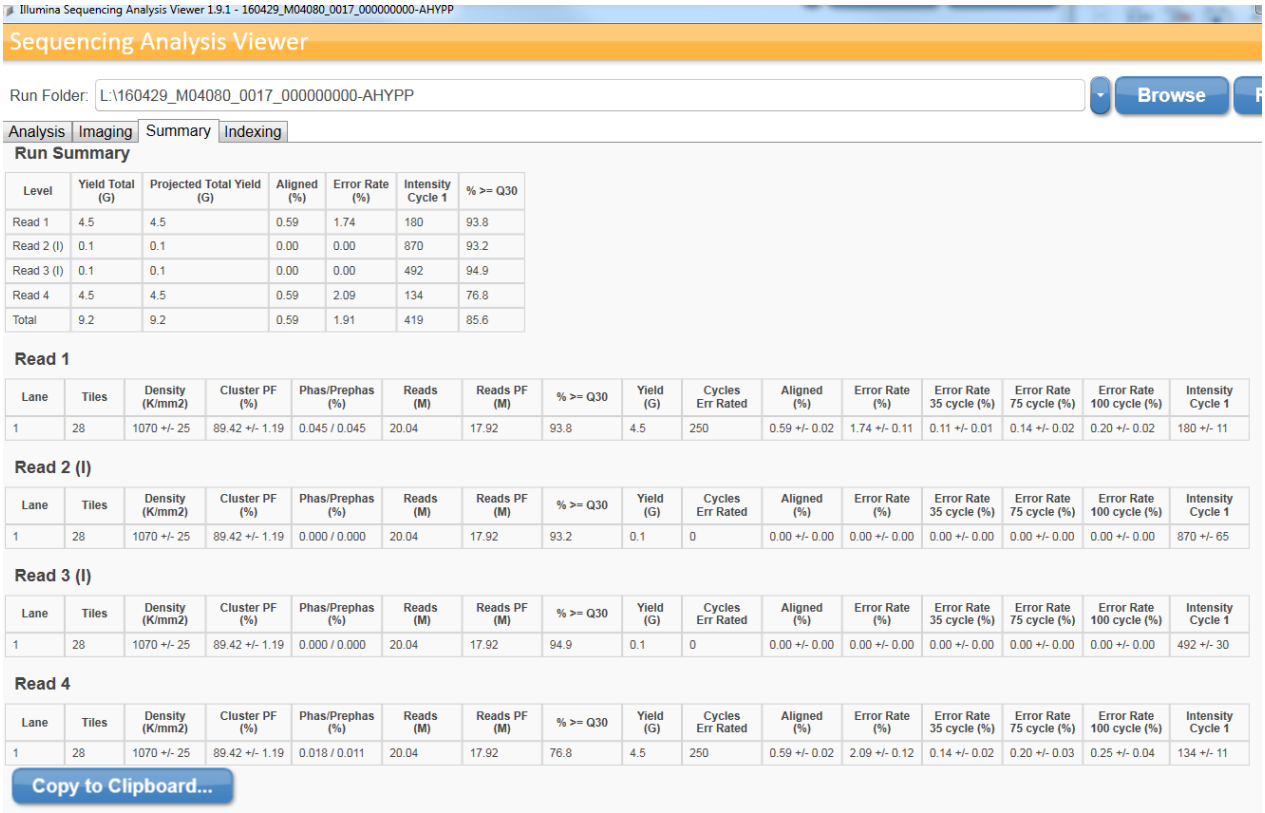


Figure A-3. Sequence Analysis Viewer 1.9.1 Summary Tab

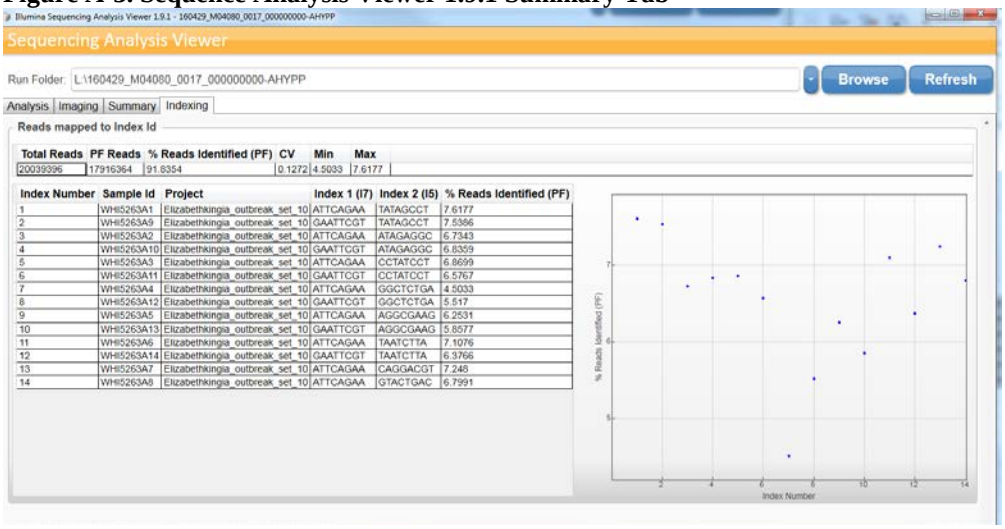


Figure A-4. Sequence Analysis Viewer 1.9.1 Indexing Tab

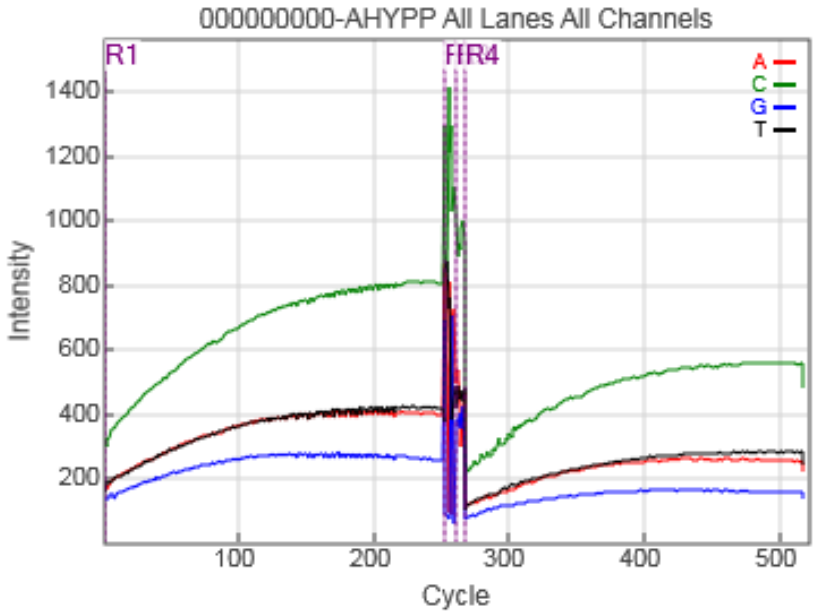


Figure A-5. Sequence Analysis Viewer 1.9.1 Intensity Plot

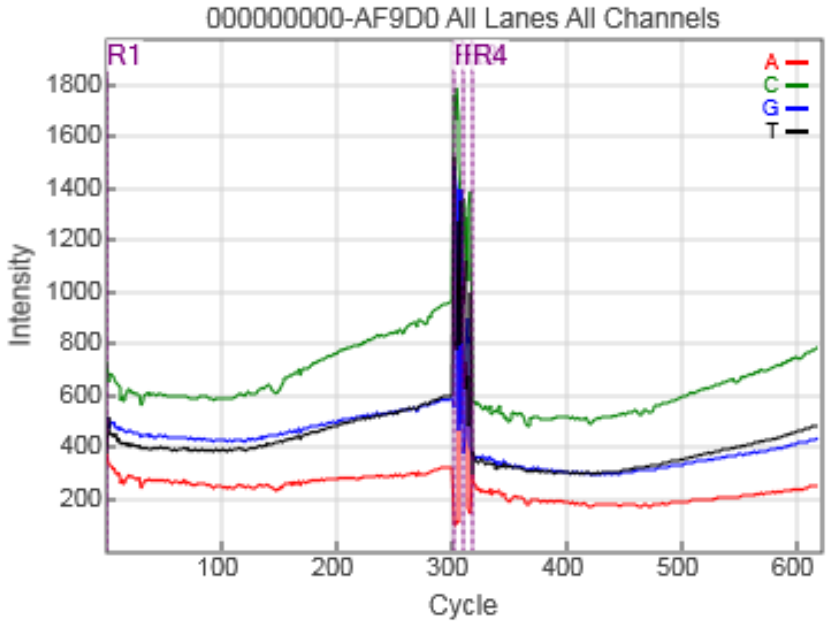


Figure A-6. Sequence Analysis Viewer 1.9.1 Intensity plot (600 cycle v3 run)

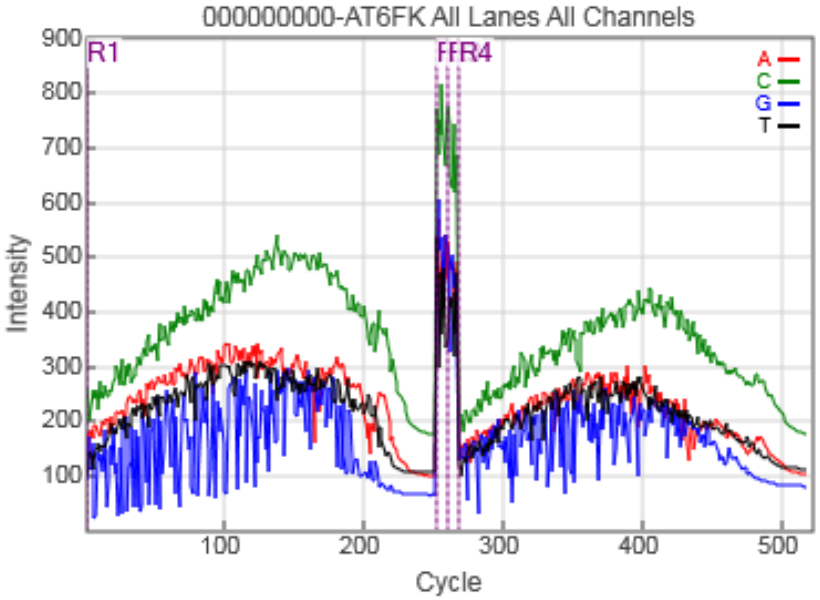


Figure A-7. Sequence Analysis Viewer 1.9.1 Intensity plot (for Amplicons)

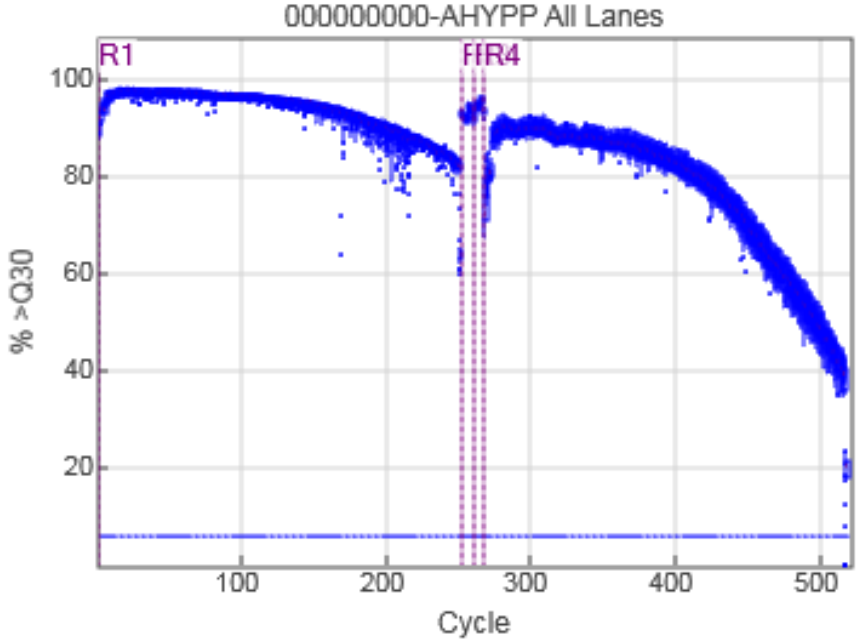


Figure A-8. Sequence Analysis Viewer 1.9.1 %Q30 plot (600 cycle v3) (Good Run)

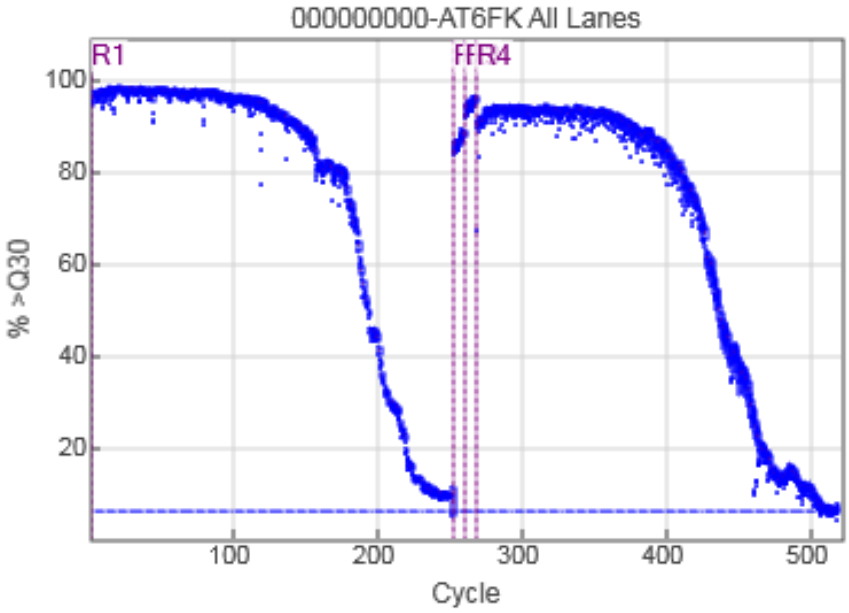


Figure A-9. Sequence Analysis Viewer 1.9.1 %Q30 plot (Low Quality Run)

%Q30 plot for 600 cycle v3 run:

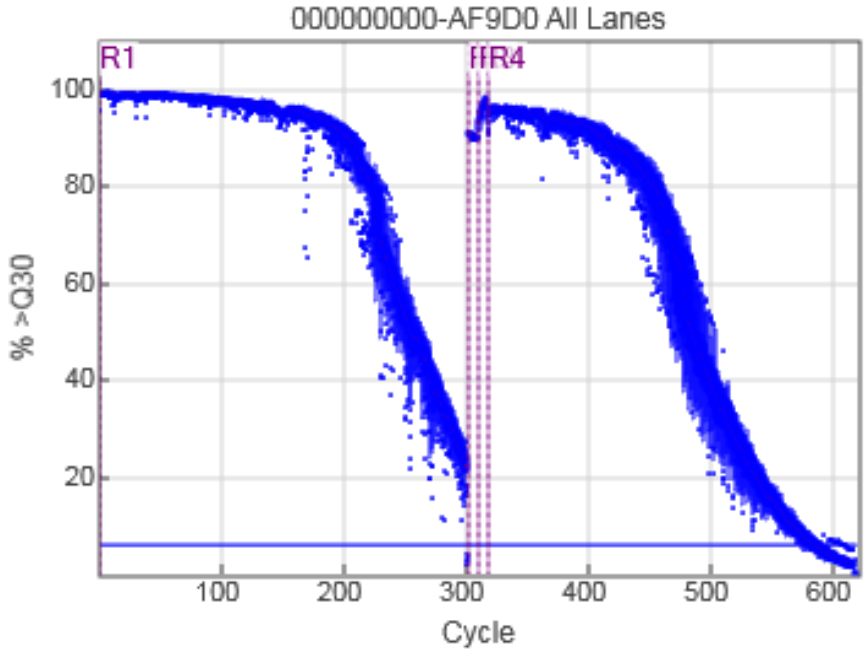


Figure A-10. Sequence Analysis Viewer 1.9.1 %Q30 plot (600 cycle v3)

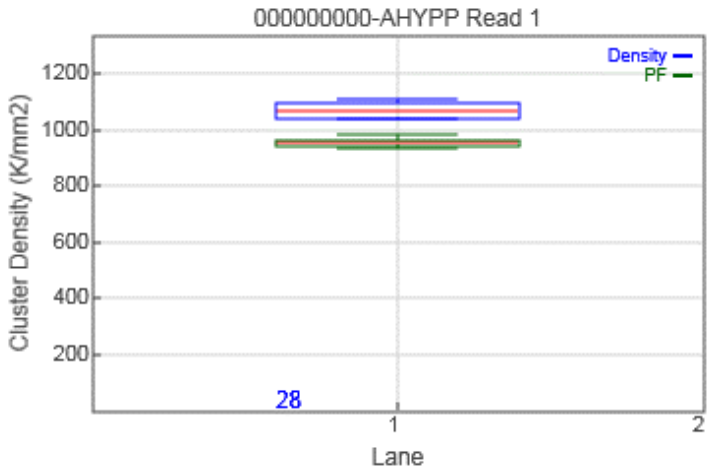


Figure A-11. Sequence Analysis Viewer 1.9.1 Cluster Density plot (Blue bar – Cluster Density, Green Bar – Cluster Pass filter)

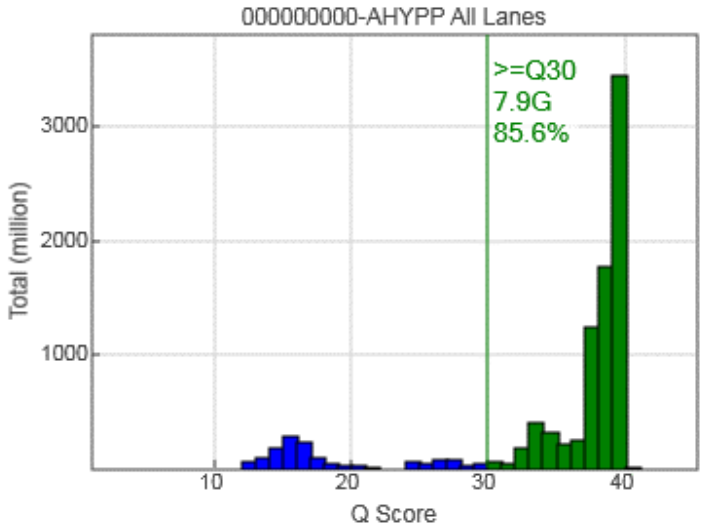


Figure A-12. Sequence Analysis Viewer 1.9.1 QScore Distribution plot

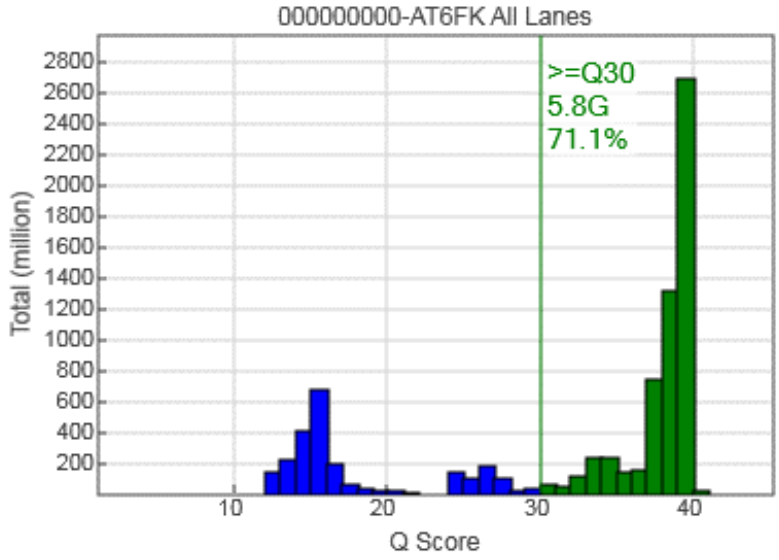


Figure A-13. Sequence Analysis Viewer 1.9.1 QScore Distribution plot (Low Quality)

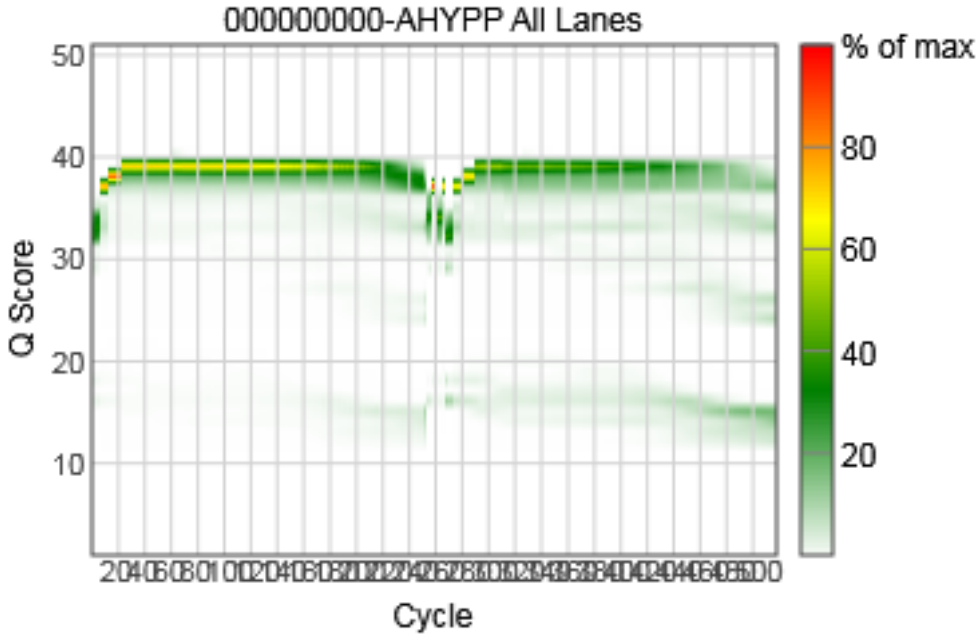


Figure A-14. Sequence Analysis Viewer 1.9.1 Qscore Heatmap

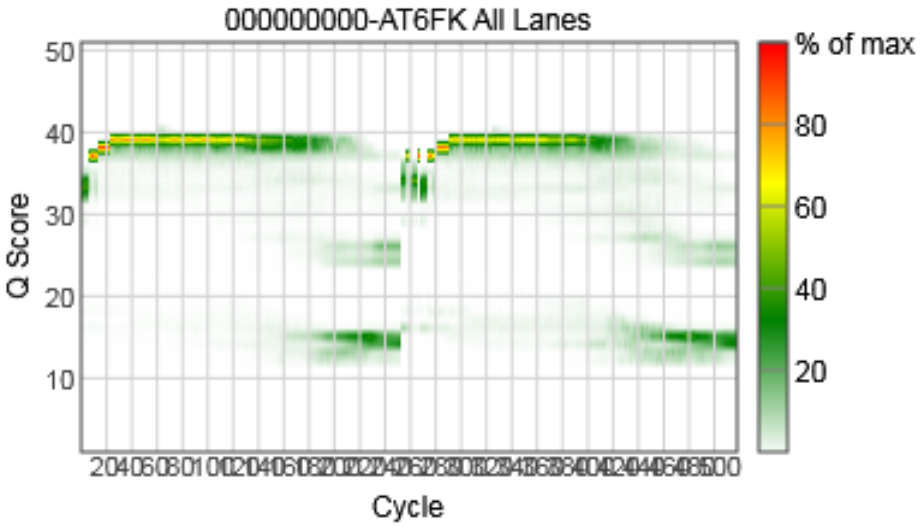


Figure A-15. Sequence Analysis Viewer 1.9.1 Qscore Heatmap (Low Quality)

Table A-1. Example of Pertussis Laboratory Expected Sample Cutoff Values and Ranges for the metrics in the Summary table.

Metric	Expected Value	Sample Values		
Organism	*(Specify the organism here.)*	**Pertussis (Reagent Kit v3)**		**E.Coli K12 MG1655[1]**
Range	*(Specify the range here.)*	**Cutoff**	**Ideal**	
Level	The sequencing read level			
Yield Total	MiSeq Reagent Kit V2: Output Max: 7.5 Gb 2.25 Gb at 2 x 300 bp 4.5 Gb at 2 x 150 bp 7.5 Gb at 2 x 250 bp Kit V3: Up to 15 Gb at 2 x 300 bp Up to 3.75 Gb at 2 x 75 bp	< 9 Gb	12-16 Gb	
Projected Total Yield	The projected number of bases expected to be sequenced at the end of the run, which is updated as the run progresses.			
Aligned	The percentage that aligned to the PhiX genome i.e. if 1% PhiX was the initial input quantity, the aligned value should be equal to 1% or below.		< 1%	
Error Rate	The calculated error rate of the reads that aligned to PhiX.			
Intensity Cycle 1	The average of the A channel intensity measured at the first cycle averaged over filtered clusters.			

%Q >/= 30	MiSeq Reagent Kit V2: > 90% bases higher than Q30 at 1 × 36 bp > 90% bases higher than Q30 at 2 × 25 bp > 80% bases higher than Q30 at 2 × 150 bp > 75% bases higher than Q30 at 2 × 250 bp MiSeq Reagent Kit V3: > 85% bases higher than Q30 at 2 × 75 bp > 70% bases higher than Q30 at 2 × 300 bp	Read 1 > 80% Read 2 > 70%	Read 1 > 85% Read 2 > 75%	MiSeq : 89.7 % HiSeq: 87.7%

Table A-2. Example of Pertussis Laboratory Expected Sample Cutoff Values and Ranges for the metrics in the Read table.

Metric	Expected Value	Sample Values	
Organism	*(Specify the organism here.)*	**Pertussis**	
Range	*(Specify the range here.)*	**Cutoff**	**Range**
Tiles	Standard Flow Cell in MiSeq Reagent Kit v3 (38 tiles) PGS Flow Cell in MiSeq Reagent Kit v3 (38 tiles) Standard Flow Cell in MiSeq Reagent Kit v2 (28 tiles) Micro Flow Cell in MiSeq Reagent Micro Kit v2 (8 tiles) Nano Flow Cell in MiSeq Reagent Nano Kit v2 (4 tiles)		
Density	Kit V2: Loading Concentration: 10-15 pM Cluster Density:1000-1200 k/mm^2 Kit V3: Loading Concentration: 15 pM Cluster Density:1200-1400 k/mm^2	< 800/mm^2 or > 1500/ mm^2	1200-1400k/mm^2
Clusters PF	80-95%	< 75%	80-95%
Phas./Prephas.	<0.25		
Reads	Kit V3: 25M Kit V2 : 15M Micro Kit V2: 4M Nano Kit V2: 1M	< 18M or > 28M	
Reads PF	Kit V2: Single Reads: 12-15M Paired End Reads: 24-30M Kit V3: Single Reads 22-25M Paired End Reads: 44-50M	Single Reads: < 15M Paired End Reads: < 30M	Single Reads: 22-25M Paired End Reads: 44-50M

%Q >/= 30	Kit V2: > 90% bases higher than Q30 at 1 × 36 bp > 90% bases higher than Q30 at 2 × 25 bp > 80% bases higher than Q30 at 2 × 150 bp > 75% bases higher than Q30 at 2 × 250 bp Kit V3: > 85% bases higher than Q30 at 2 × 75 bp > 70% bases higher than Q30 at 2 × 300 bp	< 70% bases higher than Q30 at 2 × 300 bp	Read1 > 75 % bases higher than Q30 at 2 × 300 bp Read2 > 70% bases higher than Q30 at 2 × 300 bp
Yield	Kit V2: Output Max: 7.5 Gb 2.25 Gb at 2 x 300 bp 4.5 Gb at 2 x 150 bp 7.5 Gb at 2 x 250 bp Kit V3: Up to 15 Gb at 2 x 300 bp Up to 3.75 Gb at 2 x 75 bp		
Cycles Err Rated	The number of cycles that have been error-rated using PhiX, starting at cycle 1.		
Aligned	The percentage that aligned to the PhiX genome.		
Error Rate	The calculated error rate, as determined by the PhiX alignment. Subsequent columns display the error rate for cycles 1–35, 1–75, and 1–100.		
Intensity Cycle 1	The average of the A channel intensity measured at the first cycle averaged over filtered clusters.		

Table A-3. Example of Pertussis Laboratory Cutoff Values and Ranges for the metrics in the Indexing Tab Table.

Metric	**Expected Value**	**Example Sample Values**	
Organism	*(Specify the organism here.)*	**Pertussis**	
Range	*(Specify the range here.)*	**Cutoff**	**Range**
Index Number	A unique number assigned to each index by SAV for display purposes.		
Sample ID	The sample ID assigned to an index in the sample sheet.		
Project	The project assigned to an index in the sample sheet.		
Index 1 (I7)	The sequence for the first Index Read.		
Index 2 (I5)	The sequence for the second Index Read.		

% Reads Identified (PF)	The number of reads (only includes Passing Filter reads) mapped to this index.	**< 95%**	**98-99.5%**

24.0 Revision History

Rev #	DCR #	Change Summary	Date

25.0 Approval

Approval Signature: _____ Date: _____

Bioinformatics QC: Pre-Analysis Quality Control (FASTQ) Procedure

1.0 Purpose

This document provides quality control (QC) guidance for the analysis of nucleic acid next generation sequencing (NGS) data. Following the generation of a FASTQ file, this guidance should be utilized to assess the quality of sequence data prior to assembly and further analysis. The guidance takes into account specific QC checkpoints between computational processes to ensure each step is completed correctly, with high confidence, and to generate quality data metrics that yield an informative study.

QC checkpoints are necessary at several stages of bioinformatics analysis including raw read sequence filtering, each alignment and characterization stage. These steps ensure the sequence data meets standards for analysis, allows removal of low quality reads, and reduces false negatives and positives. This guidance also aims to promote standardized best practice measures in order to improve reproducibility of results.

2.0 Scope

This document provides information on post-sequencing, pre-analysis QC: quality control steps to be performed on NGS data in the form of a FASTQ file prior to assembly or further analysis.

3.0 Related Documents

Title	Document Control Number
Bioinformatics QC Workflows	

4.0 Responsibilities

Position	Responsibility
All Laboratory Staff	• Follow documented procedures
Team Lead	• Ensure documented procedures for data quality checks are established • Ensure documented procedures are followed
Quality Manager	• Ensure documented procedures are available to the end user • Review records of data quality checks as required

5.0 Definitions

Term	Definition
FASTQC	• A quality control tool for high throughput sequence data
PrinSeq	• A quality control software for filtering, reformatting and trimming sequence data.
Trimmomatic	• A flexible read trimming tool for Illumina

6.0 Sample Information / Processing

Upon completion of the NGS run, transfer data to Isilon. *(Specify your laboratory data storage location here.)*

7.0 FASTQC

7.1 Once the sequencing run and initial Sequencing QC (SOP1) has been completed, a FASTQ file can be exported from Illumina's RTA (Real-Time Analysis Software).

a. A quality control software such as FASTQC should be used to assess the quality of the sequence data. While most sequencers will generate their own quality reports, these reports are generally more useful for identifying issues that originate with the sequencer. FASTQC can be effective in detecting problems with either the sequencer and/or the starting library material.

b. The output from FASTQC will provide several statistics (in HTML format) including: per-base sequence quality and per-sequence quality scores, per-base N content, per-sequence GC content, overrepresented sequences, adapter content and K-mer content etc. (see table below). By default FASTQC will provide a green check validating these metrics a red X signifying a failed test or a yellow exclamation mark to indicate potential areas of concern. While these tests may appear to give a pass/fail indication, these should be taken in the context of what is expected from your library. Please review these test results below before continuing to further analysis. (Please see Figures A-1 and A-2 for examples of good and bad reports)

Metric	Description
Total Sequences	A count of the total number of sequences processed. There are two values reported, actual and estimated. At the moment these will always be the same. In the future it may be possible to analyze just a subset of sequences and estimate the total number, to speed up the analysis, but since we have found that problematic sequences are not evenly distributed through a file we have disabled this for now.
Filtered Sequences	If running in Casava mode sequences flagged to be filtered will be removed from all analyses. The number of such sequences removed will be reported here. The total sequences count above will not include these filtered sequences and will include the number of sequences actually used for the rest of the analysis.
Sequence Length	Provides the length of the shortest and longest sequence in the set. If all sequences are the same length only one value is reported.
% GC	The overall %GC of all bases in all sequences.
Per base sequence quality	For each position a Box and Whisker type plot is drawn. The elements of the plot are as follows: • The central red line is the median value • The yellow box represents the inter-quartile range (25-75%) • The upper and lower whiskers represent the 10% and 90% points • The blue line represents the mean quality
Per tile sequence quality	This graph will only appear in your analysis results if you're using an Illumina library which retains its original sequence identifiers. The plot shows the deviation from the average quality for each tile. The colors are on a cold to hot scale, with cold colors being positions where the quality was at or above the average for that base in the run, and hotter colors indicate that a tile had worse qualities than other tiles for that base. In the example below you can see that certain tiles show consistently poor quality. A good plot should be blue all over.
Per sequence quality scores	The per sequence quality score report allows you to see if a subset of your sequences have universally low quality values. It is often the case that a subset of sequences will have universally poor quality, often because they are poorly imaged (on the edge

	of the field of view etc.), however these should represent only a small percentage of the total sequences.
Per base sequence content	Plots out the proportion of each base position in a file for which each of the four normal DNA bases has been called
Per sequence GC content	This module measures the GC content across the whole length of each sequence in a file and compares it to a modelled normal distribution of GC content
Per base N content	If a sequencer is unable to make a base call with sufficient confidence, then it will normally substitute an N rather than a conventional base call. This module plots out the percentage of base calls at each position for which an N was called.
Sequence Length Distribution	Some high throughput sequencers generate sequence fragments of uniform length, but others can contain reads of wildly varying lengths. Even within uniform length libraries some pipelines will trim sequences to remove poor quality base calls from the end. This module generates a graph showing the distribution of fragment sizes in the file which was analyzed
Sequence Duplication Levels	In a diverse library most sequences will occur only once in the final set. A low level of duplication may indicate a very high level of coverage of the target sequence, but a high level of duplication is more likely to indicate some kind of enrichment bias (e.g. PCR over amplification). This module counts the degree of duplication for every sequence in a library and creates a plot showing the relative number of sequences with different degrees of duplication.
Overrepresented sequences	A normal high-throughput library will contain a diverse set of sequences, with no individual sequence making up a tiny fraction of the whole. Finding that a single sequence is very overrepresented in the set either means that it is highly biologically significant, or indicates that the library is contaminated, or not as diverse as you expected.
Adapter Content	The Kmer Content module will do a generic analysis of all of the Kmers in your library to find those which do not have even coverage through the length of your reads. This can find a

	number of different sources of bias in the library which can include the presence of read-through adapter sequences building up on the end of your sequences.
Kmer Content	The analysis of overrepresented sequences will spot an increase in any exactly duplicated sequences, but there are a different subset of problems where it will not work.

c. FASTQ files should then be processed through a read trimming and filtering software of your choice such as PrinSeq or Trimmomatic. The recommended initial trimming cutoff is Q=5. This value may vary but most bases below a quality score of 4 or lower have been shown to be erroneous. *Note that NextSeq data usually need to be filtered with an increased quality cutoff of Q<15.

d. Subsequent levels of trimming/filtering with increased stringency might be needed for your data. After each round of trimming, use your FASTQ QC software to determine whether this will be needed.

e. Once reports indicate that a satisfactory level of trimming and filtering has been completed, proceed to SOP3 for assembly and further analysis.

8.0 Method Performance Specifications

N/A

9.0 Calculations

N/A

10.0 Reference Values, Alert Values

N/A

11.0 Interpretation of Results

Of the metrics shown above, key metrics that should be considered are listed below. Please keep in mind that these results are variable based on several factors including organism and workflows and thus should be interpreted within the context of expected values based on historical results.

Per Base Sequence Quality – This plot reflects the Q-score of raw reads as a box-plot for each cycle. Higher values are always better and generally a decay of quality can be observed in most runs.

Per Base Sequence Content – This plot reflects the proportion of each base at each cycle. Generally in a random fragment library from a genome you would expect to see all four bases equally represented. However, some genomes can be very GC biased and thus, this information should be compared against historical data.

Duplicate Sequences – This plot reflects the number of times the same sequence is seen in a 200,000 read subset of your sample data. Ideally, one should expect to see <10% duplicate reads. A high amount of duplicate sequences might suggest over-amplification or poor library-prep.

12.0 Results Review and Approval

Document the data quality metrics on the appropriate form or test record and obtain applicable reviews and approvals. *(Update this section to specify your laboratory's applicable form/record and processes.)*

13.0 Reporting Results; Guidelines for Notification

N/A

14.0 Sample Retention and Storage

Store data in compliance with all applicable regulations, CDC records retention policy, and laboratory data storage procedures. *(Update to specify your laboratory's data retention and storage policy)*

15.0 References

15.1 Illumina Sequence Analysis Viewer v1.11 Part # 15066069 v03 February 2017

16.0 Appendices *(Include example screen shots of good and poor quality data applicable to your laboratory methods)*

FASTQC Screenshots:

Figure A-1. FASTQC (Sample Good Report)

FastQC Report

Summary

- Basic Statistics
- Per base sequence quality
- Per tile sequence quality
- Per sequence quality scores
- Per base sequence content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Adapter Content
- Kmer Content

Basic Statistics

Measure	Value
Filename	good_sequence_short.txt
File type	Conventional base calls
Encoding	Illumina 1.5
Total Sequences	250000
Sequences flagged as poor quality	0
Sequence length	40
%GC	45

Per base sequence quality

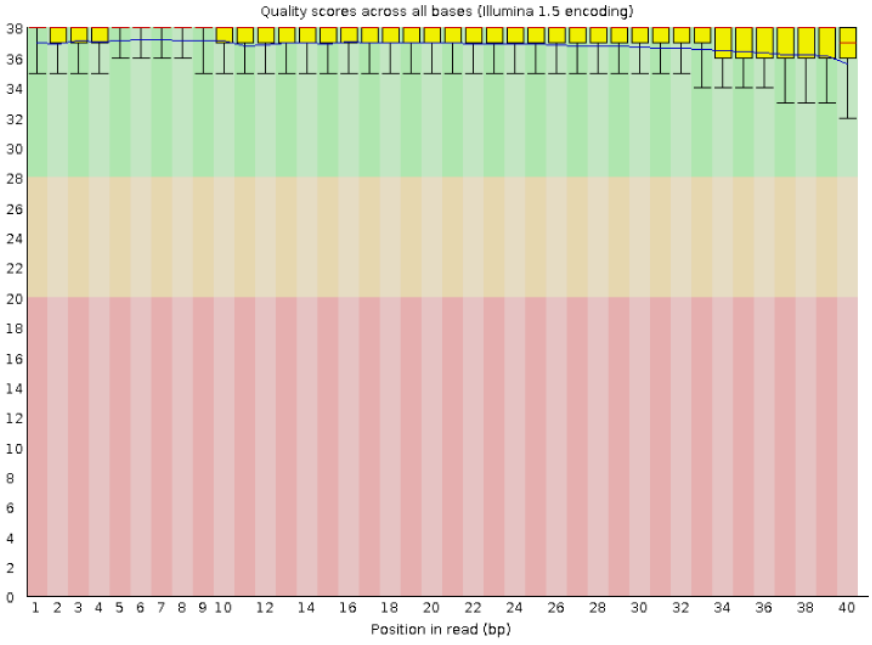


Figure A-2. FASTQC (Sample Bad Report)

FastQC Report

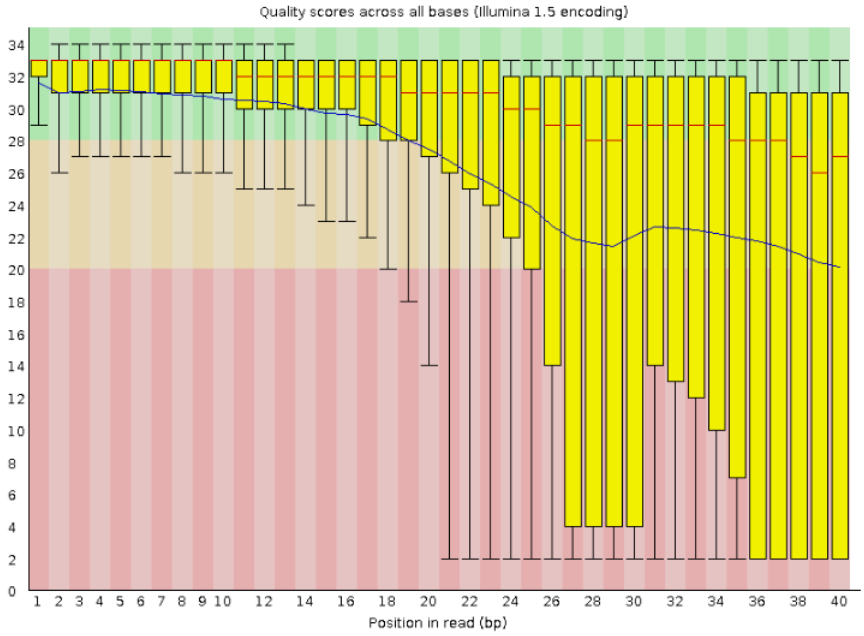
Summary

- Basic Statistics
- Per base sequence quality
- Per tile sequence quality
- Per sequence quality scores
- Per base sequence content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Adapter Content
- Kmer Content

Basic Statistics

Measure	Value
Filename	bad_sequence.txt
File type	Conventional base calls
Encoding	Illumina 1.5
Total Sequences	395288
Sequences flagged as poor quality	0
Sequence length	40
%GC	47

Per base sequence quality



17.0 Revision History

Rev #	DCR #	Change Summary	Date

18.0 Approval

Approval Signature:_____ Date: _____

1.0 Purpose/Principle

This document provides quality control (QC) guidance for the analysis of nucleic acid next generation sequencing (NGS) data. Following the Pre-Analysis QC of a FASTQ file, this guidance should be utilized to perform assembly of sequence data (generate a FASTA file from a raw FASTQ) and evaluation of the assembly prior to further analysis. This guidance will provide steps and key metrics to track for performing two types of assembly. (1) Reference-based assembly and (2) *De novo* assembly (assembly without a reference). The guidance takes into account specific QC checkpoints between computational processes to ensure each step is completed correctly, with high confidence, and to generate quality data metrics that yield an informative study.

QC checkpoints are necessary at several stages of bioinformatics analysis including raw read sequence filtering, all alignment, and characterization stages. These steps ensure the sequence data meets standards for analysis, allows removal of low quality reads, and reduces false negatives and positives. This guidance also aims to promote standardized best practice measures in order to improve reproducibility of results.

2.0 Scope

This document provides information on post-sequencing, post- initial raw read QC filtering and trimming: quality control steps to be performed on NGS data in the form of a FASTQ, generating a FASTA file and evaluating the quality of an assembly.

3.0 Related Documents

Title	Document Control Number
Bioinformatics QC Workflows	

4.0 Responsibility

Position	Responsibility
All Laboratory Staff	• Follow documented procedures
Team Lead	• Ensure documented procedures for data quality checks are established • Ensure documented procedures are followed
Quality Manager	• Ensure documented procedures are available to the end user • Review records of data quality checks as required

5.0 Definitions and Terms

Term	Definition
FASTQC	A quality control tool for high throughput sequence data
PrinSeq	A quality control software for filtering, reformatting and trimming sequence data.
Trimmomatic	A flexible read trimming tool for Illumina
SPAdes	Assembly tool for single-cell and standard (multi-cell) assembly
ABySS	*De novo* parallel, paired-end sequence assembler for short reads
Edena	*De novo* short reads assembly tool
SMALT	Reference genome based assembly tool
Bowtie2	Reference-based alignment tool for genome assembly
QUAST	Quality assessment tool for genome assemblies

6.0 Equipment

N/A

7.0 Reagents and Media

N/A

8.0 Supplies, Other Materials

N/A

9.0 Safety Precautions

N/A

10.0 Sample Information / Processing

Upon completion of the NGS run, transfer data to Isilon. *(Specify your laboratory data storage location here.)*

11.0 Quality Control

N/A

12.0 Workflow Chart

N/A

13.0 Test Procedure

13.1 Assembly

a. Assembly can be performed using an assembly software of your preference. Some examples of *de novo* assembly tools include: SPAdes, Edena, ABySS. Similarly, reference based assembly (using one or more reference genomes) can be conducted using a tool such as SMALT or Bowtie2.

b. **Choosing a reference genome** Curated reference genomes are available for some species and should be utilized when possible. These are high quality sequence data, often closed or finished genomes. Reference sequences also satisfy these requirements:

- Genome sequences with less than 1 error per 100,000 base pairs
- Each replicon is assembled into a single contiguous sequence with a minimal number of possible exceptions documented in the submission record
- All sequences are complete and have been reviewed and edited
- All known misassemblies have been resolved
- Repetitive sequences have been ordered and correctly assembled

13.2 Assembly QC

a. Once you have assembled your genome and generated a FASTA file from your pre-processed FASTQ file, your assembly quality should be evaluated using a tool such as QUAST, which uses aggregated metrics and can work with or without a reference genome to measure assembly quality. Additionally, this quality measure can be used to compare assembly results from multiple assemblers to determine the optimal tool for your workflow.

b. After running QUAST on your FASTA file, please review the following values in the text report that is generated:

Metric	Description
Number of Contigs	Total number of contigs of length
Total Length	Total number of bases in the assembly
Largest Contig	Length of the largest contig in the assembly
Reference Length (Ref-based assembly only)	Total number of bases in the reference genome
GC %	Total number of G and C nucleotides in the assembly, divided by the total length of the assembly
Reference GC % (Ref-based assembly only)	The percentage of G and C nucleotides in the reference genome (see above)
N50	The length for which the collection of all contigs of that length or longer cover at least half the assembly
NG50 (Ref-based assembly only)	The length for which the collection of all contigs of that length or longer covers at least half the reference genome.
N75/NG75	Similar to N50/NG50, but using 75% of the assembly covered
L50	The number of contigs equal to or longer than N50 (N75, NG50, NG75) or the minimal number of contigs that cover half the assembly.

c. Once these steps are completed, please proceed to the next analysis step (SOP4).

14.0 Method Performance Specifications

N/A

15.0 Calculations

N/A

16.0 Reference Values, Alert Values

N/A

17.0 Interpretation of Results

These values will vary depending on sample and organism type and should be evaluated based on your expected values and historical results. Please note that in general terms, better assemblies will have a lower **Number of Contigs**, greater **Total Length** and larger **N50** scores. Note however, that if total assembly length is much greater than expected, this can be a sign of contamination or a mixture of isolates.

18.0 Results Review and Approval

Document the data quality metrics on the appropriate form or test record and obtain applicable reviews and approvals. *(Update this section to specify your laboratory's applicable form/record and processes.)*

19.0 Reporting Results; Guidelines for Notification

N/A

20.0 Sample Retention and Storage

Store data in compliance with all applicable regulations, CDC records retention policy, and laboratory data storage procedures. *(Update to specify your laboratory's data retention and storage policy)*

21.0 References

21.1 Illumina Sequence Analysis Viewer v1.11 Part # 15066069 v04 February 2018

22.0 Appendix *(Include example screen shots of good and poor quality data applicable to your laboratory methods)*

Table A-1. Example of Pertussis Laboratory Expected Sample/Cutoff Values for Assembly QC Metrics

Metric	Description	Sample Values (Pertussis)
Number of Contigs	Total number of contigs of length	**<= 400**
Total Length	Total number of bases in the assembly	**~4.1 Mb**
Largest Contig	Length of the largest contig in the assembly	
Reference Length (Ref-based assembly only)	Total number of bases in the reference genome	
GC %	Total number of G and C nucleotides in the assembly, divided by the total length of the assembly	**~67.7%**
Reference GC % (Ref-based assembly only)	The percentage of G and C nucleotides in the reference genome (see above)	
N50	The length for which the collection of all contigs of that length or longer cover at least half the assembly	**>= 19kb**
NG50 (Ref-based assembly only)	The length for which the collection of all contigs of that length or longer covers at least half the reference genome.	
N75/NG75	Similar to N50/NG50, but using 75% of the assembly covered	
L50	The number of contigs equal to or longer than N50 (N75, NG50, NG75) or the minimal number of contigs that cover half the assembly.	

23.0 Revision History

Rev #	DCR #	Changes Made to Document	Date

24.0 Approval Signature

Approved By: _____ Date: _____

1.0 Purpose

This procedure provides general guidance for approaching validation of Next Generation Sequencing (NGS) laboratory methods that establishes and documents the performance specifications of a test method. This document was primarily developed for isolate sequencings, but included are other elements to consider for broader intended use and additional specimen types.

2.0 Scope

2.1 This procedure should be used as guidance when the laboratory is validating Laboratory Developed Tests (LDT).

2.2 This procedure is not intended to, nor does it meet, the regulatory requirements of FDA for approval to market an in vitro diagnostic device (i.e., 510(k), PMA).

2.3 It is expected that each program will examine this guidance in light of their assay and intended use and add any additional studies or study questions to ensure the full support of their intended use and resulting diagnoses with documented scientific evidence.

2.4 The lifecycle of NGS method adoption includes Assay Development, Assay Validation, and continuous Quality Management. This SOP describes the steps necessary to fulfill assay validation. It is expected that Assay Development is a precursor to the use of the SOP and that Quality Management will occur continuously throughout the useful life of the method.

3.0 Related Documents

Title	Document Control Number
NGS Method Validation Plan	
NGS Method Validation Summary Report	
NGS QC Guidance for (Illumina/MinION/Ion Torrent) Workflows	
Bioinformatics QC Workflows	
Training SOP *(include the SOP your laboratory uses for training on this method)*	

4.0 Responsibilities

Position	Responsibility
Laboratory Staff	• Perform method validation according to an approved plan

Team Lead/Designee	• Prepare a method validation plan for the test • Ensure the method validation is performed as planned and documented • Approve the validation results
Quality Manager / Designee	• Ensure the method validation documentation, including all data worksheets and records, are completed, approved, and retained
Unit Chief / Designee	• Review and approve the validation plan as described in *Method Validation Plan* prior to initiating the study • Ensure the method validation is coordinated with appropriate SME • Review and approve the summary results as described in the *Method Validation Summary Report* prior to placing the test in use • Ensure tests are validated and receive applicable approvals prior to placing the test in use
CLIA Laboratory Director	• Approve completed method validations for assays utilized to perform testing to which CLIA regulations apply

5.0 Definitions/Terms

NGS methods are used to answer many different questions. The application of performance metrics may change based on the specific use case. Below are definitions used for the purposes of this document for the most common performance metric terms in NGS.

Term	Definition
Accuracy	True value or closeness of agreement between a test result and the accepted reference value; verification that the assay is measuring what it proposes to measure
Precision (reproducibility)	Closeness of agreement between independent test results obtained under stipulated conditions
Sensitivity	The test method's ability to obtain positive results in concordance with positive results obtained by the reference method or with known positive samples

Specificity	The ability of the test method to obtain negative results in concordance with a negative result obtained by the reference method or with known negative samples; also, ability of the method to measure only the component it purports to measure and not interfering substances
Limit of detection	The lowest amount of the analyte in a sample that can be consistently detected 95% of the time
Reference Value or Normal value	For qualitative tests, the normal value is the reference value among healthy individuals in the laboratory's patient population(s). These values are made appropriate for the laboratory's patient population(s).
Applicable Genome Region	The region of the genome in which sequence of an acceptable quality can be derived by the laboratory test.
Clinical Validity (as applicable)	The ability of a diagnostic device to measure or detect the clinical condition for which the device is intended. Data and/or references should support that the method is used to analyze the correct analyte in the correct sample type(s) from the correct population(s) at the correct time point(s) with appropriate sensitivity/specificity to support he method's intended use and interpretation of results.

6.0 Applied & Extended Definition Examples

The examples of "pathogen identification" and "characterization by SNP detection" provided in this section are meant to illustrate the process a laboratory may go through in applying the definitions to a specific use case. There are many use cases for NGS and each use case has the potential to be different. Laboratorians should work with their quality manager and laboratory director to ensure the definitions of True Positive Samples and True Negative Samples are being applied appropriately.

6.1 **Accuracy:**

a. Interpretation: The overall ability of the method to both correctly identify True Positive samples and True Negative Samples (NGS WG)
b. Calculation
   i. [(TP + TN) / (TP + FN + TN + FP)] x 100%
c. Example 1. Pathogen Identification: An NGS method is used to identify 5 species of bacteria (Bacteria 1, Bacteria 2, Bacteria 3, Bacteria 4, and Bacteria 5). True Positive samples will be defined as samples containing any of these 5 bacteria that are correctly identified by the method. True

Negative samples will be defined as any samples that do not contain Bacteria 1 -5 and that are correctly identified as lacking of the 5 bacteria of interest. False Positive is defined as any sample that lacks Bacteria 1 – 5 but is identified by the method as containing one of the bacteria of interest. False Negative is defined as a sample that contains any of Bacteria 1 – 5 and is not identified by the method. Note in this example Positive and Negative are defined at the result level, not the sequence data level.

d. Example 2. Characterization by SNP Detection: An NGS method is used to detect SNPs. The SNP of interest is SNP 1 A -> C. True Positive samples will contain a C at the SNP site and the sequence data will show a C at the SNP site. True Negative samples will contain at A at the SNP site and the sequence data will show an A at the SNP site. False positives are samples that do not contain a C at the SNP site, but the sequence data shows a C at the SNP site. False negatives are samples that contain a C at the SNP site but the sequence data shows either an A, G, or T. Note in this example Positive and Negative are defined at the sequence data level.

6.2 **Precision:**

a. Interpretation: The degree to which repeated sequence analyses give the same result. (Gargis 2012)

   i. Repeatability: with-in run precision where sequencing the same samples multiple times using the same conditions (such as operator, reagent lot, and system) gives the same result

   ii. Reproducibility: Between run precision where sequencing the same sample multiple times using different conditions gives a similar result

   iii. Because reproducibility is more difficult to achieve, if reproducibility is obtained, repeatability does not need to be assessed.

   iv. Conditions to change

   - Different Operator (2 suggested)
   - Different days tested (3 suggested, spaced over 20 days)
   - Library Prep (if more than 1 library prep method used)
   - Different equipment (if applicable)
   - Different reagent lots

b. Calculation

   i. Qualitative

   - (# of results in agreement / total # of results) x 100

   ii. Quantitative: As a best practice, use a quantitative raw result to measure precision.

   - Coefficient of Variance = (Standard deviation / mean) x 100

- Example: ANI results in a numerical ANI score. Calculate the CV of the ANI score using the data from the precision runs.

c. Example on how to test:

i. Example 1: Select a subset of the samples used to establish accuracy. Test the samples on 3 or more sequencing runs using multiple operators and instruments as applicable. Calculate reproducibility using either the qualitative or the quantitative approach.

6.3 **Sensitivity:**

a. Interpretation: The likelihood that the assay will detect the targeted sequence or sequence variations, if present. (Gargis 2012, modified by NGS QW)

b. Calculation

i. [TP / (TP + FN)] x 100

6.4 **Specificity**:

a. Interpretation: The likelihood that the assay will not detect the targeted sequence or sequence variation when none are present. (Gargis 2012, modified by NGS QW)

b. Calculation

i. [TN / (TN + FP)] x 100

6.5 **Limit of Detection**:

a. Interpretation: The LOD is represented by two data points: the bioinformatics LOD and the biological LOD. (AMP/CAP 2017, modified by NGS QW)

i. Establishing the bioinformatics LOD requires determining the minimum required sequencing depth of coverage and consensus needed for the method's intended purpose. Depth of coverage is the number of independent reads assessed at a given base position. Additionally, establish the lower limit for quality of base calls necessary for the method to perform its intended purpose.

ii. Biological LOD involves determining the minimum amount of biological target material necessary for the assay to detect it. For example, a clinical specimen may contain differing levels of virus or bacteria along with human DNA, or a bacterial sample may contain some fraction of bacterial cells that are resistant to antimicrobials. Establish the minimum amount of target material necessary to accurately fulfill the assay's intended purpose.

b. Example on how to test:

i. Example 1: Pathogen Identification: Establish the informatics LOD by loading different dilutions of sequencing libraries or using data down sampling. Determine the minimum depth of coverage and minimum base call quality score necessary to accurately identify the bacteria. Establish the biological LOD by preparing clinical matrices spiked with decreasing amounts of the target of interest. Determine the minimum amount of target material present in the clinical matrix necessary for the assay to perform as intended.

6.6 **Reference Value/Normal Value:**

a. Interpretation: Reportable sequence or sequence variations the assay can detect that are expected to occur in an unaffected population, the range of values typically found in individuals who do not have the disease or condition that is being assayed by the test. A reference value may not be applicable if testing is performed to qualitatively characterize (e.g., serotyping) clinical isolates.

b. Example on how to test:

i. Example 1: Normal, healthy individuals would be expected to test negative for Bacteria 1 – 5, which are infectious diseases. Test samples from normal healthy individuals (True Negatives) to document that the samples are correctly identified as lacking one of the 5 bacteria of interest.

ii. Example 2: Normal healthy individuals do not have SNP 1 A -> C. Test samples from normal healthy individuals (True Negatives) to document that the samples are correctly sequenced as A at the SNP site.

6.7 **Applicable Genome Region:**

a. Interpretation: The region of the genome targeted for sequencing may be the entire genome (Whole genome sequencing) or targeted regions. Describe the portion of the genome from which sequence data will be used to generate results. (NGS QW)

6.8 **Clinical Validity:**

a. Interpretation: The accuracy with which a genetic test identifies a particular clinical condition (Holztman and Watson, 1999).

b. Calculation: Positive Predictive Value (PPV) and Negative Predictive Value (NPV) are important measures of clinical validity

i. PPV = TP / (TP + FP)

ii. NPV = TN / (TN + FN)

7.0 Stage 1: Planning and Development

7.1 Draft the technical procedure. The documented procedure remains in draft form through the validation process. However, record any changes in the final report.

7.2 *Method Validation Plan* serves as a record of the validation planning.

7.3 Define the **Purpose** for the validation.

a. Describe whether the validation is designed to evaluate a newly developed method, a newly modified method, or a newly expanded method.

b. Identify the comparison or recognized reference method that will be used for comparison. Include the anticipated benefit of the new method (e.g. increased accuracy, lower costs, quicker turnaround time, resource availability, new analytes).

7.4 Provide a **Summary of the Test Procedure Purpose/Principle:**

a. Describe the assay result and regulatory type. Include information on the analyte detected and how the results will be used.

7.5 As applicable, describe **limiting factors and justification for small sample size**.

7.6 Define the **Acceptance Criteria** for the validation plan.

a. Define the expected performance of the method during validation, necessary to ensure the intended use will ultimately be met.

b. Expected performance is expressed in the form of minimum acceptance criteria for the validation. Include the criteria listed in Section 5:

7.7 Define the **Sample Requirements** for the validation plan.

a. Determine the appropriate sample types, volumes, and quantities (e.g., serum, 100 µl, n = 20).

i. When appropriate, the same set of samples may be used to evaluate accuracy, sensitivity, and specificity. Precision, or measurement reproducibility, may be evaluated by repeating a sub-set of these samples on different days by different personnel. Clinical samples should be used to determine the clinical accuracy, clinical sensitivity, and clinical specificity.

ii. True positive samples should fulfill three criteria:

- Genetic diversity – select samples that are representative of the genetic diversity expected given the test method's intended use. Depending on the test method's purpose (e.g. identification, characterization), genetic diversity may be interpreted as the type of organisms expected to be identified and/or the type of sequence variations (e.g. deletions, splice sites, SNPs, %GC) expected.
- Expected submission volume – select samples representative of the expected submissions. For example, if Bacteria 1 is expected to make up 50 % of the submissions and Bacteria 2 – 5 make up the other 50 % collectively, include a greater number of Bacteria 1 samples in the True Positive set.
- Public Health Impact – select samples that cause the greatest harm or disease to individuals. For example, if Bacteria 5 or SNP1 cause the greatest disease or harm to individuals but are expected to be submitted for testing only on rare occasions, they should still be included in the True Positive set.

iii. True negative samples should fulfill three criteria:

- Genetic similarity – select samples that are genetically similar to establish the test method's ability to distinguish between negative and positive samples
- Symptom similarity – select samples that contain organisms or conditions that cause similar symptoms to the organism(s) or condition(s) of interest
- Healthy Population (if applicable)– select samples that represent a normal, healthy population

b. Record each matrix to be validated as part of the plan. Sample sets should be prepared in each matrix to be validated.

c. Provide a summary of the sample volume and total number of samples of each type that will be required to perform the plan. The required sample size and number should be based on statistical techniques and account for unique issues (i.e. technology or the biology of the condition being studied).

   i. Note: The appropriate number of samples depends on many factors, including but not limited to: expected precision, assay complexity, prevalence of the target(s) in the indicated population, sample availability, established accuracy of the reference method, data analysis, and level of statistical confidence that the user is willing to accept.

d. Provide the sample requirements for the evaluation of each of the acceptance criteria. Include the number of samples needed and an example of the calculation to be used for evaluation.

e. Select the samples and reference materials to be used during the validation, and record the source of the materials. Additional information such as catalog numbers, lot numbers, and specific purity requirements should be included in the Description/Characterization column as applicable.

f. If known positives are not available, the appropriate matrix (e.g. serum, sputum, spinal fluid) may be spiked with known levels of analyte. Interfering substances may be spiked as well.

7.8 Record the following elements within the plan to provide logistics and traceability for the proposed validation.

a. **Roles and Responsibilities:** Identify the individuals responsible for performing the validation procedure itself, as well as ancillary tasks, including document management, equipment maintenance, and approvals.

b. **Timeline:** Identify the anticipated sequence of events, including estimated time requirements and target dates for completion.

c. **Related Documents:** Provide a list of procedures, including established supporting procedures and the draft procedures under evaluation, necessary to perform the validation.

d. **Instrumentation:** Identify the equipment to be used as part of the test method and performance criteria. Include a list of equipment that is expected to be used for the validation, including serial or ID numbers and maintenance/calibration dates.

e. **Bioinformatics Pipeline:** Record the name of the pipeline, the version number for each tool within the pipeline, parameter settings used in each tool, developer, and technical support of each component of the pipeline, including the hardware, software, transmission system, backups, and networks.

f. **Training Requirements:** Identify training required to operate equipment as well as perform testing, calibration, and maintenance procedures for personnel involved in performing the validation protocol.

7.9 Prior to testing and analysis, laboratory leadership reviews and approves the *NGS Method Validation Plan*. Approval is recorded through signatures at the end of the section.

8.0 Stage 2: Testing and Analysis

8.1 Implement required training.

a. Prior to performing testing procedures, operators should complete training as required per the plan.

i. Note: Training may include both internal training on the method and supporting procedures, as well as external training provided by a supplier or manufacturer.

b. Record training, including attendees and dates completed, in the summary report.

8.2 Perform testing according to the completed plan.

8.3 Analyze data resulting from validation tests.

a. Evaluate data against the acceptance criteria established in the plan. The preliminary expectation is that all criteria will be met when testing is performed as detailed in the plan.

b. In the case where results fail one or more acceptance criteria, the underlying reason should be identified and a corrective action selected.

8.4 Perform corrective action as necessary. The action will generally be one of three options. Record the details of any corrective action in **Changes to the method validation plan** within the *NGS Method Validation Summary Report*.

a. Option 1: If there appears to be a unique, one-time occurrence that led to failing acceptance criteria, then the protocol may be performed a second time without changing the method or acceptance criteria.

i. Note: Record both sets of data. In the **Evaluation of discrepant results** section, describe the reason for the initial failure and why it is not expected to recur.

b. Option 2: If the failure requires a change to the method, then a revision to the method should be drafted and approved along with any necessary updates to the plan.

i. Note: Record both draft versions of the method and both versions of the plan. Subsequent versions of the plan should be identified as revisions of the first. In the **Changes to the method validation plan** section, describe the changes to the method and the rationale behind them.

c. Option 3: If the acceptance criteria can be changed while maintaining the intended use of the method, the acceptance criteria may be updated to reflect the performance of the method.

i. Note: Record any changes to the acceptance criteria in the **Changes to the method validation plan** section, along with a justification and an analysis of the potential impact or lack thereof.

9.0 Stage 3: Reporting and Implementation

9.1 Record the following elements in the *NGS Method Validation Summary Report*.

a. **Changes from the method validation plan**: Record any changes that were necessary during the validation process, to the method, plan, or acceptance criteria.

b. **Summary of Results:** Designate whether each of the acceptance criteria has been met, and provide a synopsis of the actual value measured for each of the acceptance criteria detailed in the plan.

c. **Interpretation of Results:**

i. **Evaluation of discrepant results**: Provide reasoning for discrepant results and evidence to support acceptance of the result or the process put in place to address the limitation.

ii. **Limitations:** Record any limitations to the method that were discovered during the validation process. Describe processes put in place to mitigate the limitations.

iii. **Disclaimers, as applicable:** Provide language used in disclaimers that will be included on the final test report provided to submitters.

iv. **Fast Tracked test QA monitoring plan, as applicable:** For tests that have been fast tracked due to urgent public health need, provide the planned QA monitoring activities to gather additional evidence of test performance.

d. **Statement of Suitability**

9.2 Submit the completed *NGS Method Validation Summary Report* for final approval to the personnel identified during the planning stage.

a. Note: If the method is subject to CLIA regulations, the *NGS Method Validation Plan* and the *NGS Method Validation Summary Report,* with all approval signatures, must be sent to the CLIA Laboratory Director to support the approval of the technical procedure.

9.3 Finalize technical procedures according to *Document Control* best practices.

9.4 Monitor test performance according to *Quality Control Program* best practices.

10.0 References

10.1 Clark, R. B., M. A. Lewinski, M. J. Loeffelholz, and R. J. Tibeets, 2009. *Cumitech 31A, Verification and Validation of Procedures in the Clinical Microbiology Laboratory.* Coordinating ed., S. E. Sharp. ASM Press, Washington, DC.

10.2 CLSI. *Evaluation of Precision Performance of Quantitative Measurement Methods*; Approved Guideline—Second Edition. CLSI document EP5-A2. Wayne, PA: Clinical and Laboratory Standards Institute; 2004.

10.3 CLSI. *Evaluation of the Linearity of Quantitative Measurement Procedures*: A Statistical Approach; Approved Guideline. CLSI document EP6-A. Wayne, PA: Clinical and Laboratory Standards Institute, 2003.

10.4 CLSI. *Interference Testing in Clinical Chemistry; Approved Guideline – Second Edition*. CLSI document EP7-A2. Wayne, PA: Clinical and Laboratory Standards Institute, 2005.

10.5 CLSI. *Method Comparison and Bias Estimation Using Patient Samples; Approved Guideline – Third Edition*. CLSI document EP9-A3. Wayne, PA: Clinical and Laboratory Standards Institute, 2013.

10.6 CLSI. *User Protocol for Evaluation of Qualitative Test Performance; Approved Guideline-Second Edition*. CLSI document EP12-A2. Wayne, PA: Clinical and Laboratory Standards Institute; 2008.

10.7 CLSI. *User Verification of Performance for Precision and Trueness*; Approved Guideline – Second Edition. CLSI document EP15-A2. Wayne, PA: Clinical and Laboratory Standards Institute; 2006.

10.8 CLSI. *Protocols for Determination of Limits of Detection and Limits of Quantitation; Approved Guideline*. CLSI document EP17-A. Wayne, PA: Clinical and Laboratory Standards Institute; 2004.

10.9 CLSI. *Estimation of Total Analytical Error for Clinical Laboratory Methods; Approved Guideline*. CLSI document EP21-A. Wayne, PA: Clinical and Laboratory Standards Institute, 2003.

10.10 CLSI. *Verification and Validation Multiplex Nucleic Acid Assays Approved Guideline. CLSI document MM17-A*. Wayne, PA: Clinical and Laboratory Standards Institute, 2003.

10.11 Westgard PhD, James O., *Basic Method Validation: Training in Analytical Quality Management for Healthcare Laboratories 3rd Edition*, Westgard QC, Inc.

10.12 Burd, Eileen M. *Validation of Laboratory-Developed Molecular Assays for Infectious Diseases*. Clinical Microbiology Reviews, July 2010 p.550-576

10.13 International Conference on Harmonization (ICH) Guidelines *Q2(R1): Validation of Analytical Procedures*, 2005

10.14 Gargis, A., et al *Assuring the quality of next-generation sequencing in clinical laboratory practice.* Nature Biotechnology, November 2012 p.1033-1036

10.15 Holtzman NA, Watson MS. Final report of the Task Force on Genetic Testing. Baltimore: Johns Hopkins University Press; 1999. Promoting safe and effective genetic testing in the United States.

10.16 FDA: Guidance on Analytical Method Validation, 2000

10.17 FDA: Guidance on Bioanalytical Method Validation, 2001

10.18 FDA: Draft Guidance on Comparability Protocols, 2003

10.19 FDA: Statistical Guidance on Reporting Results from Studies Evaluating Diagnostic Tests, 2007

11.0 Revision History

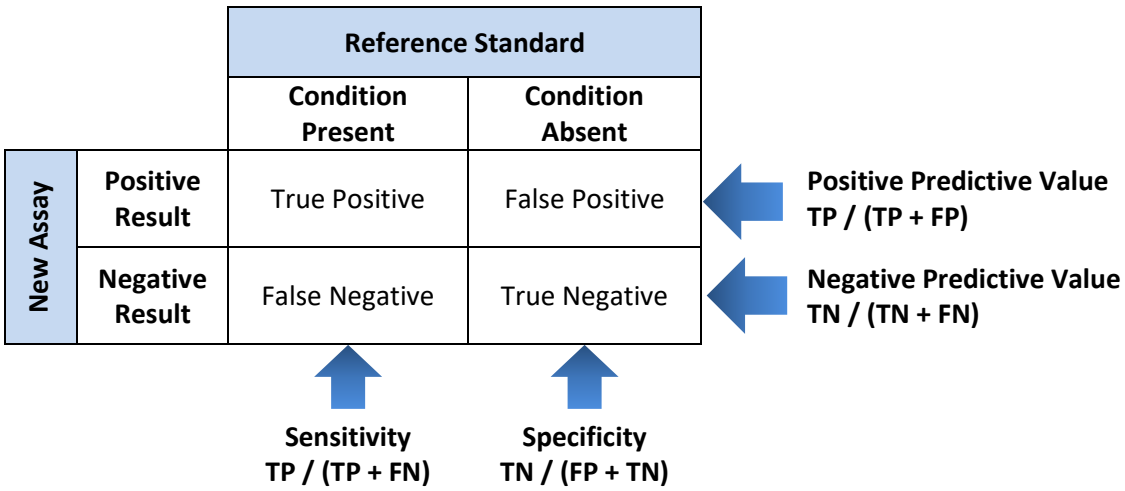
Rev #	DCR #	Change Summary	Date

12.0 Approval

This document has been approved by the CDC CLIA Laboratory Director as the standard practice for CLIA-regulated CDC Infectious Diseases Laboratories under certificates 11D0668319 and 11D2030855.

Approved By: _____ Date: _____

Appendix A: Common 2x2 table format for comparing results of a new test outcome to the reference standard outcome



Predictive values take into account the prevalence of the disease in the population being tested [e.g., the higher the prevalence, the higher the likelihood that a positive result is a True Positive].

NGS Method Validation Plan Template

Note: This document is intended to be used as a template for developing a method validation plan. Existing entries are intended as guidance and may be changed, deleted, or augmented as needed for the laboratory's specific requirements. Parentheses in blue provide specific examples for appropriate input. This document provides a record of the Planning and Development stage.

1.0 Purpose of Validation

1.1 Type of Validation:

☐Initial validation

☐Revalidation

☐Modification to procedure

☐Relocation of equipment

☐Change to instrument hardware or software

☐Change in database

☐Change in reagent formulation

☐Change in reagent manufacturer

☐Change in patient population

☐Change in intended use

☐Other_____

1.2 Statistical methods defined in this plan are designed to address the specific needs of the validation. The (comparison or reference method or material) will be used for evaluating performance. (method name) is expected to (describe improvement or how the new method differs from established testing)

2.0 Scope

2.1 **Method Validation Plan for:** (insert method name)

2.2 **Branch/Laboratory:** (branch and laboratory name)

2.3 **Test Procedure Document Number(s) and Revision Number:**

Document Name	Document Number	Revision Number	Effective Date

3.0 Summary of the Test Procedure Purpose/Principle:

3.1 **Assay result type** (choose one)**:**

☐Qualitative

☐Qualitative-titered

3.2 **Assay regulatory type, as applicable** (choose one)**:**

☐Laboratory Developed Test (LDT)

☐Modified FDA-cleared/approved

☐ Not Regulated

3.3 **Agent or analyte detected by the method:** (*insert name*)

3.4 **How test results are to be used:** *Describe how the test results will be used. Include the following elements as applicable: 1) Presumptive, screening, monitoring, confirmatory (e.g., presumptive to detect infection, screening test to rule our disease(high sensitivity, low specificity), a confirmatory test (high specificity), to monitor treatment response, to characterize or phenotype a pathogen, a research trial or surveillance activity falling under CLIA); 2) Detail if the results are used alone, or in conjunction with other assays as part of a specific testing algorithm and the extent to which interpretation needs to be in conjunction with clinical signs and symptoms. (e.g., Stand-alone test, used in conjunction with other assays (list related documents title and number). 3) Specify if result use differs among sample types or among patient populations.*

4.0 Limiting factors and justification for limited sample size:

4.1 Sample scarcity, urgent public health response.

5.0 Acceptance Criteria:

Performance Characteristic	Comparator Method Specifications (if available)	New Test Procedure: Minimum acceptable values
Sensitivity	Enter the percentage	Enter the percentage
Specificity	Enter the percentage	Enter the percentage
Accuracy	Enter the percentage	Enter the percentage
Precision/Reproducibility	Enter the percentage	Enter the percentage
Reference/Normal value	Enter whether the analyte is expected to be present or absent in the target population	
Applicable Genome Region		Enter the region of the genome in which sequence of an acceptable quality is expected to be derived

Performance Characteristic	Comparator Method Specifications (if available)	New Test Procedure: Minimum acceptable values
Clinical Validity (as applicable)	Positive Predictive Value (PPV) and Negative Predictive Value (NPV) of comparator method	Enter the prevalence in the relevant population(s). Enter the minimum acceptable PPV and NPV.
Any other performance characteristics required for this validation (e.g. cross-reactivities, interfering substances)	(row may be deleted if not applicable)	(row may be deleted if not applicable)

6.0 Sample Requirements

6.1 Sample Selection Summary

a. True Positive samples: justify True Positive sample selection according to the following three criteria: Diversity, expected sample testing volume, public health impact.
b. True negative samples: Justify True Negative Sample selection according to the following three criteria: Agent or analyte similarity, symptom similarity, and healthy population
c. Note: If revalidating after a change to data analysis methods only ("dry lab") the samples selected may be electronic data

6.2 Clinical Samples

Total positive samples	#
Total negative samples	#
Sample volume (units)	#
Sample matrix	(serum, sputum, spinal fluid, etc…)
Sample matrix	(add rows for each matrix to be validated)

6.3 Origin of Clinical Samples

Sample Material	Source	Description/ Characterization
	(internal or supplier name)	(CDC or ATCC Strain #)

6.4 Human subjects determination # for use of left-over clinical specimens, as applicable:_____

6.5 Contrived "spike-in" specimens:

Total positive samples	#
Total negative samples	#
Sample volume (units)	#
Sample matrix	(serum, sputum, spinal fluid,etc…)
Sample matrix	(add rows for each matrix to be validated)

6.6 Origin of Contrived "Spike-in" Samples:

Sample Matrix	Source	Description of "spike in" material	Description of Titrations
	(internal or supplier name)	(CDC or ATCC Strain #)	

7.0 Performance Characteristics:

Characteristic	Number of samples	Planned Calculations/Evaluations
Accuracy	(# of positive samples and # of negative samples) will be measured	$\frac{(\mathrm{TP}+\mathrm{TN})}{(\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN})} \times 100$
Precision/ Reproducibility (Qualitative)	(# of positive samples and # of negative samples) will be measured in (#) separate runs over at least (# ≥ 3) days, at least # days apart, by # different operators.	$\frac{\#\text{ of results in agreement}}{\text{total }\#\text{ of results}} \times 100$
Precision/ Reproducibility (Raw Value) (e.g. ANI score)	As above, but evaluating raw, quantitative values, if available, prior to qualitative cut-off for resulting.	$\mathrm{CV} = \frac{\text{standard deviation}}{\text{mean}} \times 100$
Clinical Sensitivity	(# of positive clinical samples) will be measured	$\frac{\mathrm{TP}}{(\mathrm{TP}+\mathrm{FN})} \times 100$

Characteristic	Number of samples	Planned Calculations/Evaluations
Analytic Sensitivity	(# of quantifiable samples at specified concentration levels near the expected detection limit) will be measured	$\frac{\text{TP}}{(\text{TP}+\text{FN})} \times 100$
Clinical Specificity	(# of negative clinical samples) will be measured	$\frac{\text{TN}}{(\text{FP}+\text{TN})} \times 100$
Analytic Specificity	(# of negative samples known to contain potentially cross-reactive agents or analytes) will be measured	$\frac{\text{TN}}{(\text{FP}+\text{TN})} \times 100$ With description of cross-reactive agents or analytes
Limit of Detection	Depth of coverage range of (# to #) will be used to determine the informatics LOD Target biological material present in the range (# to #) will be used to determine the biological LOD	Informatics LOD = (minimum depth of coverage and consensus % necessary to achieve accuracy) Biological LOD = (minimum amount of target biological material necessary to achieve accuracy)
Reference/Normal value	Determine: Literature review or to generate in-house data. In house: (# of normal population samples) will be evaluated	Summary of results List of references
Applicable genome region	(#) samples will be sequenced	Region of the genome in which sequence of acceptable quality was derived in the tested samples
Clinical validity	Determine: Literature review or to generate in-house data.	Positive Predictive Value: TP / (TP + FP) , and Negative Predictive Value : TN / (TN + FN) by sample type and/or population or List of references

8.0 Interfering Substances

8.1 If interference is observed during these studies, the interferent should be tested by serial dilutions to determine the lowest concentration that provides interference. Assay limitations may be added to the validation summary.

Potential interfering substance to generate false negative result	Concentration	Volume of Positive Clinical Specimen Diluted, or titered "spike in" specimen	Results
List here in multiple rows: e.g. Hemoglobin, human anti-mouse antibody, rheumatoid factor	High end of clinical range		# detected / total # replicates

9.0 Matrix Equivalency (if applicable)

Characteristic	Number of samples	Planned Calculations/Evaluations
Other Matrix Equivalency	Measure # of titrations of analyte (high, medium, low, and equivocal if applicable) in # replicates of negative reference matrix in parallel to # replicates of negative test matrix	$\%\ \text{Recovery} = \frac{\text{mean value Other Matrix}}{\text{mean value Reference Matrix}} \times 100$

10.0 Multiplex Assay Performance (if applicable)

10.1 For assays that detect multiple targets, it is necessary to show that high concentrations of one target do not interfere with the detection of other targets.

Characteristic	Number of samples	Planned Calculations/Evaluations
Multiplex Analytic Sensitivity	List common co-infections	Describe the analytical sensitivity in the presence of co-infections

11.0 Roles and Responsibilities

11.1 (Insert name) is responsible for preparing the Method Validation Plan

11.2 (Insert name) is responsible for performing the Method Validation

11.3 (Insert name) is responsible for Document Management

11.4 (Insert name) is responsible for review and approval of the Validation Protocol prior to testing.

11.5 (Insert name) is responsible for review and approval of the Validation Protocol upon completion.

12.0 Proposed Timeline

12.1 Identify expected timeframe for each experiment and establish a timeline for completion of the validation and approval of the method for implementation by the laboratory to report results.

13.0 Related Documents

Document Name	Document Number	Rev. #	Effective date
(Method name) Procedure, if applicable		A	TBD
(Comparison Method name) Procedure, if applicable			

14.0 Instrumentation

Name/Model	Serial/ID #	Cal Date	Cal Due Date

15.0 Bioinformatics Pipeline

Name	Version #	Parameter Settings	Developer	Tech Support POC
(list the hardware, software, transmission system, backups, networks)				

16.0 Training Requirements

16.1 Personnel performing the Method Validation are required to complete and document training in the test procedure prior to validation. The following staff are trained and documentation is complete.

Personnel	Type of Training	Date Completed

17.0 Plan Approval (as applicable, please submit concurrently to CLIA Laboratory Director for approval)

Approved By: _____ Date: _____
CLIA Technical Supervisor (as applicable)

Approved By: _____ Date: _____
Team Lead

Approved By: _____ Date: _____
Quality Manager (as applicable)

Approved By: _____ Date: _____
Branch Chief

NGS Method Validation Summary Report Template

Note: This document is intended to be used as a template for summarizing results of a method validation for approval after the validation is performed. Existing entries are intended as guidance and may be changed, deleted, or augmented as needed for the laboratory's specific requirements. Parentheses in blue provide specific examples for appropriate input.

1.0 Scope

1.1 **Method Validation Report for:** (insert method name)

1.2 **Branch/Laboratory:** (branch and laboratory name)

1.3 **Test Procedure Document Number(s) and Revision Number:**

Document Name	Document Number	Revision Number	Effective Date

2.0 Changes from the method validation plan, as applicable: If new document versions were created, provide details describing the change. Examples include changes to the test method, changes to the number or origin of samples. If new document versions were created, record them in the table.

Document Name	Document Number	Revision Number	Effective Date

3.0 Summary of Results

Characteristic	Number of samples tested	Actual Performance (complete calculations)
Accuracy	(# of positive samples and # of negative samples) were measured	$\frac{(\mathrm{TP}+\mathrm{TN})}{(\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN})} \times 100$
Precision/ Reproducibility (Qualitative)	(# of positive samples and # of negative samples) were measured in (#) separate runs over at least (# ≥ 3) days, at least (#) days apart, by (#) different operators.	$\frac{\#\text{ of results in agreement}}{\text{total }\#\text{ of results}} \times 100$

Characteristic	Number of samples tested	Actual Performance (complete calculations)
Precision/ Reproducibility (Raw Value) (e.g. ANI score), as applicable	(# of positive samples and # of negative samples) were measured in (#) separate runs over at least (# ≥ 3) days, at least (#) days apart, by (#) different operators.	$\text{CV} = \frac{\text{standard deviation}}{\text{mean}} \times 100$
Clinical Sensitivity	(# of positive clinical samples) were measured	$\frac{\text{TP}}{(\text{TP} + \text{FN})} \times 100$
Analytic Sensitivity	(# of quantifiable samples at specified concentration levels near the expected detection limit) were measured	$\frac{\text{TP}}{(\text{TP} + \text{FN})} \times 100$
Clinical Specificity	(# of negative clinical samples) were measured	$\frac{\text{TN}}{(\text{FP} + \text{TN})} \times 100$
Analytic specificity	(# of negative samples known to contain potentially cross-reactive agtents or analytes) were measured	$\frac{\text{TN}}{(\text{FP} + \text{TN})} \times 100$ With description of cross-reactive agents or analytes
Limit of Detection	Depth of coverage range of (# to #) and consensus (% to %)were used to determine the bioinformatics LOD Target biological material present in the range (# to #) were used to determine the biological LOD	Bioinformatics LOD = (minimum depth of coverage and consensus %) Biological LOD = (minimum amount of target biological material)
Reference/Normal value	Determine: Literature review or to generate in-house data. In house: (# of normal population samples) were evaluated	Summary of results Provide references here

Characteristic	Number of samples tested	Actual Performance (complete calculations)
Applicable genome region	(#) samples were sequenced	Region of the genome in which sequence of acceptable quality was derived in the tested samples
Clinical validity	Determine: Literature review or generate in-house data.	Positive Predictive Value: TP / (TP + FP) , and Negative Predictive Value : TN / (TN + FN) by sample type and/or population or Provide references here
Interfering substances	List those tested	Summary of results
Matrix Equivalency, as applicable	Measured # of titrations of analyte (high, medium, low, and equivocal if applicable) in # replicates of negative reference matrix in parallel to # replicates of negative test matrix	$\%\ \text{Recovery} = \frac{\text{mean value Other Matrix}}{\text{mean value Reference Matrix}} \times 100$
Multiplex Assay Performance, as applicable	List common co-infections	Describe the analytical sensitivity in the presence of co-infections

4.0 Interpretation of Results

4.1 **Evaluation of discrepant results:** Provide details describing any discrepant results when compared to the gold standard. Provide evidence to support whether new method is likely true or not true result. If a decision is made to recategorize/exclude the sample from validation, additional evidence needs to be provided to support this decision.

4.2 **Limitations:** Provide description of method limitations. Statements may include: 1) The method is not appropriate to determine (describe the restriction). 2) The method is not appropriate for use under (describe the condition(s) under which it should not be used).

a. **Mitigating Processes:** Decribe processes used to mitigate the assay limitations (e.g., reject certain specimen types, reflexive testing for certain results).

4.3 **Disclaimers** (as applicable): Provide pertinent method limitations to be included on the final test report.

4.4 **Fast Tracked test QA monitoring plan** (as applicable): For tests that have been fast tracked due to urgent public health need, describe the metrics to be monitored and the additional evidence of test performance characteristics to be gathered. Include the timeline for gathering this evidence.

5.0 Statement of Suitability

5.1 The method validation of (method name) has been completed according to the documented plan. The (method name) meets all of the acceptance criteria and is approved for use in the (Insert name) Laboratory.

5.2 The method validation is applicable to the documents listed within the Method Validation Plan. Subsequent revisions involving technical changes to the procedure may require additional validation.

6.0 Summary Report Approval (as applicable, please submit concurrently to CLIA Laboratory Director for approval)

Approved By: _____ Date: _____
CLIA Technical Supervisor (as applicable)

Approved By: _____ Date: _____
Team Lead

Approved By: _____ Date: _____
Quality Manager (as applicable)

Approved By: _____ Date: _____
Branch Chief

7.0 Appendices

7.1 **Summary data:** Please provide a line-by-line listing of all samples and results, and relevant summary data. Excel spreadsheets are acceptable.

7.2 Note: Please retain all validation data (worksheets, controls) and have available for review.

1.0 Purpose

1.1 This document provides quality control (QC) guidance for nucleic acid sequencing using the Oxford Nanopor MinION technology. The guidance takes into account specific QC checkpoints between laboratory processes to ensure each step is completed correctly, with high confidence, and to generate quality data metrics that are informative for downstream bioinformatics processes.

1.2 The quality of nucleic acid extraction and manipulation, fractionations and size selection, and library preparations affects fragment size uniformity and library diversity, which is important for achieving complete and even coverage of the total nucleic acid to be sequenced. Gaps resulting from poor sample preparation cannot be corrected downstream by error correction methods employed by some sequencing technologies. In addition, quality scores do not reflect errors introduced during sample preparation, as the sequencing signal will appear clean and error-free. The maximal achievable accuracy of most sequencing platforms is limited by the sample accuracy.

2.0 NGS QC Checkpoints

The following sections correspond to the process steps prior to sequencing, as outlined in Figure 1.

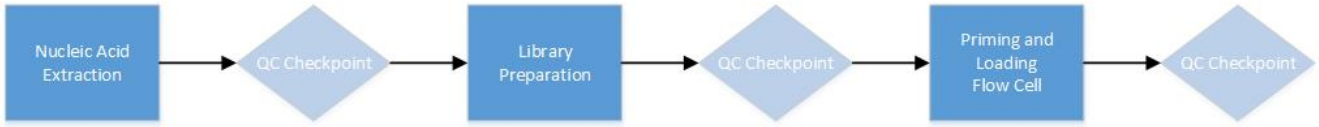


Figure 1: NGS QC Checkpoints for MinION 1D Workflows

2.1 Nucleic Acid Extraction

a. High quality nucleic acid purification is essential for obtaining accurate NGS data. The extraction method depends greatly on the sample type and matrices involved. See Appendix A for extraction methods recommended by Nanopore.
Note: Proteinase K has been known to cause pore degradation. It is recommended to use an extraction method that does not use proteinase K.

2.2 Post Extraction Nucleic Acid QC Checkpoint

It is important to check input DNA for quality before beginning library preparation. Low molecular weight, incorrectly qualified and/or contaminated DNA (e.g, salt, EDTA, protein, organic solvents) can have a significant impact on downstream processes and ultimately, your sequencing run.

a. Criteria for Input DNA

i. Purity as measured using Nanodrop – OD 260/280 of ~1.8 and OD 260/230 of 2.0-2.2. A 260/280 which is higher than ~1.8 indicates the presence of RNA. A 260/280 which is lower than ~1.8 can indicate the presence of protein or phenol. Establish the precise acceptable 260/280 range for your test during development and validation.

ii. Average fragment size >30kb. Fragment size may be measured using several methods (e.g., pulse-field, low percentage agarose gel analysis, blue pippin). This quality checkpoint is important during the development and validation of the test. Labs may elect to omit this quality check after validation if the test has proven robust and stable.

iii. Input mass, as measured by Qubit – 1 µg or 1.5 µg if carrying out a DNA repair step. In order to maximize sequencing yield, it is important that the nanopores are kept filled with DNA to minimize the time they are idle between strands. For further optimization of fragment length to improve throughput, see table 1 in section 2.6.

b. Use the configuration test cell to confirm the MinION is communicating with the computer.

i. The configuration protocol has been successfully completed when the message "Customer configuration run has completed" is displayed in the notifications panel. If configuration reports that it has failed, reinsert the flow cell and trouble shoot per manufacturer's instructions. Upon successful configuration, the MinION and MinKNOW systems are ready for platform QC of the flow cell (see Section 2.7).

2.3 DNA Fragmentation (optional): DNA fragmentation is an optional step for when experiments require specific fragment sizes.

2.4 Fragmentation QC Checkpoint

a. Determine the fragment size, quantity, and quality using the Agilent Bioanalyzer or similar instrument. Confirm the fragment size is within the expected range. If the results yield smaller fragments, this is indicative of substantial shearing/degradation of the input material and is likely to reduce the quality of the library preparation and the read length distribution

2.5 Library Preparation: Perform library preparation according to the selected protocol. It is recommended that the repaired/end-prepped DNA sample is subjected to clean-up with AMPure XP beads. This clean-up can be omitted for simplicity and to reduce library preparation time. However, it has been observed that omission of this clean-up can: reduce subsequent adapter ligation efficiency, increase the prevalence of chimeric reads, and lead to an increase in pores being unavailable for sequencing.

2.6 Library Preparation QC Checkpoint

a. In order to maximize sequencing yield, it is important that the nanopores are kept filled with DNA to minimize the time they are idle between strands. The less material goes into the flow cell, the fewer "threadable ends" will be present to be captured by the pores. Therefore, the pores will be searching for molecules for longer, and if the pores are not always sequencing, throughput could be compromised.

b. Note: During development and optimization of a method it is advisable to check the fragment size and final DNA input concentration of the library before proceeding to priming and loading the library. The below table may be used to inform optimization experiments.

Mass of extracted nucleic acid	No. of moles if library fragment length = 2kb	No. of moles if library fragment length = 8kb	No. of moles if library fragment length = 50 kb
10 µg	7.7 pmol	1.9 pmol	308 fmol
5 µg	3.9 pmol	963 fmol	154 fmol
3.5 µg	2.7 pmol	674 fmol	108 fmol
2 µg	1.5 pmol	385 fmol	62 fmol
1.5 µg	1.2 pmol	289 fmol	46 fmol
1 µg	770 fmol	193 fmol	31 fmol

500 ng	385 fmol	96 fmol	15 fmol
400 ng	308 fmol	77 fmol	12 fmol
200 ng	154 fmol	39 fmol	6.2 fmol
100 ng	77 fmol	19 fmol	3.1 fmol
30 ng	23 fmol	5.8 fmol	0.9 fmol
10 ng	7.7 fmol	1.9 fmol	0.3 fmol
10 pg	0.0077 fmol	0.009 fmol	0.0003 fmol

Table 1: Fragment Length

i. In order to keep the pores full, the current R9.4.1 and R9.5.1 pores require about **5-50 fmol of good quality library put into the flow cell.**

ii. Quantify 1 µL of adapter ligated DNA using a Qubit fluorometer. Expected recovery is 430 ng.

2.7 Pre-loading QC Checkpoint: As the MinKNOW script progresses, check the following:

a. Number of active pores should be 800 or greater

b. Heatsink temperatures: (34°C)

2.8 Post Loading QC Checkpoint: The library is loaded dropwise. Ensure each drop flows into the port before adding the next. Be sure to pipette slowly when adding priming mix to priming port to ensure the membrane stays intact. For further details on loading the Oxford Nanopore MinION flow cell click here.

2.9 Post Loading QC Checkpoint

a. Number of active pores should be above 800

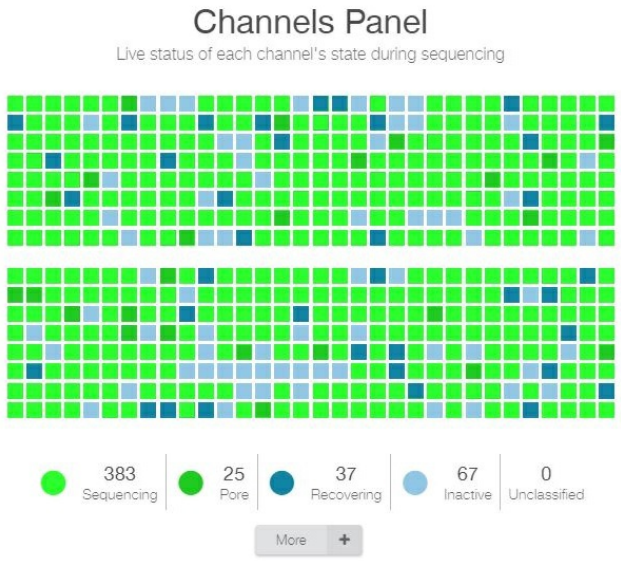
b. Development of the read histogram: Confirm the histogram reflects expected read lengths for the experimental design being used.

c. Pore occupancy: Monitor the pore occupancy for the first 30 minutes of a sequencing experiment. If you are not observing the expected percentage of pores in stand sequencing, stop the run, wash the flow cell and store it for use in another run. A good library will be indicated by a higher proportion of light green channels in Sequencing state (neon green) than are in Pore state (green). The combination of Sequencing and Pore channels indicate the number of active pores at any point in time. A low proportion of sequencing channels will reduce the throughput of the run.

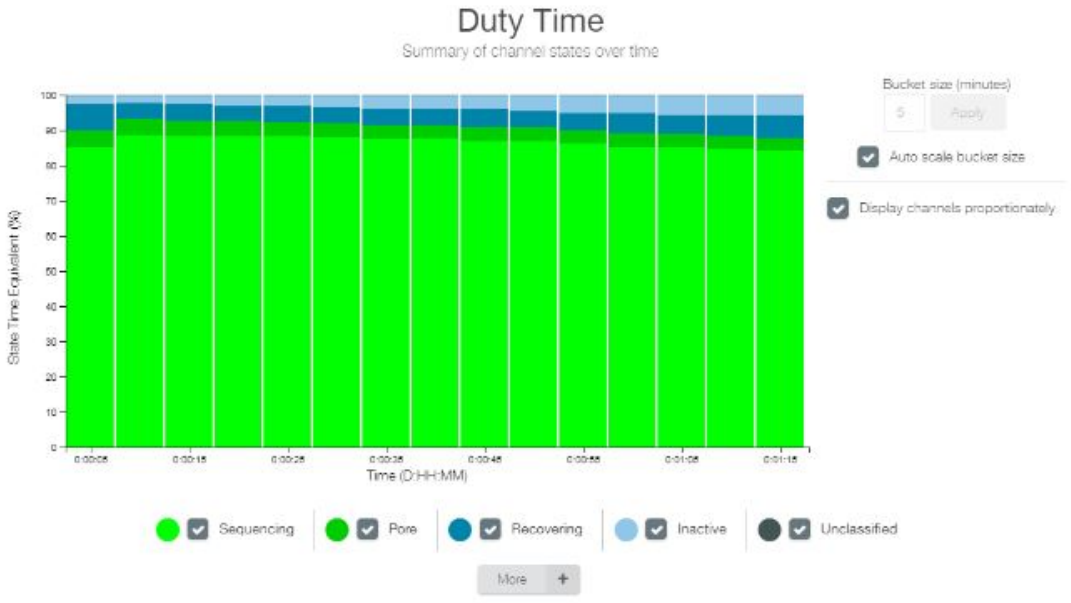
i. **Recovering** (dark blue) indicates channels that may become available for sequencing again. A high proportion of this may indicate additional clean up steps are required during your library preparation.

ii. **Inactive** (light blue) indicates channels that are no longer available for sequencing. A high proportion of these as soon as the run begins may indicate an osmotic imbalance.

iii. **Unclassified** are channels that have not yet been assigning one of the above classifications.



d. Good quality library: A good quality library will result in most of the pores being in the "Sequencing" state (neon green), and very few in "Pore" (green), "Recovering" (dark blue), or "Inactive" (light blue). A library that results in a Duty Time graph like the example below is likely to give a good sequencing throughput. The graph populates over time, and can be used as a way to assess the quality of your sequencing experiment, and make an early decision whether to continue with the experiment or to stop the run.



e. Base Calling Report: Confirm the local basecalling is being recorded in the base calling report and is within expected range *(insert laboratory specific range here).*

3.0 Appendices

3.1 Appendix A – NGS MinION Extraction Methods

4.0 Revision History

Rev #	DCR #	Change Summary	Date

5.0 Approval

Approved By: _____ Date: _____
Author

_____
Print Name and Title

Approved By: _____ Date: _____
Supervisor

_____
Print Name and Title

Approved By: _____ Date: _____
Quality Manager

_____
Print Name

Appendix A – Extraction Methods

Extraction Method	
Gram-negative bacterial DNA	Click here for protocol
Gram-positive bacterial DNA	Click here for protocol
Yeast DNA	Click here for protocol
Yeast RNA	Click here for protocol

1.0 Purpose

1.1 This document provides quality control (QC) guidance for nucleic acid sequencing using the Oxford Nanopore MinION Rapid Sequencing technology. The guidance takes into account specific QC checkpoints between laboratory processes to ensure each step is completed correctly, with high confidence, and to generate quality data metrics that are informative for downstream bioinformatics processes.

1.2 The quality of nucleic acid extraction and manipulation, fractionations and size selection, and library preparations affects fragment size uniformity and library diversity, which is important for achieving complete and even coverage of the total nucleic acid to be sequenced. Gaps resulting from poor sample preparation cannot be corrected downstream by error correction methods employed by some sequencing technologies. In addition, quality scores do not reflect errors introduced during sample preparation, as the sequencing signal will appear clean and error-free. The maximal achievable accuracy of most sequencing platforms is limited by the sample accuracy.

2.0 NGS QC Checkpoints

The following sections correspond to the process steps prior to sequencing, as outlined in Figure 1.

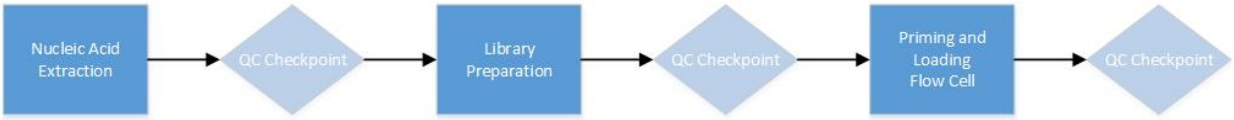


Figure 1: NGS QC Checkpoints for MinION Rapid Sequencing Workflows

2.1 Nucleic Acid Extraction

High quality nucleic acid purification is essential for obtaining accurate NGS data. The extraction method depends greatly on the sample type and matrices involved. See Appendix A for extraction methods recommended by Nanopore.

Note: Proteinase K has been known to cause pore degradation. It is recommended to use an extraction method that does not use proteinase K.

2.2 Post Extraction Nucleic Acid QC Checkpoint

It is important to check input DNA for quality before beginning library preparation. Low molecular weight, incorrectly qualified and/or contaminated DNA (e.g, salt, EDTA, protein, organic solvents) can have a significant impact on downstream processes and ultimately, your sequencing run.

a. Criteria for Input DNA

i. Purity as measured using Nanodrop – OD 260/280 of ~1.8 and OD 260/230 of 2.0-2.2. A 260/280 which is higher than ~1.8 indicates the presence of RNA. A 260/280 which is lower than ~1.8 can indicate the presence of protein or phenol. Establish the precise acceptable 260/280 range for your test during development and validation.

ii. Average fragment size >30kb. Fragment size may be measured using several methods (e.g., pulse-field, low percentage agarose gel analysis, blue pippin). This quality checkpoint is important during the development and validation of the test. Labs may elect to omit this quality check after validation if the test has proven robust and stable.

iii. Input mass, as measured by Qubit - ~400 ng. In order to maximize sequencing yield, it is important that the nanopores are kept filled with DNA to minimize the time they are idle between strands. For further optimization of fragment length to improve throughput, see table 1 in section 2.3.

b. Use the configuration test cell to confirm the MinION is communicating with the computer.

i. The configuration protocol has been successfully completed when the message "Customer configuration run has completed" is displayed in the notifications panel. If configuration reports that it has failed, reinsert the flow cell and trouble shoot per manufacturer's instructions. Upon successful configuration, the MinION and MinKNOW systems are ready for platform QC of the flow cell (see section 2.4).

2.3 Library Preparation

There are many library preparation kits available specific to the kind of sequencing and downstream application. The library preparation does not have a separate checkpoint during routine testing. Proceed to Pre-loading QC checkpoint.

2.4 Library Preparation QC Checkpoint

Note: During development and optimization of a method it is advisable to check the fragment size and final DNA input concentration of the library before proceeding to priming and loading the library. The below table may be used to inform optimization experiments.

Mass of extracted nucleic acid	No. of moles if library fragment length = 2kb	No. of moles if library fragment length = 8kb	No. of moles if library fragment length = 50 kb
10 µg	7.7 pmol	1.9 pmol	308 fmol
5 µg	3.9 pmol	963 fmol	154 fmol
3.5 µg	2.7 pmol	674 fmol	108 fmol
2 µg	1.5 pmol	385 fmol	62 fmol
1.5 µg	1.2 pmol	289 fmol	46 fmol
1 µg	770 fmol	193 fmol	31 fmol
500 ng	385 fmol	96 fmol	15 fmol
400 ng	308 fmol	77 fmol	12 fmol
200 ng	154 fmol	39 fmol	6.2 fmol
100 ng	77 fmol	19 fmol	3.1 fmol
30 ng	23 fmol	5.8 fmol	0.9 fmol
10 ng	7.7 fmol	1.9 fmol	0.3 fmol
10 pg	0.0077 fmol	0.009 fmol	0.0003 fmol

Table 1: Fragment Length

2.5 Pre-loading QC Checkpoint: As the MinKNOW script progresses, check the following:

a. Number of active pores should be 800 or greater

b. Heatsink temperatures: (34°C)

2.6 Priming and Loading: Add priming mix very slowly to ensure the membrane and protein pores are not damaged. The library is loaded dropwise. Ensure each drop flows into the port before adding the next. For further details on loading the Oxford Nanopore MinION flow cell click here.

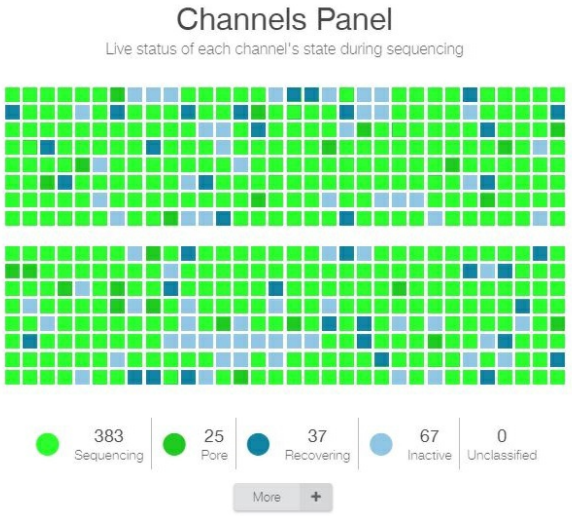
2.7 Post Loading QC Checkpoint

a. Number of active pores should be above 800

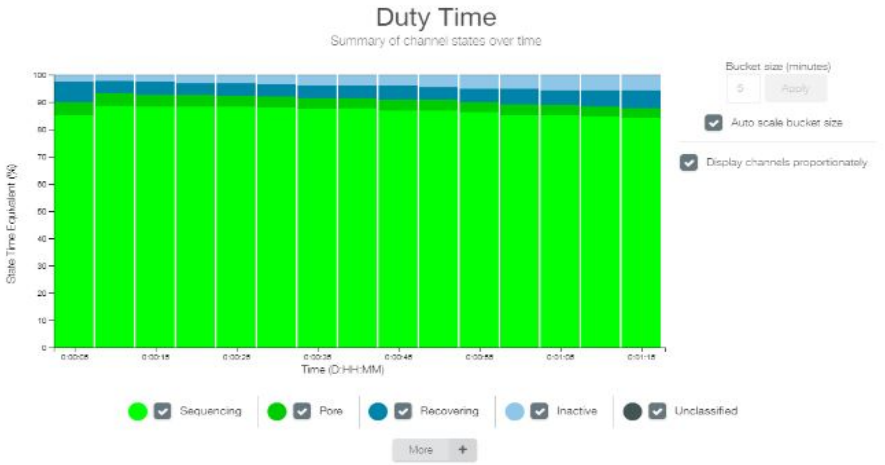
b. Development of the read histogram: Confirm the histogram reflects expected read lengths for the experimental design being used.

c. Pore occupancy: Monitor the pore occupancy for the first 30 minutes of a sequencing experiment. If you are not observing the expected percentage of pores in strand sequencing, stop the run, wash the flow cell and store it for use in another run. A good library will be indicated by a higher proportion of light green channels in Sequencing state (neon green) than are in Pore state (green). The combination of Sequencing and Pore channels indicate the number of active pores at any point in time. A low proportion of sequencing channels will reduce the throughput of the run.

- **i.** **Recovering** (dark blue) indicates channels that may become available for sequencing again. A high proportion of this may indicate additional clean up steps are required during your library preparation.
- **ii.** **Inactive** (light blue) indicates channels that are no longer available for sequencing. A high proportion of these as soon as the run begins may indicate an osmotic imbalance.
- **iii.** **Unclassified** are channels that have not yet been assigning one of the above classifications.



d. Good quality library: A good quality library will result in most of the pores being in the "Sequencing" state (neon green), and very few in "Pore" (green), "Recovering" (dark blue), or "Inactive" (light blue). A library that results a Duty Time graph like the example below is likely to give a good sequencing throughput. The graph populates over time, and can be used as a way to assess the quality of your sequencing experiment, and make an early decision whether to continue with the experiment or to stop the run.



e. Base Calling Report: Confirm the local basecalling is being recorded in the base calling report and is within expected range *(insert laboratory specific range here).*

3.0 Appendices

Appendix A- NGS MinION Extraction Methods

4.0 Revision History

Rev #	DCR #	Change Summary	Date

5.0 Approval

Approved By: _____ Date: _____
Author

_____
Print Name and Title

Approved By: _____ Date: _____
Supervisor

_____
Print Name and Title

Approved By: _____ Date: _____
Quality Manager

_____
Print Name

Appendix A – Extraction Methods

Extraction Method	
Gram-negative bacterial DNA	Click here for protocol
Gram-positive bacterial DNA	Click here for protocol
Yeast DNA	Click here for protocol
Yeast RNA	Click here for protocol