1. **PURPOSE:** All genomic sequences generated under PulseNet surveillance are uploaded in real-time to the sequence read archive (SRA) located at NCBI. PulseNet has set minimum coverage and quality score requirements for sequences to be uploaded to SRA. The purpose of this document is to describe a standardized procedure for Illumina MiSeq sequence data quality control (QC) prior to sharing the sequence files with PulseNet Central.

2. **SCOPE:** This procedure applies to all whole genome sequence data generated by PulseNet surveillance laboratories.

3. **DEFINITIONS/ACRONYMS:**
   - **3.1. BaseSpace:** Illumina cloud-based computing environment for next generation sequencing data analysis, management and storage, including data sharing
   - **3.2. Biosample**: Term used by NCBI to for the descriptive information of unique (biologically or physically) specimens/sequenced bacterial strains; it is the metadata for a sample
   - **3.3. CDC:** Centers for Disease Control and Prevention
   - **3.4. Cluster density:** Density of clusters (in thousand per square millimeter) generated on the flow cell prior to sequencing
   - **3.5. Coverage:** The average number of reads that include a given nucleotide in the reconstructed sequence
   - **3.6. FASTQ**: A text-based file format for storing both sequence and its corresponding quality scores.
   - **3.7. FTP:** File Transfer Protocol**;** a standard network protocol used to transfer computer files from one host to another host via Internet.
   - **3.8. NCBI:** National Center for Biotechnology Information, part of the National Institutes of Health (NIH). NCBI houses several databases relevant to biotechnology, including GenBank for DNA sequence assemblies and Sequence Read Archive (SRA) for raw reads.
   - **3.9. PF Reads:** Passing Filter Reads, the number of reads which passed filtering (are useable reads) for a sequencing run
   - **3.10. PHL:** Public Health Lab
   - **3.11. PN:** PulseNet
   - **3.12. QC:** Quality Control
   - **3.13. Q score:** The sequence quality score for each individual base position in a sequence. Phred scores are used, where Q = -10log (Error Probability). The higher the quality score, the more reliable the base call. A Q30 means a 1 in 1000 likelihood of an incorrect base call at that position.
   - **3.14. Q30 (%):** Refers to the sequencing run and represents the percentage of nucleotide calls in a sequencing run which have a Q score of 30 or higher
   - **3.15. RTA:** Real-time Analysis software, generates quality metrics of the sequencing run
   - **3.16. SAV:** Sequencing Analysis Viewer, an application software that allows real-time viewing of quality metrics generated by the real-time analysis (RTA) software on the Illumina sequencing systems
   - **3.17. SRA:** Sequence Read Archive, database at NCBI which stores raw sequence data and alignment information
   - **3.18. SOP:** Standard Operating Procedure

4. RESPONSIBILITIES:

4.1. **PulseNet Public Health Laboratory:**

4.1.1. Sequence isolates and perform quality check of the sequencing run and subsequent sequence data.

4.1.2. Re-sequence any isolates which do not meet quality thresholds.

4.1.3. Communicate any instrument or sequencing issue with PulseNet Central, as necessary.

4.2. **PulseNet Central:**

4.2.1. Perform additional sequence data quality analysis.

4.2.2. Notify PN Public Health laboratory if any sequences do not meet quality thresholds.

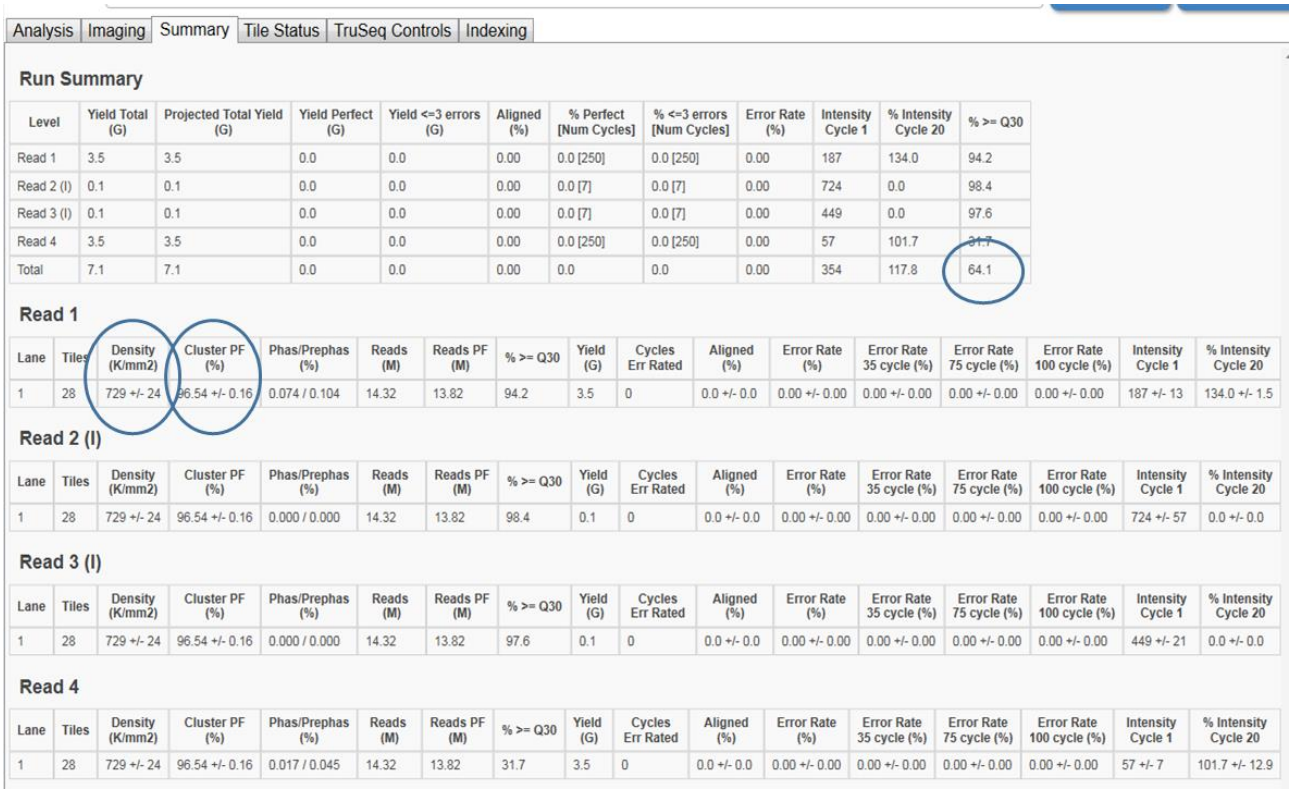
4.2.3. Assist PN Public Health laboratories with troubleshooting, as necessary.

5. PROCEDURE:

5.1. **Review Run Metrics**: Upon run completion, confirm that the sequencing run meets the basic quality metrics (Q30, Cluster Density and Clusters Passing Filter). See Table 1 below. These run metrics may be found on the Summary Tab of SAV, see Figure 1. **NOTE*: If the Q30 for a run is below the threshold listed below, the run must be repeated. If the Cluster Density & Clusters Passing Filter metrics do not meet the threshold below, but the Q30 is sufficient, the run may or may not need to be repeated. Further analysis of sequence data is required. Contact PulseNetNGSlab@cdc.gov for assistance.***

Kit Chemistry	Q30 (%)	Cluster Density (K/mm^2)	Clusters Passing Filter (%)
v3, 600 cycle	≥ 70	1200-1400	~ 80 or higher
v2, 500 cycle	≥ 75	600-1200	~ 80 or higher
v2, 300 cycle	≥ 80	600-1200	~ 80 or higher
Nano, v2 500 cycle	≥ 75	600-1200	~ 80 or higher
Micro, v2 300 cycle	≥ 80	600-1200	~ 80 or higher

Table 1. Run Metric Thresholds



Analysis | Imaging | Summary | Tile Status | TruSeq Controls | Indexing

Run Summary

Level	Yield Total (G)	Projected Total Yield (G)	Yield Perfect (G)	Yield <=3 errors (G)	Aligned (%)	% Perfect [Num Cycles]	% <=3 errors [Num Cycles]	Error Rate (%)	Intensity Cycle 1	% Intensity Cycle 20	% >= Q30
Read 1	3.5	3.5	0.0	0.0	0.00	0.0 [250]	0.0 [250]	0.00	187	134.0	94.2
Read 2 (I)	0.1	0.1	0.0	0.0	0.00	0.0 [7]	0.0 [7]	0.00	724	0.0	98.4
Read 3 (I)	0.1	0.1	0.0	0.0	0.00	0.0 [7]	0.0 [7]	0.00	449	0.0	97.6
Read 4	3.5	3.5	0.0	0.0	0.00	0.0 [250]	0.0 [250]	0.00	57	101.7	31.7
Total	7.1	7.1	0.0	0.0	0.00	0.0	0.0	0.00	354	117.8	64.1

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	% Intensity Cycle 20
1	28	729 +/- 24	96.54 +/- 0.16	0.074 / 0.104	14.32	13.82	94.2	3.5	0	0.0 +/- 0.0	0.00 +/- 0.00	0.00 +/- 0.00	0.00 +/- 0.00	0.00 +/- 0.00	187 +/- 13	134.0 +/- 1.5

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	% Intensity Cycle 20
1	28	729 +/- 24	96.54 +/- 0.16	0.000 / 0.000	14.32	13.82	98.4	0.1	0	0.0 +/- 0.0	0.00 +/- 0.00	0.00 +/- 0.00	0.00 +/- 0.00	0.00 +/- 0.00	724 +/- 57	0.0 +/- 0.0

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	% Intensity Cycle 20
1	28	729 +/- 24	96.54 +/- 0.16	0.000 / 0.000	14.32	13.82	97.6	0.1	0	0.0 +/- 0.0	0.00 +/- 0.00	0.00 +/- 0.00	0.00 +/- 0.00	0.00 +/- 0.00	449 +/- 21	0.0 +/- 0.0

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	% Intensity Cycle 20
1	28	729 +/- 24	96.54 +/- 0.16	0.017 / 0.045	14.32	13.82	31.7	3.5	0	0.0 +/- 0.0	0.00 +/- 0.00	0.00 +/- 0.00	0.00 +/- 0.00	0.00 +/- 0.00	57 +/- 7	101.7 +/- 12.9

Figure 1. Screen-shot of SAV "Summary Tab" highlighting key run metrics to be reviewed (in blue circles).

5.2. **<u>Calculate Coverage for Individual Sequences:</u>** Calculate average coverage for each isolate to determine if the sequence data passes coverage requirements (see Table 2 below for coverage requirements). Coverage may be calculated using one of the following methods:

- Using the ReadMetrics tab in the Nextera XT library prep workbook (Appendix PNL32-4 and PNL32-5), using values from BaseSpace or SAV: See Section 5.2.1 below.

OR

- Manually using values from SAV/BaseSpace: See Section 5.2.2 below.

OR

- Manually using FastQC: See Section 5.2.3 below

OR

- Using Illumina BaseSpace Sequence Hub

  NOTE: *BaseSpace iCredits will need to be purchased to use the cloud computing capacity for this analysis. Refer to Illumina's Help section on BaseSpace for information on how to analyze runs and data. Refer to this SOP for quality metric information and thresholds.*

5.2.1. **Using the ReadMetrics tab of the Nextera XT library prep workbook (Appendix PNL32-4 and PNL32-5):**

NOTE1: *The following fields must be filled in correctly in the workbook in order for the coverage to be accurately calculated:*

- *Initial Dilution tab: Sample ID and Genome Size Estimate*
- *Read Metrics tab: Sample ID (matching the Initial Dilution tab), % Reads Identified (PF), and PF Reads*
- *Read Metrics tab: Correct number of cycles - kit selected from the "Number of Cycles" drop-down menu*

NOTE2: *If coverage for all isolates passes the thresholds listed in the table below, proceed to assess sequence quality. If the coverage does not pass the required threshold, the isolate will need to be re-sequenced.*

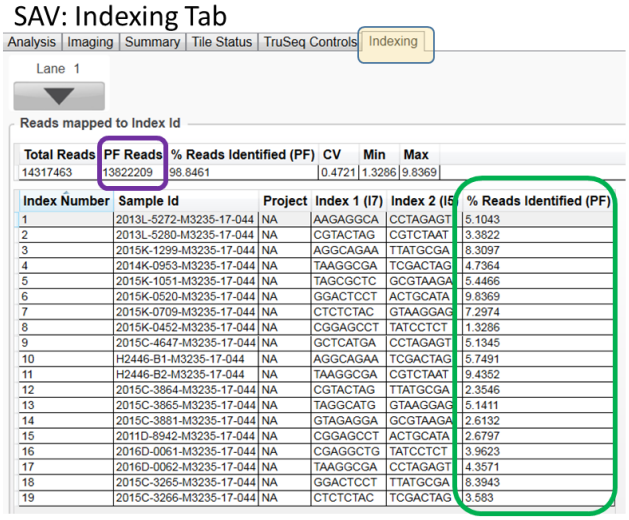
5.2.1.1. Open the "Read Metrics" tab of the workbook.

5.2.1.2. Enter the "PF Reads" value for the run into column G of the workbook. In SAV or BaseSpace: This value is found on the Indexing (SAV) or the Indexing QC tab (BaseSpace) as the "**PF Reads**" value (See Figures 2 & 3)

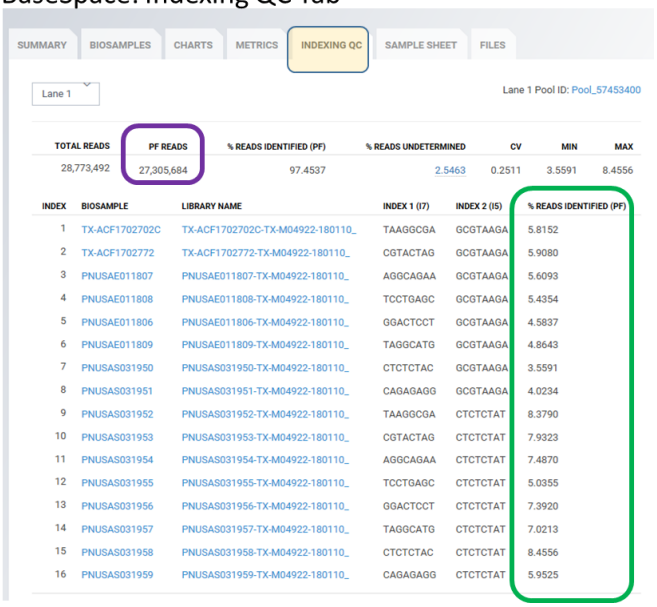
5.2.1.3. Enter the value of the "% Reads Identified (PF)" into column F of the workbook. This value represents the percentage of the PF reads that have been assigned to a particular set of indices and will vary for each isolate. In SAV or BaseSpace this value is found on the Indexing (SAV) tab or the Indexing QC tab (BaseSpace) as "**% Reads Identified (PF)**" (See Figures 2 & 3)

5.2.1.4. Estimated coverage and the total number of reads (bp) for each isolate should now be displayed in Column I and H of the workbook. (See Figure 4) See Table 2 for passing coverage requirements.

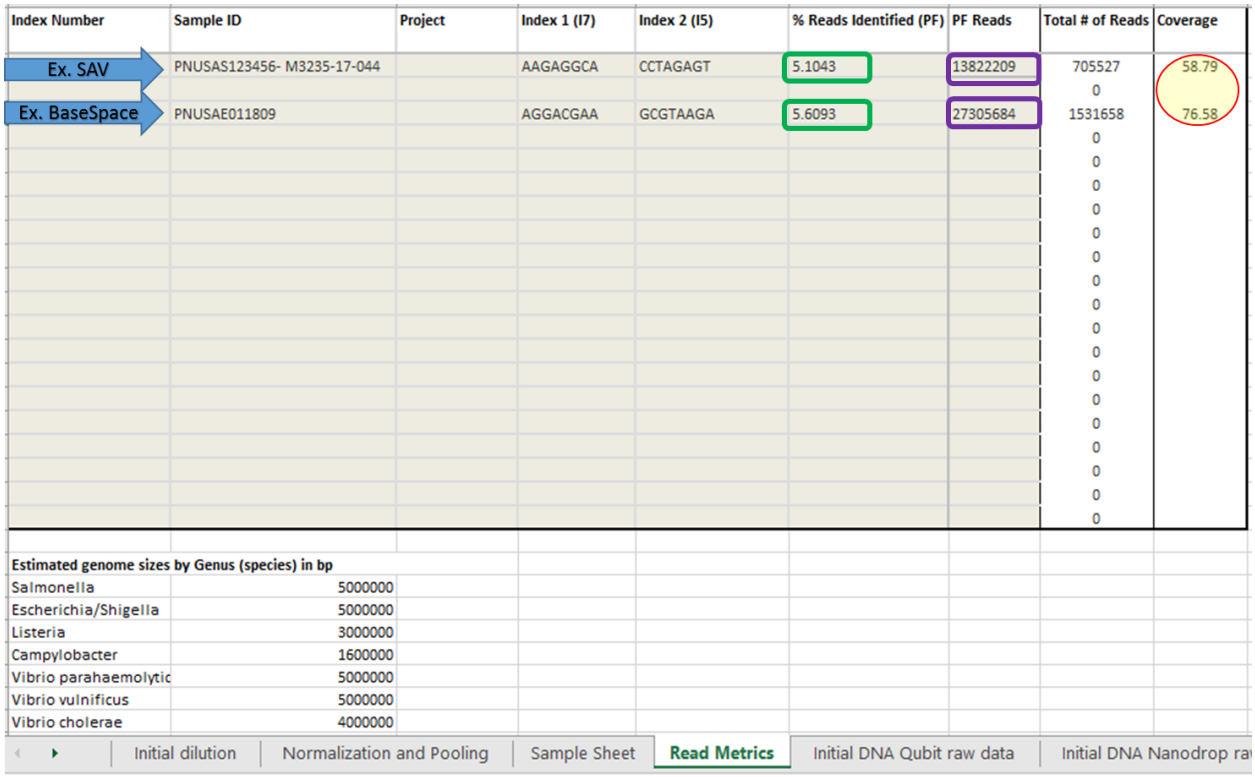
SAV: Indexing Tab



BaseSpace: Indexing QC Tab



Figures 2 & 3. Screen-shot of the Indexing QC tab in SAV (left) and BaseSpace (right).



Index Number	Sample ID	Project	Index 1 (I7)	Index 2 (I5)	% Reads Identified (PF)	PF Reads	Total # of Reads	Coverage
Ex. SAV	PNUSAS123456- M3235-17-044		AAGAGGCA	CCTAGAGT	5.1043	13822209	705527	58.79
							0	
Ex. BaseSpace	PNUSAE011809		AGGACGAA	GCGTAAGA	5.6093	27305684	1531658	76.58
							0	
							0	
							0	
							0	
							0	
							0	
							0	
							0	
							0	
							0	
							0	
							0	
							0	
							0	
							0	
							0	
							0	
							0	
							0	
							0	

Estimated genome sizes by Genus (species) in bp	
Salmonella	5000000
Escherichia/Shigella	5000000
Listeria	3000000
Campylobacter	1600000
Vibrio parahaemolytic	5000000
Vibrio vulnificus	5000000
Vibrio cholerae	4000000

Initial dilution | Normalization and Pooling | Sample Sheet | **Read Metrics** | Initial DNA Qubit raw data | Initial DNA Nanodrop ra

Figure 4. Example of Nextera XT workbook, ReadMetrics Tab

5.2.2. **To calculate coverage manually using data from SAV or BaseSpace:**

((PF Reads × % Reads Identified PF) × Maximum read length) = Coverage (x)

(Estimated genome size)

- PF Reads & % Reads Identified (PF): From the Indexing tab (SAV) or the Indexing QC tab (BaseSpace)
  NOTE4: *Use the percentage, NOT the percentage value – see Example below*
- Maximum read length = ½ the number of cycles in the run
  - *Ex: For a 500 cycle run, the Maximum read length = 250*
- Estimated genome size = See Table 2 below

- **Example:**

Using the data from **Figure 2** to determine coverage for 2013L-5272-M3235-17-044, and assuming that this was a 500 cycle run:

PF (for the run): 13822209
% Reads Identified (PF) for 2013L-5272-M3235-17-044: 5.1043 % = 0.051043
Maximum read length: ½ of 500 = 250
Estimated genome size (*Listeria*): 3000000

- Using the formula, coverage for isolate 2013L-5272-M3235-17-044 is:

((13822209 × 0.051043) × 250) / 3000000 = 58.79x

Organism	Estimated Genome size (bp)	Target Quality Thresholds		Passing but with additional qualifications – Scenario 1		Passing but with additional qualifications – Scenario 2	
		Q scores	Coverage	Q scores	Coverage	Q scores	Coverage
Listeria monocytogenes	3000000	≥ 30	≥ 20x	29.0-29.99	≥ 30x	28.0-28.99	≥ 40x
E. coli/Shigella spp.	5000000	≥ 30	≥ 40x	29.0-29.99	≥ 50x	28.0-28.99	≥ 60x
Salmonella spp.	5000000	≥ 30	≥ 30x	29.0-29.99	≥ 40x	28.0-28.99	≥ 50x
Campylobacter spp.	1600000	≥ 30	≥ 20x	29.0-29.99	≥ 30x	28.0-28.99	≥ 40x
Vibrio spp.	5000000	≥ 30	≥ 40x	29.0-29.99	≥ 50x	28.0-28.99	≥ 60x

Table 2. Coverage Requirements for PulseNet Organisms, including minimum combined coverage and Q score acceptance.
If Q scores are below 30, more coverage is required for sequence data to pass.

5.2.3. **To calculate coverage manually using FastQC**

5.2.3.1 Open FastQC and choose File on the main screen.

5.2.3.2 Choose Open and then select one read file (.fastq file), either R1 or R2, of the sequence to be analyzed.

5.2.3.3 Using the **Total Sequences** value and the maximum number depicted for **Sequence Length**, calculate coverage using the formula: **(Total Sequences x Maximum Sequence Length x 2)/Estimated genome length = isolate coverage. NOTE**: *This will give the coverage for the isolate. Multiplying by 2 is required for determining coverage for paired-end reads, and accounts for R1 and R2. If determining coverage for one sequence read (not for the isolate), do not multiply by 2.* See Figure 5 and the Example below.

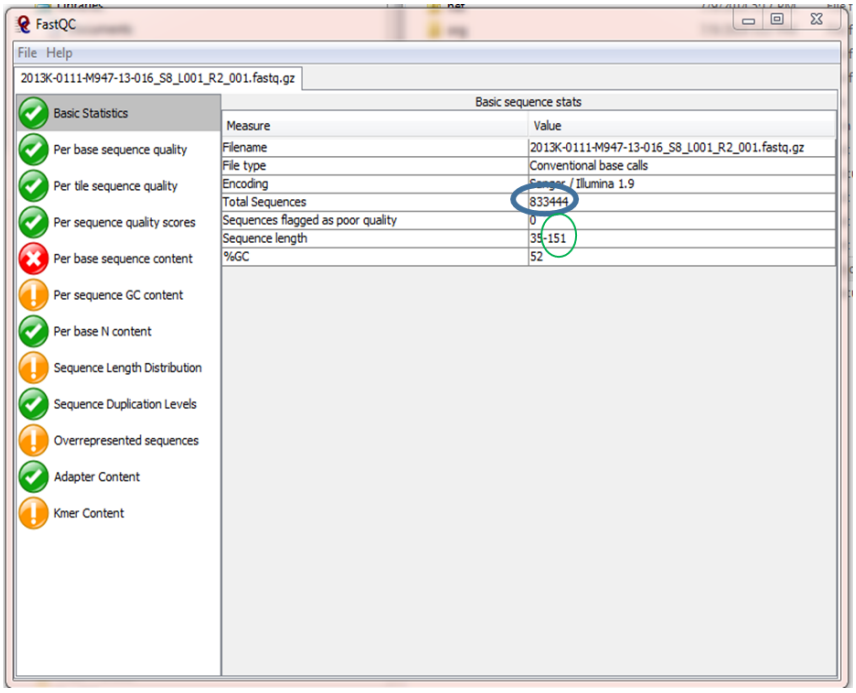


Figure 5. Basic Statistics tab in FastQC

- Example: Coverage calculation for this isolate (assuming this is a *Salmonella* isolate):

  (833444 × 151 × 2)/5000000 = 50.34x

  NOTE: *See Table 2 in Section 5.2.2 of the SOP for coverage thresholds and estimated genome sizes for PulseNet organisms and Section 5.3 for additional information on interpretation of FastQC graphs.*

5.3. **Review Sequence Data Quality:** This is a basic overview to evaluate general sequence quality using some of the graphs generated by the FastQC software. The graphs are open to subjective interpretation, and will not provide actual numeric quality score values. Additional analyses will be performed by PulseNet Central to obtain numeric average quality score values. If minimum required quality scores and coverage are not met, the isolate will need to be re-sequenced. Note that, if quality scores are below 30 (i.e. 28-29.99), the sequence data may still be accepted but will require additional coverage (see Table 2). **Any sequence data with an average Q score < 28.0 will need to be re-sequenced regardless of the amount of coverage.**
NOTE: *FastQC may be downloaded, free of charge, at:* www.bioinformatics.babraham.ac.uk/projects/fastqc

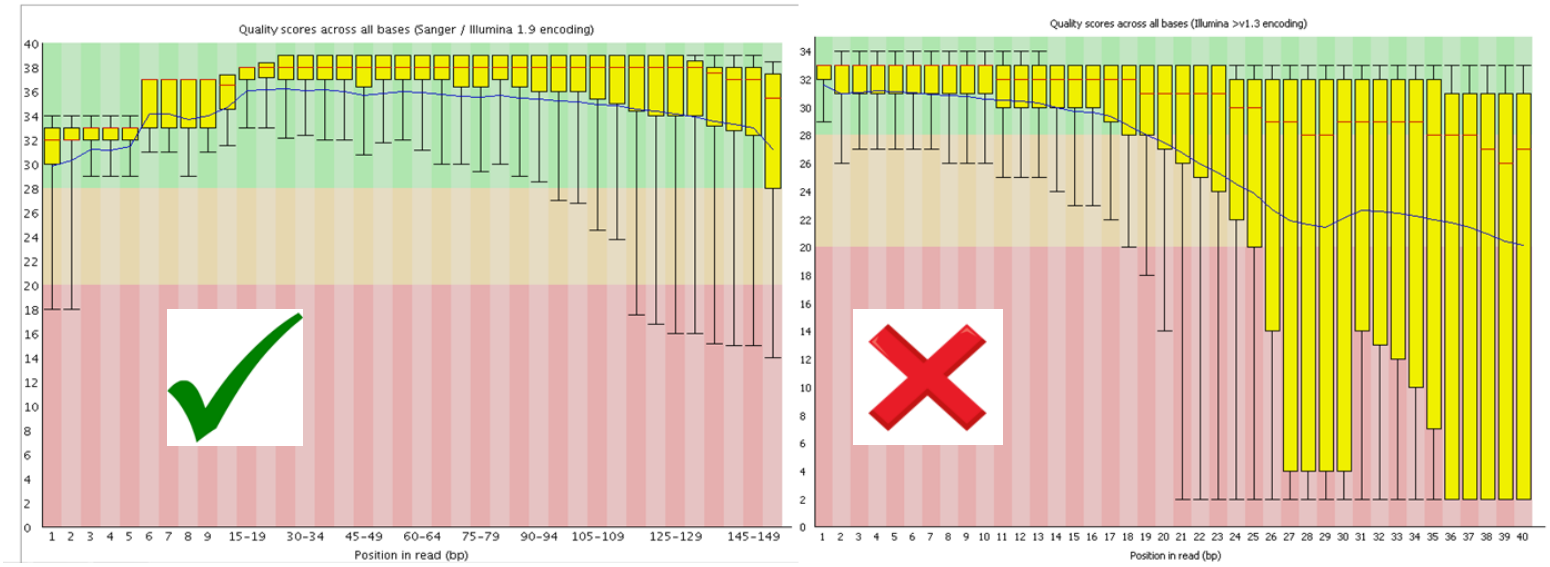
5.3.1. Using Fast QC:

5.3.1.1. Open FastQC and select "File" from the toolbar on the main screen.

5.3.1.2. Choose "Open" and select a read file for analysis (.fastq file).

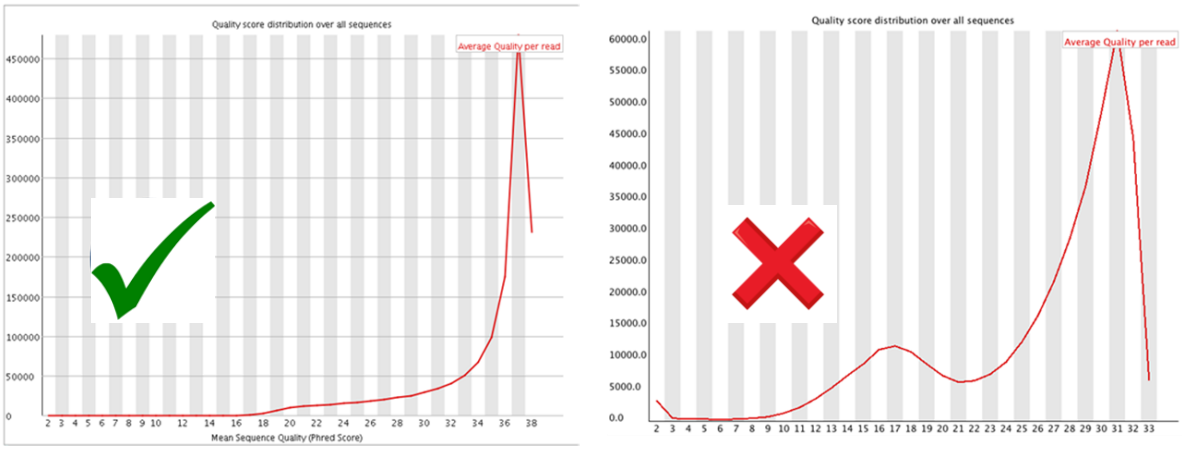
NOTE1: *More than one sequence read file may be open at a time.*
NOTE2: *It may be more helpful to analyze R2 data. Generally, R2 will have slightly decreased quality compared to R1. Therefore if R2 passes initial quality assessment, it may be assumed that R1 will pass as well.*

5.3.1.3. Per Base Sequence Quality: Assess the "Per base sequence quality" graph. The length of the read (in bp) is along the x-axis and the quality score (Q score) is along the y-axis. The yellow box plots indicate the 25th/75th inter-quartile (extremes of the boxes), and the whiskers the10th and 90th percentage points (ends of whiskers). For this metric, the majority of the length of the reads (i.e. greater than half) should have a quality score ≥ 30. Therefore, most of the yellow box plots should be within the green area of the graph.



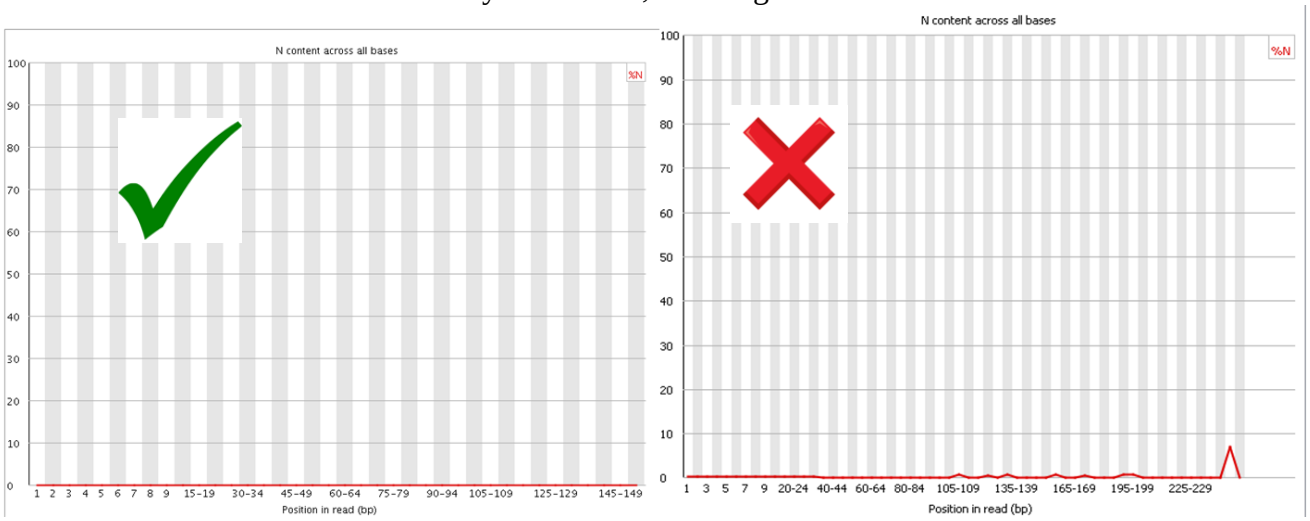
Figures 6 & 7. Example of sequences with passing per base sequence quality (left) and non-passing per base sequence quality (right).

5.3.1.4. Per Sequence Quality Scores: Assess the "Per sequence quality scores" graph. This graph provides a view of the quality score (Q score) along the x-axis by number of reads, along the y-axis. The quality for the majority of the reads should be ≥ 30. See Figures 8 & 9.



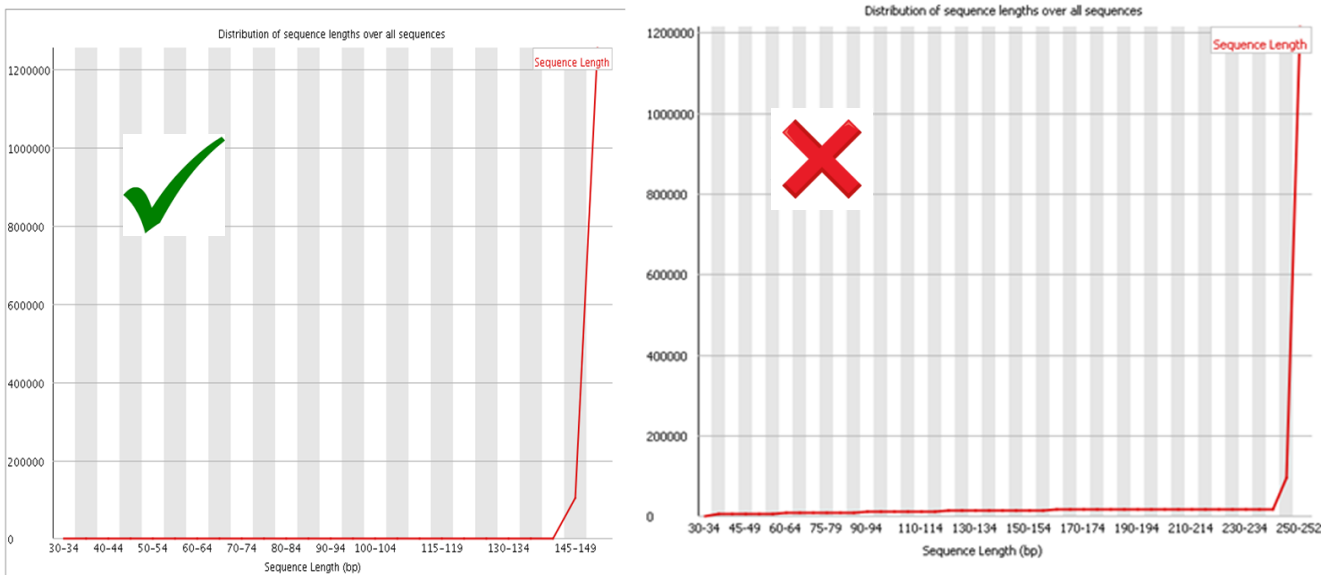
Figures 8 & 9. Examples of sequences with passing quality score distribution over all sequences (left,) and non-passing quality score distribution over all sequences (right).

5.3.1.5. Per Base N Content: Assess the "Per base N content". This graph depicts the number of Ns (bases that could not be accurately called, along the y-axis) identified for position in the read (bp, along the x-axis). This value should be 0% across the entirety of the read, as in Figures 10 & 11 below.



Figures 10 & 11. Example of "Per base N content" graph for a sequence with ideal N content across all bases (left) and an example of a graph with a higher percentage of N content (right).

5.3.1.6. Sequence Length Distribution: Assess the "Sequence Length Distribution" graph. This graph depicts varying sequence lengths (in bp, along x-axis) detected. This graph should be flat until reaching the maximum read length, indicating that the fragments are of sufficient lengths. If the line is above baseline prior to the desired read length, this is indicative of short inserts in the library. See Figures 12 & 13 for examples of each. Note that even a small increase above baseline may lead to a short average read length and the isolate may need to be re-sequenced.



Figures 12 & 13. Examples of the "Sequence Length Distribution" graphs: A sequence with ideal distribution of sequence length (left), and the graph for a sequence with less than ideal sequence length distribution (right).

5.4. **Sharing sequence data:**

5.4.1. If the sequence data passes the target coverage and appears to pass quality, the files (fastq.gz files) are ready to be shared using either Illumina BaseSpace or the PulseNet1 ftp-site (refer to SOP PND19 for instructions).

5.4.2. For troubleshooting assistance, contact PulseNetNGSLab@cdc.gov.

6. FLOW CHART: N/A

7. REFERENCES:

7.1. *Sequencing Analysis Viewer Software Guide v.2.4*. Illumina. 15066069 v03. November 2017

7.2. *www*.bioinformatics.babraham.ac.uk/projects/fastqc FastQC. Babraham Bioinformatics

7.3. https://support.illumina.com/sequencing/sequencing_software/basespace/documentation.html BaseSpace Documentation and Literature. Illumina

8. CONTACTS:

8.1. CDC PulseNet NGS lab Inbox: PulsenetNGSlab@cdc.gov

8.2. Eija Trees: EHyytia-Trees@cdc.gov

8.3. Ashley Sabol: asabol@cdc.gov

9. AMENDMENTS:

12/22/2015: Added coverage calculation instructions for using the Read Metrics tab in the Nextera XT library prep workbook, and included image in new appendix PNQ07-3.
04/29/2016: Cluster density range corrected from 600-1300 to 600-1200.
06/27/2016:

- Attached updated image for Appendix PNQ07-3

10/13/2016:

- Updated formatting
- Updated information concerning running FastQC within BaseSpace Sequence Hub using iCredits
- Added quality metrics for v3 600 cycle chemistry
- Changed clusters passing filter values so that values were uniform across reagent kits
- Included basic quality guidance and graphical examples for fastq file assessment using FastQC.

01/20/2017:

- Corrected numbering of steps within procedure.
- Reformatted document layout according to new layout (removed footer, updated header, added “Approvals Signatures”).
- Added PF to “Definitions”.
- Updated formula and wording within step 5.2.2.

04/11/2018:

- Updated Purpose, Responsibilities, Definitions, clarified process for data analysis, added tables and diagrams for assessing quality and references.
- Updated document to include quality metrics for *Vibrio spp.*
- Updated PNQ07-3 to include *Vibrio spp.* Coverage.