1. **PURPOSE:** All genomic sequences generated by PulseNet participating laboratories are uploaded in real-time to the sequence read archive (SRA) located at NCBI and the analyzed data are uploaded to the organism-specific surveillance databases located on servers at CDC. In order to ensure the integrity and comparability of the data across the network, PulseNet has set minimum quality requirements for sequences to be uploaded to the national databases and SRA. The purpose of this document is to describe a standardized procedure for Illumina sequence data quality control (QC) prior to upload to the national database and SRA.

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

3. **DEFINITIONS/ACRONYMS:**
   - **3.1. Allele:** One of two or more alternative forms of a gene that arise by mutation and are found at the same place on a chromosome.
   - **3.2. Analysis Certified:** An individual who is certified for checking the quality, performing analysis, and uploading WGS data to the PulseNet National Database and NCBI using BioNumerics.
   - **3.3. ANI:** Average Nucleotide Identity.
   - **3.4. BaseSpace:** Illumina cloud-based computing environment for next generation sequencing data analysis, management, and storage, including data sharing.
   - **3.5. BioNumerics:** Analysis software used by PulseNet, developed by Applied Maths (Sint-Martens-Latem, Belgium).
   - **3.6. Bp:** Base Pair.
   - **3.7. CDC:** Centers for Disease Control and Prevention.
   - **3.8. Core genome:** Genes shared by all strains of the same species.
   - **3.9. Coverage:** The average number of reads that include a given nucleotide in the reconstructed sequence.
   - **3.10. Critical Quality Metrics:** Average denovo coverage, average quality (Q score), assembly length, secondary species abundance (contamination detected by MIDAS) and percent core present. Failing to meet the minimum thresholds/acceptable range for any one of these metrics will result in rejection of the sequence.
   - **3.11. De Novo Assembly:** A sequence assembly generated from the short raw reads without the use of a reference genome.
   - **3.12. FASTQ**: A text-based file format for storing both sequence and its corresponding quality scores.
   - **3.13. FastQC**: Raw sequence quality control tool available free-of-charge at www.bioinformatics.babraham.ac.uk/projects/fastqc.
   - **3.14. Insert:** The stretch of sequence in fragmented target DNA that is between the paired-end adapters. Includes the sequenced portion (reads), as well as the un-sequenced portion. Insert size is determined bioinformatically by mapping the reads back to the reference sequence.
   - **3.15. MIDAS:** Metagenomic Intra-species Diversity Analysis System. An integrated computational pipeline for quantifying bacterial species abundance.
   - **3.16. 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.

3.17. Organism-specific Database: A BioNumerics database, v 7.6 or higher, used for comparing isolates for surveillance. Part of the standard PulseNet workflow.

3.18. PF Reads: Passing Filter Reads, the number of reads which passed filtering (are useable reads) for a sequencing run. % Reads Identified (PF) represents the percentage of the PF reads that have been assigned to an index pair. This is displayed for the entire run (% of reads assigned to an index pair) as well as for a particular set of indices and will vary for each index pair.

3.19. PHL: Public Health Laboratory

3.20. PN: PulseNet

3.21. PulseNet Central: PulseNet team at CDC comprising of the Database Unit (PulseNet@cdc.gov) and the Next Generation Subtyping Methods Unit (PulseNetNGSlab@cdc.gov).

3.22. QC: Quality Control

3.23. Q score: The sequence quality score for each individual base position in a sequence, indicating the accuracy of the base call. 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.

3.24. Read: A unit of continuous DNA sequence (based pairs) derived by sequencing a part of the fragmented target DNA.

3.25. RefID Database: A BioNumerics database, v 7.6 or higher, used for quality control of raw sequence data, assembly of sequences, contamination detection, and species identification. Part of the standard PulseNet workflow.

3.26. 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.

3.27. Size selection: A step in the sequencing library preparation which involves the targeted capture of DNA fragments of a specific size or a size range.

3.28. SOP: Standard Operating Procedure.

3.29. SRA: Sequence Read Archive, database at NCBI which stores raw sequence data and alignment information.

3.30. Tagmentation: The first step of the Illumina Nextera sequencing library preparation where the target DNA is enzymatically cleaved and tagged with universal adapters at both ends of the fragment.

3.31. WGS: Whole Genome Sequencing

3.32. wgMLST: Whole Genome Multi-Locus Sequence Typing

4. RESPONSIBILITIES:

4.1. **Certified PulseNet public health laboratory personnel**

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

4.1.1.1. BioNumerics analysis-certified personnel:

4.1.1.1.1. Use the BioNumerics RefID database workflow to determine average denovo coverage, average read quality, sequence assembly length, secondary species abundance (contamination), species ID, and average read length.

4.1.1.1.2. Use the BioNumerics organism-specific databases to obtain percentage (%) values of core genome present.

4.1.1.2. Personnel not BioNumerics analysis-certified:

4.1.1.2.1. Coverage can be **estimated** in semi-automated manner using SAV/BaseSpace and the metrics tab in Nextera XT (PNL34.W1) or Nextera DNA Flex (PNL35.W1&2) workbooks or manually using values derived from SAV/BaseSpace or FastQC.

4.1.1.2.2. Quality and average read length can be **roughly** evaluated using FastQC.

4.1.2. Re-sequence any isolates which do not meet the minimum thresholds/acceptable ranges for critical quality metrics.

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 public health laboratory if any sequences do not meet quality thresholds.

4.2.3. Assist public health laboratories with troubleshooting, as necessary.

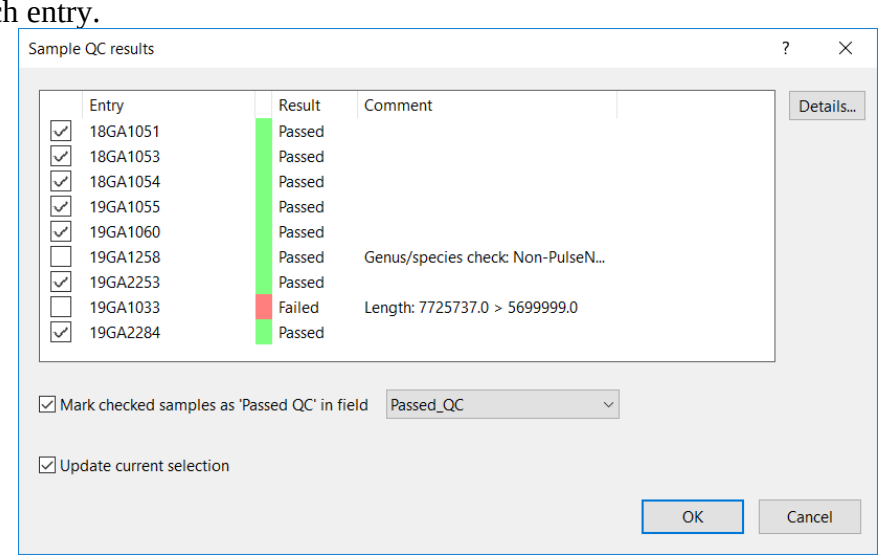
5. PROCEDURE:

5.1. Sequence quality assessment using BioNumerics 7.6 or higher RefID database workflow

5.1.1. Process the fastq-files in the RefID database following the workflow outlined in the SOP PND20 (PulseNet Standard operating procedure for the BioNumerics Reference Identification database).

5.1.2. Review sequence quality:

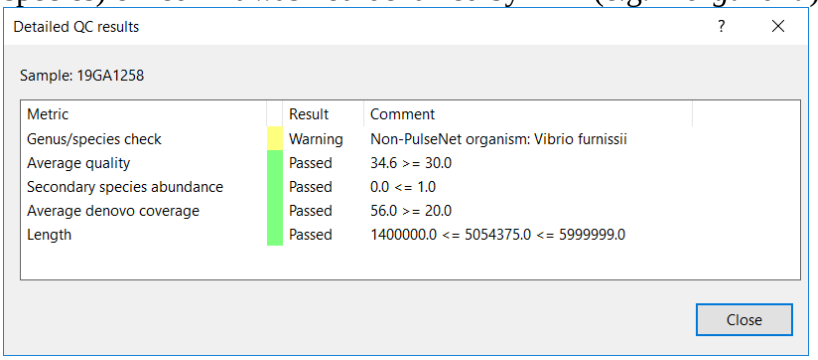
5.1.2.1. The “Sample QC results” window lists the pass/fail results for the critical quality metrics (except % core present) and comments for each selected entry. Select an entry and click on “Details” to see specific metrics that pass/fail for each entry.



5.1.2.1.1. If the sequence is flagged green and is a PulseNet organism, the sequence passes quality, remains checked, and the comment is blank.

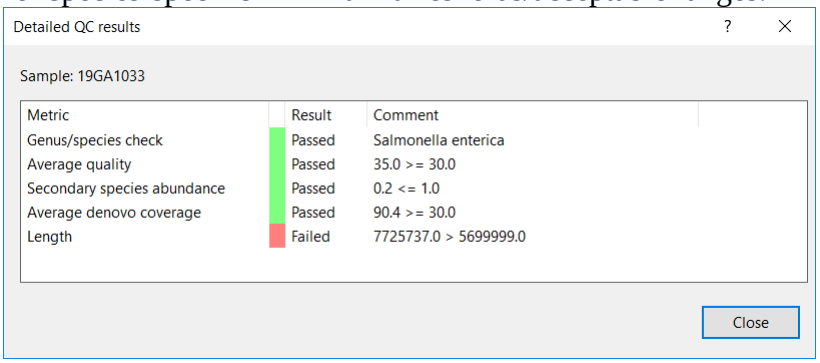
5.1.2.1.2. If the sequence is flagged green and is not a PulseNet organism, the sequence passes quality, becomes unchecked, and the comment indicates "Genus/species check: Non-PulseNet organism".

5.1.2.1.2.1. Click on "Details" to pull up a "Detailed QC results" window for all sequences that have comments listed. In the "Detailed QC results" window, the Genus/species check is flagged yellow if it was identified by ANI (e.g. non-PulseNet *Vibrio* or *Campylobacter* species) or red if it was not identified by ANI (e.g. *Morganella*).



5.1.2.1.3. If the sequence is flagged red, the sequence fails quality, becomes unchecked, and the reason for failed QC is in the "Comment" column.

5.1.2.1.3.1. Failing to meet any of the critical quality metrics will result in the rejection of the sequence. Refer to appendix PNQ07-1 Table 3 for species-specific minimum thresholds/acceptable ranges.



5.1.2.2. Review average read length:

5.1.2.2.1. For 500 cycle chemistry, an average read length ≥ 225 bp is acceptable.

5.1.2.2.2. For 300 cycle chemistry, an average read length ≥ 135 bp is acceptable.

NOTE1: Average read length is a good proxy for estimating the insert size. Shorter than expected read lengths are an indication of sub-optimal library preparation indicating that DNA may be over-tagmented or size selection is not being performed correctly.

NOTE2: Average read length is **not** a critical quality metric for routine sequence submission, i.e. shorter than expected read length does not result in an automatic rejection of a sequence. However, shorter than expected read lengths may result in fragmented assemblies and a low percentage of core alleles detected which can result in sequence rejection. Therefore,

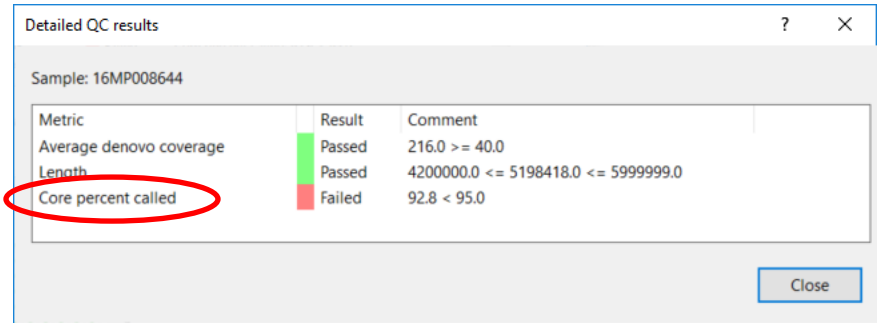
when shorter than expected read lengths are detected, the root cause should always be investigated and addressed.

NOTE3: Insert size (and hence the read length) **is** a critical quality metric for wetlab (fastq) certification only. Sequences with read lengths < 225 bp for 500 cycle chemistry and <135 bp for 300 cycle chemistry will not pass certification. These read lengths correspond to approximately 300 bp insert size which is the PulseNet minimum threshold for insert size. For proficiency tests, shorter than expected read lengths will not result in automatic rejection but will result in points deduction.

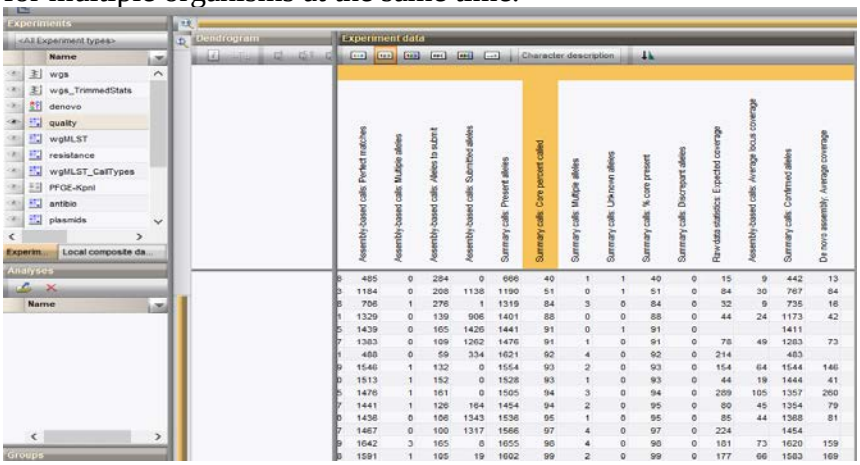
5.1.3. Transfer sequences that belong to PulseNet organisms and pass the critical quality metrics to organism-specific databases.

5.1.3.1. Follow the workflow (SOP under development) in the organism-specific database to obtain the quality metrics for the allele calls, including % of core present (a critical quality metric).

5.1.3.1.1. You can review the percent core present values by selecting the desired entries in the database and opening the “Detailed QC results” window for one isolate at a time.



5.1.3.1.2. Alternatively, you can create a comparison and review all QC metrics for multiple organisms at the same time.



5.1.4. Re-sequence PulseNet organisms and non-PulseNet organisms identified by ANI that did not pass critical quality metrics. For non-PulseNet organisms not identified by ANI, use another validated method (for example phenotypic identification or 16S sequencing) for identification.

5.2. **Estimating coverage without BioNumerics**: Coverage may be estimated manually by using metrics from FastQC (Section 5.2.1.), semi-automatically by using the PN

Workbook and values derived from BaseSpace/SAV (Section 5.2.2.) or manually using the values derived from BaseSpace/SAV (Section 5.2.3.). **NOTE:** This is a mathematical estimate of coverage – actual coverage values may differ from the calculations below due to shorter than expected read lengths, variation of genome size from the estimated value used for coverage determination, etc. True coverage values are obtained using BioNumerics.

5.2.1. **Manually, using FastQC:** Coverage may be calculated using the formula below and values derived from FastQC: **NOTE:** FastQC may be downloaded, free of charge, at: www.bioinformatics.babraham.ac.uk/projects/fastqc.

5.2.1.1. Open FastQC and choose "File" on the main screen.

5.2.1.2. Choose "Open" and then select one read file (.fastq file), either R1 or R2, for the sample to be analysed.

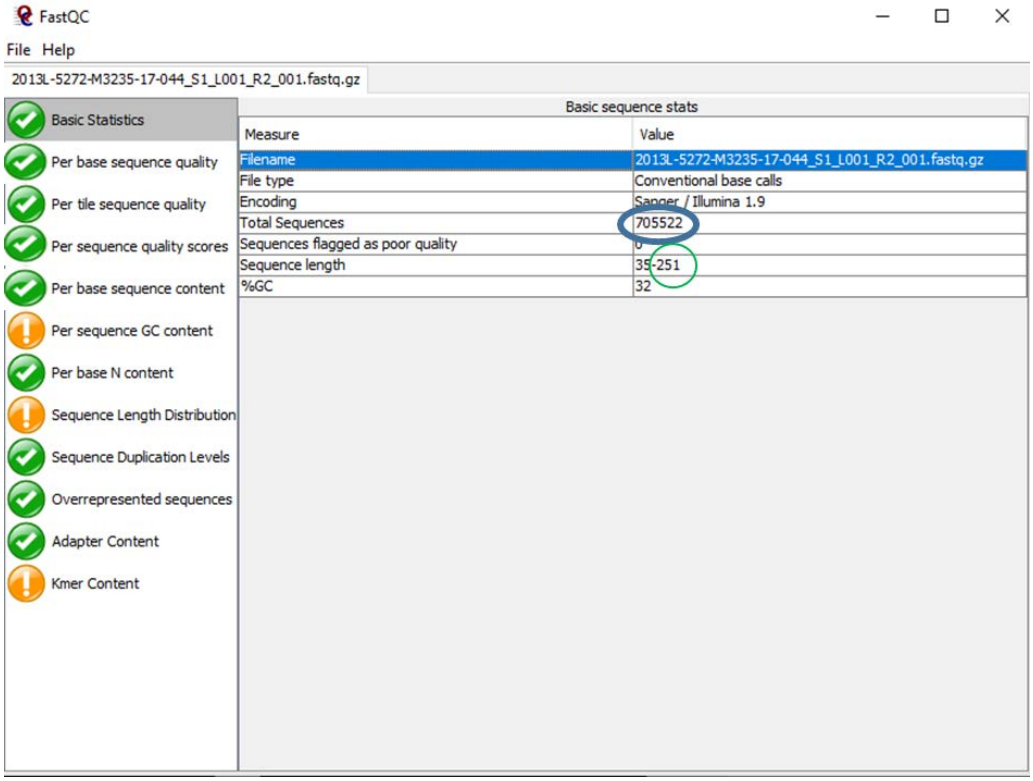
5.2.1.3. Using the **Total Sequences** value and the maximum number depicted for **Sequence Length**, calculate coverage using the formula:
(Total Sequences x Maximum Sequence Length x 2)/Estimated genome length = Isolate coverage.
NOTE: multiplying by 2 is required for determining coverage for paired-end reads.

5.2.1.3.1. Refer to Table 1 for estimated genome sizes.

Organism	Estimated Genome size (bp)
Listeria monocytogenes	3000000
E. coli/Shigella spp.	5000000
Salmonella spp.	5000000
Campylobacter spp.	1600000
Vibrio spp.	5000000

Table 1.Estimated genome sizes to be used in the coverage calculations

5.2.1.4. Refer to Appendix PNQ07-1 Table 3 for minimum coverage thresholds



Example: Coverage calculation for this *Listeria* isolate:

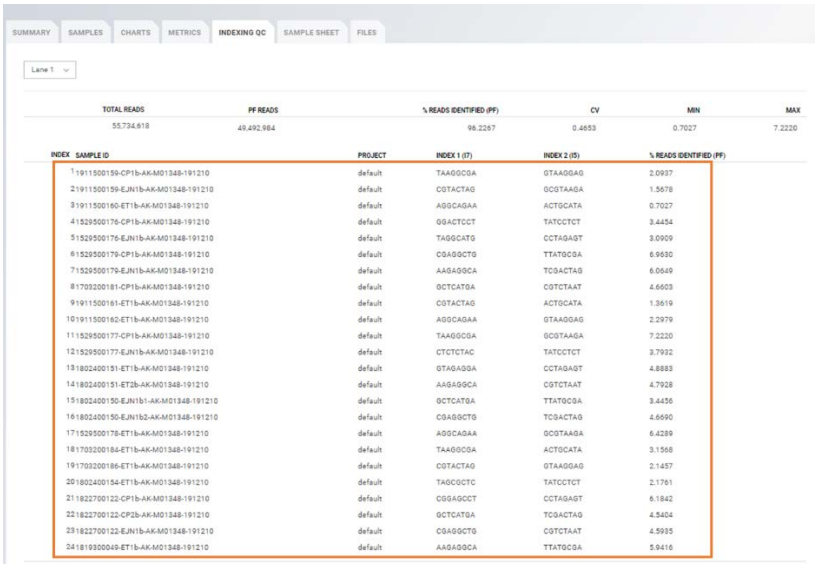
(705522 × 251 × 2)/3000000 = 118x

5.2.2. **Semi-automatically, using the PN Workbook:** Coverage may be calculated using the "Metrics tab" of the library prep workbooks and values derived from SAV/BaseSpace. **NOTE:** The following fields must be filled in correctly in the workbook in order for the coverage to be accurately calculated:

Workbook	Tab	Field
Nextera XT (PNL 34.W1)	Initial Dilution	Sample ID Genome Size Estimate
	Metrics	Sequencing Kit Type/Chemistry
DNA Flex (PNL 35.W1 & W2)	Library Prep	State Key Genome Size Estimate Sequencing Kit Type/Chemistry

Table 2. Fields that need to be correctly filled out in PulseNet sequence library workbooks for accurate coverage estimations

5.2.2.1. Select and copy the columns for Index Number, Sample Id, Project, Index 1 (i7), Index 2 (i5) and % Reads Identified (PF) columns **NOTE:** In SAV these values are found in the Indexing tab. On BaseSpace they are found in the Indexing QC tab.



5.2.2.2. Paste these values into the corresponding columns on the "Metrics" tab of the workbook. **NOTE**: **Ensure** that the order of the isolates in the "Metrics" tab is the same as the order of the isolates on the previous workbook tabs.

5.2.2.3. Enter the "PF Reads" value for the run into column G of the workbook. The total number of reads (bp) for each isolate should now be displayed in Column H of the workbook and the estimated coverage either in Column I or J depending on which MiSeq software version is in use. **NOTE**: Depending on the software version in use, SAV/BaseSpace will either display the **total** number of reads that passed filter for the run, or only **half** of the total reads that passed filter. It is advised to first calculate coverage estimate for one isolate on a run using FastQC to determine which value of PF Reads is displayed. This will allow users to know which column of the "Metrics" tab may be used to automatically generate coverage estimates for the run.

5.2.2.4. See Appendix PNQ07-1 Table 3 for passing coverage requirements.

SAV: Indexing Tab

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

Lane 1

Reads mapped to Index Id

Total Reads	PF Reads	% Reads Identified (PF)	CV	Min	Max
14317463	13822209	96.8461	0.4721	1.3286	9.8369

Index Number	Sample Id	Project	Index 1 (I7)	Index 2 (I5)	% Reads Identified (PF)
1	2013L-5272-M3235-17-044	NA	AAGAGGCA	CCTAGAGT	5.1043
2	2013L-5280-M3235-17-044	NA	CGTACTAG	CGTCTAAT	3.3822
3	2015K-1299-M3235-17-044	NA	AGGCAGAA	TTATGCGA	8.3097
4	2014K-0953-M3235-17-044	NA	TAAGGCGA	TCGACTAG	4.7364
5	2015K-1051-M3235-17-044	NA	TAGCGCTC	GCGTAAGA	5.4466
6	2015K-0520-M3235-17-044	NA	GGACTCCT	ACTGCATA	9.8369
7	2015K-0709-M3235-17-044	NA	CTCTCTAC	GTAAGGAG	7.2974
8	2015K-0452-M3235-17-044	NA	CGGAGCCT	TATCCTCT	1.3286
9	2015C-4647-M3235-17-044	NA	GCTCATGA	CCTAGAGT	5.1345
10	H2446-B1-M3235-17-044	NA	AGGCAGAA	TCGACTAG	5.7491
11	H2446-B2-M3235-17-044	NA	TAAGGCGA	CGTCTAAT	9.4352
12	2015C-3864-M3235-17-044	NA	CGTACTAG	TTATGCGA	2.3546
13	2015C-3865-M3235-17-044	NA	TAGGCATG	GTAAGGAG	5.1411
14	2015C-3881-M3235-17-044	NA	GTAGAGGA	GCGTAAGA	2.6132
15	2011D-8942-M3235-17-044	NA	CGGAGCCT	ACTGCATA	2.6797
16	2016D-0061-M3235-17-044	NA	CGAGGCTG	TATCCTCT	3.9623
17	2016D-0062-M3235-17-044	NA	TAAGGCGA	CCTAGAGT	4.3571
18	2015C-3265-M3235-17-044	NA	GGACTCCT	TTATGCGA	8.3943
19	2015C-3266-M3235-17-044	NA	CTCTCTAC	TCGACTAG	3.583

BaseSpace: Indexing QC Tab

SUMMARY | BIOSAMPLES | CHARTS | METRICS | INDEXING QC | SAMPLE SHEET | FILES

Lane 1 | Lane 1 Pool ID: Pool_57453400

TOTAL READS	PF READS	% READS IDENTIFIED (PF)	% READS UNDETERMINED	CV	MIN	MAX
28,773,492	27,305,684	97.4537	2.5463	0.2511	3.5591	8.4556

INDEX	BIOSAMPLE	LIBRARY NAME	INDEX 1 (I7)	INDEX 2 (I5)	% READS IDENTIFIED (PF)
1	TX-ACF1702702C	TX-ACF1702702C-TX-M04922-180110_	TAAGGCGA	GCGTAAGA	5.8152
2	TX-ACF1702772	TX-ACF1702772-TX-M04922-180110_	CGTACTAG	GCGTAAGA	5.9080
3	PNUSAE011807	PNUSAE011807-TX-M04922-180110_	AGGCAGAA	GCGTAAGA	5.6093
4	PNUSAE011808	PNUSAE011808-TX-M04922-180110_	TCCTGAGC	GCGTAAGA	5.4354
5	PNUSAE011806	PNUSAE011806-TX-M04922-180110_	GGACTCCT	GCGTAAGA	4.5837
6	PNUSAE011809	PNUSAE011809-TX-M04922-180110_	TAGGCATG	GCGTAAGA	4.8643
7	PNUSAS031950	PNUSAS031950-TX-M04922-180110_	CTCTCTAC	GCGTAAGA	3.5591
8	PNUSAS031951	PNUSAS031951-TX-M04922-180110_	CAGAGAGG	GCGTAAGA	4.0234
9	PNUSAS031952	PNUSAS031952-TX-M04922-180110_	TAAGGCGA	CTCTCTAT	8.3790
10	PNUSAS031953	PNUSAS031953-TX-M04922-180110_	CGTACTAG	CTCTCTAT	7.9323
11	PNUSAS031954	PNUSAS031954-TX-M04922-180110_	AGGCAGAA	CTCTCTAT	7.4870
12	PNUSAS031955	PNUSAS031955-TX-M04922-180110_	TCCTGAGC	CTCTCTAT	5.0355
13	PNUSAS031956	PNUSAS031956-TX-M04922-180110_	GGACTCCT	CTCTCTAT	7.3920
14	PNUSAS031957	PNUSAS031957-TX-M04922-180110_	TAGGCATG	CTCTCTAT	7.0213
15	PNUSAS031958	PNUSAS031958-TX-M04922-180110_	CTCTCTAC	CTCTCTAT	8.4556
16	PNUSAS031959	PNUSAS031959-TX-M04922-180110_	CAGAGAGG	CTCTCTAT	5.9525

	v2, Nano, Micro	v3	v3, 600c	v2, Nano	300c v2, Micro	Filter	(Gb)
	600-1200	1200 - 1400	70	75	80	> 80%	
Run Metrics:	0			0.00%		0.00%	0

Escherichia/Shigella	Salmonella	Listeria	Campylobacter	Vibrio
40X	30X	20X	20X	40X

State Key	Project	Index 1 (I7)	Index 2 (I5)	% Reads Identified (PF)	PF Reads	Total # of Reads	Coverage (if MCS v 2.6 or earlier)	Coverage (if LRM/MCS 3.0 or later)
2013L-5272-M3235-17-044		AAGAGGCA	CCTAGAGT	5.1043	13822209	705527	58.79	117.59
						0	#DIV/0!	#DIV/0!

Library Prep | SampleSheet | LRM_SampleSheet | Indexes | Metrics

5.2.3. **Manually, using BaseSpace or SAV**: Coverage may be calculated using the formula below and values derived from SAV or BaseSpace:

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

5.2.3.1. Get the values for PF Reads & % Reads Identified (PF) from the SAV Indexing tab or the BaseSpace Indexing QC tab.

NOTE1: Use the percentage, NOT the percentage value – see Example below.

NOTE2: Whether or not the PF Reads value for the run, displayed in SAV or BaseSpace, is the **total** for the entire run, or only **half** of the total number of reads must be established in order to use the values for manual mathematical estimation of coverage. This may be determined by estimating coverage in FastQC first.

Maximum read length = ½ the number of cycles in the run. For a 500 cycle run, the Maximum read length = 250.
Estimated genome size = See Table 1

5.2.3.2. See Appendix PNQ07-1Table 3 for passing coverage requirements.

Example:
Using the values from the SAV example above to determine coverage for 2013L-5272-M3235-17-044, and assuming that this was a 500 cycle run:

PF Reads (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

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

****HOWEVER, in this example the PF Reads displayed in SAV is only HALF of all of the reads on the run. This is confirmed by noting that (1382209 x 0.051043 = 70552, which corresponds with the Total Sequences value in FastQC for only ONE fastq file. Therefore, the estimated coverage value above (58.79) should be doubled, to find total coverage for the isolate in this instance (117.58). Thus, if the workbook were to be used for semi-automated calculations for this run, the second column (pink header) will yield the most accurate coverage estimate.

5.3. **Assess sequence data quality and read length using FastQC**

NOTE: The graphs generated by FastQC are open to subjective interpretation and will not provide actual numeric quality scores and read lengths (which are derived using BioNumerics or command line tools). The graphs depicted below have been correlated with numeric values derived from the CDC in-house command line QC tool (https://github.com/lskatz/CG-Pipeline).

5.3.1. Open FastQC and select “File” from the toolbar on the main screen.
5.3.2. Choose “Open” and select a read file for analysis (.fastq file).

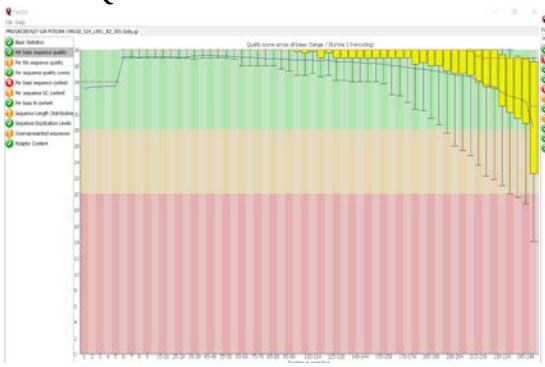
NOTE1: More than one sequence read file may be opened at a time.
NOTE2: It may be more helpful to analyze R2 data. Generally, R2 will have slightly decreased quality compared to R1. Therefore, if R2 passes initial quality assessment, it may be assumed that R1 will pass as well.

5.3.3. Assess the “Per base sequence quality” graph:
5.3.3.1. The length of the read (in bp) is along the x-axis and the quality score (Q score) is along the y-axis.
5.3.3.2. The yellow box plots indicate the 25th/75th inter-quartile (extremes of the boxes).

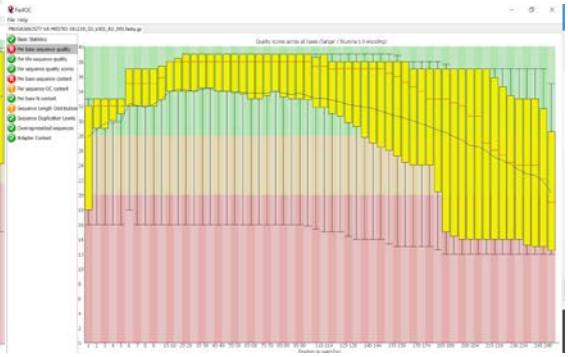
5.3.3.3. The whiskers indicate the 10th and 90th percentage points (ends of whiskers).

5.3.3.4. For this graph, 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.

R2 Q score 36.18



R2 Q score 30.69



R2 Q score 29.56

R2 Q score 27.08

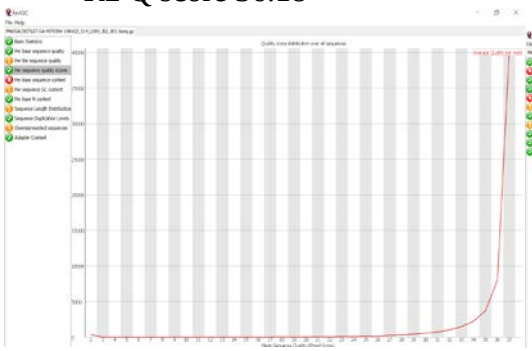
5.3.4. Assess the "Per sequence quality scores" graph:

5.3.4.1. This graph provides a view of the quality score (Q score) along the x-axis by number of reads, along the y-axis.

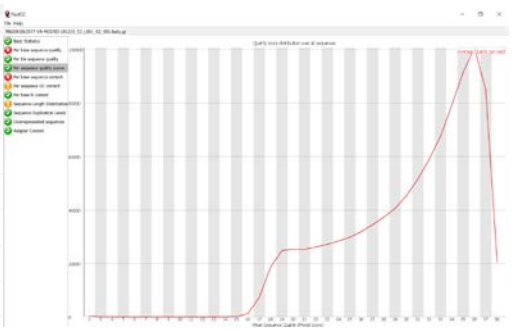
5.3.4.2. The peak should be sharp and > 30.

5.3.4.3. The peak of the graph being < 30 or having a pronounced shoulder indicates the presence of low quality reads.

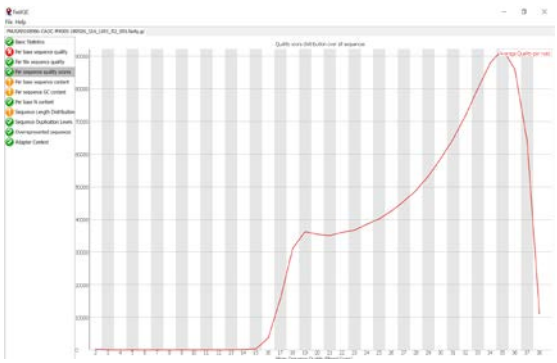
R2 Q score 36.18



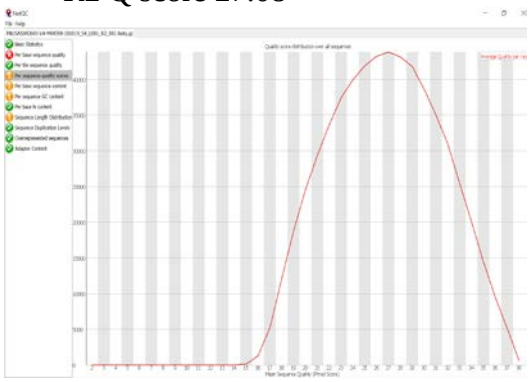
R2 Q score 30.69



R2 Q score 29.56



R2 Q score 27.08



5.3.5. Assess the "Sequence Length Distribution" graph: This graph depicts the number of sequences (y-axis) with varying lengths (in bp, along x-axis).

5.3.5.1. The graph should be flat until reaching the maximum read length, indicating that the reads are of sufficient lengths.

5.3.5.1.1. Acceptable average read length for 500 cycle chemistry is ≥ 225 bp.

5.3.5.1.2. Acceptable average read length for 300 cycle chemistry is ≥ 135 bp.

5.3.5.2. If the line is above baseline prior to the desired read length, this is indicative of short inserts in the library.

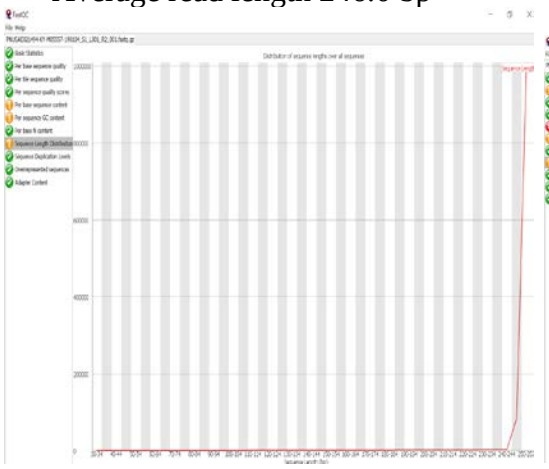
NOTE1: Shorter than expected read lengths are an indication of sub-optimal library preparation; indicating that DNA may be over-tagmented or size selection is not being performed correctly.

NOTE2: Average read length is **not** a critical quality metric for routine sequence submission, i.e. shorter than expected read length does not result in an automatic rejection of a sequence. However, shorter than expected read lengths may result in fragmented assemblies and a low percentage of core alleles detected which can result in sequence rejection. Therefore, when shorter than expected read lengths are detected, the root cause should always be investigated and addressed.

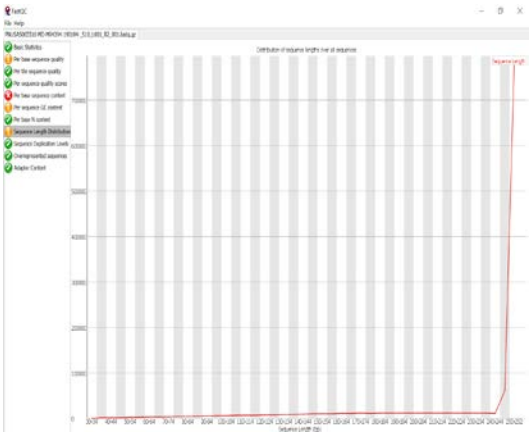
NOTE3: Inset size (and hence the read length) **is** a critical quality metric for wetlab (fastq) certification, only. Sequences with read lengths < 225 bp for 500 cycle chemistry and <135 bp for 300 cycle chemistry will not pass

certification. These read lengths correspond to approximately 300 bp insert size which is the PulseNet minimum threshold for insert size. For proficiency tests, shorter than expected read lengths will not result in automatic rejection but will result in points deduction.

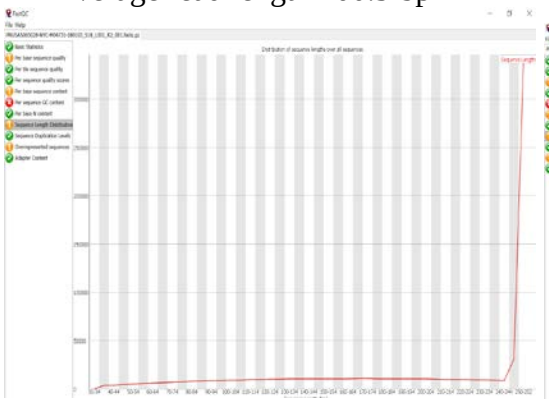
Average read length 246.0 bp



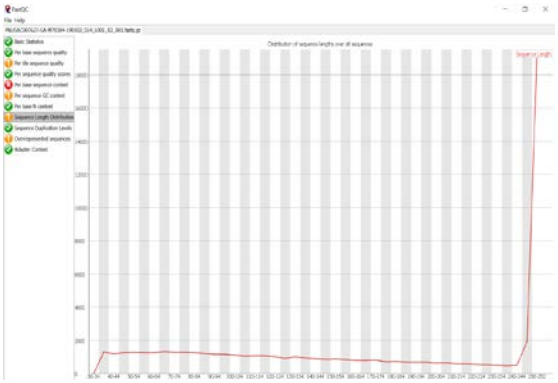
Average read length 225.2 bp



Average read length 200.9 bp

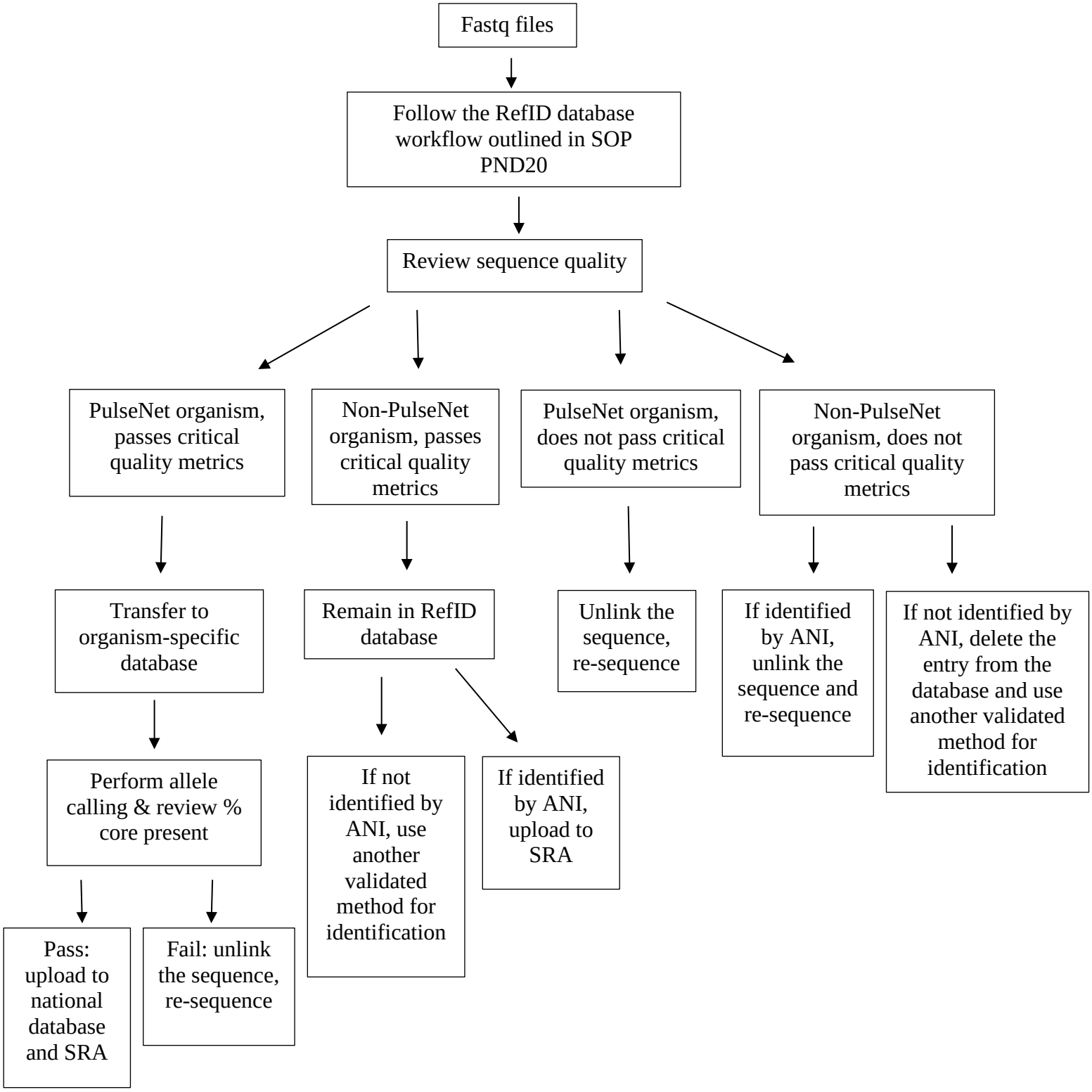


Average read length 170.0 bp

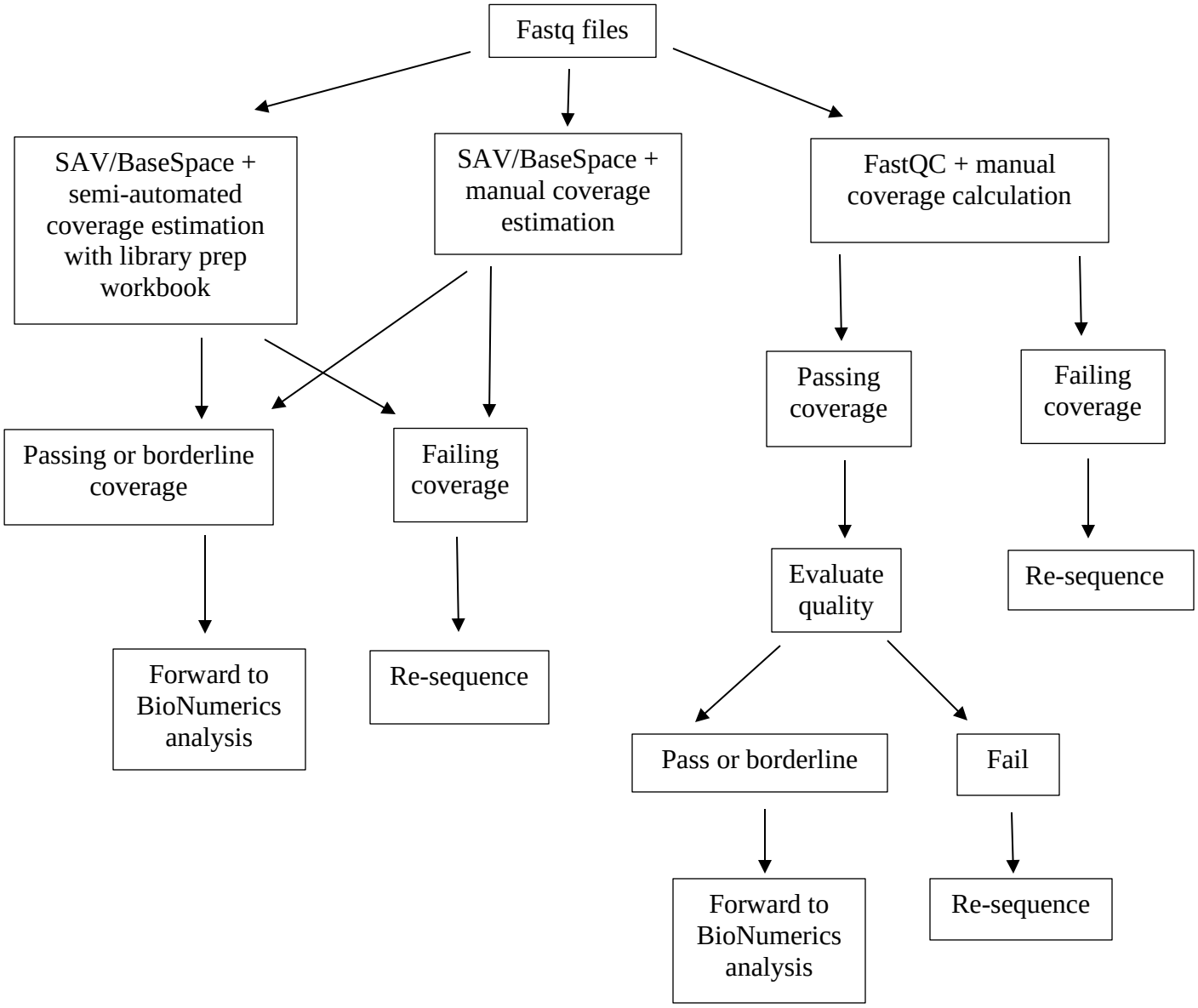


6. FLOW CHART:

Sequence data QC using BioNumerics 7.6 or higher RefID database workflow (analysis certified individuals)



Sequence data QC without BioNumerics (individuals not certified for analysis)



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 Laboratory Inbox: PulsenetNGSlab@cdc.gov
8.2. CDC PulseNet Database Inbox: PulseNet@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

03/16/2020

- Removed the review of run quality metrics section from this SOP and moved it to PNL38
- Added BioNumerics 7.6 RefID database workflow as the primary option for raw data QC
- Added additional examples on how to interpret FastQC graphs
- Added workflow diagrams for raw data QC using BioNumerics and without BioNumerics
- Average quality score (Q score) minimum threshold increased from 28.0 to 30.0. Sequences with Q scores <30.0 will be rejected regardless of coverage.

10. APPROVAL SIGNATURES:

Approved By: Jennifer Adams Digitally signed by Jennifer Adams Date: 2020.03.23 14:21:58 -04'00' Date: _____
PulseNet QA/QC Personnel

Approved By: Kelley B Hise Digitally signed by Kelley B Hise Date: 2020.03.20 18:24:28 -04'00' Date: _____
PulseNet Outbreak Detection and Surveillance Unit Chief

Approved By: N/A Date: N/A
PulseNet PFGE Reference Unit Chief

Approved By: Rebecca L. Lindsey -S Digitally signed by Rebecca L. Lindsey -S Date: 2020.03.20 16:19:22 -04'00' Date: _____
PulseNet Next Generation Subtyping Methods Unit Chief

Approved By: Efrain M. Ribot -S Digitally signed by Efrain M. Ribot -S Date: 2020.03.18 10:49:00 -04'00' Date: _____
PulseNet Reference Outbreak Surveillance Team Lead

Appendix PNQ07-1. Requirements for Critical Quality Metrics for PulseNet Organisms for Routine Sequence Submissions.

Organism	Average denovo coverage	Average quality (Q score)	Assembly length (MB)	Secondary species abundance	% core present[1]
Listeria monocytogenes	≥ 20x	≥ 30	2.8-3.2	≤ 1.0	≥ 95
***E. coli* (most serotypes)**	≥ 40x	≥ 30	4.9-6.0	≤ 1.0	≥ 85
Shigella spp./*Rare *E. coli	≥ 40x	≥ 30	4.2-4.9	≤ 1.0	≥ 85
Salmonella spp.	≥ 30x	≥ 30	4.4-5.7	≤ 1.0	≥ 85
Campylobacter spp.	≥ 20x	≥ 30	1.4-2.2	≤ 1.0	≥ 85[2]
Vibrio cholerae	≥ 40x	≥ 30	3.8-4.3	≤ 1.0	NA
Vibrio parahaemolyticus	≥ 40x	≥ 30	4.9-5.5	≤ 1.0	NA
Vibrio vulnificus	≥ 40x	≥ 30	4.7-5.3	≤ 1.0	NA

Table 3. Requirements for critical quality metrics for PulseNet organisms

- All critical quality metrics, except % core present can be found in the RefID database (SOP PND20). % core present is determined in the organism-specific databases.
- Rows shaded with the same color represent organisms transferred to the same organism-specific database.

[1] The goal for percent core present is 95% or above. Routine *Escherichia, Shigella, Salmonella* and *C jejuni* sequences that fall between 85%-94% and pass other quality thresholds can be uploaded to the national database and SRA. The percent core present value has not been evaluated for *Vibrio* spp. because the allele scheme is still under development.

[2] Core scheme applies to *C. jejuni* only. Other *Campylobacter* species may still pass quality with % core present <85%; please contact PulseNet@cdc.gov before resequencing.

.