

1. **PURPOSE:** to describe a standardized procedure for Illumina MiSeq data quality control (QC) before upload to PulseNet Central

2. **SCOPE:** This procedure applies to all clinical isolates that are whole genome sequenced under PulseNet surveillance.

3. **DEFINITIONS:**
 - 3.1. **CDC:** Centers for Disease Control and Prevention
 - 3.2. **NCBI:** National Center for Biotechnology Information
 - 3.3. **SRA:** Sequencing Read Archive
 - 3.4. **SOP:** Standard Operating Procedure
 - 3.5. **QC:** Quality Control
 - 3.6. **FTP:** File Transfer Protocol
 - 3.7. **Biosample:** a description of biologically or physically unique specimens (=sequenced bacterial strains)
 - 3.8. **FASTQ:** a text-based format for storing both a biological sequence and its corresponding quality scores

4. **RESPONSIBILITIES:**
 - 4.1. 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 requirements for sequences to be uploaded to SRA.
 - 4.2. Laboratory personnel performing the sequencing are responsible for performing the QC.

5. **PROCEDURE:**
 - 5.1. Upon run completion, confirm that the sequencing run meets the basic quality metrics (steps 5.1.1.-5.1.2.). If the run did not complete within the specifications below it may be necessary to repeat the run. Contact pfge@cdc.gov to determine if the run needs to be repeated.
 - 5.1.1. For a 500 cycle kit: Q30>75, cluster density between 600-1300, Pass Filter > 75%.
 - 5.1.2. For a 300 cycle kit: Q30>85, cluster density between 600-1300, Pass Filter > 80%.
 - 5.2. After sequencing is complete, calculate basic coverage numbers to determine if data passes the target coverage for each sample (see Appendix PNQ07-1 to do the analysis using a standalone version of the software FastQC; see Appendix PNQ07-2 to do this analysis using data from BaseSpace or FastQC in Illumina BaseSpace)
 - 5.2.1. Determine the total number of base pairs generated per isolate; this can be calculated by using the following equation: total number of reads * longest read length * sequencing chemistry (1 for single-end; 2 for paired-end)
 - 5.2.1.1. **NOTE:** For Illumina 2x150 chemistry, the longest read length would be 150bp.
 - 5.2.2. Calculate the coverage

VERSION:	REPLACED BY:	AUTHORIZED BY:	Page 1 of 6
-----------------	---------------------	-----------------------	--------------------

- 5.2.2.1. *Listeria monocytogenes* - total number of base pairs (calculated in 5.2.1)/3000000; target coverage is 20x or above
 - 5.2.2.2. *E. coli* and *Shigella spp.* – total number of base pairs (calculated in 5.2.1)/5000000; target coverage is 40x or above
 - 5.2.2.3. *Salmonella spp.* – total number of base pairs (calculated in 5.2.1)/5000000; target coverage is 30x or above
 - 5.2.2.4. *Campylobacter spp.* – total number of base pairs (calculated in 5.2.1)/1600000; target coverage is 20x or above
- 5.3. If the sequence data passes the target coverage threshold, submit the biosample to NCBI (refer to SOP PND18 for instructions) and transfer the raw data (fastq.gz files) using either Illumina BaseSpace or the PulseNet1 ftp-site (refer to SOP PND19 for instructions). If the sequence data does not pass target coverage threshold, repeat sequencing.

6. FLOW CHART:

7. BIBLIOGRAPHY:

8. CONTACTS:

- 8.1. CDC PFGE Inbox: PFGE@cdc.gov
- 8.2. Eija Trees: EHyytia-Trees@cdc.gov
- 8.3. Darlene Wagner: DWagner3@cdc.gov
- 8.4. Heather Carleton: hcarleton@cdc.gov

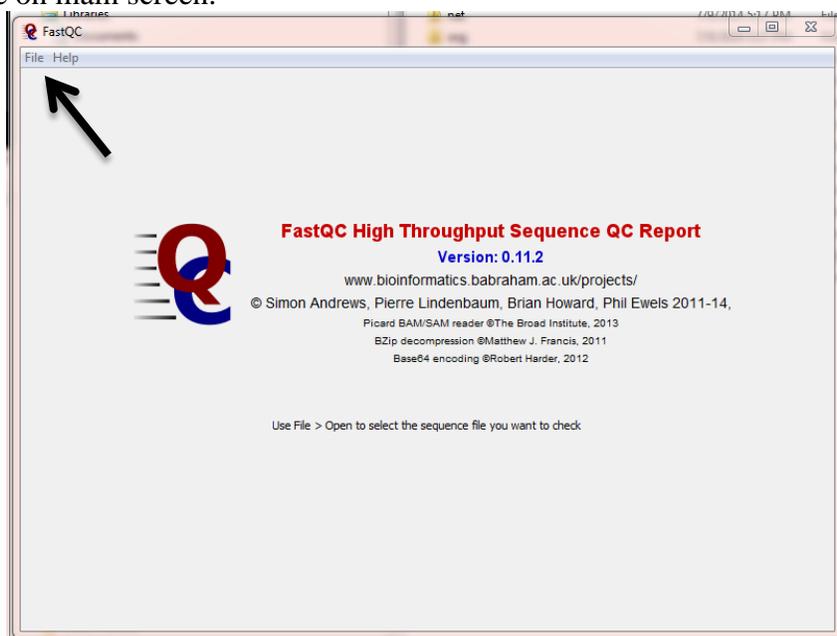
9. AMENDMENTS:

VERSION:	REPLACED BY:	AUTHORIZED BY:	Page 2 of 6
-----------------	---------------------	-----------------------	--------------------

Appendix PNQ07-1

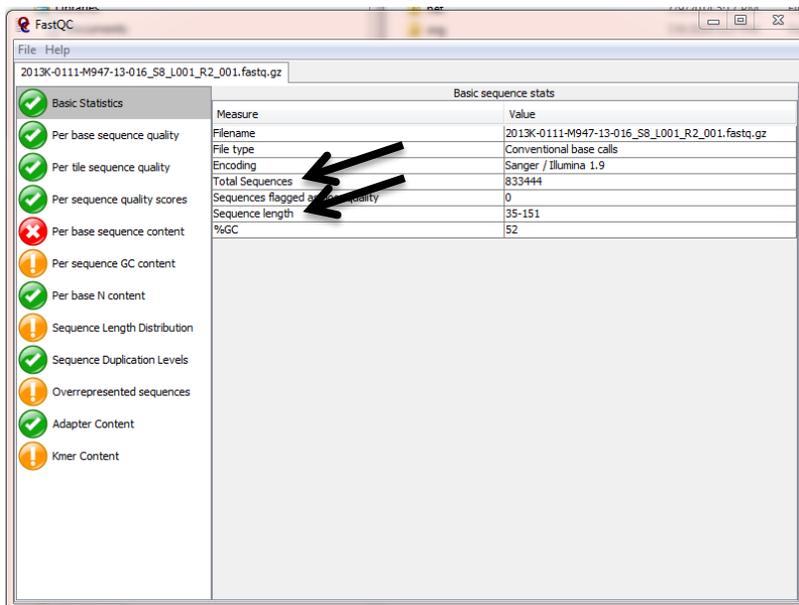
Installing FastQC software and analysing data

1. Go to www.bioinformatics.babraham.ac.uk/projects; click on the FastQC link on the page. Click on “Download Now” on the next page.
2. Unzip the FastQC file.
3. In the FastQC folder select run_fastqc (double click).
 - 3.1. **NOTE:** If the FastQC window does not open you need to install or update Java (<http://www.java.com/en/>).
4. Choose File on main screen.



5. Choose Open and then select one read of your paired-end sequence.
6. Use total sequences (number of reads) multiplied by largest read length in range then divide that number by the genome size of your sequence to determine coverage (see the screenshot below).
 - 6.1. *Listeria monocytogenes* - ((total number of reads)*(average read length) *2(if doing paired-end sequencing)/3000000; target coverage is 20x or above
 - 6.2. *E. coli* and *Shigella spp.* – ((total number of reads)*(average read length) *2(if doing paired-end sequencing))/5000000; target coverage is 40x or above.
 - 6.3. *Salmonella spp.* – ((total number of reads)*(average read length) *2 (if doing paired-end sequencing))/5000000; target coverage is 30x or above
 - 6.4. *Campylobacter spp.* – ((total number of reads)*(average read length) *2 (if doing paired-end sequencing))/1600000; target coverage is 20x or above

VERSION:	REPLACED BY:	AUTHORIZED BY:	Page 3 of 6
----------	--------------	----------------	-------------



Appendix PNQ07-2

Using BaseSpace FastQC in BaseSpace and analysing data

1. Login to BaseSpace; either stream your sequence data to BaseSpace as the run is being performed or upload the data afterwards (this only works with Illumina data by selecting Projects > Import > Sample; then drag and drop sequences into the displayed window)
2. To determine coverage, click on the sample in the sample list. The metrics for the isolate are listed including the Number of Reads. Use that number to calculate the total number of base pairs for the coverage calculation.

Projects : 218110 : Samples List : PNUSAE000217

PNUSAE000217	
Sample name	15MN00010
Date created	Jan 25, 2015
Genome	
Paired end	Paired
Number of Reads	961,560
Read 1 length	251
Read 2 length	251
Size	353.70 MB

3. To run FastQC in BaseSpace, go to the Apps menu. Select Quality (1.) from the Categories panel and then select the Launch button for the free FastQC (2.) application.

The screenshot shows the BaseSpace interface. At the top, there is a navigation bar with icons for Dashboard, Prep, Runs, Projects, Apps, Public Data, and Help. The 'Apps' menu is selected, displaying 'BaseSpace Applications'. On the left, there are three application cards: 'DeepChek®-HIV ABL S.A.', 'FastQC BASESPACE LABS', and 'Variant Calling Assessment Tool BASESPACE LABS'. The 'FastQC' card is highlighted with a '2.' next to it. On the right, there is a 'Categories' panel with a grid of application categories. The 'Quality' category is highlighted with a '1.' next to it. The 'FastQC' application card has a 'Launch' button. The 'Quality' category in the 'Categories' panel also has a 'Launch' button.

VERSION:	REPLACED BY:	AUTHORIZED BY:	Page 5 of 6
-----------------	---------------------	-----------------------	--------------------

- Select your input sample, leaving the default kmer settings, and click the acknowledgement button. Then click “Continue” on the right hand side panel.

Output

App Session Name:

Output Project:

Input

Input Sample:

Options

Kmer Size:

Use Contaminant Filter:

BaseSpace Labs Disclaimer

BaseSpace Labs: I acknowledge and agree that (i) this is a BaseSpace Labs App, (ii) I am using it AS-IS without any warranty of any kind, (iii) Illumina has no obligation to provide any technical support for this App, and (iv) Illumina has no liability for my use of this App, including without limitation, any loss of data, incorrect results, or any costs, liabilities, or damages that result from use of this App.

This app is free.

- Once analysis is complete (this can take a few minutes per isolate) a report is generated. Only the Basic Statistics are needed for calculating coverage (refer to Appendix PNQ07-1 step 6 for details on coverage calculation).

Basic Statistics

Measure	Value
Filename	PNUSAL000905_S2_L001_R1_001.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	1844166
Filtered Sequences	0
Sequence length	35-250
%GC	38