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:** Sequence 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 manually or using the Read Metrics tab in the Nextera XT library prep workbook.

5.2.3. **Manually** calculate coverages using the formulas below:

5.2.3.1. *Listeria monocytogenes* - total number of base pairs (calculated in 5.2.1)/3000000; target coverage is 20x or above

5.2.3.2. *E. coli* and *Shigella spp*. – total number of base pairs (calculated in 5.2.1.)/5000000; target coverage is 40x or above

5.2.3.3. *Salmonella spp.* – total number of base pairs (calculated in 5.2.1)/5000000; target coverage is 30x or above

5.2.3.4. *Campylobacter spp.* – total number of base pairs (calculated in 5.2.1)/1600000; target coverage is 20x or above

5.2.4. **Automatically** calculate coverages using the Read Metrics tab in the Nextera XT library prep workbook (see Appendix PNQ07-3).

5.2.4.1. Copy and paste the contents of the Indexing QC table from the Sequencing Analysis Viewer (SAV) or BaseSpace into the corresponding columns in the table found in the workbook.

5.2.4.2. Enter the PF Reads value in the appropriate cell for each isolate in the run.

5.2.4.3. Select the correct reagent kit from the MiSeq Reagent Kit dropdown menu in the Read Metrics tab. This is important since the number of cycles in this cell is used for coverage calculation.

5.2.4.4. The coverage should be automatically generated in the Coverage column of the Read Metrics tab.

5.2.4.4.1. **NOTE:** The following fields must be filled in order for the coverage to be accurately calculated:

5.2.4.4.1.1. Initial Dilution tab: Sample ID and Genome Size Estimate

5.2.4.4.1.2. Read Metrics tab: Sample ID, % Reads Identified (PF), and PF Reads

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. Heather Carleton: hcarleton@cdc.gov

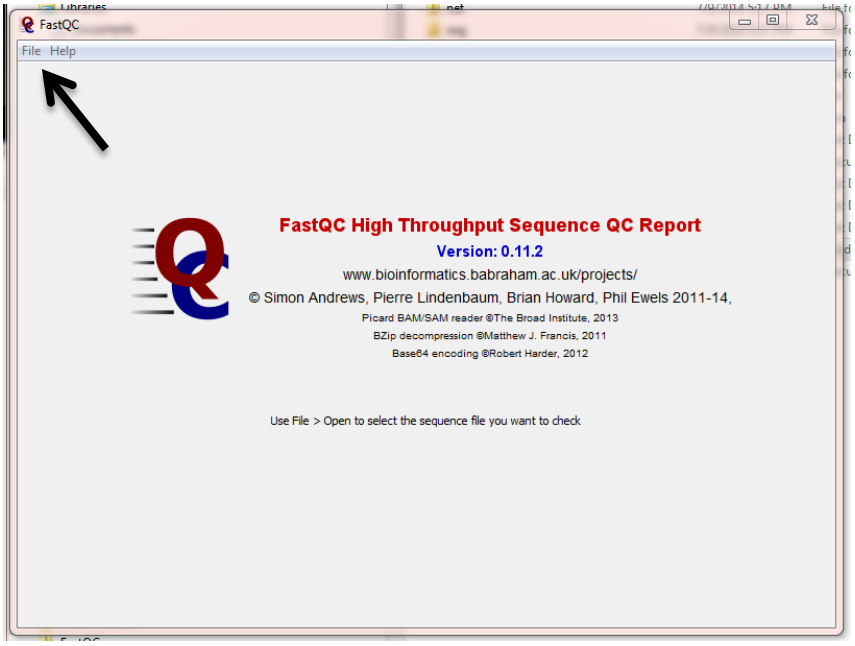
9. AMENDMENTS:

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.

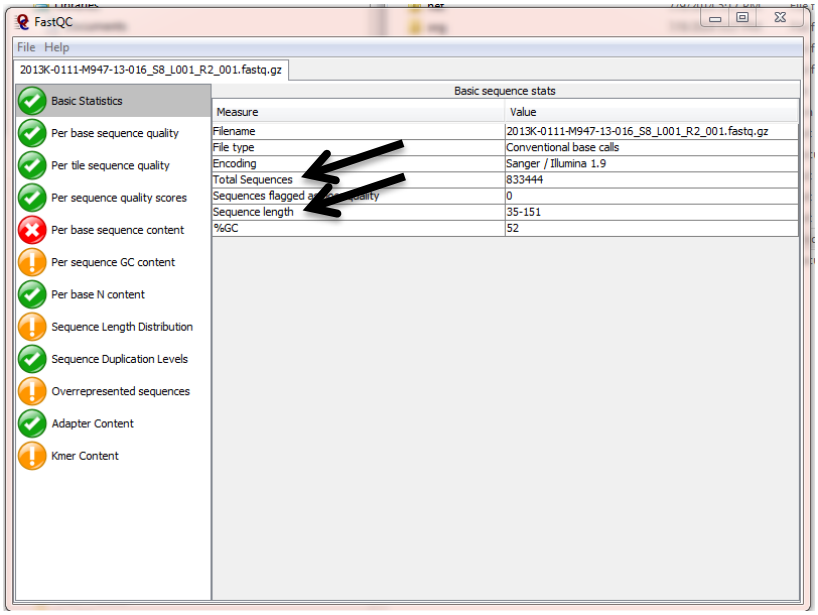
Appendix PNQ07-1

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.



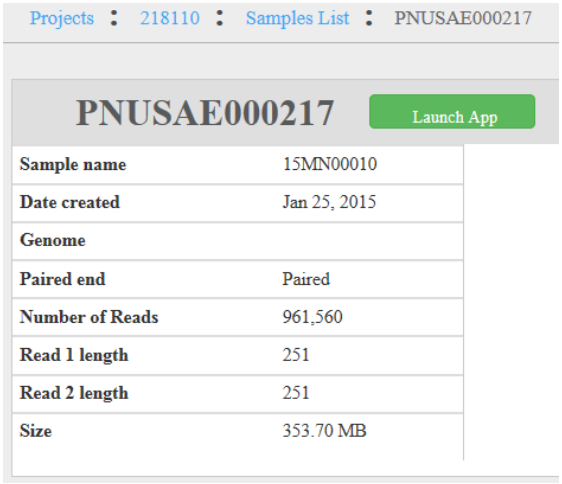
5. Choose Open and then select one read of your paired-end sequence.
6. Use total sequences (number of reads) multiplied by largest read length in range then divide that number by the genome size of your sequence to determine coverage (see the screenshot below).
   - **6.1.** *Listeria monocytogenes* - ((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


FastQC
File Help
2013K-0111-M947-13-016_S8_L001_R2_001.fastq.gz
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 sequence stats
Measure
Value
Filename
2013K-0111-M947-13-016_S8_L001_R2_001.fastq.gz
File type
Conventional base calls
Encoding
Sanger / Illumina 1.9
Total Sequences
833444
Sequences flagged as poor quality
0
Sequence length
35-151
%GC
52

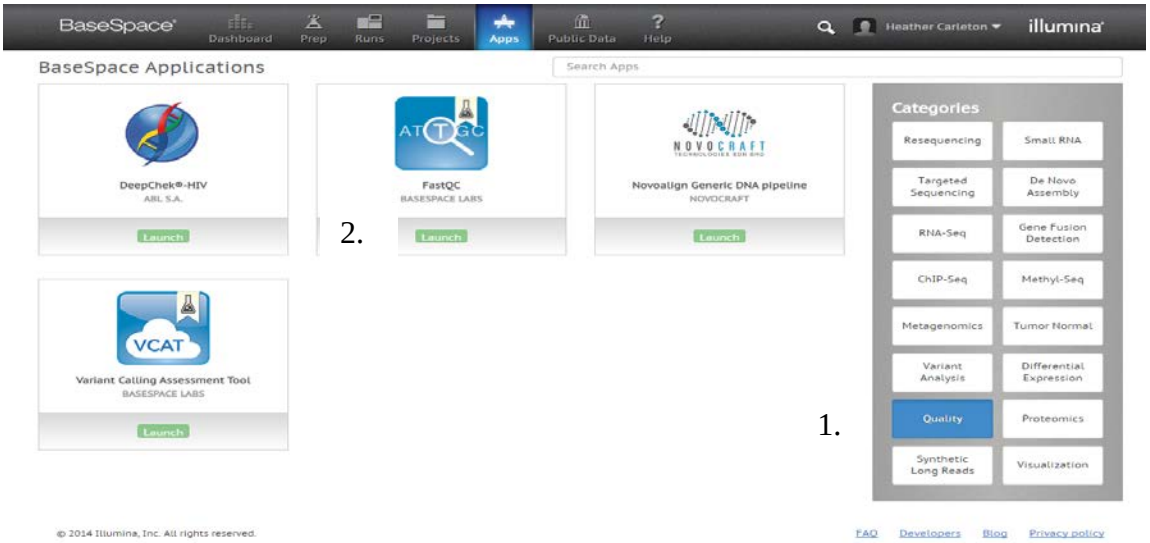
Appendix PNQ07-2

Using BaseSpace FastQC in BaseSpace and analysing data

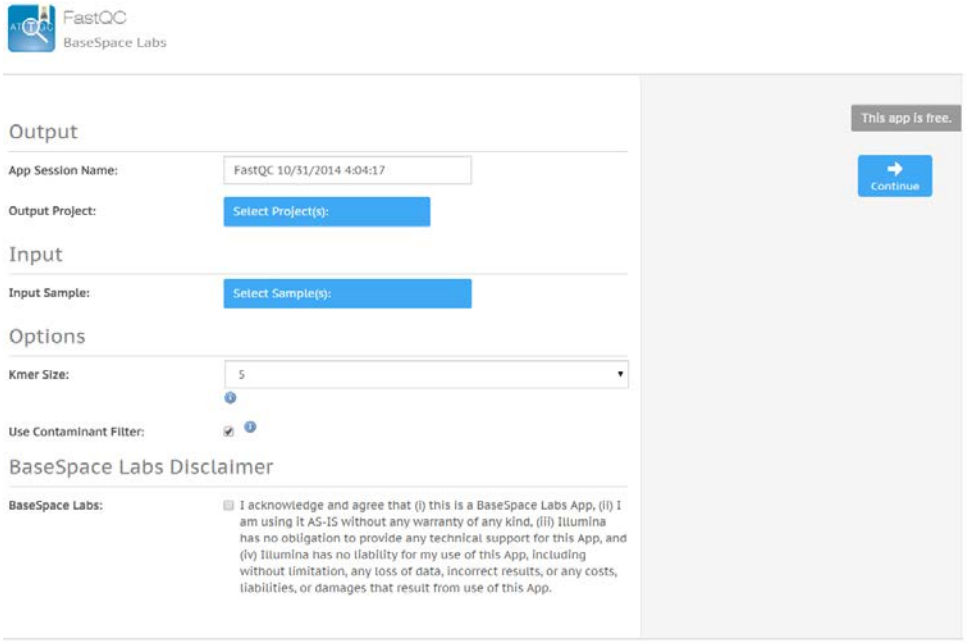
1. Login to BaseSpace; either stream your sequence data to BaseSpace as the run is being performed or upload the data afterwards (this only works with Illumina data by selecting Projects > Import > Sample; then drag and drop sequences into the displayed window)
2. To determine coverage, click on the 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.



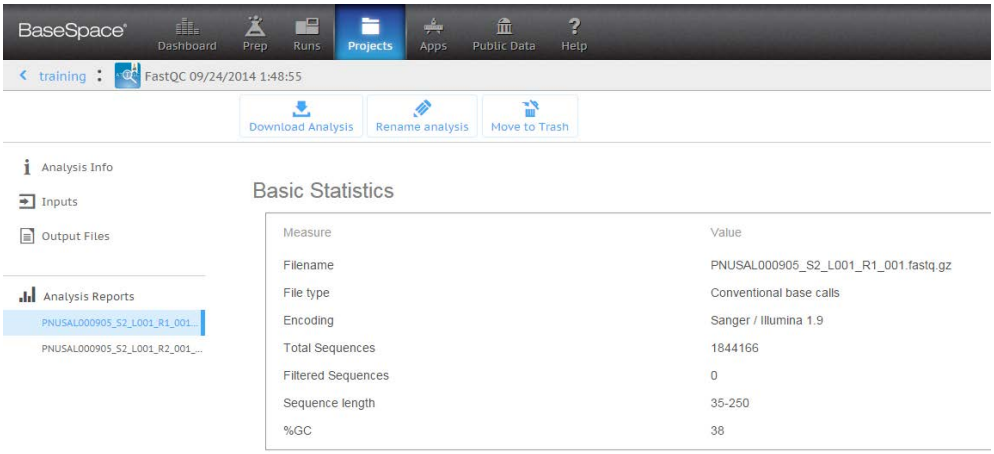
3. To run FastQC in BaseSpace, go to the Apps menu. Select Quality (1.) from the Categories panel and then select the Launch button for the free FastQC (2.) application.



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



5. 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).


BaseSpace
Dashboard
Prep
Runs
Projects
Apps
Public Data
Help
training
FastQC 09/24/2014 1:48:55
Download Analysis
Rename analysis
Move to Trash
Analysis Info
Inputs
Output Files
Analysis Reports
PNUSAL000905_S2_L001_R1_001...
PNUSAL000905_S2_L001_R2_001_...
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

Appendix PNQ07-3

Run Name LabID_MXXXX_YYMMDD
MiSeq Run Start Date 10/5/2015
MiSeq Reagent Kit 500 cycle V2 kit

Post-Run Metrics:	Q30	Cluster Density (K/mm^2)	Clusters Passing Filter	Estimated Yield (Gb)
FDA				
CDC	> 75%	600 - 1300	> 80%	7.5 - 8.5
Actual Values				

Minimum Coverage:	Escherichia/Shigella	Salmonella	Listeria	Campylobacter
FDA				
CDC	40X	30X	20X	20X

Index Number	Sample ID	Project	Index 1 (I7)	Index 2 (I5)	% Reads Identified (PF)	PF Reads	Total # of Reads	Coverage
1	PNUSAS123456				7.7379	17862215	1382160	138.22
2	PNUSAV123456				5.4	17862215	964560	96.46
							0	
							0	
							0	
							0	
							0	
							0	
							0	
							0	
							0	
							0	
							0	
							0	
							0	