1. **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.

1. **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.

1. **Related Documents**

|  |  |
| --- | --- |
| **Title** | **Document Control Number** |
| Bioinformatics QC Workflows |  |

1. **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 |

1. **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 |

1. **Equipment**

N/A

1. **Reagents and Media**

N/A

1. **Supplies, Other Materials**

N/A

1. **Safety Precautions**

N/A

1. **Sample Information / Processing**

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

1. **Quality Control**

N/A

1. **Workflow Chart**

N/A

1. **Test Procedure** 
   1. **Assembly**
      1. 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.
      2. **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
  1. **Assembly QC**
     1. 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.
     2. 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. |

* + 1. Once these steps are completed, please proceed to the next analysis step (SOP4).

1. **Method Performance Specifications**

N/A

1. **Calculations**

N/A

1. **Reference Values, Alert Values**

N/A

1. **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** andlarger **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.

1. **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.)*

1. **Reporting Results; Guidelines for Notification**

N/A

1. **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)*

1. **References**
   1. Illumina Sequence Analysis Viewer v1.11 Part # 15066069 v04 February 2018
2. **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. |  |

1. **Revision History**

|  |  |  |  |
| --- | --- | --- | --- |
| **Rev #** | **DCR #** | **Changes Made to Document** | **Date** |
|  |  |  |  |

1. **Approval Signature**

Approved By: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Date: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_