1. **Purpose**

This document provides quality control (QC) guidance for the analysis of nucleic acid next generation sequencing (NGS) data using bioinformatics. Following the generation of this NGS data, this guidance should be utilized with the analytical techniques used to process this data. 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 evaluation of run metrics, filtering of raw sequences, alignment/assembly, 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. Due to the diverse and rapidly advancing number of pipelines used in bioinformatics, this guidance document will describe the general steps that should undergo QC.

1. **Related Documents**

NOTE: Always check for updates and recent versions of guides.

|  |  |
| --- | --- |
| **Title** | **Document Control Number** |
| Sequencing QC SOP |  |
| Pre-Analysis QC SOP |  |
| Assembly QC SOP |  |

1. **Bioinformatics QC Checkpoints**

The following sections correspond to the process steps involved in bioinformatics for NGS, as outlined in the figure below (see Appendix A for a detailed process map).

****

 **Figure 1. Bioinformatics Checkpoints**

* 1. **NGS Output QC- Initial Filter and Sequencing Run QC:** Run metrics from the sequencer are evaluated using Illumina Sequencing Analysis Viewer (SAV). The key metrics involved in this step include the following:
		+ 1. **Cluster Density:** The density of clusters for each tile (in thousands per mm2).
			2. **% Clusters PF:** The percentage of clusters passing filter for each tile.
			3. **Yield Total:** Check that the read length is what is expected for the NGS platform and chemistry used for the sequenced organism.
			4. **% Aligned (PhiX):** The percentage of the passing filter clusters that aligned to the PhiX genome.
			5. **Accuracy of base (%Q>/=30):** Base calling accuracy describes the probability that the sequencer incorrectly assigned a nucleotide base. This is most commonly given as Q score which is calculated as: Q = - 10 log10 P, where P is the probability of error. Data with low Q scores may mean the data is unusable for further analysis. Q >30 is a standard threshold, which corresponds to 99.9% base calling accuracy.
	2. **Pre-Analysis QC – Trimming, Filtering and Quality Assessment using FASTQC:** This stage of QC follows after the generation of a FASTQ file. This guidance should be utilized to assess the quality of sequence data prior to assembly and further analysis. This step includes the use of a quality based trimming and filtering tool such as PrinSeq and quality assessment using a tool such as FASTQC. The key metrics assessed at this stage include:
		+ 1. **Total Sequences -** A count of the total number of sequences processed. There are two values reported, actual and estimated. At the moment, these will always be the same. In the future, it may be possible to analyze just a subset of sequences and estimate the total number, to speed up the analysis, but since we have found that problematic sequences are not evenly distributed through a file we have disabled this for now.
			2. **Filtered Sequences -** If running in Casava mode, sequences flagged to be filtered will be removed from all analyses. The number of such sequences removed will be reported here. The total sequences count above will not include these filtered sequences and will be the number of sequences actually used for the rest of the analysis.
			3. **Sequence Length -** Provides the length of the shortest and longest sequence in the set. If all sequences are the same length only one value is reported.
			4. **% GC -** The overall %GC of all bases in all sequences
			5. **Per base sequence quality -** For each position a BoxWhisker type plot is drawn. The elements of the plot are as follows: the central red line is the median value, the yellow box represents the inter-quartile range (25-75%), the upper and lower whiskers represent the 10% and 90% points, and the blue line represents the mean quality.
	3. **Alignment and Assembly QC:** At this stage, overlapping reads are aligned to create contigs and scaffolds for paired end reads. Homologous samples (i.e. bacterial isolate) are mapped to preexisting consensus genomes. Novel or heterogenous samples (i.e. isolates with no reference genome and/or metagenomics sample like stool) require *de novo* assembly. The quality of these assemblies is then evaluated using QUAST. The key metrics involved in the assessment stage for your assembly can be referenced below. Assembly joins reads that overlap into contigs (contiguous sequences). This is controlled by establishing minimum coverage, **N50**, **L50** and **minimum length of contigs** cutoff values that must be met. N50 describes a contig (contiguous sequence) length whereas L50 describes a number of contigs.
		+ 1. **Number of Contigs:** Total number of contigs of length
			2. **Total Length:** Total number of bases in the assembly
			3. **Minimum Coverage:** The minimum average depth of coverage and uniformity of coverage necessary for good assembly.
			4. **N50 length:** A statistical measure of average length of a set of sequences.N50 is the length (in basepairs) of the smallest contig that takes the sum length of all contigs – when summing from longest to shortest – past 50% of the total size (in basepairs) of the assembly.
			5. **L50 count:** The number of contigs evaluated at the point when the sum length exceeds 50% of the assembly size.
			6. **Minimum length of contig:** For very large assemblies the number of contigs can be over a million and mapping reads back to contigs will take a long time. Set a minimum contig length to reduce the number of contigs that have to be incorporated into the data structure.
	4. **Reference-based Assembly QC:** The metrics below are relevant to the evaluation of reference-based assembly quality only and should be used in tandem with the other metrics for assembly outlined above:
		+ 1. **Percentage of Genome Covered**: Assessed by calculating genome coverage or the average number of reads that align to the reference genome. How well the reads map to the reference genome indicates a certain level of confidence that any conclusions made downstream are reliable.
			2. **Uniformity of coverage:** This refers to the distribution of coverage within specific targeted regions. Although the average coverage may meet the laboratory established threshold, the depth of coverage will vary across the genome resulting in variable accuracy across the genome. Check that there is uniformity of coverage across the regions that are sequenced. This is calculated by the variance in sequencing depth across the genome post mapping. Non-uniformity can increase rate of false positives.
			3. **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. **De Novo Contig Assembly QC:** A *de novo* assembly joins reads that overlap into contigs (contiguous sequences). *De novo* assembly by definition lacks a reference sequence to use as a basis, therefore the quality of such an assembly should be evaluated using the aforementioned metrics in 2.3. which include **Minimum Coverage**, **N50**, **L50** and **Minimum length of contig**.
	2. **Analysis QC**
		1. **Variant calling QC:** The quality of variant calling is controlled by establishing the following parameters and thus, variant calling only occurs at positions that meet these requirements.
			1. A non-reference base (a variant) is detected.
			2. **Allele call score**: The allele call score refers to the probability of an incorrect base call (e.g. a score of 3 is equivalent to a Phred score of Q30 meaning the likelihood of an incorrect call for the base is 1 in 1000). The default Illumina setting for allele call score is >/=10.
			3. **Minimum coverage**: Illumina recommends a mean coverage of 30x for DNA sequencing assuming a Phred score of Q30. Additionally, the depth at the SNPs position should be no greater than three times the chromosomal mean.
			4. **Heterozygous calls:** Both alleles should have an allele call score >/= 10, and the ratio of their scores should be </=3.
1. **Appendices**

Appendix A – Bioinformatics QC Checkpoints Process Map

1. **References**
	1. Andrews, S. (2010). FastQC: a quality control tool for high throughput sequence data.
	2. Alexey Gurevich, Vladislav Saveliev, Nikolay Vyahhi and Glenn Tesler, QUAST: quality assessment tool for genome assemblies, Bioinformatics (2013) 29 (8): 1072-1075.
	3. Schmieder, R., & Edwards, R. (2011). Quality control and preprocessing of metagenomic datasets. *Bioinformatics*, *27*(6), 863-864.
	4. Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics, btu170. Chicago.
	5. Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., ... & Pyshkin, A. V. (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of computational biology*, *19*(5), 455-477.
	6. Simpson, J. T., Wong, K., Jackman, S. D., Schein, J. E., Jones, S. J., & Birol, I. (2009). ABySS: a parallel assembler for short read sequence data. *Genome research*, *19*(6), 1117-1123.
	7. *De novo* bacterial genome sequencing: millions of very short reads assembled on a desktop computer. Hernandez D, François P, Farinelli L, Osterås M, Schrenzel J. *Genome Research.* 18:802-809, 2008.
	8. Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature methods*, *9*(4), 357-359.
	9. Ponstingl, H. (n.d.). SMALT. Retrieved May 15, 2017, from http://www.sanger.ac.uk/science/tools/smalt-0
	10. MiSeq Specifications. (n.d.). Retrieved May 16, 2017, from https://www.illumina.com/systems/sequencing-platforms/miseq/specifications.html
2. **Revision History**

|  |  |  |  |
| --- | --- | --- | --- |
| **Rev #** | **DCR #** | **Change Summary** | **Date** |
|  |  |  |  |

1. **Approval**

This document has been approved by the CDC CLIA Laboratory Director as the standard practice for CLIA-regulated CDC Infectious Diseases Laboratories under certificates 11D0668319 and 11D2030855.

Approved:

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Appendix A – Bioinformatics QC Checkpoints Process Map**

