Sequencing quality

The quality of sequencing data can be measured from raw reads and assembled sequences.

1. Raw reads

Among several available software options, FASTQC is among the most popular tool for calculating and presenting sequence statistics for NGS sequencing reads and can be used for estimating quality. The tool provides a simple way to perform quality control checks on raw sequence, providing a flexible set of QC parameters such as

1.1. *Per base sequence quality*. It shows an overview of the range of quality, measured as phred scores across all bases at each position in the FastQ file (raw reads). The score can be divided into very good quality calls (green, more than 30), calls of reasonable quality (orange, between 20-30), and calls of poor quality (red, less than 20).

1.2. *Per base sequence content*. It plots out the proportion of each base position in raw read file for which each of the four normal DNA bases has been called. In a random library you would expect that there would be little to no difference between the different bases of a sequence run, so the lines in this plot should run parallel with each other. The relative amount of each base should reflect the overall amount of these bases in your genome, but in any case they should not be hugely imbalanced from each other.

1.3. *Per sequence GC content*. It measures the GC content across the whole length of each sequence in a raw read file and compares it to a modelled normal distribution of GC content. In a normal random library, you would expect to see a roughly normal distribution of GC content where the central peak corresponds to the overall GC content of the underlying DNA sequenced.

1.4. *Duplicate sequences*. In a diverse library most sequences will occur only once in the final set. A low level of duplication may indicate a very high level of coverage of the target sequence, but a high level of duplication is more likely to indicate enrichment bias (e.g. PCR over amplification). This module counts the degree of duplication for every sequence in a library and creates a plot showing the relative number of sequences with different degrees of duplication.

After the evaluation of the quality of the raw data, reads usually undergo a preprocessing step during which they are “cleaned” from low quality data including adaptor sequences inserted during library preparation. Several read processing

software options are available, among the most popular is TRIMMOMATIC as it is a flexible and efficient preprocessing tool, which can handle paired-end data.

2. De novo assembly

In WGS, the genomes will frequently be assembled in larger continuous sequences defined as 'contigs', after read processing. Genome draft assembly is performed on sequence reads, often by using one of the following assembly strategies: overlap, layout, consensus (OLC) or de Bruijn graph. Depending of the platform used for producing the reads. Complete assemblies are compared to draft assemblies rare, as they require additional resources. Table 1 shows a non-exhaustive review of assemblers and their applications. The QC parameters that should be measured from assembled genomes are summarized in the table 2.

Alternatively, the pilot proficiency test (PT) from Global Microbial Identifier (GMI) is a testing platform to assist in evaluating the reliability of laboratory results of consistently good quality within the area of whole genome sequencing. Specifically, the PT will evaluate the consistency and robustness of a laboratory's ability to perform deoxyribonucleic acid (DNA) extraction, library preparation, WGS, assembly and phylogenetic analysis following different laboratory protocols, software tools, and sequence platforms for the reliability of submitted sequence data.

Table 1 List of assembler used in foodborne pathogen WGS pipelines

Assembler	**Platforms**
Velvet	Illumina reads
SPAdes	Illumina or IonTorrent reads. It is capable of providing hybrid assemblies using PacBio, Oxford Nanopore and Sanger reads
MIRA	Sanger, 454, Illumina and IonTorrent reads. Can perform hybrid assemblies.
Canu	PacBio and Oxford Nanopore reads. A fork of the Celera Assembler designed for high-noise single-molecule sequencing

Table 2 List of QC parameters and description

QC parameters	**Description**
Number of reads	The number of reads refers to the sequence yield, how much was sequenced.

Average read length	The average length of all the reads and is measured in bp.
Depth of coverage, total DNA sequence	Number of bps sequenced divided by the total size (both chromosome and plasmids) of the closed genome (same strain). This number can be rounded to the nearest integer. In essence this number describes the number of times the sequenced bps covers the reference DNA and is often ended with an "x" (e.g. 30x). Coverage greater than 30X indicates good quality.
Size of assembled genome	The total size of all the contigs in bp. The total bp should be corresponded to the size of sequenced genome.
Total number of contigs greater than 500 bp	The total number of contigs greater than 500 bp. A number of contigs less than 500 contigs normally recommend as good quality.
N50	the N50 length is defined as the length for which the collection of all contigs of that length or longer contains at least half of the sum of the lengths of all contigs, and for which the collection of all contigs of that length or shorter also contains at least half of the sum of the lengths of all contigs. A N50 more than 30 Kb normally indicate good quality.