# Quality control of High-Throughput Sequence (HTS) data

# Aim

# The aim of this standard operating procedure (SOP) is to provide guidance on quality control of sequences used for diagnostic purpose and generated by high-throughput sequencer. The report from this procedure is necessary for evaluating whether the data have any problems which should be aware before doing any further analysis.

The analysis is performed by the bioinformatician from Norwegian High-Throughput Sequencing Centre (NSC).

# Tools

The quality control is performed by the bioinformatics tool: FastQC (version v10.0). It is a quality control tool for high throughput sequence data.

# Input

The input file is the sequence file in FASTQ format, which is delivered from NSC.

Each entry in a FASTQ file consists of four lines. Line 1 begins with a '@' character and is followed by a sequence identifier. Line 2 is the raw sequence letters. Line 3 is quality score identifier line, consisting of a ‘+’. Line 4 encodes the quality values for the sequence in Line 2, and must contain the same number of symbols as letters in the sequence.

A quality score (or Q-score) expresses an error probability. In particular, it serves as a convenient and compact way to communicate very small error probabilities. Given an assertion, A, the probability that A is not true, P(~A), is expressed by a quality score, Q(A), according to the relationship:

Q(A) =-10 log (P(~A)) 10

where P(~A) is the estimated probability of an assertion A being wrong. The quality scores are then encoded into a compact form in FASTQ files which uses only one byte per quality value. In this encoding the quality score is represented as the character with an ASCII code equal to its value + 33. (See details about FASTQ format in Appendix 1 – CASAVA *1*8\_2\_UG\_15011196C.pdf, page 39 - 41)

# Procedure

The control is performed by the following command:

fastqc input\_fastq\_file

(See Appendix 3 for fastqc option detail).

**Output**

The final report is in PDF format with all associated information extracted from the html file generated by the above command (See one example on Appendix 2). The following qualities of the sequences are described in the final report:

1. **The number of fragments**: The number of fragments is the number of sequences in the input file.
2. **Per Base Sequence Quality**: The graph shows an overview of the range of quality scores across all based at each position in the FASTQ file. The y-axis shows quality scores and the x-axis shows the read position. For each read position, a boxplot is used to show the distribution of quality scores for all reads. The yellow boxes represent quality scores within the inter-quartile range (25% - 75%). The upper and lower whiskers represent 10% and 90% point. The central red line shows the median of the quality values and the blue line shows the mean of the quality values. The higher the score the better the base call. The minimum requirement for Per Base Sequence Quality is that the first 36 bases should have a median and mean quality score over 20.
3. **Per Sequence Quality Scores**: The graph is generated by computing the average quality score of a read (by averaging across read positions) and then plotting the distribution of this average quality. The y-axis shows the number of reads and the x-axis shows the mean quality score.
4. **Per Base Sequence Content:** The graph plots the percentage of each base type (each line) in each read position. The y-axis shows the percentage of a base type and the x-axis shows the read position. In a whole exome sequencing sample, all lines should be roughly flat, and the GC content should be between 40% - 50%.
5. **Per Base N content:** The graph shows the percentage of base calls at each position for which an N was called. If a sequencer is unable to make a base call with sufficient confidence then it will normally call an N rather than A, T, G or C. The y-axis shows percentage of Ns among all reads and the x-axis shows the read position. It is common to see a very low percentage of Ns appearing near the end of a sequence.
6. **Sequence Duplication Levels**: The graph shows the number of sequences with different degrees of duplication (indicated on the x-axis) relative to the number of unique sequences (which is set to 100%). In a diverse library (e.g. exome sequencing sample), most sequences will occur only once in the final set and the graph will show a peak in the unique category. However, some sequences may be present in more than one copy (for example, as the result of PCR amplification), in which the graph may show high numbers of sequences in the other categories (2 copies, 3 copies, etc). The last category is for 10 copies or more. Therefore, it is normal to see a small rise in this category. This graph is useful to indicate whether the sample contains a large amount of PCR duplicates.

All the original results from fastqc command and PDF report will be delivered from NSC to AMG.

**Storage**

See “SOP for storage and security of high-throughput sequencing data”.

**Reference**

1, <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>

Appendix

1, CASAVA *1*8\_2\_UG\_15011196C.pdf

2, 121114\_SN586\_0138\_BC1CG7ACXX.5.Sample\_Diag-excap12-D12-3429.Read1.qc.pdf

3, fastqcOptions.txt