FASTQ files – Answers Part 1

1. What are Fastq files?

FASTQ format is a text-based format for storing both a biological sequence (nucleotide sequence) and its corresponding quality scores. Both the sequence letter and quality score are each encoded with a single ASCII character

2. Types of FASTQ files?

Single read FASTQ – provided as one file

Paired FASTQ – provided as two files, the main part of the filename (before the extension) for both paired files must be exactly the same except for the R1 and R2 designations (or just a 1 or 2 in the filename instead of R1 or R2). The filename of the file containing the forward sequence should include R1, and reverse sequence is marked by R2.

Concatenated FASTQ – in which all forward reads are followed by all reverse reads

3. Main parts?

A FASTQ file normally uses four lines per sequence. First line begins with a '@' character and is followed by a sequence identifier and an optional description (like a FASTA title line). Second line is the raw sequence data (letters A, C, G, T and N). Third line begins with a '+' character and is optionally followed by the same sequence identifier (already displayed in the first line). Fourth line encodes the quality values for the sequence, base call quality scores, and must contain the same number of symbols as letters in the sequence. These are Phred +33 encoded, using ASCII characters to represent the numerical quality scores.

4. How are they created?

Illumina sequencing technology uses cluster generation and sequencing by synthesis (SBS) chemistry to sequence millions or billions of clusters on a flow cell. During SBS chemistry, for each cluster, base calls are made and stored for every cycle of sequencing by the Real-Time Analysis (RTA) software on the instrument. RTA stores the base call data in the form of individual base call (or BCL) files. When sequencing completes, the base calls in the BCL files must be converted into sequence data. This process is called BCL to FASTQ conversion. If samples were multiplexed, the first step in FASTQ file generation is demultiplexing. Demultiplexing assigns clusters to a sample, based on the cluster's index sequence(s). After demultiplexing, the assembled sequences are written to FASTQ files per sample. If samples were not multiplexed, the demultiplexing step does not occur, and, for each flow cell lane, all clusters are assigned to a single sample.

5. Import FQ files in IGV and observe it.

IGV does not support importing FQ files since they are not alignments and there is no possibility to represent it on reference genome. The information can be displayed if possible by converting FASTQ to BAM.

6. Explain different distance between FASTQ pairs.

In Illumina paired-end sequencing, chromosomes are sheared into small fragments and size selected such that most fragment lengths are within the interval [lmin,lmax], normal distribution with given average and standard deviation. Each fragment is sequenced from both ends from opposite DNA strands; thus one read will originate from the forward (R1) strand and one from the reverse (R2) strand. Paired reads are aligned to the reference genome. If we present each aligned read as x = (x1, x2, r) where x1 is leftmost position, x2 is rightmost position and r is R1 or R2 for orientation. For read-pair [x, y] (where x has smaller starting position) distance between aligned reads is dist = y2-x1. If lmin<dist<lmax and x has orientation R1 and y has orientation R2 than read-pair is properly paired or concordant.