NGI-Stockholm -- Science For Life Laboratory

Best-practice analysis for quality checking report A.Blomberg_15_20 -- P3722_201

For sample P3722_201 belonging to the project A.Blomberg_15_20 NGI-Stockholm best-practice analysis for quality checking has been performed. For mate pair libraries produced with Nextera, best-practice analysis described at this address has been performed:
http://res.illumina.com/documents/products/technotes/technote_nextera_matepair_data_processing.pdf

The following tools have been employed (tools are listed in order of execution):

i trimmomatic : /sw/apps/bioinfo/trimmomatic/0.32/irma/trimmomatic.jar
ii fastqc : /sw/apps/bioinfo/fastqc/0.11.2/irma/fastqc
iii abyss : /sw/apps/bioinfo/abyss/1.3.5/irma/bin/

The results from each tool is reported in the following sections. Moreover you will find all the results and commands that have been run in the delivery folder on Uppmax

Trimmomatic

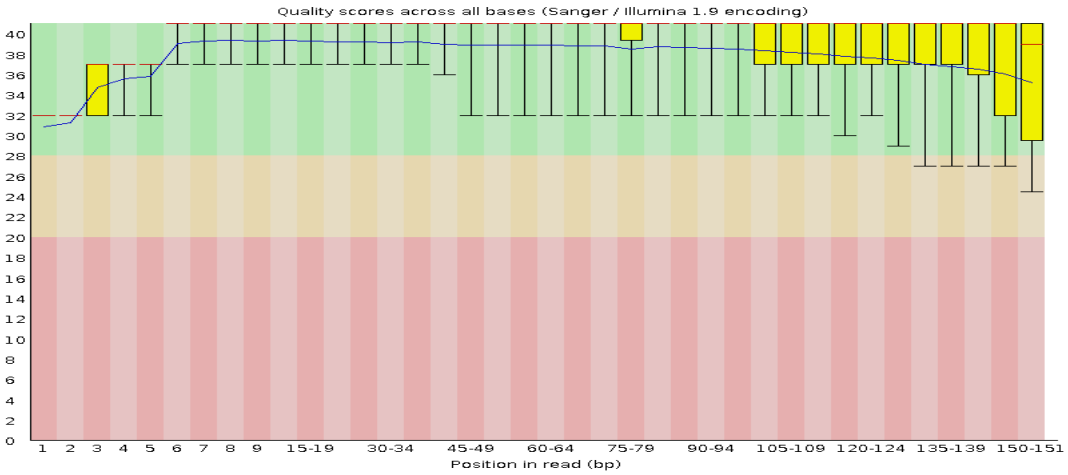
Reads (both paired and mate pairs) can contain parts of the adapter sequence or, in the case of mate pairs, part of the linker sequence. Illumina recommends to remove the adapter before use of the reads in any downstream analysis (this is mandatory for mate pairs).
Adapter sequences removed are:

i CTGTCTCTTATACACATCTAGATGTGTATAAGAGACAG
ii CTGTCTCTTATACACATCT
iii AGATGTGTATAAGAGACAG

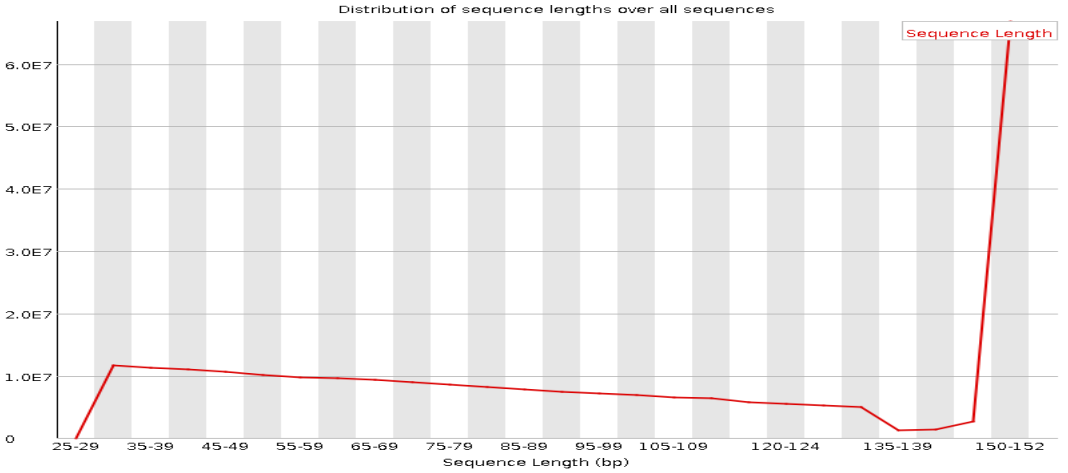
P3722_201	#orig_pairs	#survived_pairs
P3722_201_S1_L001_R1_001_trimmomatic.stdErr	435321273	247076038 (57%)
total	435321273	247076038 (57%)

P3722_201	#survived_fw_only	#survived_rv_only	#discarded
P3722_201_S1_L001_R1_001_trimmomatic.stdErr	99184311 (23%)	83052205 (19%)	6008719 (1%)
total	99184311 (23%)	83052205 (19%)	6008719 (1%)

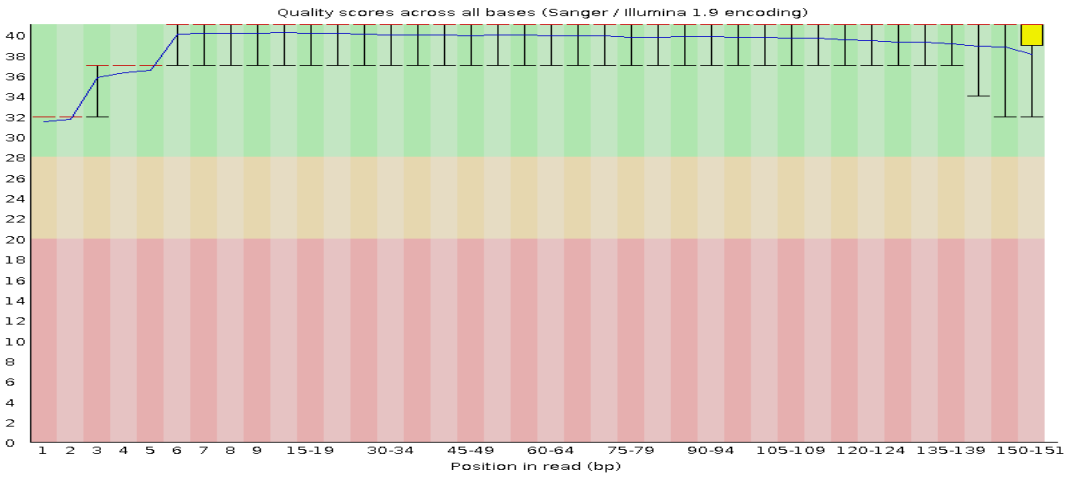
Fastqc



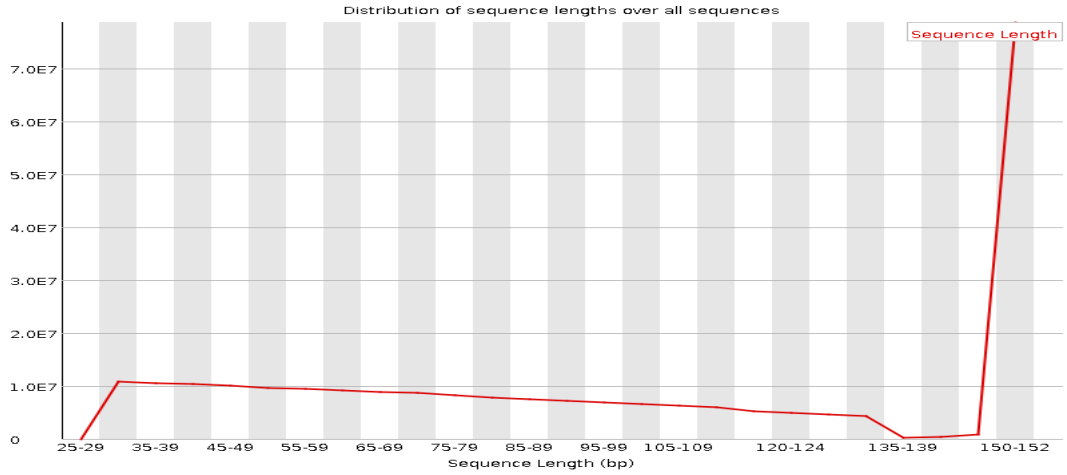
P3722_201_S1_L001_R2_001_fastqc -- Per Base Quality



P3722_201_S1_L001_R2_001_fastqc -- Sequence Length Distribution



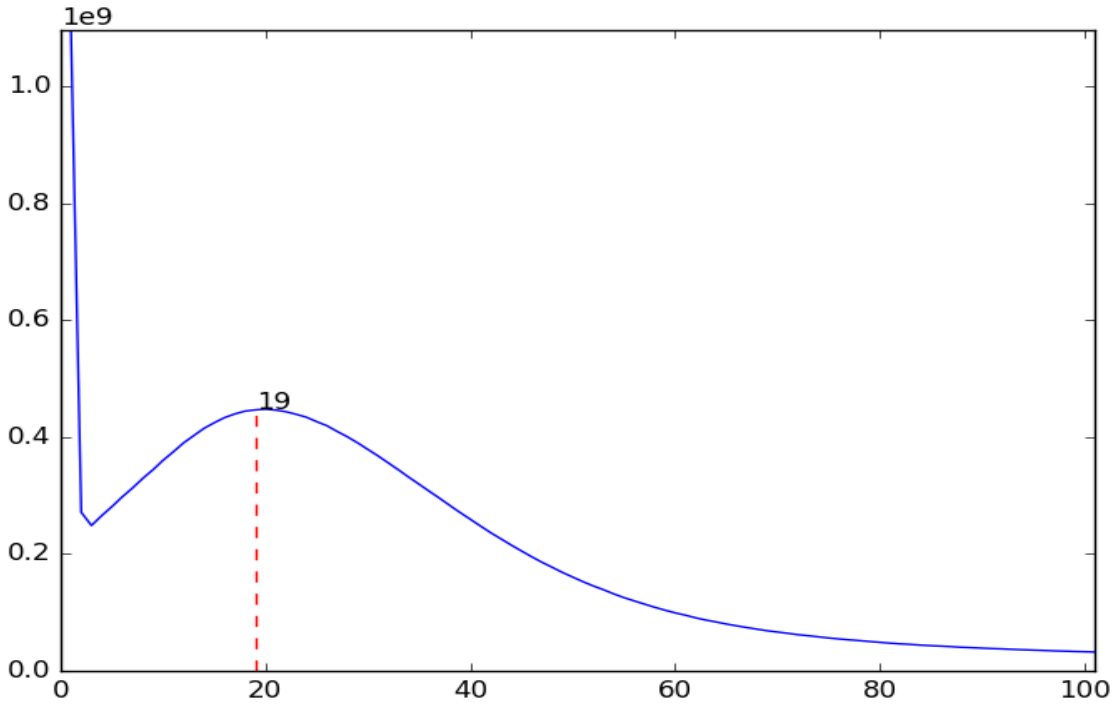
P3722_201_S1_L001_R1_001_fastqc -- Per Base Quality



P3722_201_S1_L001_R1_001_fastqc -- Sequence Length Distribution

Abyss

A possible way to assess the complexity of a library even in absence of a reference sequence is to look at the kmer profile of the reads. The idea is to count all the kmers (i.e., sequence of length k) that occur in the reads. In this way it is possible to know how many kmers occur 1,2,..., N times and represent this as a plot. This plot tell us for each x, how many k-mers (y-axis) are present in the dataset in exactly x-copies. In an ideal world (no errors in sequencing, no bias, no repeating regions) this plot should be as close as possible to a gaussian distribution. In reality we will always see a peak for x=1 (i.e., the errors) and another peak close to the expected coverage. If the genome is highly heterozygous a second peak at half of the coverage can be expected.



kmer profile with k=35.