Sample: CLLR146

This report presents a detailed view of the processing steps and controls performed on the sample J25_rep3 from the project S18148 sequenced on the flowcell HS463.

Contents

1. General information

Sample platform ID: CLLR146

Sample name: J25_rep3

Organism: Homo sapiens

Library number: P1382

Index used to demultiplex: ACAGTG

Flowcell: HS463

2. Results

2.1. Statistics

2.1.1. Adapter dimers

Adapter dimer reads were removed using DimerRemover [1]. Table 1 provides the corresponding number and percentage of reads.

Table 1: Adapter dimers count. *The "Raw reads" column represents the number of sequenced reads, the "Adapter dimers" column respresents the number of sequenced adapter dimers and the "Total read" column represents the final set of reads that is delivered in the fastq file (i.e. raw reads without adapter dimer reads). "% dimer" is the ratio (dimer / raw reads).*

Sample ID	Sample name	Raw reads	Adapter dimers	% dimer	Total read
CLLR146	J25_rep3	19,198,961	2	0.0	19,198,959

Thereafter, the set of reads used to assess data quality corresponds to reads without adapter dimers, *i.e.* all reads available in the delivered fastq file.

2.1.2. Alignment

Sequenced reads were mapped to the Homo sapiens genome assembly hg19 using Bowtie [2]. Alignment statistics are provided in Table 2.

Table 2: Sequencing and mapping statistics. *The "Uniquely mapped" column represents the number of reads uniquely aligned onto the reference genome. "% mapped" is the ratio (uniquely mapped / total reads). The "Unique positions" column contains the number of different positions in the genome to which uniquely aligned reads are. "% positions" is the ratio (unique positions / uniquely mapped). The unique aligned reads and unique position numbers are for information purpose only.*

Sample ID	Sample Name	Total reads	Uniquely mapped	% mapped	Unique positions	% positions
CLLR146	J25_rep3	19,198,959	319	0.0	316	99.06

2.2. Quality controls

2.2.1. Base quality

Table 3 provides the percentage of bases with a quality score above 30 (Q30) and figure 1 shows the average base quality at each position along the read.

Table 3: Percentage of bases above Q30.

Sample ID	Sample name	% of bases above Q30
CLLR146	J25_rep3	86.96

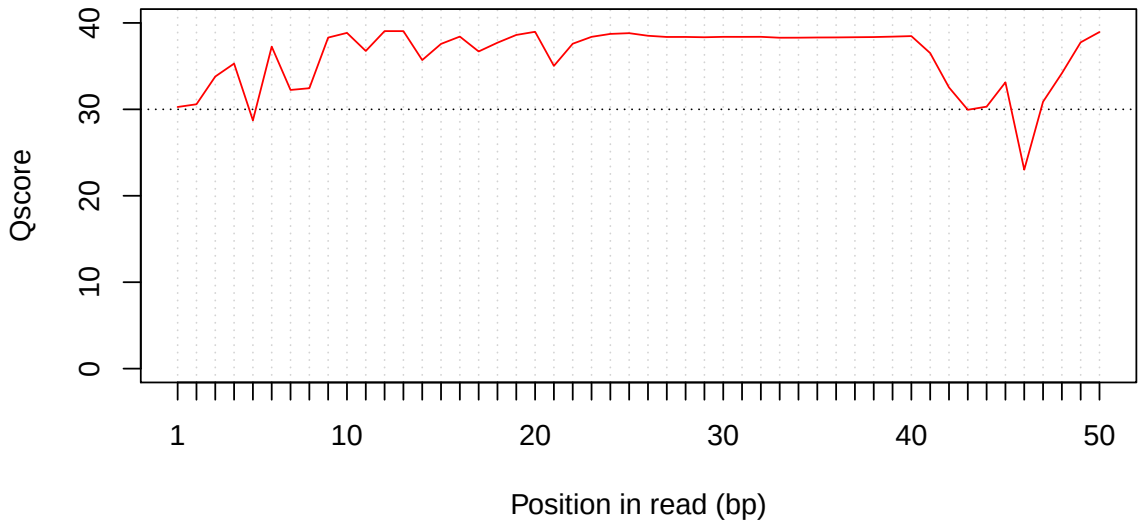

Figure 1: Average base quality at each position along the read. *The x-axis represents the base pair position in the read and the y-axis the Qscore. The dotted line represents a quality score of 30 corresponding to an error probability of 1‰.*

2.2.2. Base sequence content

Figure 2 provides the percentatge of each base (A, C, G and T) for each position along the read and figure 3 the percentage of undetermined (N) bases for each position along the read.

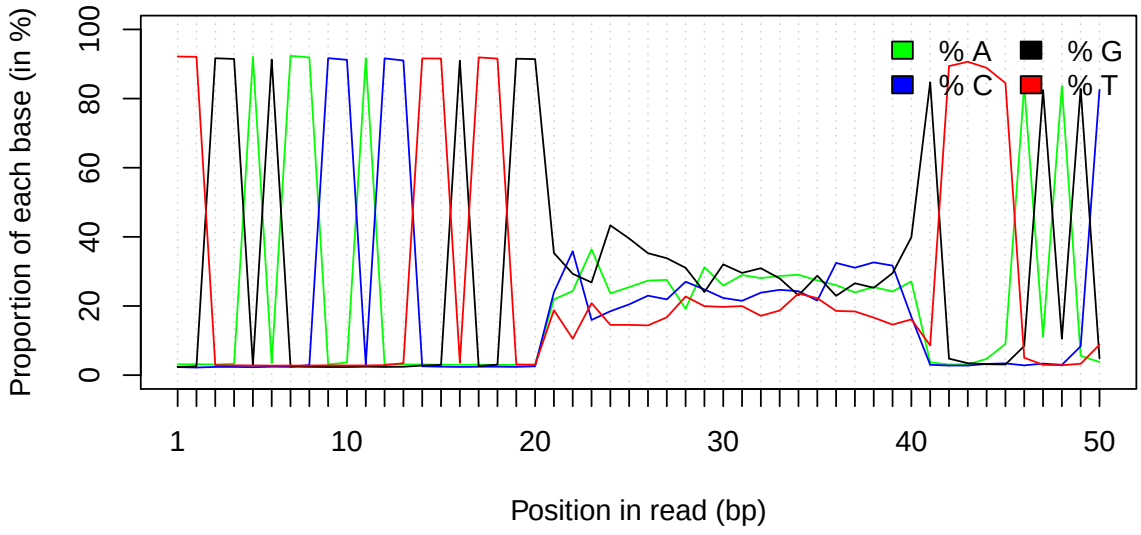

Figure 2: Proportion of each nucleotide (A, C, G and T) along the read. *The x-axis represents the base position along the read and the y-axis the percentage of the corresponding bases.*

2.2.3. Duplicated reads

Figure 4 represents the percentage of duplicated reads within the first 100,000 reads in function of their duplication level. 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 some kind of enrichment bias (*e.g.* PCR over amplification).

2.2.4. FastQScreen

FastQScreen [4] maps a subset of reads to selected DNA sequences in order to assess the presence of contamination within the sample. Results are displayed in Figure 5. Most of the reads should map to the reference genome of the sample. Some reads may be mapped to more than one DNA sequence (in red), this can result from reads mapped in region conserved across different species.

3. Methods

3.1. Sequencing and data preprocessing

The library was sequenced on Illumina Hiseq 4000 sequencer as Single-Read 50 base reads following Illumina's instructions. Image analysis and base calling were performed using RTA 2.7.3 and bcl2fastq 2.17.1.14. Adapter dimer reads were removed using DimerRemover [1].

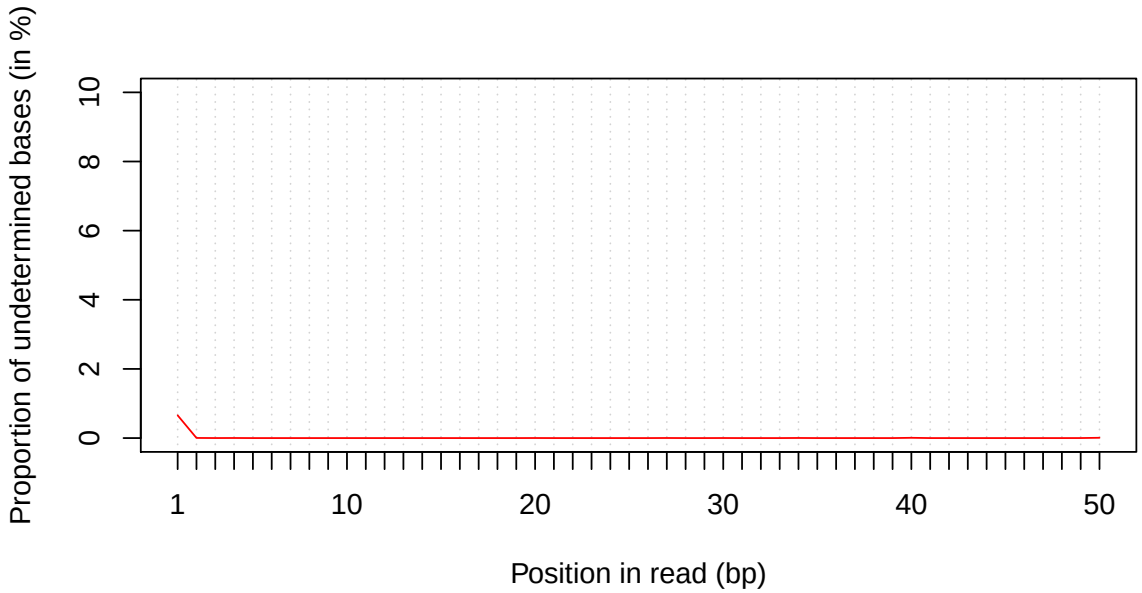

Figure 3: Proportion of undetermined nucleotide (N) along the read positions. *The x-axis represents the base position along the read and the y-axis the percentage of undetermined bases.*

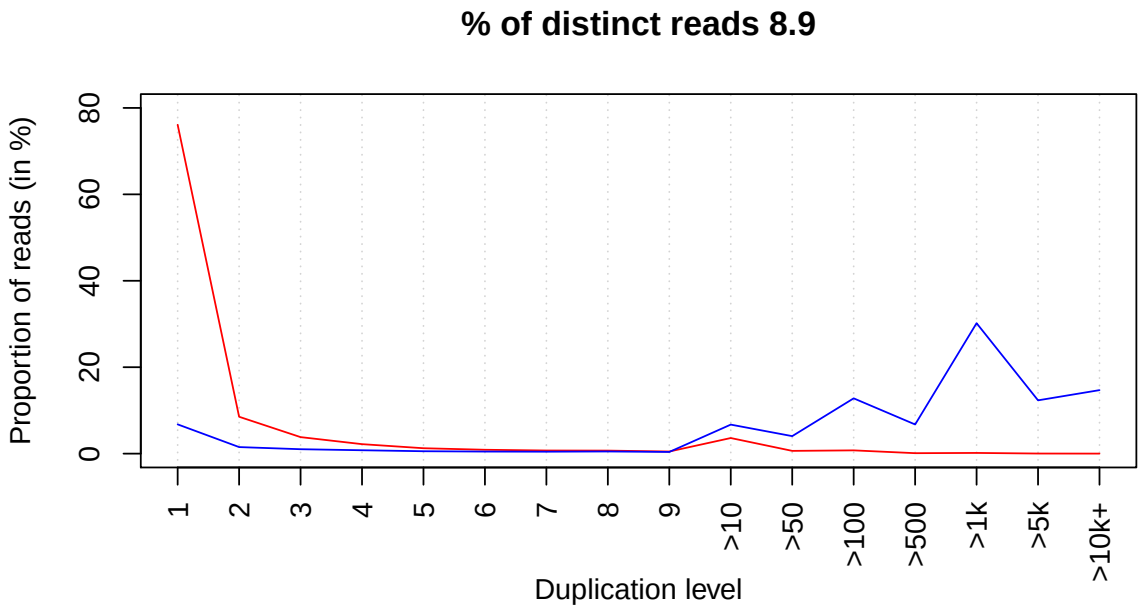

Figure 4: Duplicated reads. *The x-axis represents the duplication level and the y-axis the percentage of reads in the first 100,000 reads. A duplication level of 1 means that the read is only seen one time in the set of reads, i.e. it represents the proportion of unique reads; a duplication level of 2 means that a given read is seen two times in the set of reads, and so on. For a given duplication level, the blue (resp. red) line represents the proportion of all sequences relative to the total number of reads (resp. to the total number of distinct sequences).*

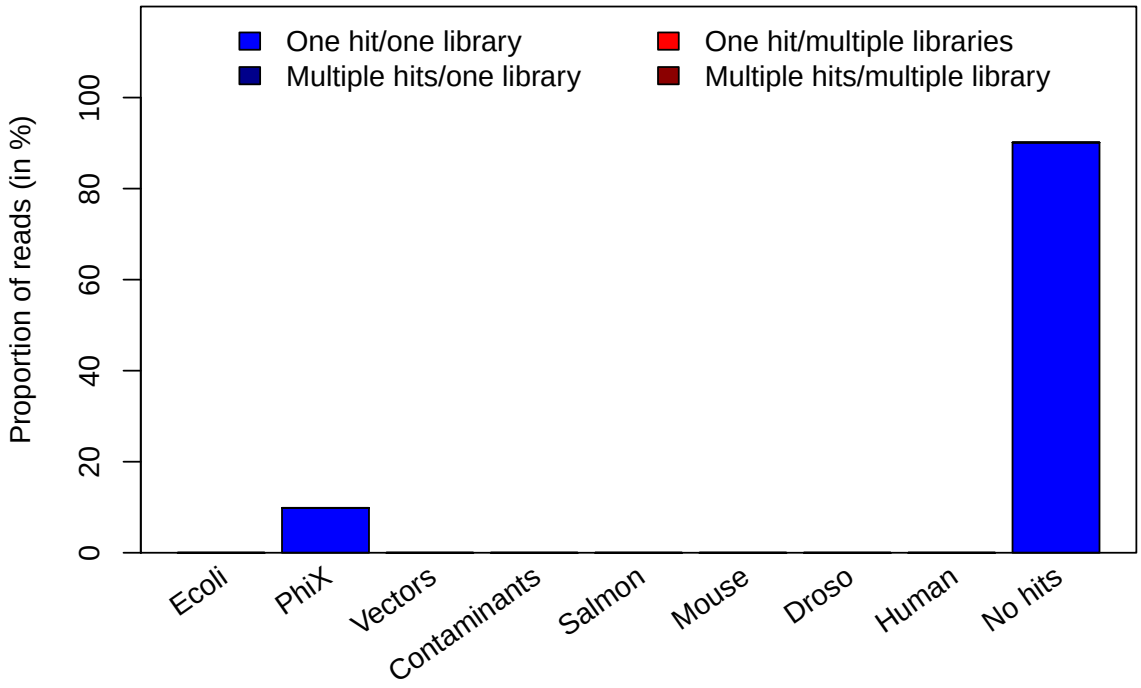

Figure 5: Proportion of reads aligned to each DNA sequences. *The x-axis represents the DNA sequence where a subset of reads has been aligned to. The y-axis represents the % of reads that have been aligned to the corresponding DNA sequence. The blue (red) colors indicate reads which can be assigned to only (more than) one DNA sequence. The ligth (dark) colors indicate reads that can be aligned to only (more than) one position onto the DNA sequence(s).*

3.2. Alignment

Alignement was performed using Bowtie 1.0.0 with the following arguments: `-m 1 --strata --best -y -S -l 40 -p 8`.

Detailed description of the arguments:

`-m <int>` Suppress all alignments for a particular read or pair if more than `<int>` reportable alignments exist for it. Reportable alignments are those that would be reported given the `-l`, `--best`, and `--strata` options.

`--strata` If many valid alignments exist and are reportable and they fall into more than one alignment "stratum", report only those alignments that fall into the best stratum.

`--best` Make Bowtie guarantee that reported singleton alignments are "best" in terms of stratum and in terms of the quality values at the mismatched position(s).

`-y/--tryhard` Try as hard as possible to find valid alignments when they exist, including paired-end alignments. This mode is generally much slower than the default settings, but can be useful for certain problems.

`-l/--seedlen <int>` The "seed length"; *i.e.*, the number of bases on the high-quality end of the read to which the `-n` ceiling applies (default $n = 2$).

3.3. Quality control

3.3.1. FastQC

Version 0.11.2 was run using the following arguments `--nogroup --casava` to produce base quality, base sequence content and duplicated reads data represented on figures 1, 2, 3 and 4.

3.3.2. FastQScreen

Version 0.5.1 was run using the following arguments: `--subset 10000000 --aligner bowtie --bowtie '-p 2'`.

4. Contributors

- Wet lab operator: Customer
- Data processing and report operator: Matthieu Jung

5. References

[1] DimerRemover : `https://sourceforge.net/projects/dimerremover/`

[2] Bowtie: Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10:R25.

[3] FastQC: `http://www.bioinformatics.babraham.ac.uk/projects/fastqc/`

[4] FastQScreen: `http://www.bioinformatics.babraham.ac.uk/projects/fastq_screen/`