segemehlQC

A quality control tool for mapped NGS data

Marcel Winter

Bioinf Leipzig

Preface

SAM format: text-based format for storing biological sequences aligned to a reference sequence[1]

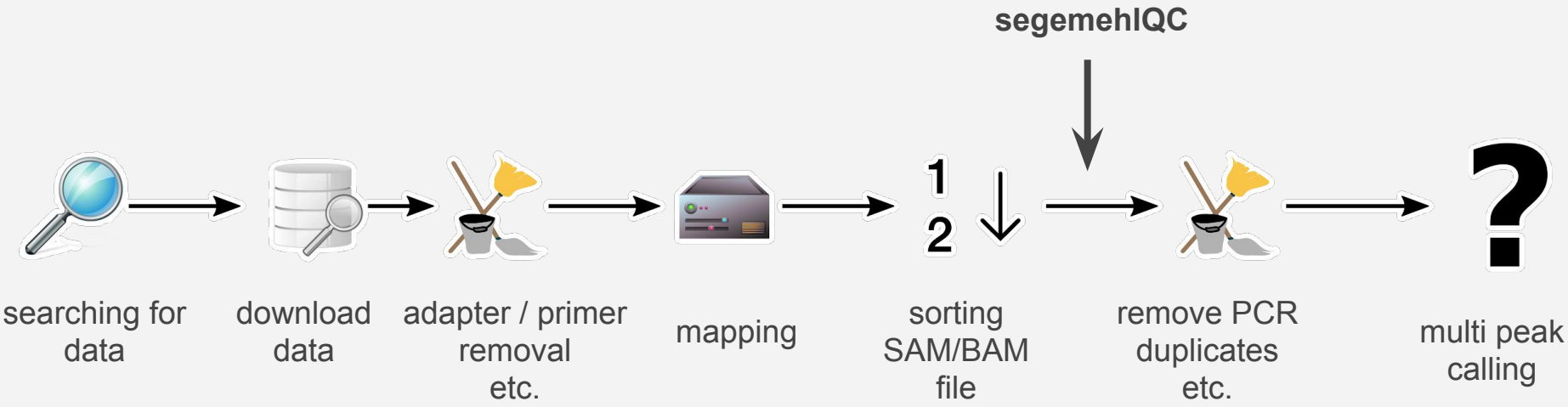
BAM format: compressed binary representation of SAM[2]

segemehl: software to map short sequencer reads to reference genomes[3]

[1] https://en.wikipedia.org/wiki/SAM_(file_format) (last visited on 15.02.2018, 12:20)
[2] https://en.wikipedia.org/wiki/Binary_Alignment_Map (last visited on 15.02.2018, 12:20)
[3] http://www.bioinf.uni-leipzig.de/Software/segemehl/ (last visited on 15.02.2018, 12:20)

The goal of this work

Analyzing the data generated by the mapping process so that
the quality of the data can be improved further in the second clean-up step.



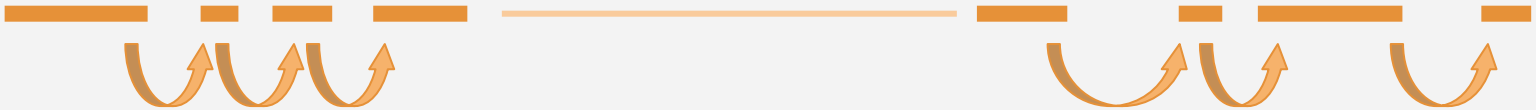
What does segemehlQC do differently?

Providing **one single tool** which covers a set of **visualizations specific to segemehl's strengths**:

- different mapping operations (matches, insertions, deletions, …) are visualised by proportion

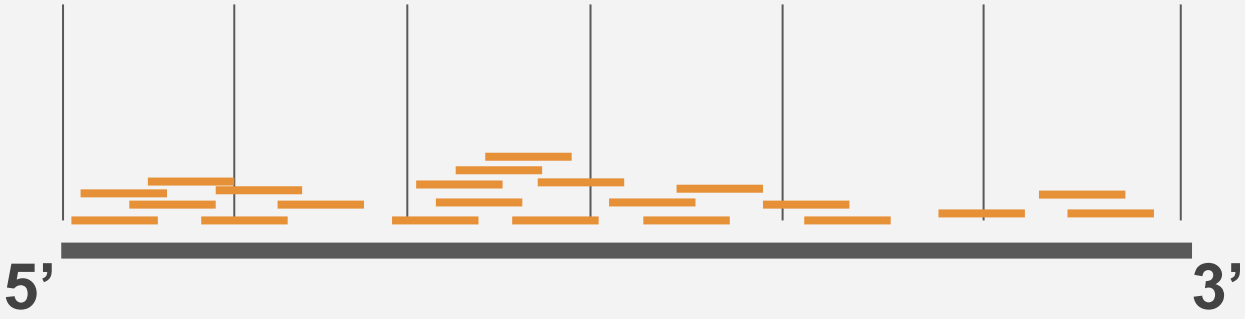
- gap length analysis of paired-end reads (how far are read pairs apart)

- split reads are reassembled to analyse them



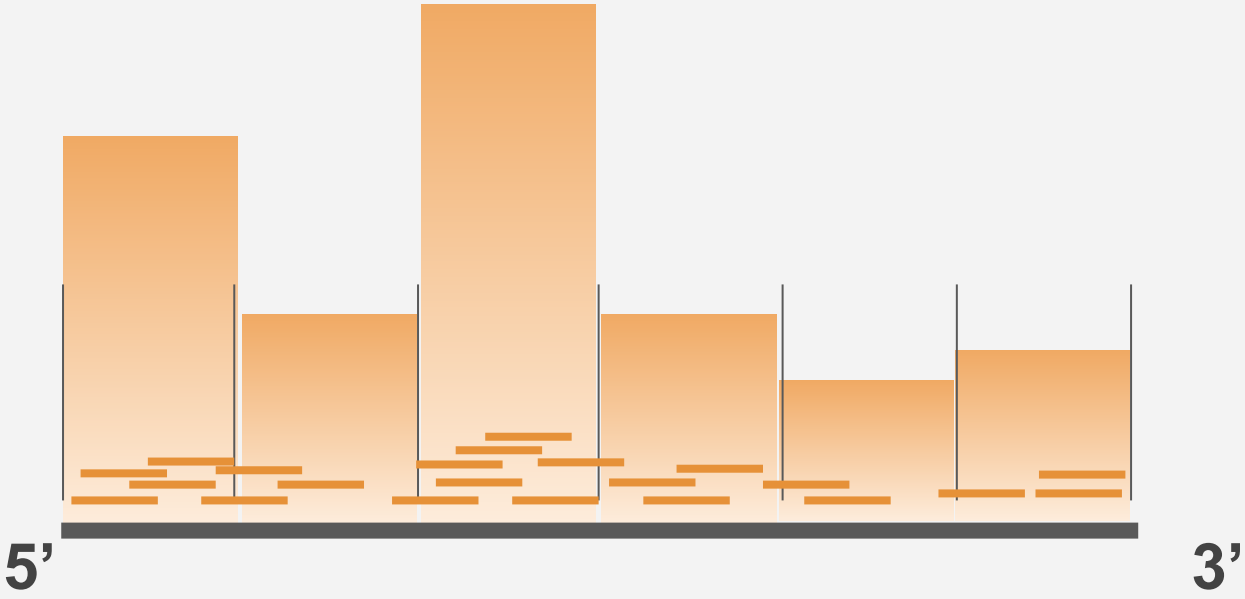
The read coverage plot

Divide the chromosome into bins and add the reads at their respective spots.



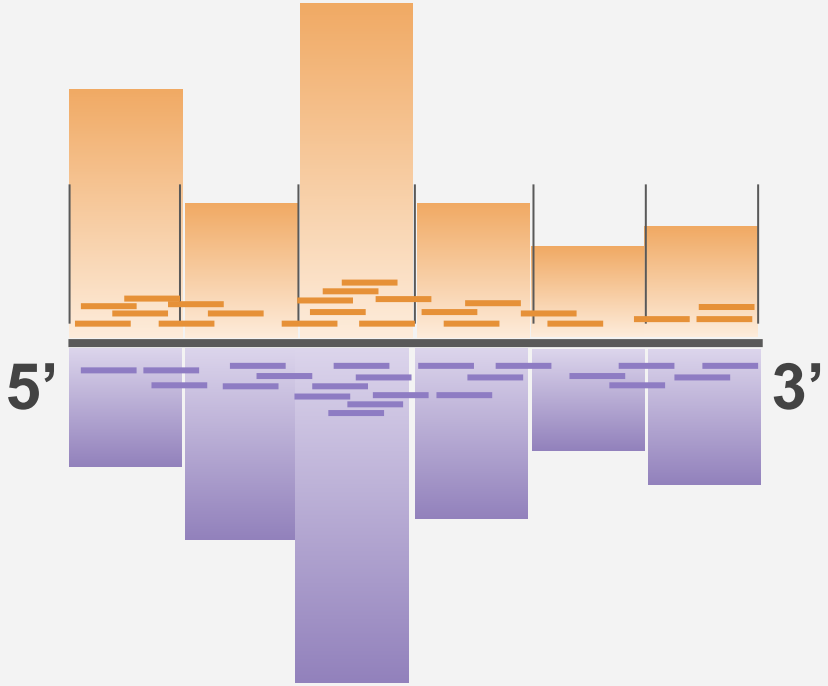
The read coverage plot

The higher the coverage of reads, the higher the bar.



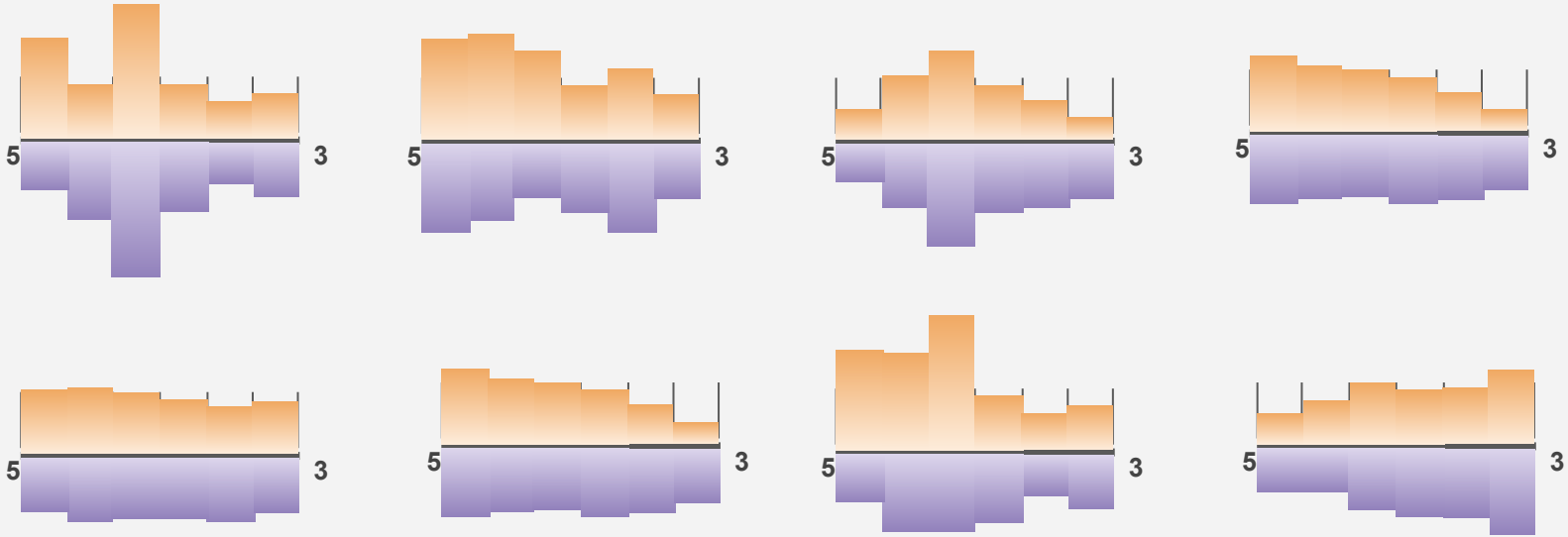
The read coverage plot

Also do this for the reads mapped to the negative strand.



The read coverage plot

Repeat for every chromosome.



The per base CIGAR operation summary

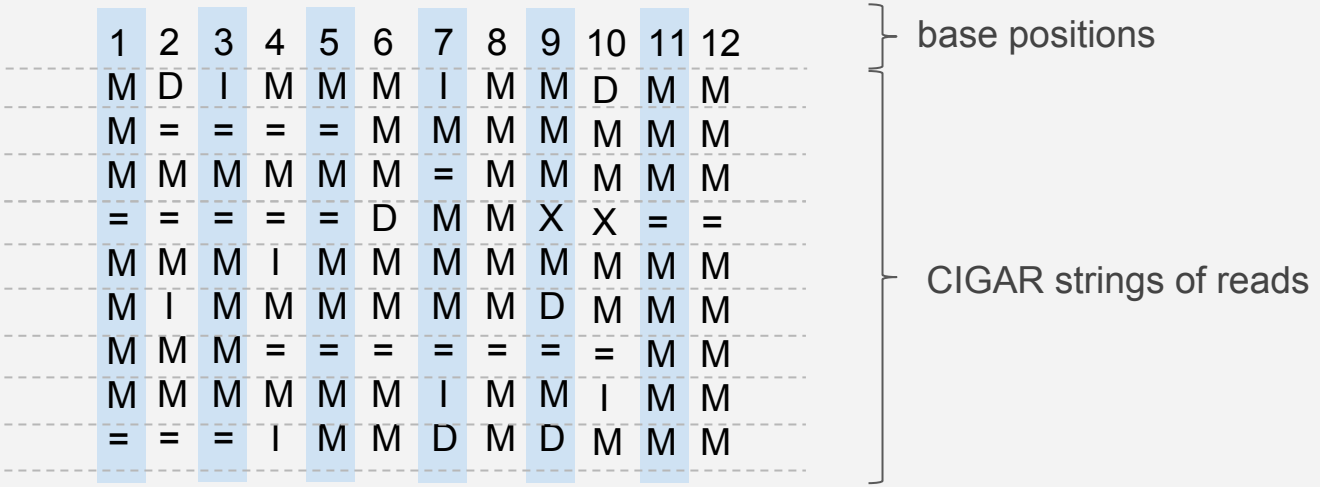
M	Alignment match (can be a sequence match or mismatch)
I	Insertion to the reference
D	Deletion from the reference
N	Skipped region from the reference
S	Soft clipping (clipped sequences present in SEQ)
H	Hard clipping (clipped sequences NOT present in SEQ)
D	Skipped region from the reference
P	Padding (silent deletion from padded reference)
=	Sequence match
X	Sequence mismatch

Inspired by The SAM/BAM Format Specification Working Group (**2017**). Sequence Alignment/Map Format Specification. p. 6

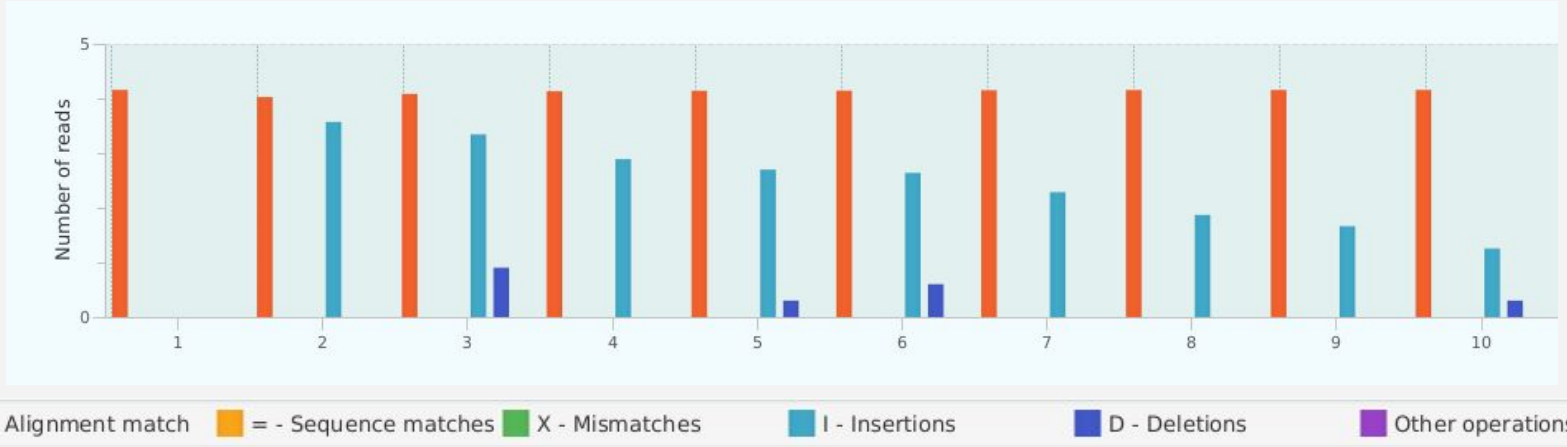
The per base CIGAR operation summary

Normal CIGAR string: 3M1I3M1D5M

Extended CIGAR string: MMMIMMMDMMMMM



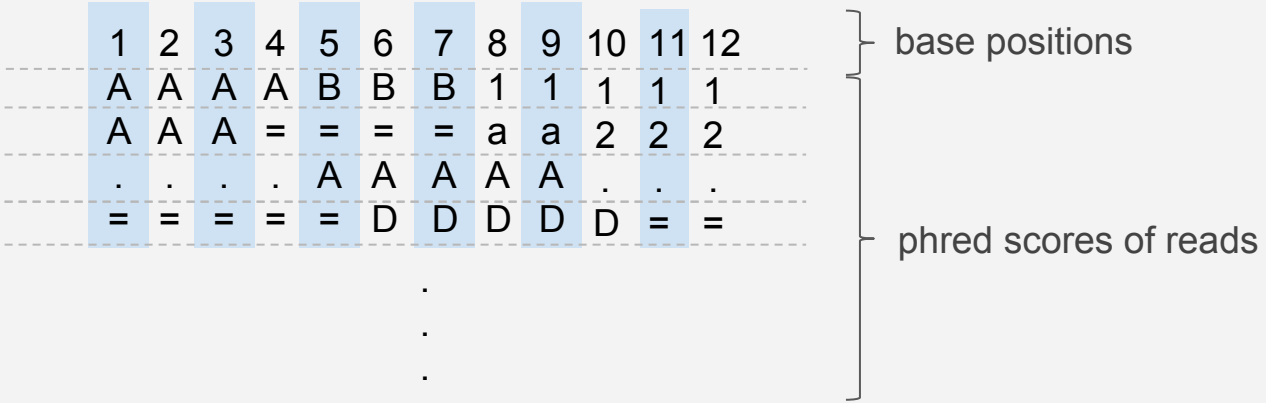
The per base CIGAR operation summary



The per base quality plot

Quality encoding:

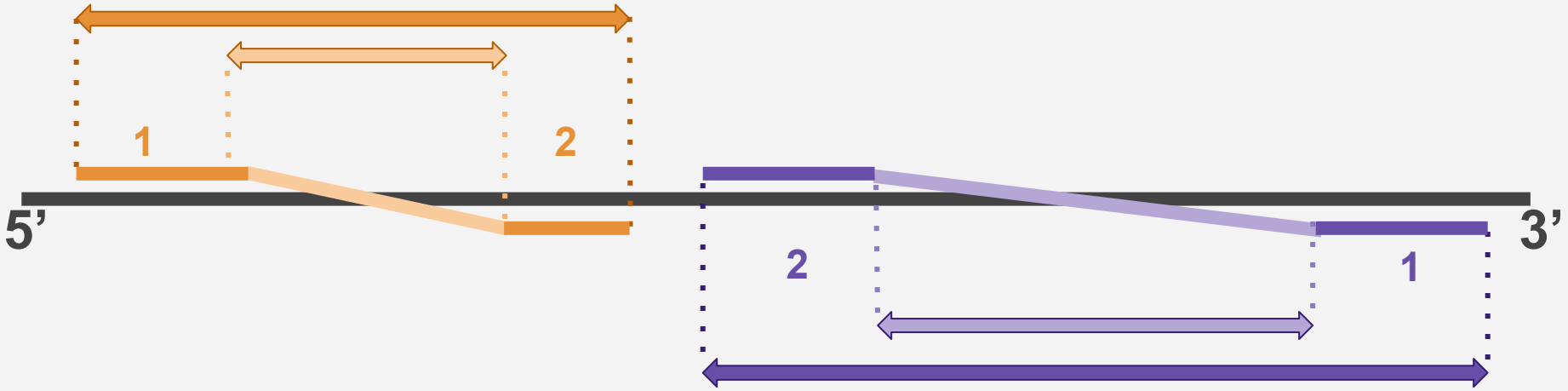
!"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~



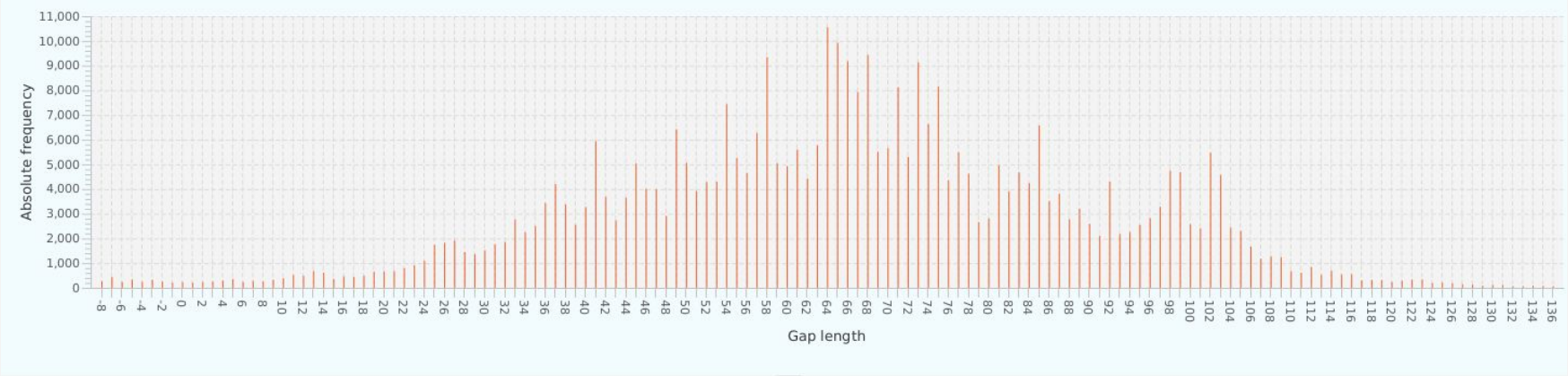
The per base quality plot

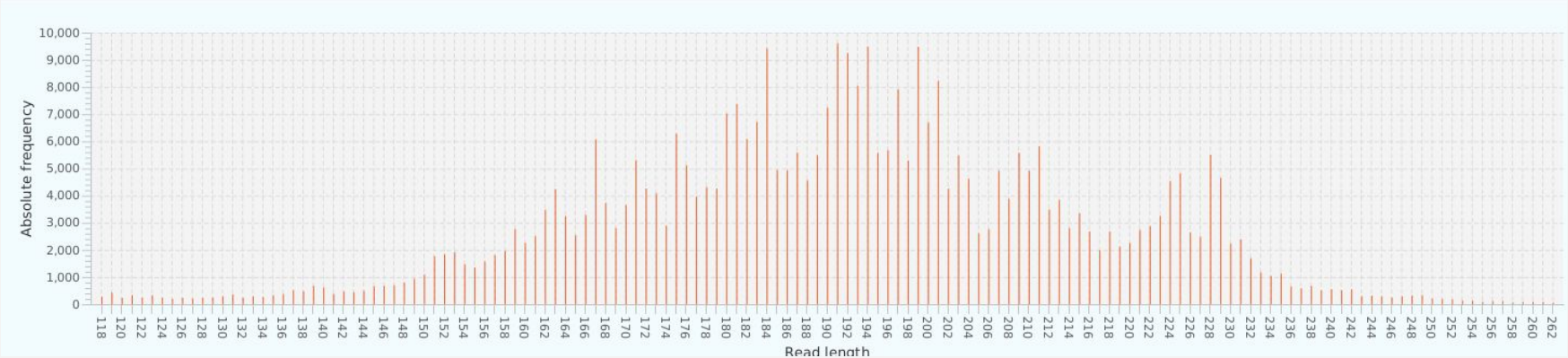
The gap and read length distribution plots

1. measure lengths for both gaps and whole reads
2. count the frequency for both separately



The gap and read length distribution plots



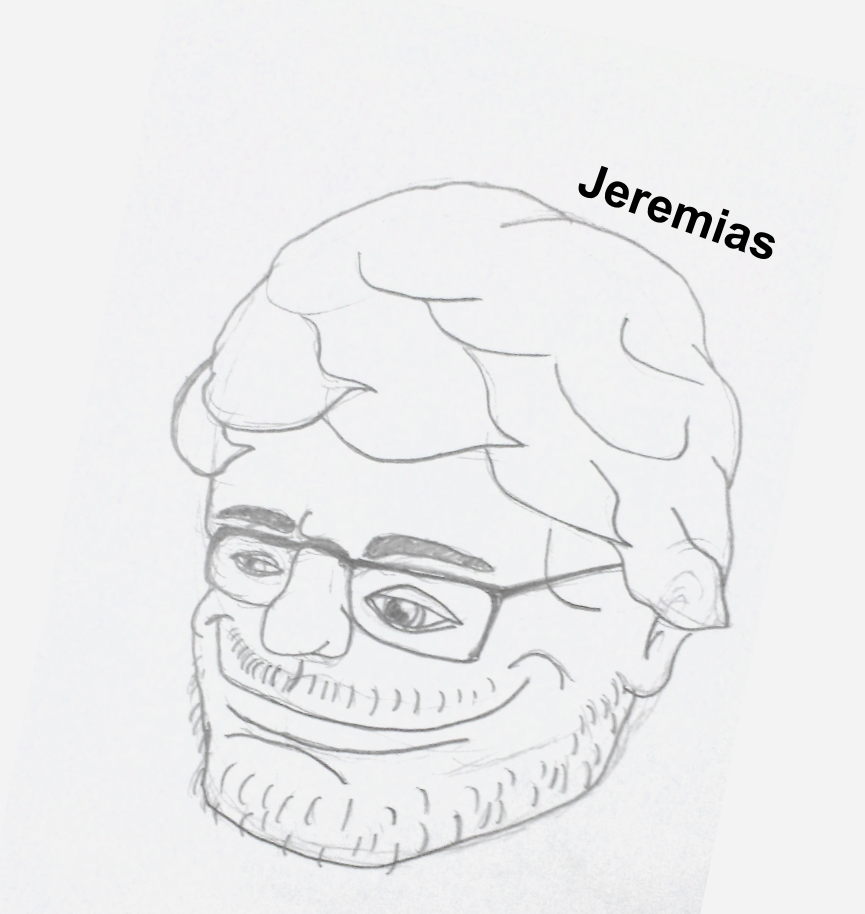


Next steps

- Split read statistics - how are splits distributed within in the dataset?
- DNA-methylation analysis
- Giving some thanks to some people

Some thanks

Some thanks



Some thanks