Link to my analysis: https://cgc.sbgenomics.com/u/zoranastaka/gi-druga-vezbs/analysis/cruncher/secondary\_dna\_analysis/

Explanation:

In this exercise, we used interactive analysis to get information about DNK. As an input file, we used the .bam file, which is the binary version of a SAM file. The SAM file

is a file that contains sequence alignment data. The BAM file needs to be sorted and indexed.

When I tried to fetch data from .bam file, without .bam.bai file there was an error. After I've included the .bam.bai file the error was not showing anymore.

How I understood this analysis: This is Data Analysis that was shown to us during the first lecture (https://www.youtube.com/watch?v=womKfikWlxM since 4:23).

The reads are the object of the AlignedSegment class. Reads represent information about the sequence of base pairs (A, T, C, G). First read in the analysis I've conducted had a length of 76.

Other information regarding read is: is it paired, is it proper pair, if the mate is reversed, is it duplicate, and many other pieces of information.

The compact way of getting some information is the flag. Information that is contained in the flag:

* read paired (0x1)
* read mapped in proper pair (0x2)
* read unmapped (0x4)
* mate unmapped (0x8)
* read reverse strand (0x10)
* mate reverse strand (0x20)
* first in pair (0x40)
* second in pair (0x80)
* not primary alignment (0x100)
* read fails platform/vendor quality checks (0x200)
* read is PCR or optical duplicate (0x400)
* supplementary alignment (0x800)

Flag works like a binary number length 12, where 1 or 0 on a specific position is giving information about the particular property of reading. The function that does mapping from the number (max 4095) to set of properties is a bijection, meaning one number corresponds to a unique set of properties and vice versa.

Flag property in the first read was 118710 = 0100101000112. This means read has these properties:

* read paired (0x1)
* read mapped in proper pair (0x2)
* mate reverse strand (0x20)
* second in pair (0x80)
* read is PCR or optical duplicate (0x400)

Mapped reads refer to those reads from the sequenced sample that align directly to a single region on the reference genome. Unmapped reads are those that map nowhere on the reference genome. In the analysis, there were 17 765 unmapped reads of 2 921 629. This means there is approximately 0.61% of unmapped reads. The percentage of mapped reads, in our case 99.39%, is a global indicator of overall sequencing accuracy.

Mapping quality is the measure of the confidence that a read comes from the position it is aligned to by a mapping algorithm.