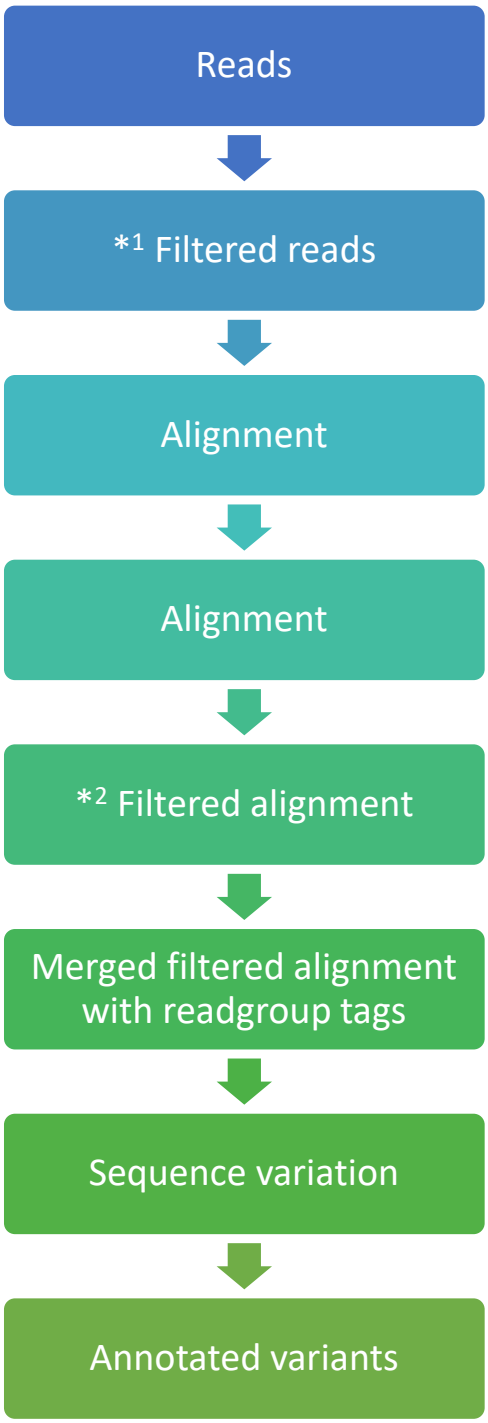
Student: _____

PRACTICAL – EXIT TICKET

We applied this workflow in the practical. Fill it with the format file and software used in each step:



FORMAT: Fastq

SOFTWARE: Galaxy/fastqc

FORMAT: Fastq

SOFTWARE: bwa/fastqc trim

FORMAT: SAM

SOFTWARE: SAMtools

FORMAT: BAM

SOFTWARE: BAMtools

FORMAT: BAM

SOFTWARE: Picard

FORMAT: BAM

SOFTWARE: Galaxy

FORMAT: VCF

SOFTWARE: Variant Annotator

FORMAT: VCF

Why did you detect SNPs within each of the two samples (within mum and within daughter)? Why the two samples contain a different set of SNPs?

We detected SNPs in the 2 samples in case heteroplasmy happened.

And because since the daughter was born her mitochondrial DNA(and the mother's) might have mutated.

The 2 samples might contain different sets of SNPs because of the reasons listed above (heteroplasmy and mutations since birth), furthermore,

every individual has its unique set of variants that form its genetic profile (inheritance, mutations, etc)

Why initial data for each of the two samples come in the form of two separate files (two for the mum and two for the daughter)? What is the nature of the initial data?

Because there is a file for the forward strand and a file for the reverse strand for each sample, due to using paired-end sequencing.

This format is used for better data integrity and more flexibility in the data analysis.

The nature of the initial data is a fastq format, which presents sequencing reads with their associated quality scores for each base.

Do initial FASTQ files encode quality scores? If so, how are they encoded?

Fastq files do encode quality scores for each sequenced base, this quality score represents the confidence level of each base read.

The quality scores are encoded using ascii characters, with one character per score, ecoded using the following formula:

Quality score = ascii character value - 33

There are two steps in the workflow where we filter the data (see *1 and *2). What is being filtered in each case and why?

*1

We filter the sequence data straight from the fastq file, this is to remove low-quality reads and wrongly mapped pairs

We filter the data in this step to ensure that the errors do not carry over the next processes, because it would cause a snowball effect

*2

After aligning the reads to the reference genome we filter to remove reads that have been wrongly aligned.

We filter the data in this step to ensure that the variant annotations that come in the later steps use good quality reads, so that there is a smaller

error margin in the results.