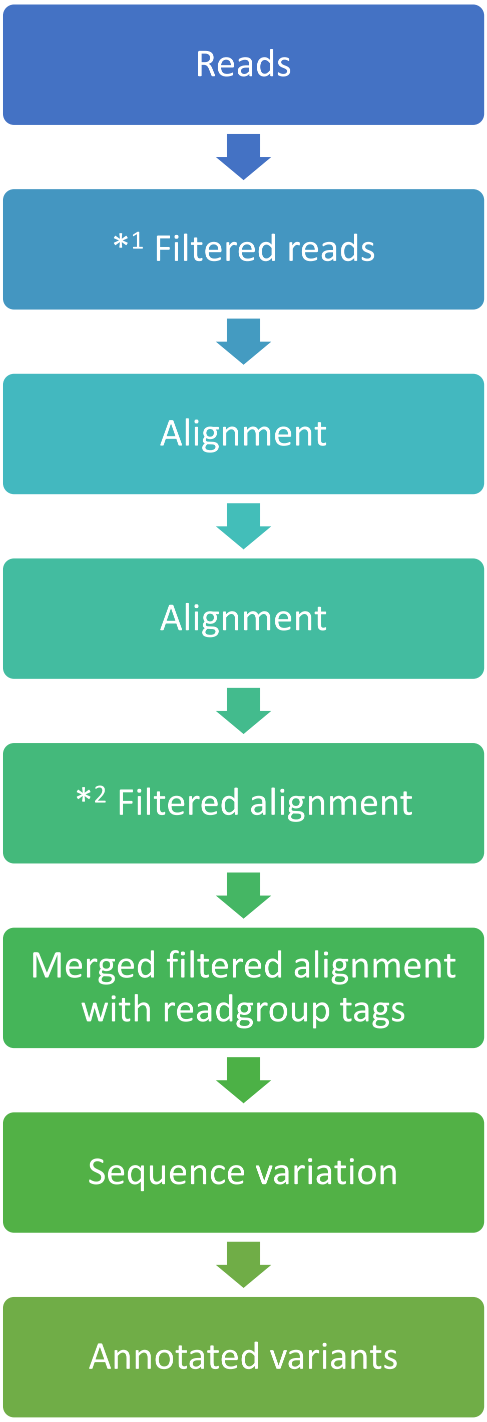
Student: Martí Díez Macià

PRACTICAL – EXIT TICKET

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



Reads

FORMAT: FASTQ

SOFTWARE: Galaxy

*1 Filtered reads

FORMAT: FASTQ

SOFTWARE: bwa

Alignment

FORMAT: SAM

SOFTWARE: SAMtools

Alignment

FORMAT: BAM

SOFTWARE: BAMtools

*2 Filtered alignment

FORMAT: BAM

SOFTWARE: Picard

Merged filtered alignment with readgroup tags

FORMAT: BAM

SOFTWARE: Galaxy

Sequence variation

FORMAT: VCF

SOFTWARE: Variant annotator

Annotated variants

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 can expect SNPs since every individual has its unique set of variants/mutations that build their genetic profile

Although the mum and daughter share a lot of their genetic material, the daughter also has genes from the father and new

mutations may have occured. The samples contain different SNPs due to inheritance, mutations or genetic recombination

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?

The initial data come in the form of 2 files because the sequencing method we use is paired-end sequencing. With this

method, we sequence both ends of the DNA fragment. This technique is useful since it improves alignment to the reference

genome and thus provides more information. The files represent the forward and reverse reads of the fragments. The nature

of the data is raw sequence data, since they are in FASTQ format

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

Yes, initial FASTQ files do encode quality scores. They are encoded as ASCII characters, each character representing

a quality score, ranging from 0 to 93. It can be calculated by subtracting 33 from the ASCII value of the caracter.

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

In this step, we are filtering the raw sequence data. Our goal is to remove low-quality reads that may have been produced

during the sequencing process (for example, high proportion of unknown N nucleotides, low-quality score reads or too short reads)

We must filter the data in this step to reduce the likelihood of wrong base calls, which could lead to a wrong downstream analysis

*2

When we have aligned the reads to the reference genome, we filter again by removing duplicate reads, discarding reads that

align to multiple loci in the reference genome and filtering out reads that have a low mapping quality score. This filtering step

reassures that only high-quality unique alignments are used in the variant calling process coming next.