Omics Techniques   **Topics 1&2. DNA-seq techniques and applications**
Bachelor's Degree in Bioinformatics   **Practical – Exit ticket**
Jan Carreras Boada 107765

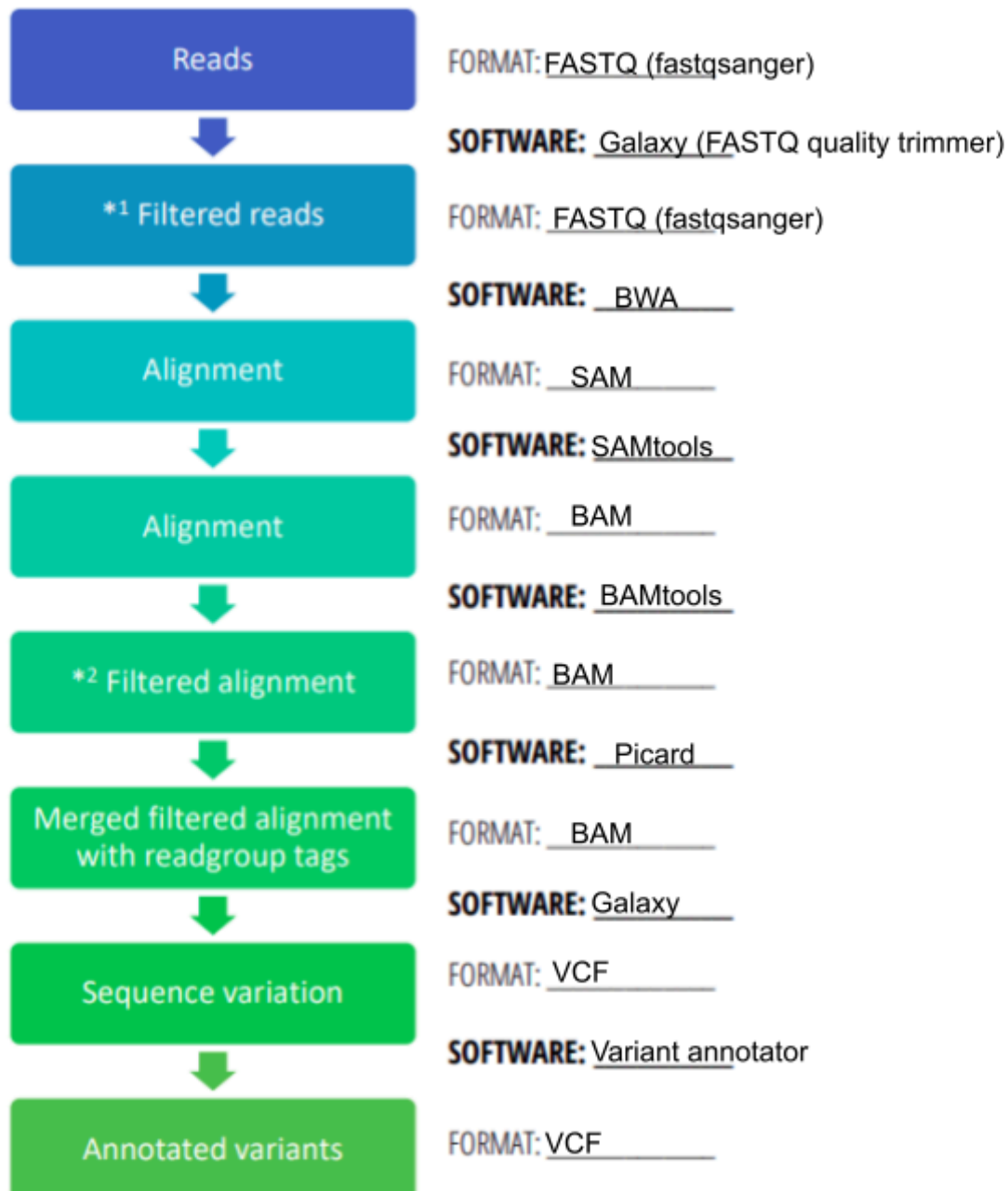PRACTICAL EXIT TICKET

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

Reads — FORMAT: FASTQ (fastqsanger)

SOFTWARE: Galaxy (FASTQ quality trimmer)

*1 Filtered reads — FORMAT: FASTQ (fastqsanger)

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?

The two samples (mother and daughter) have SNPs that we found due to mitochondrial inheritance and heteroplasmy.
The term "heteroplasmy" describes the coexistence of many organellar genome types, such as mitochondrial DNA, in a single person. The identification of unique SNPs in every sample may be impacted by this circumstance.
The exclusive transfer of mitochondrial DNA from the mother to the daughter is known as mitochondrial inheritance.
We can say that there exist many reasons why the two samples have different sets of SNPs.

First, only a portion of SNPs, which can differ from those in the mother, may be transmitted down to the progeny as a result of the bottleneck effect during transmission.

Also, de novo mutations, which are spontaneous mutations that happen during DNA replication and cell division, also contribute to the diversity of SNPs seen in the mother and daughter samples.

Last, differences in SNPs between the samples may potentially be explained by genetic drift.

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?

Illumina-generated paired-end RNA-seq reads make up the first set of data. We keep two distinct files for each sample, one for the forward read and one for the reverse read. These data are in the FASTQ format, which offers information on each read's quality in addition to sequence information.

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

Yes, quality scores are encoded in the first FASTQ files.The Phred quality score system, which assigns a unique quality score to each base, is the basis for the quality scores seen in FASTQ files. ASCII characters are used to represent these scores.

ASCII_BASE=33 Illumina, Ion Torrent, PacBio and Sanger

Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII
0	1.00000	33 !	11	0.07943	44 ,	22	0.00631	55 7	33	0.00050	66 B
1	0.79433	34 "	12	0.06310	45 -	23	0.00501	56 8	34	0.00040	67 C
2	0.63096	35 #	13	0.05012	46 .	24	0.00398	57 9	35	0.00032	68 D
3	0.50119	36 $	14	0.03981	47 /	25	0.00316	58 :	36	0.00025	69 E
4	0.39811	37 %	15	0.03162	48 0	26	0.00251	59 ;	37	0.00020	70 F
5	0.31623	38 &	16	0.02512	49 1	27	0.00200	60 <	38	0.00016	71 G
6	0.25119	39 '	17	0.01995	50 2	28	0.00158	61 =	39	0.00013	72 H
7	0.19953	40 (	18	0.01585	51 3	29	0.00126	62 >	40	0.00010	73 I
8	0.15849	41)	19	0.01259	52 4	30	0.00100	63 ?	41	0.00008	74 J
9	0.12589	42 *	20	0.01000	53 5	31	0.00079	64 @	42	0.00006	75 K
10	0.10000	43 +	21	0.00794	54 6	32	0.00063	65 A			

ASCII_BASE=64 Old Illumina

Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII
0	1.00000	64 @	11	0.07943	75 K	22	0.00631	86 V	33	0.00050	97 a
1	0.79433	65 A	12	0.06310	76 L	23	0.00501	87 W	34	0.00040	98 b
2	0.63096	66 B	13	0.05012	77 M	24	0.00398	88 X	35	0.00032	99 c
3	0.50119	67 C	14	0.03981	78 N	25	0.00316	89 Y	36	0.00025	100 d
4	0.39811	68 D	15	0.03162	79 O	26	0.00251	90 Z	37	0.00020	101 e
5	0.31623	69 E	16	0.02512	80 P	27	0.00200	91 [	38	0.00016	102 f
6	0.25119	70 F	17	0.01995	81 Q	28	0.00158	92 \	39	0.00013	103 g
7	0.19953	71 G	18	0.01585	82 R	29	0.00126	93]	40	0.00010	104 h
8	0.15849	72 H	19	0.01259	83 S	30	0.00100	94 ^	41	0.00008	105 i
9	0.12589	73 I	20	0.01000	84 T	31	0.00079	95 _	42	0.00006	106 j
10	0.10000	74 J	21	0.00794	85 U	32	0.00063	96 ~			

Omics Techniques **Topics 1&2. DNA-seq techniques and applications**
Bachelor's Degree in Bioinformatics **Practical – Exit ticket**
Jan Carreras Boada 107765

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 used a filtering method in the second stage because the quality of the reads significantly decreased toward the ends. In order to save just the high-quality reads for further examination, we therefore removed the low-quality bases at the 3' ends. This filtering process is essential because it ensures that the subsequent analyses are performed on accurate and reliable data, which improves the overall caliber of our results.

*2
During the fifth stage, we noticed that certain pairs of readings were not correctly mapping. We had to improve our alignment as a consequence of this. We used BAMTools for this purpose. We only kept the pairs that had correctly mapped during the filtering phase. Only reliable and correctly aligned readings will be employed in subsequent studies thanks to this critical step in accurate variation recognition.