GULBENKIAN CIÊNCIA

Data formats

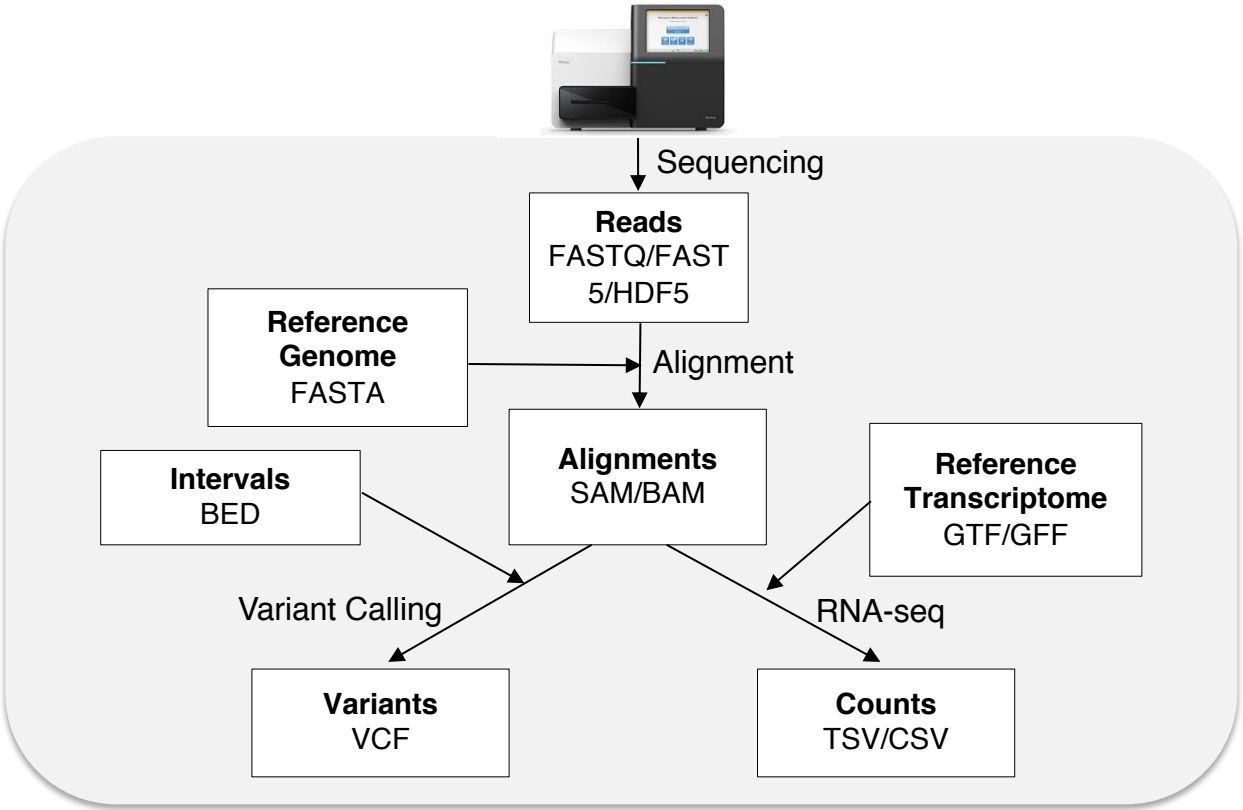
Precision Oncology Course

cnio SPANISH NATIONAL CANCER RESEARCH CENTRE

BU CNIO **BIOINFORMATICS** UNIT

Coral Fustero Torre
Bioinformatics Unit,
Structural Biology Programme.
cfustero@cnio.es | bioinformatics.cnio.es

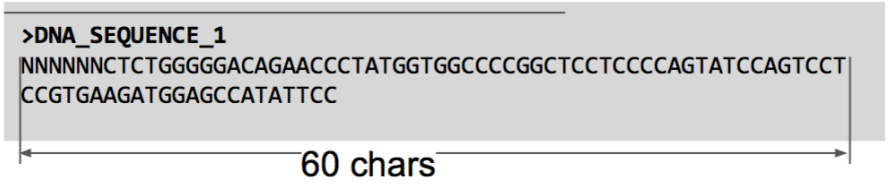
Formats outline



Data formats

Reference Genome: FASTA format

- Typical extensions: .fasta, .fas, .fa, .fna, .fsa
- Each sequence is composed by at least two consecutive lines:
  - ">" Sequence name and optional description (space separated)
  - Line(s) with the whole sequence




```
>DNA_SEQUENCE_1
NNNNNNCTCTGGGGGACAGAACCCTATGGTGGCCCCGGCTCCTCCCCAGTATCCAGTCCT
CCGTGAAGATGGAGCCATATTCC
>DNA_SEQUENCE_2
GGGGGACAGAACCCTATGGTGGCCCCGGCTCCTCCCCAGTATCCAGTCCT
>DNA_SEQUENCE_3
CTCCTCCCCAGTATCCAGTCCTGGGGGACAGAACCCTATGGTGGCCCCGCCCAGTATCCA
```

We can have multiple sequences in the same file: **multifasta**

Data formats

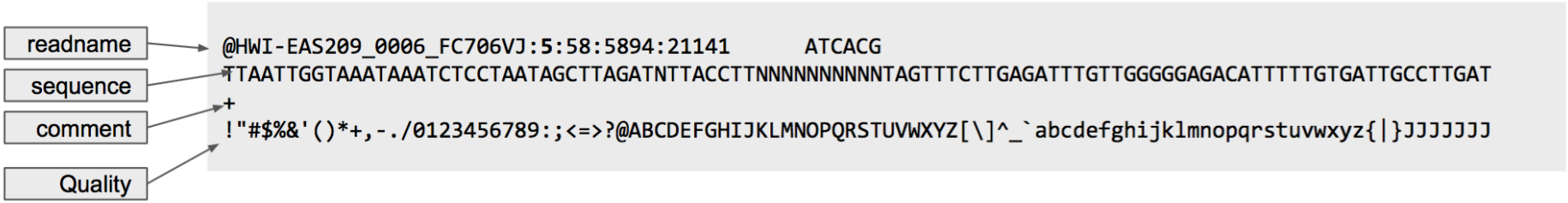
Nucleotides codes: IUPAC

IUPAC nucleotide code	Base
A	Adenine
C	Cytosine
G	Guanine
T (or U)	Thymine (or Uracil)
R	A or G
Y	C or T
S	G or C
W	A or T
K	G or T
M	A or C
B	C or G or T
D	A or G or T
H	A or C or T
V	A or C or G
N	any base
. or -	gap

Data formats

Reads: FASTQ

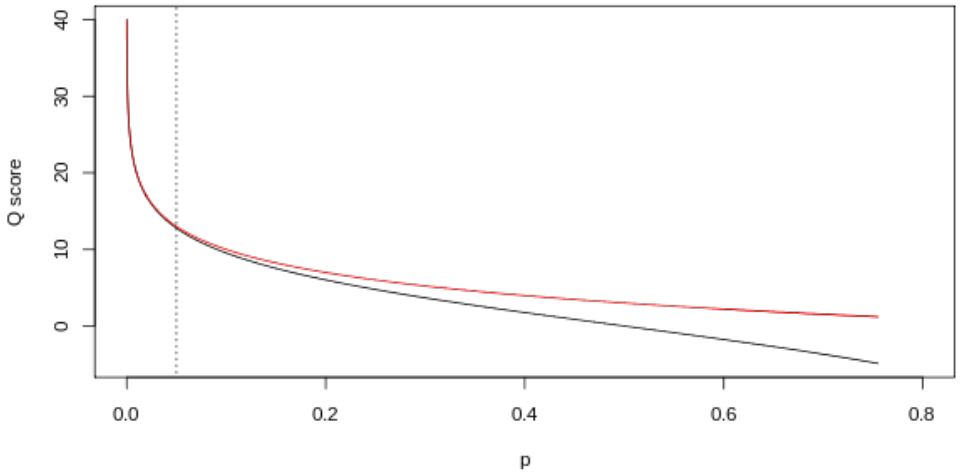
- Typical extensions: .fq, .fastq
- Each read is composed by 4 lines:
  - "@" Read name and optional description (space separated)
  - Sequence
  - "+" (optionally: repeat the read name)
  - Base Quality Score



Data formats

Reads: FASTQ – Quality Score

- Phred quality scores *QPhred* are defined as a property which is logarithmically related to the base-calling **error** probabilities *p*
  QPhred = -10 log10(p)
- The score is written as the character whose ASCII code is: QPhred + 33
- The higher the the *QPhred* , the lower the probability that the base calling is erroneous



Data formats

Reads: FASTQ – ASCII code

Dec	Hx	Oct	Char	Dec	Hx	Oct	Html	Chr	Dec	Hx	Oct	Html	Chr	Dec	Hx	Oct	Html	Chr
0	0	000	NUL (null)	32	20	040		Space	64	40	100	@	@	96	60	140	`	`
1	1	001	SOH (start of heading)	33	21	041	!	!	65	41	101	A	A	97	61	141	a	a
2	2	002	STX (start of text)	34	22	042	"	"	66	42	102	B	B	98	62	142	b	b
3	3	003	ETX (end of text)	35	23	043	#	#	67	43	103	C	C	99	63	143	c	c
4	4	004	EOT (end of transmission)	36	24	044	$	$	68	44	104	D	D	100	64	144	d	d
5	5	005	ENQ (enquiry)	37	25	045	%	%	69	45	105	E	E	101	65	145	e	e
6	6	006	ACK (acknowledge)	38	26	046	&	&	70	46	106	F	F	102	66	146	f	f
7	7	007	BEL (bell)	39	27	047	'	'	71	47	107	G	G	103	67	147	g	g
8	8	010	BS (backspace)	40	28	050	(	(	72	48	110	H	H	104	68	150	h	h
9	9	011	TAB (horizontal tab)	41	29	051	)	)	73	49	111	I	I	105	69	151	i	i
10	A	012	LF (NL line feed, new line)	42	2A	052	*	*	74	4A	112	J	J	106	6A	152	j	j
11	B	013	VT (vertical tab)	43	2B	053	+	+	75	4B	113	K	K	107	6B	153	k	k
12	C	014	FF (NP form feed, new page)	44	2C	054	,	,	76	4C	114	L	L	108	6C	154	l	l
13	D	015	CR (carriage return)	45	2D	055	-	-	77	4D	115	M	M	109	6D	155	m	m
14	E	016	SO (shift out)	46	2E	056	.	.	78	4E	116	N	N	110	6E	156	n	n
15	F	017	SI (shift in)	47	2F	057	/	/	79	4F	117	O	O	111	6F	157	o	o
16	10	020	DLE (data link escape)	48	30	060	0	0	80	50	120	P	P	112	70	160	p	p
17	11	021	DC1 (device control 1)	49	31	061	1	1	81	51	121	Q	Q	113	71	161	q	q
18	12	022	DC2 (device control 2)	50	32	062	2	2	82	52	122	R	R	114	72	162	r	r
19	13	023	DC3 (device control 3)	51	33	063	3	3	83	53	123	S	S	115	73	163	s	s
20	14	024	DC4 (device control 4)	52	34	064	4	4	84	54	124	T	T	116	74	164	t	t
21	15	025	NAK (negative acknowledge)	53	35	065	5	5	85	55	125	U	U	117	75	165	u	u
22	16	026	SYN (synchronous idle)	54	36	066	6	6	86	56	126	V	V	118	76	166	v	v
23	17	027	ETB (end of trans. block)	55	37	067	7	7	87	57	127	W	W	119	77	167	w	w
24	18	030	CAN (cancel)	56	38	070	8	8	88	58	130	X	X	120	78	170	x	x
25	19	031	EM (end of medium)	57	39	071	9	9	89	59	131	Y	Y	121	79	171	y	y
26	1A	032	SUB (substitute)	58	3A	072	:	:	90	5A	132	Z	Z	122	7A	172	z	z
27	1B	033	ESC (escape)	59	3B	073	;	;	91	5B	133	[	[	123	7B	173	{	{
28	1C	034	FS (file separator)	60	3C	074	<	<	92	5C	134	\	\	124	7C	174			|
29	1D	035	GS (group separator)	61	3D	075	=	=	93	5D	135	]	]	125	7D	175	}	}
30	1E	036	RS (record separator)	62	3E	076	>	>	94	5E	136	^	^	126	7E	176	~	~
31	1F	037	US (unit separator)	63	3F	077	?	?	95	5F	137	_	_	127	7F	177	&#127;	DEL

Source: www.LookupTables.com

Data formats

Reads: FASTQ – Single-end or Paired-end?

One unique sample can have 1 or 2 files:

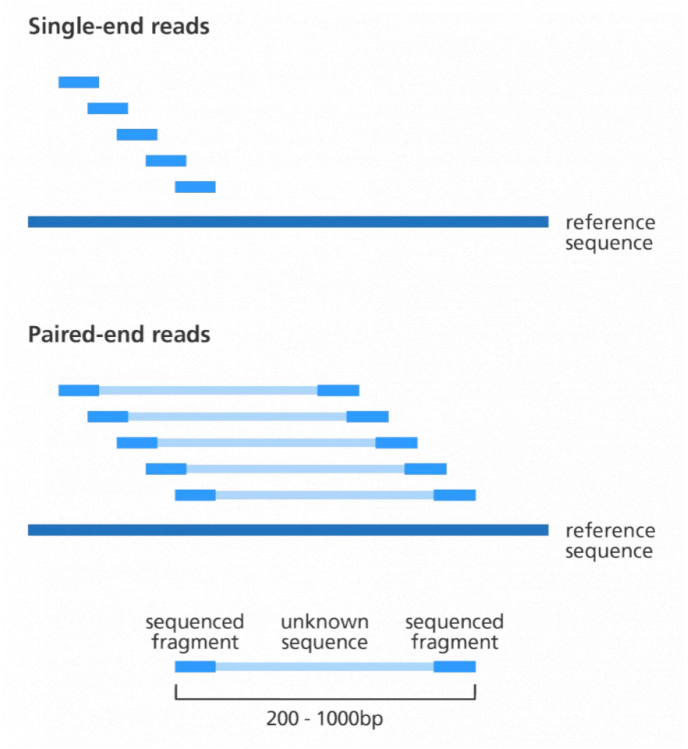
- If **single-end** seq → 1 file
  filename: *.fastq
- If **paired-end** seq → 2 files
  filenames: *_R1.fastq AND *_R2.fastq

_R1

```
@SN1083:379:H8VA1ADXX:2:1101:1248:2144 1:N:0:12
CCTAAATGGTGCCATGCTAGGAGGCCGTGCCCTTCTTGAAAAGTTGTATGTGAA
+
BBBFFFFFFBFFFIIIIFI<FFIIIIIFIIIIFBFIIIIIIIIFFFIIIIFIII
```

_R2

```
@SN1083:379:H8VA1ADXX:2:1101:1248:2144 2:N:0:12
CATTTTCGACGTTGTTAATAAGCTCTGCGTACTTGCAAGCTATCTGCGCGAACG
+
BBBFFFFFFFFFFFIIIIIIIIIIIIIIIIIFIIIIIIIIIIIIIIIIIIIFFF
```



Exercise: FastQC

1. Download data: https://fundacioncnio-my.sharepoint.com/:u:/g/personal/cfustero_cnio_es/Eb1S_Y-D8SpBspZK3MFZQiUBcbncd8huI6_c0YPOihnzwQ?e=AQeHGu

2. Go to the terminal and open the file

```
$ more WEx_Normal_R1.fastq
```

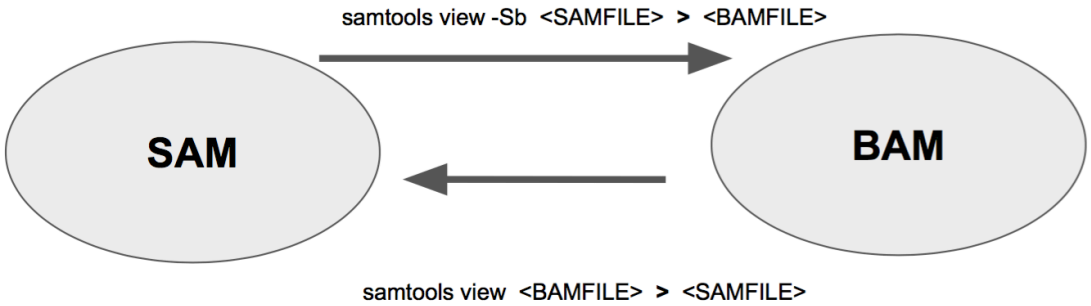
3. How would you detect the number of sequences present in the file?

Useful commands
`grep`: searches plain-text data sets for lines that match a regular expression
`wc`: counts words
`wc -l`: counts lines
`|`: lets you send the output of one command to another.

Data formats

Aligment: SAM or BAM

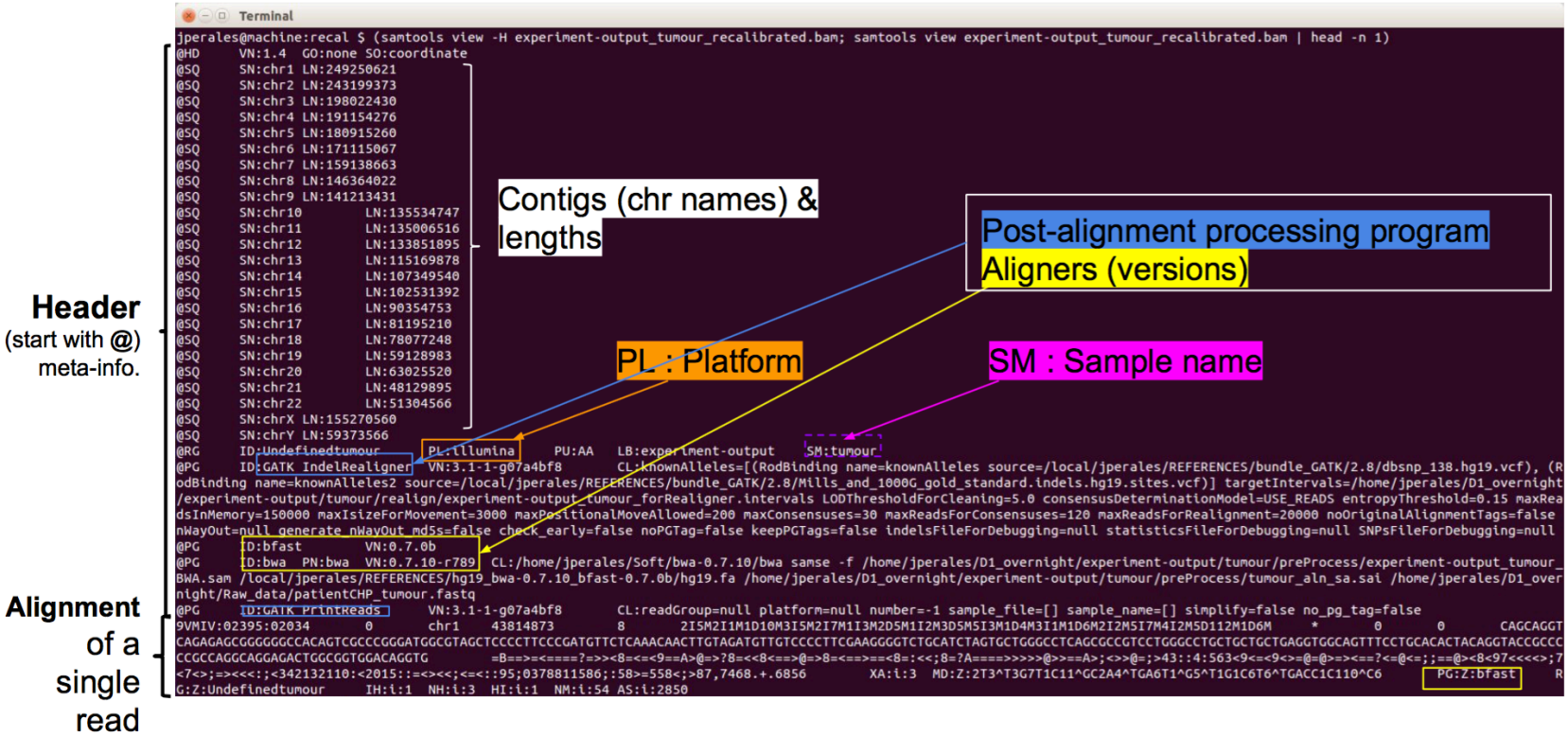
- **SAM** is the human readable text format (.sam extension)
- **BAM** is the binary, machine efficient format (.bam extension)
- Both contains exactly the same information and are interconvertible (samtools)



File specifications https://samtools.github.io/hts-specs/SAMv1.pdf

Data formats

Aligment: SAM or BAM - Header



Data formats

Aligment: SAM or BAM - Aligments

#1	#2	#3	#4	#5	#6	#7	#8	#9	#10	#11	#12
ReadName	99	chr10	2	30	3MD2M1I1M	=	14	20	CATCTG	jjjjjjj	z:Aligner

If single-end:

7. reference sequence name of the alignment of the next read in sequence
8. position in the alignment of the next read in sequence
9. number of bases covered by reads from the same fragment. Plus/minus means the current read is the leftmost/rightmost read

Data formats

Aligment: SAM or BAM - Aligments

```
#1          #2    #3      #4   #5    #6                #7    #8    #9    #10        #11          #12
ReadName    99    chr10   2    30    3MD2M1I1M         =     14    20    CATCTG     jjjjjjj      z:Aligner
```

FIELDS

#Col	Field	Description
1.	QNAME	read name
2.	FLAG	bitwise FLAG* (unmapped, pair unmapped, properly mapped, …)
3.	RNAME	Reference sequence name (e.g. chr1).
4.	POS	1-based leftmost position.
5.	MAPQ	Mapping Quality (Phred-scaled). Scale 0 to 255.
6.	CIGAR	extended CIGAR string.
7.	MRNM	Paired-end: Mate Reference sequence Name (= if same as RNAME).
8.	MPOS	Paired-end: 1-based Mate position.
9.	TLEN	Paired-end: Insert size
10.	SEQ	Read sequence
11.	QUAL	Base Quality Score from the Read sequence.
12.	OPT	Optional Tags.

SAM FLAGS

Decimal	Description of read
1	Read paired
2	Read mapped in proper pair
4	Read unmapped
8	Mate unmapped
16	Read reverse strand
32	Mate reverse strand
64	First in pair
128	Second in pair
256	Not primary alignment
512	Read fails platform/vendor quality checks
1024	Read is PCR or optical duplicate
2048	Supplementary alignment

One of the reads is unmapped:
73, 133, 89, 121, 165, 181, 101, 117, 153, 185, 69, 137

Both reads are unmapped:
77, 141

Mapped within the insert size and in correct orientation:
99, 147, 83, 163

Mapped within the insert size but in wrong orientation:
67, 131, 115, 179

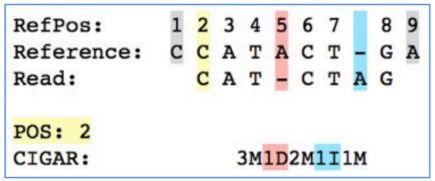
Mapped uniquely, but with wrong insert size:
81, 161, 97, 145, 65, 129, 113, 177

Data formats

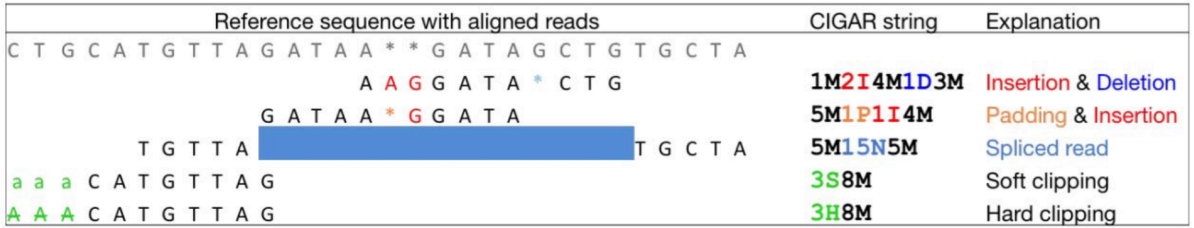
Aligment: SAM or BAM - CIGAR

Concise Idiosyncratic Gapped Alignment Report

- Compressed representation of an alignment
- **Format**: A CIGAR string is made up of <integer><op> pairs
- Where "op" is an operation specified as a single character, usually an upper-case letter (see table)



Op	Description
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)
P	padding (silent deletion from padded reference)
=	sequence match
X	sequence mismatch



Data formats

Intervals: BED

- Typical extension: .bed
- The first three are required BED fields, the rest are optional

1. **chrom** - The name of the chromosome (e.g. chr3, chrY, chr2)
2. **chromStart** - The starting position of the feature in the chromosome
   0-based (The first base in a chromosome is numbered 0)
3. **chromEnd** - The ending position of the feature in the chromosome or scaffold

Additionally, 9 optional fields:

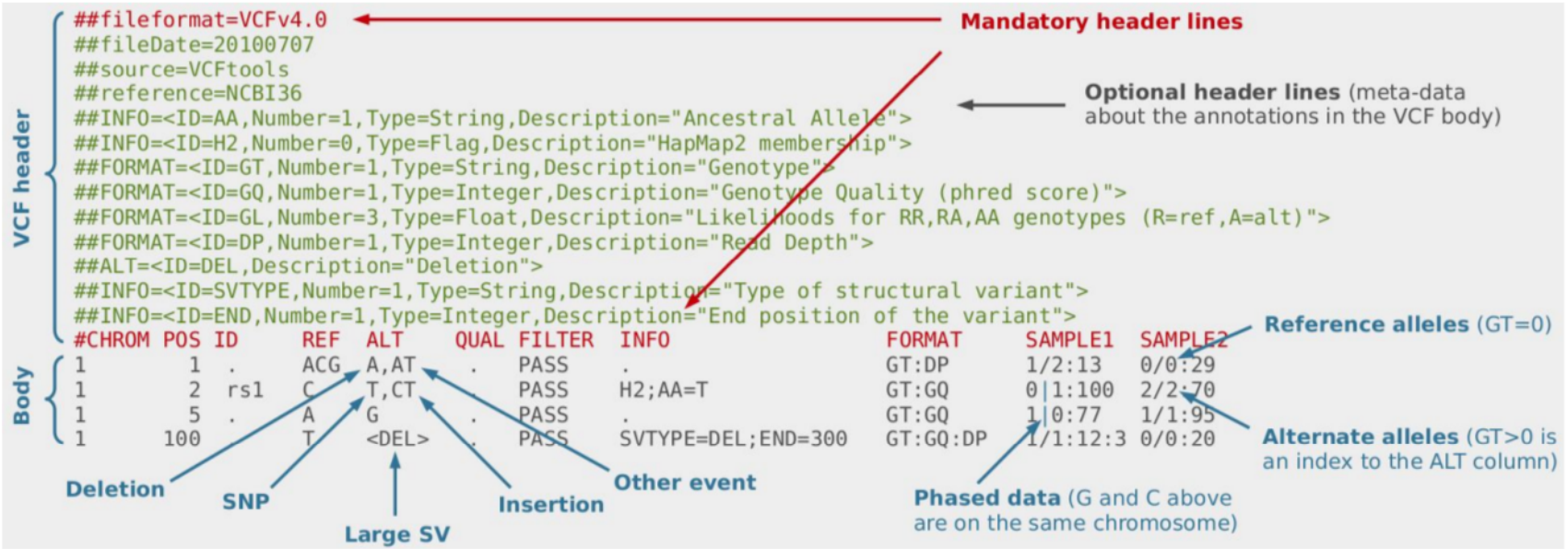
4. name - Defines the name of the BED line
5. Score (. or a number between 0 and 1000)
6. strand (+ forward, - reverse)
7. thickStart
8. thickEnd
9. itemRgb (255,0,0)
10. blockCount
11. blockSizes
12. blockStarts

```
track name="CHP2_Designed" description="Amplicon_Insert_CHP2" type=bedDetail ionVersion=4.0
chr1   43814968    43815086    CHP2_MPL_1    .   GENE_ID=MPL
chr1   115252185   115252269   CHP2_NRAS_3   .   GENE_ID=NRAS
chr1   115256504   115256584   CHP2_NRAS_2   .   GENE_ID=NRAS
chr1   115258689   115258774   CHP2_NRAS_1   .   GENE_ID=NRAS
chr2   29432572    29432680    CHP2_ALK_2    .   GENE_ID=ALK
chr2   29443607    29443729    CHP2_ALK_1    .   GENE_ID=ALK
chr2   209113103   209113206   CHP2_IDH1_1   .   GENE_ID=IDH1
chr2   212288904   212288990   CHP2_ERBB4_8  .   GENE_ID=ERBB4
chr2   212530051   212530180   CHP2_ERBB4_7  .   GENE_ID=ERBB4
```

Data formats

Variants: vcf

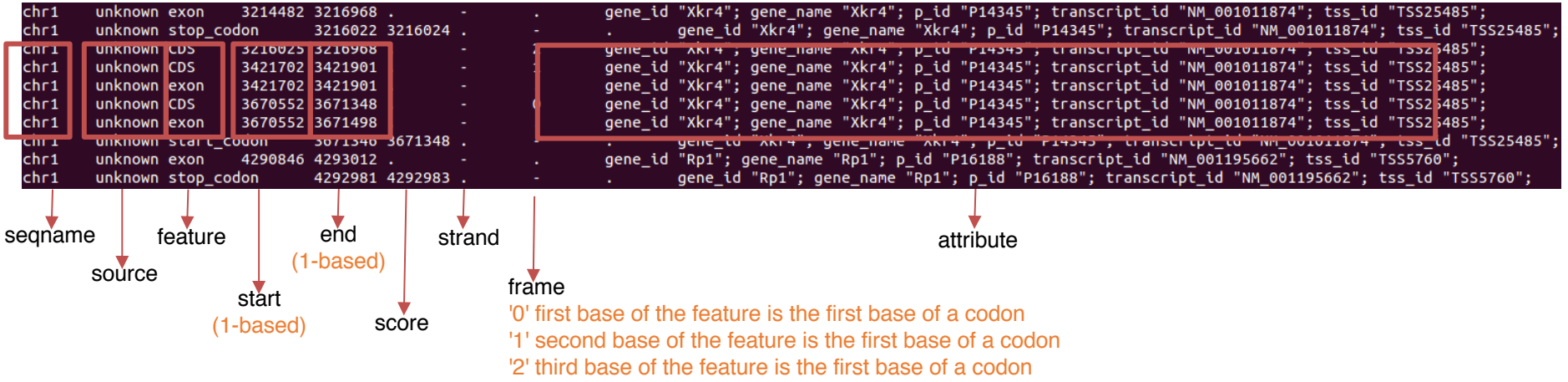
- Typical extension: .vcf (.bcf binary counterpart)
- Not all records in a VCF are true calls, the FILTER column specifies those which passed the calling
- QUAL is the score assigned to a given call. The greater QUAL is, the more reliable is. It is in log-scale



Data formats

Reference transcriptome: GTF/GFF

- Typical extensions: .gtf, .gff
- General Feature Format / Gene Transfer Format
- Annotation file for features
- One line per feature. 9 columns. Tab separated
- Used in: Reference transcriptomes (RNAseq) or to upload features to Genomic Browsers



Data formats

Counts: CSV/TSV

Comma-separated Text

```
Gene,Sample1,Sample2,Sample3
DDR1,0,13,21
TP53, 12,1,2
KRAS, 50, 5,5
```

Tab-separated Text

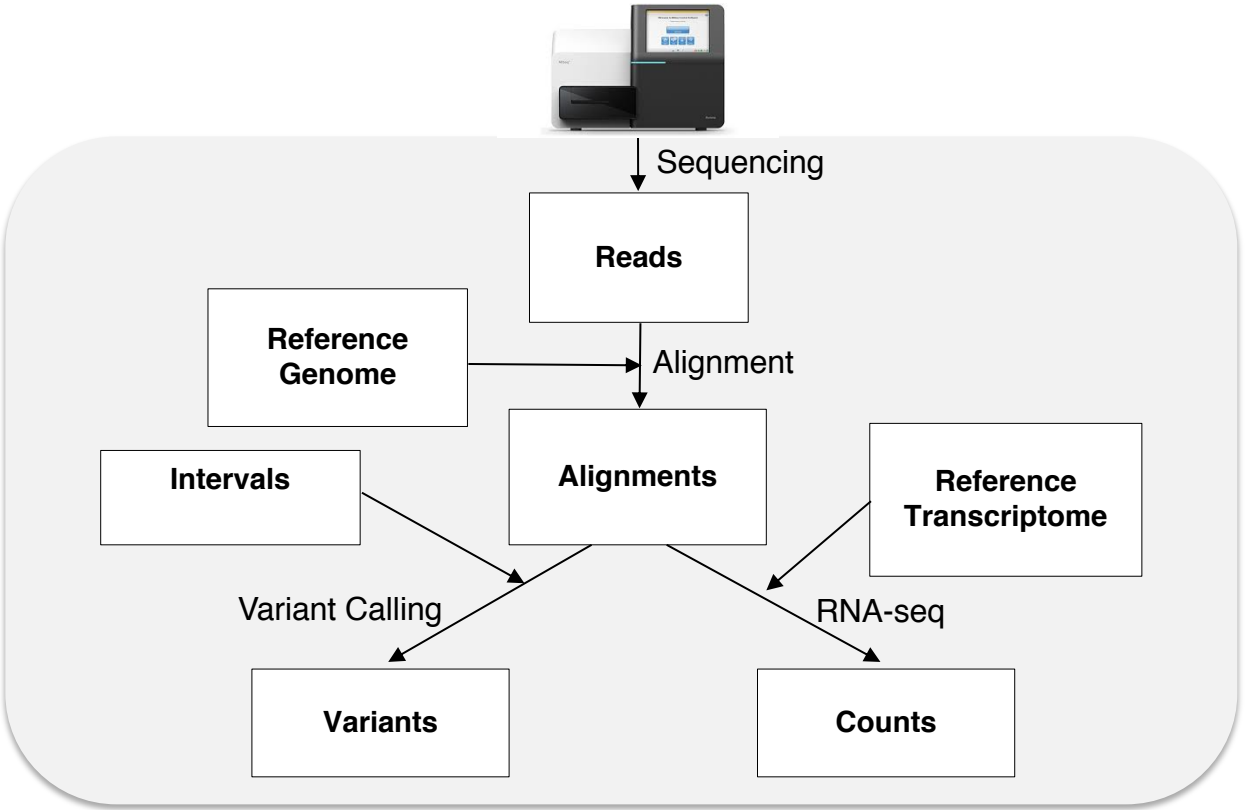
```
SRA_SAMPLE          SAMPLE_NAME
SRS1307733  GSM2069823
SRS1307731  GSM2069824
```

Excel

No format specifications.
There are no limits in terms of the matrix dimension (rows X columns).
However, there are good practices in Bioinformatics:

1. Use intuitive column names without spaces or rare chars. Instead of those, use “.” or “_”.
2. Data dimension: higher number of rows than columns.
3. Rows → individual observations or records. E.g. genomic positions.
4. Columns→ Individual variables for each individual observation. E.g. Chromosome, position, Score, Gene mutated, etc.
5. Do NOT mix data from different observations

Formats outline



Data formats cheat sheet

Format	Uses	Example	File type	Software Management	File Extension
Fasta	**Human genome** Define biological sequences (DNA, RNA, cDNA, proteins).	human_genome.fa	Plain text	samtools, picard-tools	.fa; .fasta
FastQ	**Raw sequencing data** Single-end sequencing → 1 file Paired-end sequencing → 2 files (R1 and R2 for each end, respectively)	DNAseq_raw_data.fastq (DNAseq_R1.fastq and DNAseq_R2.fastq)	Plain text	samtools, picard-tools **Aligners**	.fq; .fastq
SAM	Define read alignments. Store alignment meta-info (reference, methods, one- or multi-sample).	mapped_reads.sam	Plain text	samtools, picard-tools	.sam
BAM	**VISUALIZE ALIGNMENTS (IGV)** The same as SAM, but compressed and indexed. Also to store UNMAPPED reads (compressed).	mapped_reads.bam unmapped_reads.bam	Binary	samtools, bcftools, picard-tools, **IGV** (Integrative Genome Viewer)	.bam
VCF	**SNV & Indels calls** Indicates genomic variations. Store Variant calling meta-info (reference, methods,one- or multi-sample).	point_variants.vcf	Plain text	bcftools, Unix	.vcf
BED	**Intervals** Delimit genomic regions (i.e. intervals) w or w/o annotations.	targeted_regions.bed intervals.bed	Plain text	bedtools, Unix **GATK, picard-tools**	.bed
TSV or CSV	Create data matrix (rows X Columns)	annotated_variants.tsv	Plain text	Unix, Microsoft Excel, OpenOffice	.tsv; .csv; .txt

Thanks!

Credits for many class materials to:

Héctor Tejero: htejero@cnio.es
Elena Piñeiro: epineiro@cnio.es
Javier Perales-Patón: jperales@cnio.es

CNIO **BIOINFORMATICS** UNIT