



PO: Precision Oncology Course

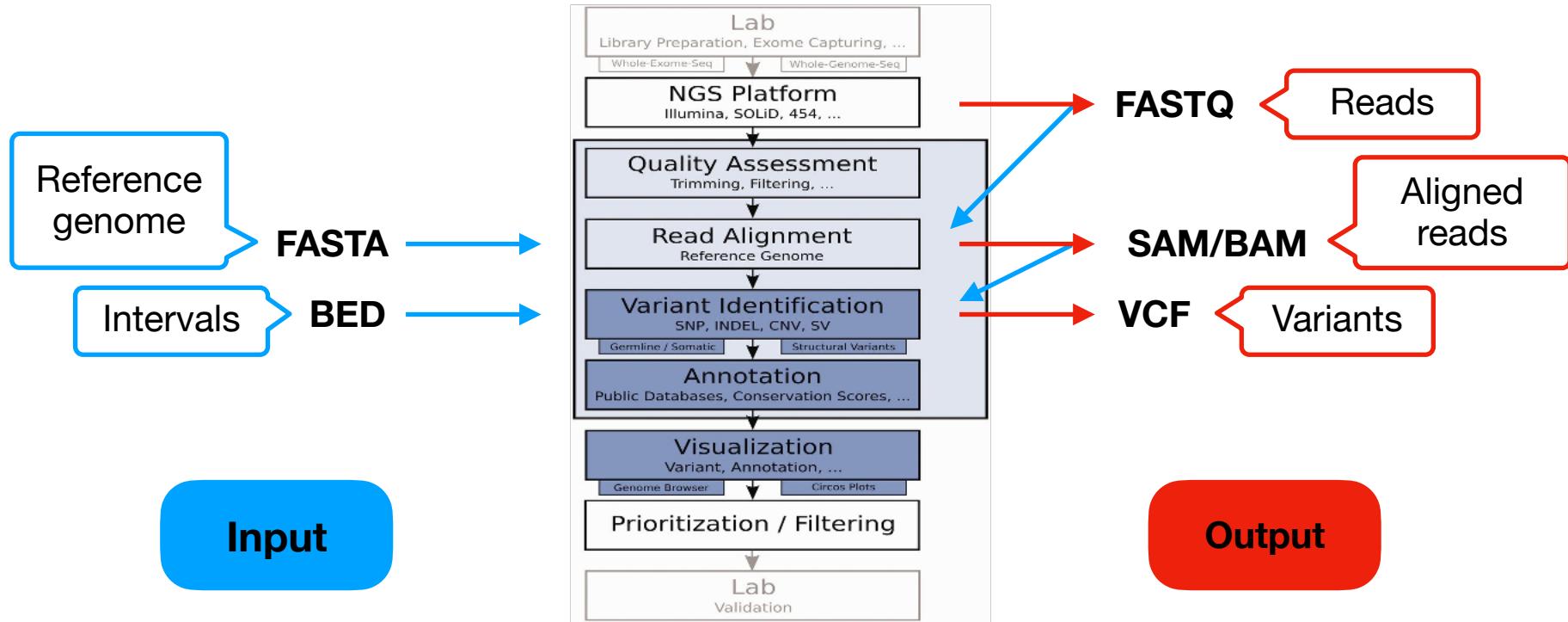
Data Formats



CNIO BIOINFORMATICS UNIT

cnio stop cancer

Formats Outline



Sequencing

Output: Reads (FASTQ)

FASTQ format

Typical extensions: **.fq, .fastq**

Each read is composed by **4 lines**:

- "@" Read ID and optional description (space separated)
- Sequence
- "+" (optionally: repeat the read ID)
- Encoded Base Quality Score

Identifier	• @SRR566546.970 HWUSI-EAS1673_11067_FC7070M:4:1:2299:1109 length=50
Sequence	• TTGCCTGCCTATCATTAGTCAGGTGGAGATGTGAGGATCAGT
'+' sign	• +
Quality scores	• hhhhhhhhhghhhhhfhhhhfffffe'ee['X]b[d[ed' [Y[^Y
Identifier	• @SRR566546.971 HWUSI-EAS1673_11067_FC7070M:4:1:2374:1108 length=50
Sequence	• GATTGTATGAAAGTATAACAACTAAAAGTCAGGTGGATCAGAGTAAGTC
'+' sign	• +
Quality scores	• hhggfhhcghghggfcffdhfehhhhcehdchhdhaehffffde'bVd

FASTQ format

Nucleotide 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

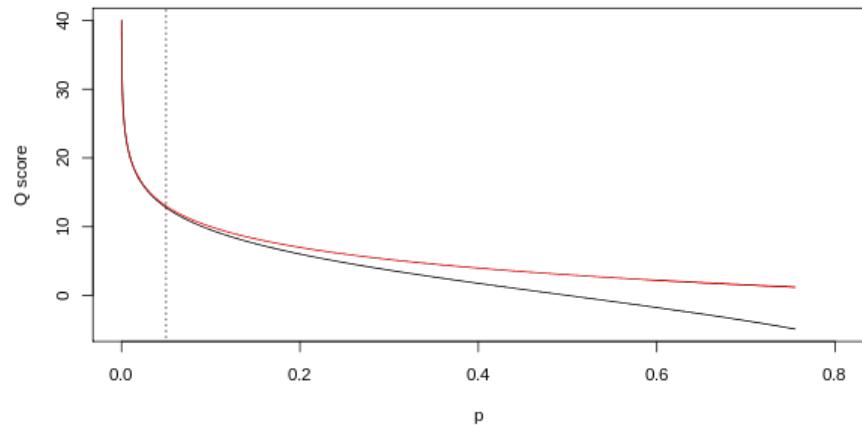
FASTQ format

Base Quality Score

Phred quality score (Q): Measure of the quality of the DNA sequencing for each base.

P: Base calling error probability.

$$Q = -10 \cdot \log_{10}(P)$$



The higher the Q, the lower the P.

[FASTQ format. Wikipedia.](#)

FASTQ format

ASCII code

Q is encoded in a
single ASCII
character for brevity.

$$ASCII = Q + 33$$

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	@	Ø		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		DEL	

Source: www.LookupTables.com

FASTQ format

Single-end vs paired-end

One unique sample can have 1 or 2 files:

- **1 file:** Single-end sequencing
(`*.fastq`)
- **2 files:** Paired-end sequencing
(`*_R1.fastq` and `*_R2.fastq`)

R1

```
@SN1083:379:H8VA1ADXX:2:1101:1248:2144 1:N:0:12
CCTAAATGGTGCATGCTAGGAGGCCGTGCCCTCTTGAAAAGTTGTATGTGAA
+
BBBFFFFFFFBFFFFIIIFI<FFIIIIIFIIIFBFIIIIIIIIFFIIIIIFI
```

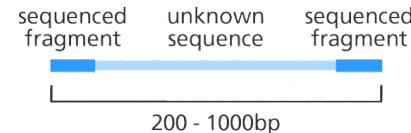
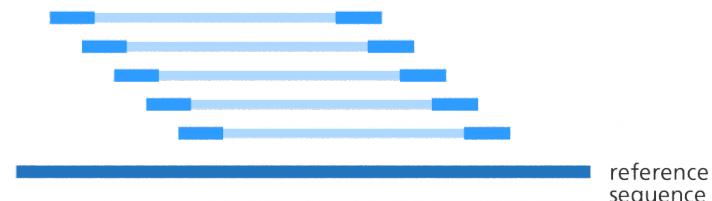
R2

```
@SN1083:379:H8VA1ADXX:2:1101:1248:2144 2:N:0:12
CATTTCGACGTTGTTAATAAGCTCTGCGTACTTGCAAGCTATCTGCGCGAACG
+
BBBFFFFFFFIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIFFF
```

Single-end reads



Paired-end reads



NGS overlapping reads. Biostars. Post 241139.

FASTQ format

Exercise

1. Download the data

<https://drive.google.com/file/d/15OxI4O8GO3I4e8WJFwNZet8DqOoR94y2/view?usp=sharing>

2. Go to the terminal and open the file

```
$ more WEx_Normal_R1.fastq
```

FASTQ format

5 min

Question

How would you detect the number of sequences present in the file? How much reads are there?

Useful commands

- **grep**: Searches plain-text data sets for lines that match a regular expression.
- **wc**: Counts words
- **wc -l**: Counts lines
- **| (Pipe)**: Lets you send the output of one command to another.

Alignment

Input: Reads (FASTQ) and Reference Genome (FASTA)

Output: Aligned reads (BAM/SAM)

FASTA format

Typical extensions: **.fasta, .fas, .fa, .fna, .fsa**

Each sequence is composed by at least **2 lines**:

- ">" Sequence ID and optional description (space separated)
- Line(s) with the whole sequence

Header	• >VIT_201s0011g03530.1
Sequence	• AATTAAGCATAAATACTCACTCTTACCCCTTATTTCTTATCTCTCATCACTTTGGTGCGAAG
Header	• GACCATGAGAACAGCTGCAATGGGTGAGGGTCTTCGCAAGGCATGCAGCCAAGACTGCATCA
Sequence	• >VIT_201s0011g03540.1
Header	• CAGGTAGCGTGAAGTTAACCCCTAGCGCTTAGACAAACAGCTGTAGTCACCGCCCACAAACACC
Sequence	• AGCCTCTGAGACACCACCTCAAACCTTCCACTTAAATACACATCCCTCACACCCTTTCAATT
Header	• >VIT_201s0011g03550.1
Sequence	• CATGCAAAGCTGAACGCGATGCTGTGATTGGTGGTAAGTGGTAGTTGAGTAAATTGACAGTGAA
	• GCCGAAATGGTAAAAGACTAAGGCTAGAAGTAGAATACCACTGTTCTCATCACGTGGGCCA

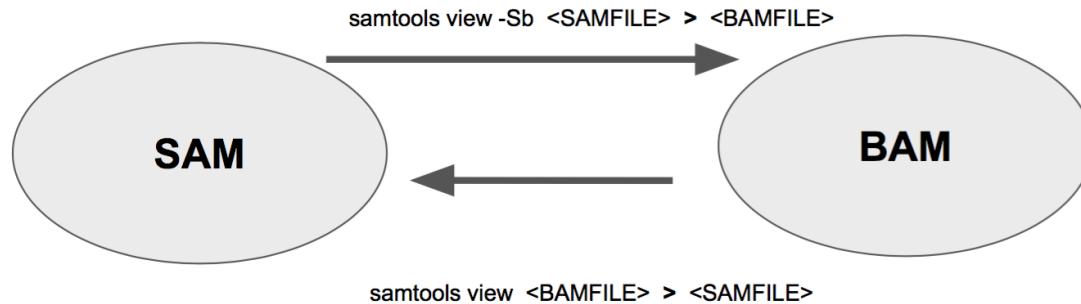
We can have
multiple
sequences in the
same file:
multifasta

SAM/BAM format

SAM is the human readable text format ([.sam extension](#)).

BAM is the binary, machine efficient format ([.bam extension](#)).

Both contain exactly the **same information** and are interconvertible using **samtools**.



[File specifications](#)

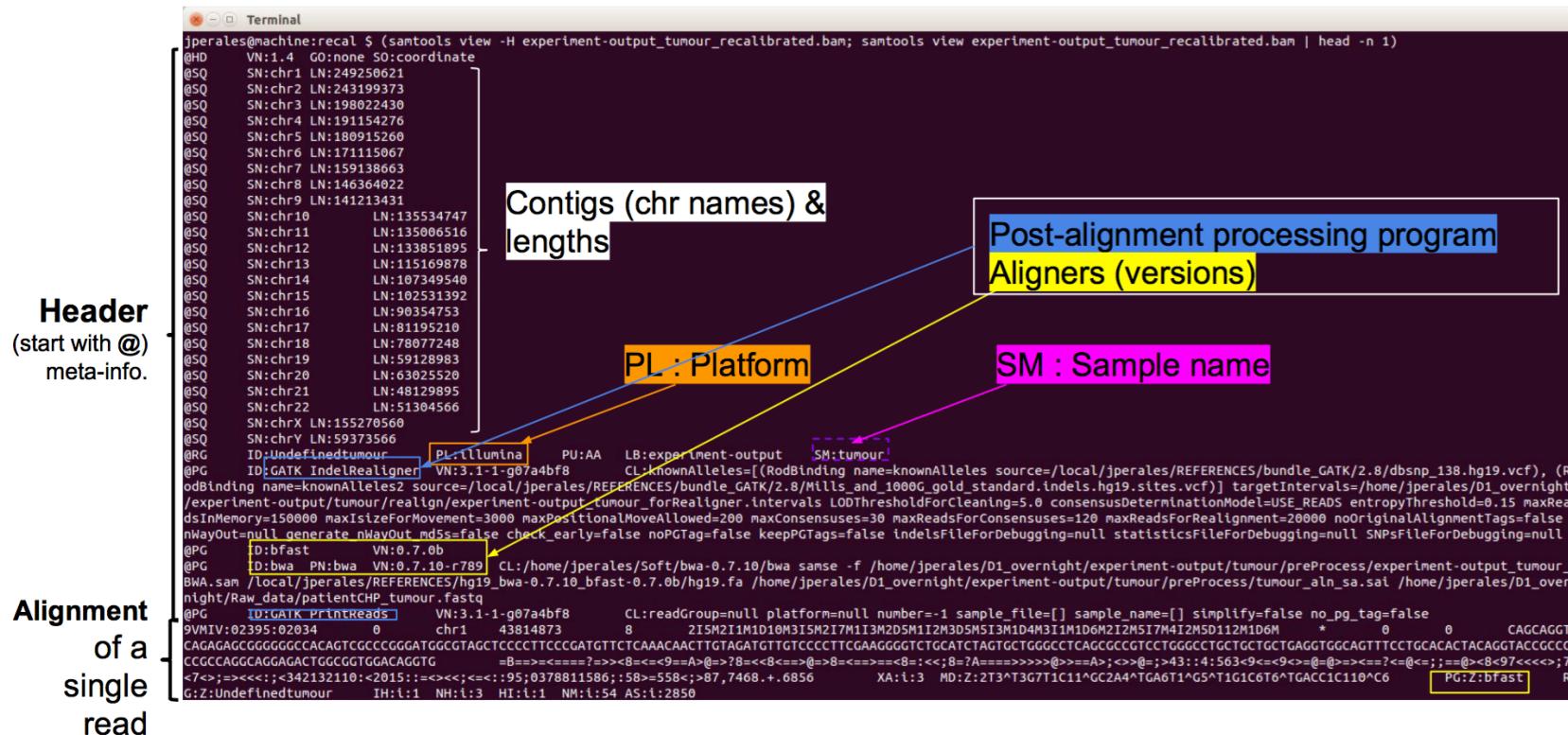
SAM/BAM format

It has two different parts

- **Header:** All lines that start with “@“. Contains information about the alignment.
- **Alignments:** The rest of the lines, with 12 tab separated columns. Information relative to each aligned read.

SAM/BAM format

Parts



SAM/BAM format

Alignment

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

#Col	Field	FIELDS
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

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.

SAM/BAM format

Alignment - FLAG

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

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

SAM/BAM format

Alignment - CIGAR

CIGAR: Concise Idiosyncratic Gapped Alignment Report

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

RefPos:	1 2 3 4 5 6 7 8 9
Reference:	C C A T A C T - G A
Read:	C A T - C T A G
POS:	2
CIGAR:	3M1D2M1I1M

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

Reference sequence with aligned reads	CIGAR string	Explanation
C T G C A T G T T A G A T A A * * G A T A G C T G T G C T A A A G G A T A * C T G G A T A A * G G A T A T G T T A [REDACTED] T G C T A	1M2I4M1D3M	Insertion & Deletion
a a a C A T G T T A G A A A C A T G T T A G	5M1P1I4M	Padding & Insertion
	5M1N5M	Spliced read
	3S8M	Soft clipping
	3H8M	Hard clipping

Variant Identification

Input: Alignments (SAM/BAM) and Intervals (BED)

Output: Called Variants (VCF)

BED format

Intervals: Regions of interest (important in targeted sequencing and WES). **More info here:** <https://gatk.broadinstitute.org/hc/en-us/articles/360035531852-Intervals-and-interval-lists>

Typical extension: .bed

Tab delimited file with **up to 12 different columns. The first 3 are mandatory:**

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.

```
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
```

BED format

Typical extension: .bed

Tab delimited file with **up to 12 different fields. The rest are optional:**

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
```

VCF format

Typical extension: .vcf

It has two different parts

- **Header:** All lines that start with “##”. Contains information about the variant annotation process.
- **Alignments:** The rest of the lines, with at least 8 tab separated fields. Information relative to each aligned read.

VCF format

- **QUAL** is the **score** assigned to a given call **in log-scale**.
- The **FILTER** column specifies which records **passed the calling**.

The diagram illustrates the structure of a VCF file, divided into **VCF header** and **Body**.

Mandatory header lines: These are the first few lines starting with `##`, such as `##fileformat=VCFv4.0` and `##fileDate=20100707`. They provide metadata about the file.

Optional header lines (meta-data about the annotations in the VCF body): These lines start with `#`, such as `#CHROM`, `POS`, `ID`, `REF`, `ALT`, `QUAL`, `FILTER`, `INFO`, `FORMAT`, and `SAMPLE`. They define the schema for the data rows.

Body: This section contains the actual variant data.

Annotations for the Body:

- Reference alleles (GT=0):** The `REF` column, which is "ACG".
- Alternate alleles (GT>0 is an index to the ALT column):** The `ALT` column, which includes "A, AT", "T, CT", "G", and "".
- Phased data (G and C above are on the same chromosome):** The `FORMAT` column, which includes "GT:DP", "GT:GQ", "GT:GQ", and "GT:GQ:DP".
- Deletion:** The first row has an empty `REF` and `ALT` column.
- SNP:** The second row has `rs1` in the `ID` column.
- Large SV:** The third row has `` in the `ALT` column.
- Insertion:** The fourth row has "T" in the `REF` column.
- Other event:** The fifth row has "G" in the `ALT` column.

#	CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	SAMPLE1	SAMPLE2
1		1	.	ACG	A, AT	.	PASS	.	GT:DP	1/2:13	0/0:29
1		2	rs1	C	T, CT	.	PASS	H2;AA=T	GT:GQ	0 1:100	2/2:70
1		5	.	A	G	.	PASS	.	GT:GQ	1 0:77	1/1:95
1		100		T		.	PASS	SVTYPE=DEL;END=300	GT:GQ:DP	1/1:12:3	0/0:20

VCF format

- **INFO:** List with **further info** about the variations. The fields are **separated by**
“,”.

The diagram illustrates the structure of a VCF file, divided into two main sections: the **VCF header** and the **Body**.

Mandatory header lines (red arrows):

```
##fileformat=VCFv4.0
##fileDate=20100707
##source=VCFtools
##reference=NCBI36
```

Optional header lines (meta-data about the annotations in the VCF body) (black arrow):

```
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality (phred score)">
##FORMAT=<ID=GL,Number=3,Type=Float,Description="Likelihoods for RR,RA,AA genotypes (R=ref,A=alt)">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##ALT=<ID=DEL,Description="Deletion">
##INFO=<ID=SVTYPE,Number=1,Type=String,Description="Type of structural variant">
##INFO=<ID=END,Number=1,Type=Integer,Description="End position of the variant">
```

Body (grey background):

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	SAMPLE1	SAMPLE2
1	1	.	ACG	A,AT	.	PASS	.	GT:DP	1/2:13	0/0:29
1	2	rsl1	C	T,CT	.	PASS	H2;AA=T	GT:GQ	0 1:100	2/2:70
1	5	.	A	G	.	PASS	.	GT:GQ	1 0:77	1/1:95
1	100		T		.	PASS	SVTYPE=DEL;END=300	GT:GQ:DP	1/1:12:3	0/0:20

Annotations for the Body:

- Deletion**: Points to the first row (ID=rsl1).
- SNP**: Points to the second row (ID=rsl1).
- Large SV**: Points to the fourth row (ID=100).
- Insertion**: Points to the third row (ID=5).
- Other event**: Points to the fifth row (ID=100).
- Phased data** (G and C above are on the same chromosome): Points to the second row (ID=rsl1), showing the '|'. This indicates that the G allele at position 2 is on the same chromosome as the C allele at position 1.
- Reference alleles (GT=0)**: Points to the first row (ID=rsl1), where GT is 1/2:13 (GT=1 is reference).
- Alternate alleles (GT>0 is an index to the ALT column)**: Points to the second row (ID=rsl1), where GT is 0|1:100 (GT=0 is reference, 1 is alternate).

VCF format

- **FORMAT** (optional): List with fields for describing the samples.
- **SAMPLEs** (optional): A column per sample with the values defined in FORMAT.

The diagram illustrates the structure of a VCF file, divided into **VCF header** and **Body**.

Mandatory header lines: These are at the top of the header section and include metadata like fileformat, fileDate, source, reference, and various INFO and FORMAT entries.

Optional header lines: These provide meta-data about annotations in the VCF body.

Body: This section contains the main variant data.

Annotations:

- Deletion:** An arrow points from the first row of the body to the 'ALT' column, which shows ''.
- SNP:** An arrow points from the second row to the 'ALT' column, which shows 'C'.
- Large SV:** An arrow points from the third row to the 'ALT' column, which shows 'T'.
- Insertion:** An arrow points from the fourth row to the 'ALT' column, which shows 'G'.
- Other event:** An arrow points from the fifth row to the 'ALT' column, which shows ''.

Phased data: A note states that 'G and C above are on the same chromosome'.

Reference alleles (GT=0): The 'REF' column shows 'ACG' for the first variant.

Alternate alleles (GT>0 is an index to the ALT column): The 'ALT' column shows 'A, AT' for the first variant, with 'AT' being the alternate allele.

FORMAT: The 'FORMAT' column specifies the data types for each sample.

SAMPLE1: The 'SAMPLE1' column shows genotype data for each variant across three samples.

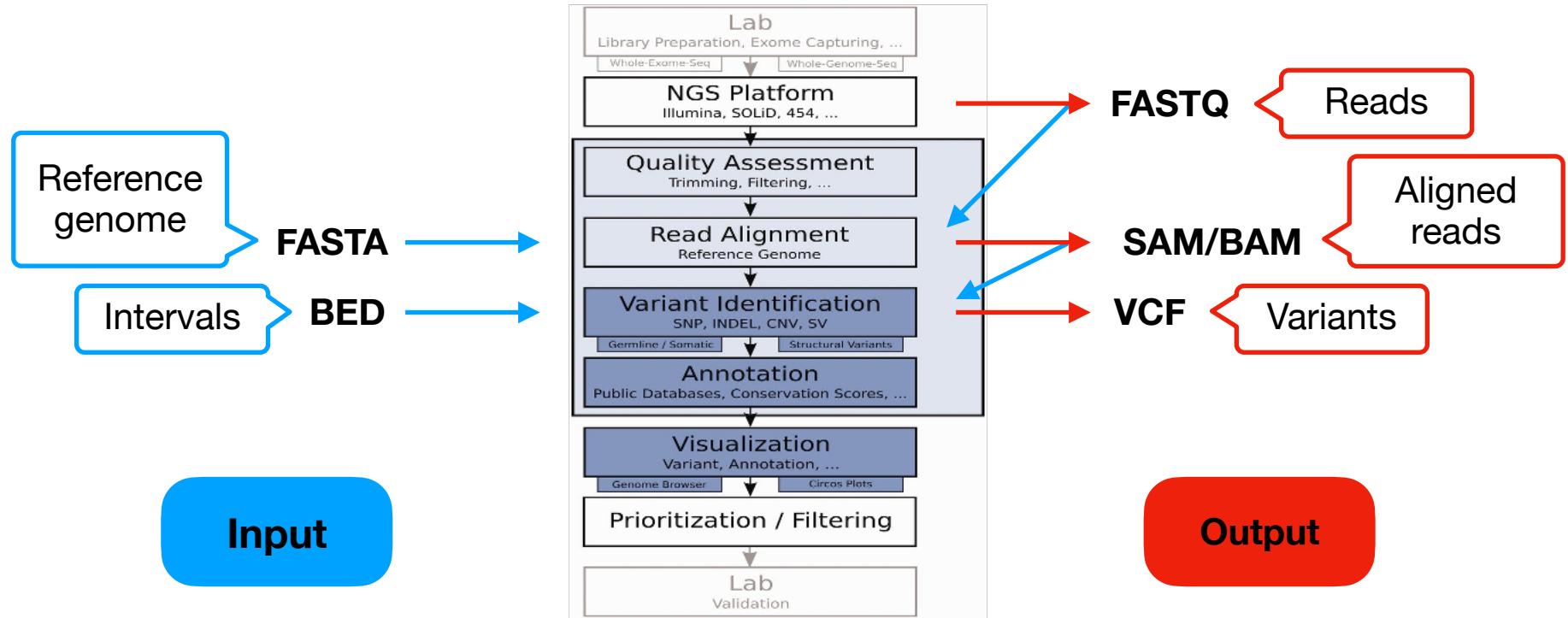
SAMPLE2: The 'SAMPLE2' column shows genotype data for each variant across three samples.

#CHROM								POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	SAMPLE1	SAMPLE2
1								1	.	ACG	A, AT	.	PASS	.	GT:DP	1/2:13	0/0:29
1								2	rsl1	C	T, CT	.	PASS	H2:AA=T	GT:GQ	0 1:100	2/2:70
1								5	.	A	G	.	PASS	.	GT:GQ	1 0:77	1/1:95
1								100		T		.	SVTYPE=DEL;END=300		GT:GQ:DP	1/1:12:3	0/0:20

Data formats cheat sheet

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

Formats Outline





Thanks!



CNIO BIOINFORMATICS UNIT

cnio stop cancer