GULBENKIAN CIÊNCIA

PO: Precision Oncology Course

Data Formats

BU CNIO BIOINFORMATICS UNIT

cnio stop cancer

Formats Outline

Stephan Pabinger et al. Brief Bioinform 2013;bib.bbs086

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	`TTGCCTGCCTATCATTTTAGTGCCTGTGAGGTGGAGATGTGAGGATCAGT`
'+' sign	`+`
Quality scores	`hhhhhhhhhhghhghhhhhfhhhhhfffffe‘ee[‘X]b[d[ed‘[Y[^Y`
Identifier	`@SRR566546.971 HWUSI-EAS1673_11067_FC7070M:4:1:2374:1108 length=50`
Sequence	`GATTTGTATGAAAGTATACAACTAAAACTGCAGGTGGATCAGAGTAAGTC`
'+' sign	`+`
Quality scores	`hhhhgfhhcghghggfcffdhfehhhhcehdchhdhahehffffde‘bVd`

Hosseini, M.; Pratas, D.; Pinho, A.J. A Survey on Data Compression Methods for Biological Sequences. *Information* **2016**, *7*, 56. doi: 10.3390/info7040056

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	@	@	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

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
CCTAAATGGTGCCATGCTAGGAGGCCGTGCCCTTCTTGAAAAGTTGTATGTGAA
+
BBBFFFFFFBFFFIIIIFI<FFIIIIIFIIIIFBFIIIIIIIIFFFIIIIFIII
```

_R2

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

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
Sequence
Header
Sequence
Header
Sequence

```
>VIT_201s0011g03530.1
AATTAAGCATAAATACTCACTCTTACCCCCTTATTTTCTTATCTCTCATCACTTTTGGTGCGAAG
GACCATGAGAACAAGCTGCAATGGGTGTAGGGTTCTTCGCAAGGCATGCAGCCAAGACTGCATCA
>VIT_201s0011g03540.1
CAGGTAGCGTGAAGTTAAACCCTAGCGCTTTAGACAAACAGCTGTAGTCACCGCCCACAAACACC
AGCCTCTGAGACACCACCTCAAACCTTTCCACTTAAATACACATCCCTCACACCCTTTTCAATTC
>VIT_201s0011g03550.1
CATGCAAAGCTGAACGCGATGCTGTGATTGGTGGTAAGTGGTAGTTGAGTAAATTTGACAGTGAA
GCCGAAATGGTAAAAGACTAAGGCTAGAAGTAGAATACCACTGTTCTTCTCATCACGTGGGCCCA
```

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

Hosseini, M.; Pratas, D.; Pinho, A.J. A Survey on Data Compression Methods for Biological Sequences. *Information* **2016**, *7*, 56. doi: 10.3390/info7040056

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

Header
(start with @)
meta-info.

Alignment
of a
single
read

```
Terminal
jperales@machine:recal $ (samtools view -H experiment-output_tumour_recalibrated.bam; samtools view experiment-output_tumour_recalibrated.bam | head -n 1)
@HD     VN:1.4  GO:none SO:coordinate
@SQ     SN:chr1 LN:249250621
@SQ     SN:chr2 LN:243199373
@SQ     SN:chr3 LN:198022430
@SQ     SN:chr4 LN:191154276
@SQ     SN:chr5 LN:180915260
@SQ     SN:chr6 LN:171115067
@SQ     SN:chr7 LN:159138663
@SQ     SN:chr8 LN:146364022
@SQ     SN:chr9 LN:141213431
@SQ     SN:chr10        LN:135534747
@SQ     SN:chr11        LN:135006516
@SQ     SN:chr12        LN:133851895
@SQ     SN:chr13        LN:115169878
@SQ     SN:chr14        LN:107349540
@SQ     SN:chr15        LN:102531392
@SQ     SN:chr16        LN:90354753
@SQ     SN:chr17        LN:81195210
@SQ     SN:chr18        LN:78077248
@SQ     SN:chr19        LN:59128983
@SQ     SN:chr20        LN:63025520
@SQ     SN:chr21        LN:48129895
@SQ     SN:chr22        LN:51304566
@SQ     SN:chrX LN:155270560
@SQ     SN:chrY LN:59373566
@RG     ID:Undefinedtumour      PL:illumina     PU:AA   LB:experiment-output     SM:tumour
@PG     ID:GATK IndelRealigner  VN:3.1-1-g07a4bf8       CL:knownAlleles=[(RodBinding name=knownAlleles source=/local/jperales/REFERENCES/bundle_GATK/2.8/dbsnp_138.hg19.vcf), (R
odBinding name=knownAlleles2 source=/local/jperales/REFERENCES/bundle_GATK/2.8/Mills_and_1000G_gold_standard.indels.hg19.sites.vcf)] targetIntervals=/home/jperales/D1_overnight
/experiment-output/tumour/realign/experiment-output_tumour_forRealigner.intervals LODThresholdForCleaning=5.0 consensusDeterminationModel=USE_READS entropyThreshold=0.15 maxRea
dsInMemory=150000 maxIsizeForMovement=3000 maxPositionalMoveAllowed=200 maxConsensuses=30 maxReadsForConsensuses=120 maxReadsForRealignment=20000 noOriginalAlignmentTags=false
nWayOut=null generate nWayOut md5s=false check_early=false noPGTag=false keepPGTags=false indelsFileForDebugging=null statisticsFileForDebugging=null SNPsFileForDebugging=null
@PG     ID:bfast        VN:0.7.0b
@PG     ID:bwa  PN:bwa  VN:0.7.10-r789  CL:/home/jperales/Soft/bwa-0.7.10/bwa samse -f /home/jperales/D1_overnight/experiment-output/tumour/preProcess/experiment-output_tumour_
BWA.sam /local/jperales/REFERENCES/hg19_bwa-0.7.10_bfast-0.7.0b/hg19.fa /home/jperales/D1_overnight/experiment-output/tumour/preProcess/tumour_aln_sa.sai /home/jperales/D1_over
night/Raw_data/patientCHP_tumour.fastq
@PG     ID:GATK PrintReads      VN:3.1-1-g07a4bf8       CL:readGroup=null platform=null number=-1 sample_file=[] sample_name=[] simplify=false no_pg_tag=false
9VMIV:02395:02034       0       chr1    43814873        8       2I5M2I1M1D10M3I5M2I7M1I3M2D5M1I2M3D5M5I3M1D4M3I1M1D6M2I2M5I7M4I2M5D112M1D6M     *       0       0       CAGCAGGT
CAGAGAGCGGGGGGCCACAGTCGCCCGGGATGGCGTAGCTCCCCTTCCCGATGTTCTCAAACAACTTGTAGATGTTGTCCCCTTCGAAGGGGTCTGCATCTAGTGCTGGGCCTCAGCGCCGTCCTGGGCCTGCTGCTGCTGAGGTGGCAGTTTCCTGCACACTACAGGTACCGCCC
CCGCCAGGCAGGAGACTGGCGGTGGACAGGTG        =B==>=<====?=>><8=<=<9==A>@=>?8=<<8<===>@=>8=<==>==<8=:<<;8=?A==========@>>==A>;<>>@=;>43::4:563<9<=<9<>=@=@>=>><==?<=@<=;;==@><8<97<<<<>;7
<7<>;=>><<<:;<342132110:<2015::=<>><;<=<:95;0378811586;:58>=558<;>87,7468.+.6856        XA:i:3  MD:Z:2T3^T3G7T1C11^GC2A4^TGA6T1^G5^T1G1C6T6^TGACC1C110^C6     PG:Z:bfast      R
G:Z:Undefinedtumour     IH:i:1  NH:i:3  HI:i:1  NM:i:54 AS:i:2850
```

Contigs (chr names) & lengths

Post-alignment processing program
Aligners (versions)

PL : Platform

SM : Sample name

SAM/BAM format

Alignment

```
#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.

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  jjjjjjj  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	1M2I4M1D3M	Insertion & Deletion
G A T A A * G G A T A	5M1P1I4M	Padding & Insertion
T G T T A [redacted] T G C T A	5M15N5M	Spliced read
a a a C A T G T T A G	3S8M	Soft clipping
A A A C A T G T T A G	3H8M	Hard clipping

Variant Identification

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

Output: Called Variants (VCF)

BED format

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.**

VCF format

- **INFO:** List with **further info** about the variations. The fields are **separated by ";"**.

VCF format

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

Data formats cheat sheet

Format	Uses	Example	Example	Software Management	File Extension
Fasta	**Reference genome** Define biological sequences (DNA, RNA, cDNA, proteins).	human_genome.fa	Plain text	samtools, picard-tools	.fasta; .fas; .fa; .fna; .fsa
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

Formats Outline

Stephan Pabinger et al. Brief Bioinform 2013;bib.bbs086

GULBENKIAN CIÊNCIA

Thanks!

BU CNIO BIOINFORMATICS UNIT

cnio stop cancer