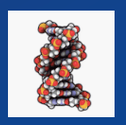
ALIGNMENT FILE FORMATS

SAM FORMAT

The **SAM Format** (**S**equence **A**lignment/**M**ap) is a text format for storing sequence alignment data in a series of tab delimited ASCII columns.
The file has two parts:

1. **Header** - Each line starts with a “@“.
   @HD, @SQ, @RG, @PG

2. **Alignments** - One line for each entry.

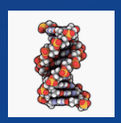


ALIGNMENT FILE FORMATS

SAM FORMAT

Example of SAM Header

@HD VN:1.0 **SO:unsorted**
@SQ SN:chr1 LN:195471971
@SQ SN:chr2 LN:182113224
@SQ SN:chr3 LN:160039680
@SQ SN:chr4 LN:156508116
@SQ SN:chr5 LN:151834684
@SQ SN:chr6 LN:149736546
@SQ SN:chr7 LN:145441459
@SQ SN:chr8 LN:129401213
@SQ SN:chr9 LN:124595110
@SQ SN:chr10 LN:130694993
@SQ SN:chr11 LN:122082543
@SQ SN:chr12 LN:120129022
@SQ SN:chr13 LN:120421639
@SQ SN:chr14 LN:124902244
@SQ SN:chr15 LN:104043685
@SQ SN:chr16 LN:98207768
@SQ SN:chr17 LN:94987271
@SQ SN:chr18 LN:90702639
@SQ SN:chr19 LN:61431566
@SQ SN:chrX LN:171031299
@SQ SN:chrY LN:91744698
@SQ SN:chrM LN:16299
@PG ID:bowtie2 PN:**bowtie2 VN:2.2.9** CL:"/usr/local/apps/bowtie/2-2.2.9/bowtie2-align-s --wrapper basic-0 -x /fdb/bowtie2.DELETE/mm10 -q jun_minus_dex_rep1a -S jun_minus_dex_rep1a_mm10.sam -p8"



ALIGNMENT FILE FORMATS

SAM FORMAT

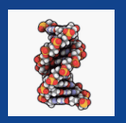
Example of Alignment portion to the file

Single Ended Sequences

```
D00537:377:HCCMGBCX3:1:2206:6809:3485 0   chr1   10453 255    22M    *  0  0
CCCTAACCCTAACCCTCGCGGT
HHIIIIIIIIIIIIIIIIIII    XA:i:0 MD:Z:22   NM:i:0 XM:i:2
```

Paired Ended Sequences

```
8_100_10000_12419       163      chrVII  271183  255      40M      =       271294  151
TGGTGTATTATACGCTACCGTGCGGTGCCGGGGGCAACCG
bbbabbbbbbbbbbbbbbcbbbbcbbbbbbbbbbbbbb          XA:i:0  MD:Z:40 NM:i:0
```

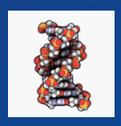


ALIGNMENT FILE FORMATS

SAM FORMAT

Reference Info

- https://en.wikipedia.org/wiki/SAM_(file_format)
- https://samtools.github.io/hts-specs/SAMv1.pdf
- https://samtools.github.io
- http://broadinstitute.github.io/picard/explain-flags.html



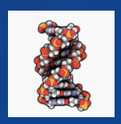
ALIGNMENT FILE FORMATS

SAM FORMAT

```
8_100_10000_12419        163      chrVII  271183  255      40M      =       271294  151
TGGTGTATTATACGCTACCGTGCGGTGCCGGGGGCAACCG
bbbabbbbbbbbbbbbbbbcbbbbcbbbbbbbbbbbbbb      XA:i:0  MD:Z:40 NM:i:0
```

8_100_10000_12	163	chr7	271183	255	40M	=	271294	151	TGGTGTA TTATACG	bbbabbbb bbbbbbbb	XA:i:0 MD:Z:40
QNAME	FLAG	RNAME	POS	MAPQ	CIGAR	MRNM	MPOS	TLEN	SEQ	QUAL	OPT

Col	Field	Description
1	QNAME	Query template/pair NAME
2	FLAG	bitwise FLAG
3	RNAME	Reference sequence NAME
4	POS	1-based leftmost POSition/coordinate of clipped sequence
5	MAPQ	MAPping Quality (Phred-scaled)
6	CIGAR	extended CIGAR string
7	MRNM	Mate Reference sequence NaMe (‘=’ if same as RNAME)
8	MPOS	1-based Mate POSistion
9	TLEN	inferred Template LENgth (insert size)
10	SEQ	query SEQuence on the same strand as the reference
11	QUAL	query QUALity (ASCII-33 gives the Phred base quality)
12+	OPT	variable OPTional fields in the format TAG:VTYPE:VALUE



ALIGNMENT FILE FORMATS

SAM FORMAT

```
8_100_10000_12419        163     chrVII  271183  255     40M     =       271294  151
TGGTGTATTATACGCTACCGTGCGGTGCCGGGGGCAACCG
bbbabbbbbbbbbbbbbbbcbbbbcbbbbbbbbbbbbbb     XA:i:0  MD:Z:40 NM:i:0
```

Understanding Flag codes

http://broadinstitute.github.io/picard/explain-flags.html

flag values

Value	Description
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

ALIGNMENT FILE FORMATS

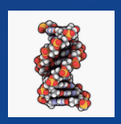
SAM FORMAT

```
8_100_10000_12419        163      chrVII  271183  255      40M      =       271294  151
TGGTGTATTATACGCTACCGTGCGGTGCCGGGGGCAACCG
bbbabbbbbbbbbbbbbbbcbbbbcbbbbbbbbbbbbbb                XA:i:0  MD:Z:40 NM:i:0
```

Understanding CIGAR codes
bases-code

code meanings

Operation	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 clip on the read (clipped sequence present in <seq>)
H	Hard clip on the read (clipped sequence NOT present in <seq>)
P	Padding (silent deletion from the padded reference sequence)



ALIGNMENT FILE FORMATS

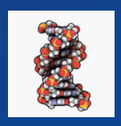
SAM FORMAT

```
8_100_10000_12419       163     chrVII  271183  255     40M     =       271294  151
TGGTGTATTATACGCTACCGTGCGGTGCCGGGGGCAACCG
bbbabbbbbbbbbbbbbbbcbbbbcbbbbbbbbbbbbbb                 XA:i:0  MD:Z:40 NM:i:0
```

Understanding TAG codes
TAG:type:value

type of value

Type	Description
A	Character
i	Integer
f	Floating decimal point
Z	String
H	Hex String



ALIGNMENT FILE FORMATS

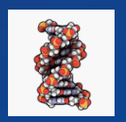
BAM/CRAM FORMAT

BAM (*.bam) is the compressed binary version of the Sequence Alignment/Map (SAM) format, a compact and index-able representation of nucleotide sequence alignments. **BAM** is compressed in the **BGZF** format that supports **random access** through the BAM file index (*.bam.bai).

HINT: Filename.bam and filename.bai always go together

The ability to randomly access portions of the file based on genomic coordinates makes it the perfect format for viewing data in IGV.

(Note: IGV and UCSC viewers can use this ability to efficiently access and display portions of the file from files housed on a remote server - no need to download the entire file and shared views.)

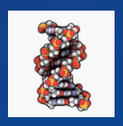


ALIGNMENT FILE FORMATS

BAM/CRAM FORMAT

CRAM (*.cram) - newer implementation of BAM like binary data.

1. Significantly better lossless compression than BAM
2. Full compatibility with BAM
3. Effortless transition to CRAM from using BAM files
4. Like BAM it has an associated index
5. Support for controlled loss of BAM data



ALIGNMENT FILE FORMATS

SAMTOOLS

Samtools is the "swiss army knife" for SAM/BAM/CRAM data

samtools help

samtools view -H aligned.bam (**display** the header info)

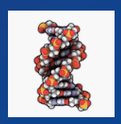
samtools view aligned.bam (**display** the read info)

samtools view -c aligned.bam (count the entries)

samtools view -F 4 aligned.bam (filter out the unaligned reads and display)

samtools index aligned.bam (generate and index aligned.bam.bai)

samtools flagstat aligned.bam



The End