Aligning reads to a genome

Analysis of Next-Generation Sequencing Data

Luce Skrabanek

Applied Bioinformatics Core

Slides at https://bit.ly/2T3sjRg[1]

11 February, 2020

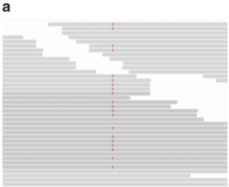
Weill Cornell Medicine

[1] https://physiology.med.cornell.edu/faculty/skrabanek/lab/angsd/schedule_2020/

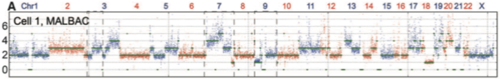
1. Why do we align?
2. What do we align to?
3. How do we align?
4. Output files
5. References

Why do we align?

What do we learn?

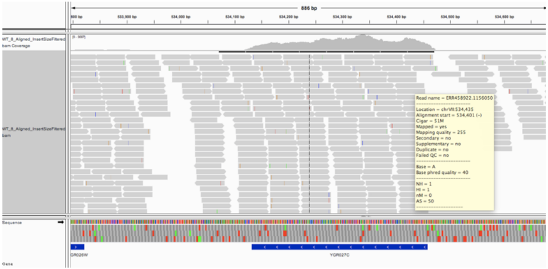


SNP identification and frequency estimation



CNV detection

identify protein-binding sites, histone marks

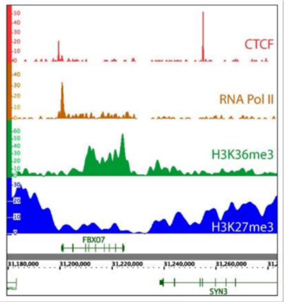


Read name = ERR458922.1156050

Location = chrVII:534,435
Alignment start = 534,401 (-)
Cigar = 51M
Mapped = yes
Mapping quality = 255
Secondary = no
Supplementary = no
Duplicate = no
Failed QC = no

Base = A
Base phred quality = 40

NH = 1
HI = 1
nM = 0
AS = 50



which genes are expressed, and how much

What do we align to?

What do we need?

- Reference sequence: the nucleotide sequence of the chromosomes of a species [2]
- Optional annotations: the gene/transcript models for a genome; includes the coordinates of the exons of a transcript on a reference genome, optionally the strand, gene name, coding portion of the transcript.

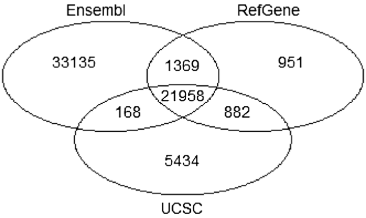
[2] see discussion on reference genomes in [Ballouz et al., 2019]

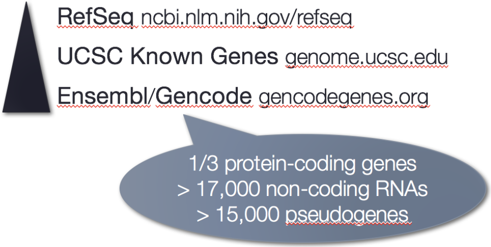
Sources for reference genomes

- **Ensembl**
  - http://www.ensembl.org
- **UCSC**
  - https://genome.ucsc.edu/
- **NCBI**
  - https://www.ncbi.nlm.nih.gov/
- **Gencode**
  - https://www.gencodegenes.org/
- **Organism-specific databases**
  - (e.g., http://toxodb.org/toxo/)

Always note the source and version of your reference genome.
Look out for chromosome naming conventions.

Annotations





RefSeq ncbi.nlm.nih.gov/refseq

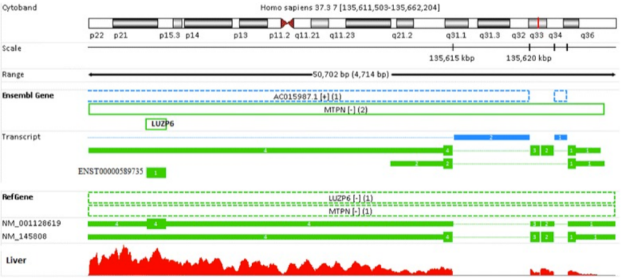
UCSC Known Genes genome.ucsc.edu

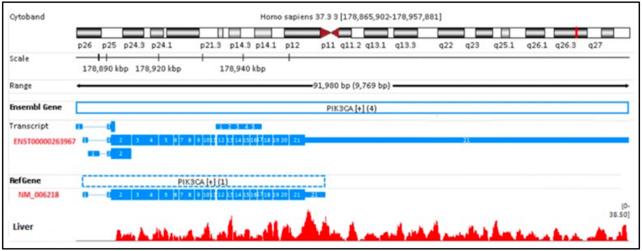
Ensembl/Gencode gencodegenes.org

1/3 protein-coding genes
> 17,000 non-coding RNAs
> 15,000 pseudogenes

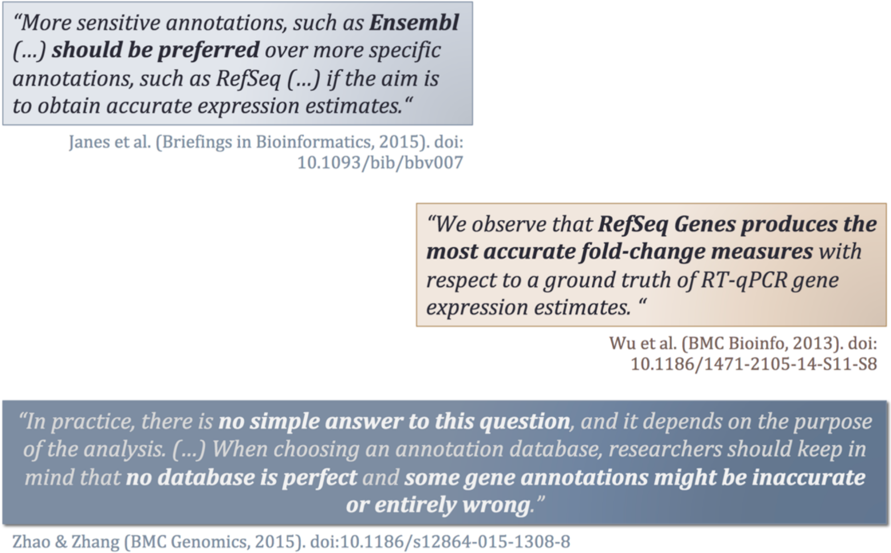
The chromosome names must match those in your reference genome; annotations must correspond to the same reference genome build as your reference genome fasta file.

Gene models can vary dramatically





Which annotation should you use?



> *"More sensitive annotations, such as **Ensembl** (...) **should be preferred** over more specific annotations, such as RefSeq (...) if the aim is to obtain accurate expression estimates."*

Janes et al. (Briefings in Bioinformatics, 2015). doi: 10.1093/bib/bbv007

> *"We observe that **RefSeq Genes produces the most accurate fold-change measures** with respect to a ground truth of RT-qPCR gene expression estimates. "*

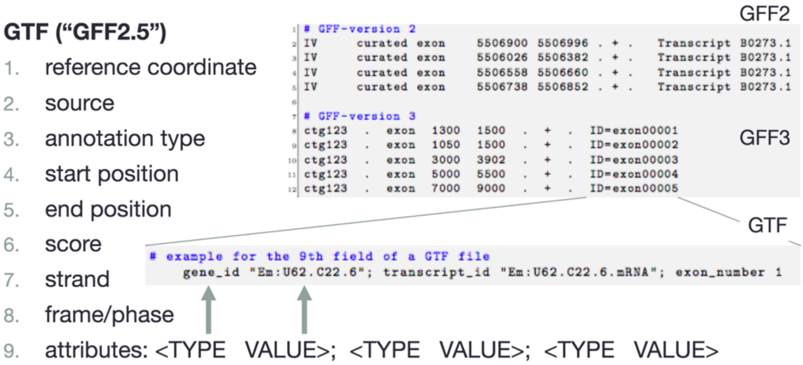
Wu et al. (BMC Bioinfo, 2013). doi: 10.1186/1471-2105-14-S11-S8

> *"In practice, there is **no simple answer to this question**, and it depends on the purpose of the analysis. (...) When choosing an annotation database, researchers should keep in mind that **no database is perfect** and **some gene annotations might be inaccurate or entirely wrong**."*

Zhao & Zhang (BMC Genomics, 2015). doi:10.1186/s12864-015-1308-8

[Jänes et al., 2015, Zhao and Zhang, 2015, Wu et al., 2013]

Storing annotation information



GTF ("GFF2.5")

1. reference coordinate
2. source
3. annotation type
4. start position
5. end position
6. score
7. strand
8. frame/phase
9. attributes: <TYPE VALUE>; <TYPE VALUE>; <TYPE VALUE>

GFF2

```
# GFF-version 2
IV      curated exon    5506900 5506996 . + .   Transcript B0273.1
IV      curated exon    5506026 5506382 . + .   Transcript B0273.1
IV      curated exon    5506558 5506660 . + .   Transcript B0273.1
IV      curated exon    5506738 5506852 . + .   Transcript B0273.1

# GFF-version 3
ctg123  .  exon  1300  1500   .  +  .  ID=exon00001
ctg123  .  exon  1050  1500   .  +  .  ID=exon00002
ctg123  .  exon  3000  3902   .  +  .  ID=exon00003
ctg123  .  exon  5000  5500   .  +  .  ID=exon00004
ctg123  .  exon  7000  9000   .  +  .  ID=exon00005
```

GFF3

GTF

```
# example for the 9th field of a GTF file
    gene_id "Em:U62.C22.6"; transcript_id "Em:U62.C22.6.mRNA"; exon_number 1
```

- Represent genome coordinates and gene descriptions/names
- multiple formats: GFF2, GFF3, GTF[3], BED, SAF...

[3] http://genome.ucsc.edu/FAQ/FAQformat#format4

How do we align?

Aligners

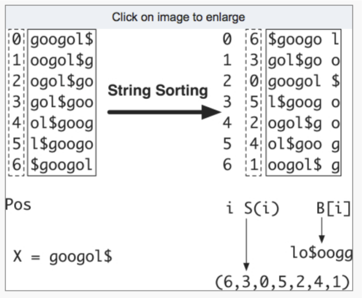
- Genomic aligners
  - BWA [Li and Durbin, 2009], Bowtie2
- Splice-aware aligners
  - STAR [Dobin et al., 2013], TopHat, HiSAT2
- Pseudo alignment
  - Salmon, kallisto, RSEM

Challenge

Mapping millions of reads accurately and in a reasonable amount of time, despite complications from sequencing errors, genomic variation and repetitive elements.

Genomic aligner: BWA

BWA uses a canonical seed-and-extend paradigm. BWA is based on the Burrows-Wheeler Transform and uses the FM-index[4] to search for exact string matches.



Pos	
0	googol$
1	oogol$g
2	ogol$go
3	gol$goo
4	ol$goog
5	l$googo
6	$googol

i	S(i)		B[i]
0	6	$googo	l
1	3	gol$go	o
2	0	googol	$
3	5	l$goog	o
4	2	ogol$g	o
5	4	ol$goo	g
6	1	oogol$	g

X = googol$

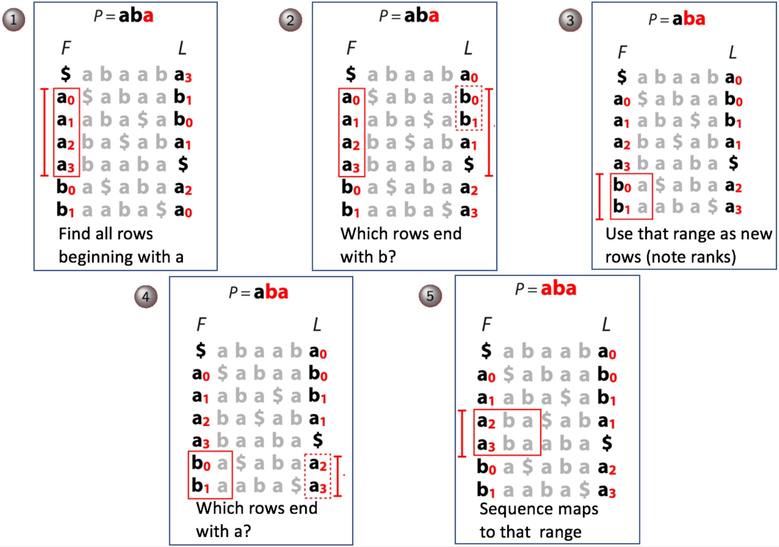
lo$oogg

(6,3,0,5,2,4,1)

This has a very small memory footprint.

[4] Full-text Minute-space, or Ferragina and Manzini [Ferragina and Manzini, 2010]

FM-index backwards search



1

$P =$ **aba**

F L

\$ a b a a b a_3
a_0 \$ a b a a b_1
a_1 a b a \$ a b_0
a_2 b a \$ a b a_1
a_3 b a a b a \$
b_0 a \$ a b a a_2
b_1 a a b a \$ a_0

Find all rows beginning with a

2

$P =$ **aba**

F L

\$ a b a a b a_0
a_0 \$ a b a a b_0
a_1 a b a \$ a b_1
a_2 b a \$ a b a_1
a_3 b a a b a \$
b_0 a \$ a b a a_2
b_1 a a b a \$ a_3

Which rows end with b?

3

$P =$ **aba**

F L

\$ a b a a b a_0
a_0 \$ a b a a b_0
a_1 a b a \$ a b_1
a_2 b a \$ a b a_1
a_3 b a a b a \$
b_0 a \$ a b a a_2
b_1 a a b a \$ a_3

Use that range as new rows (note ranks)

4

$P =$ **aba**

F L

\$ a b a a b a_0
a_0 \$ a b a a b_0
a_1 a b a \$ a b_1
a_2 b a \$ a b a_1
a_3 b a a b a \$
b_0 a \$ a b a a_2
b_1 a a b a \$ a_3

Which rows end with a?

5

$P =$ **aba**

F L

\$ a b a a b a_0
a_0 \$ a b a a b_0
a_1 a b a \$ a b_1
a_2 b a \$ a b a_1
a_3 b a a b a \$
b_0 a \$ a b a a_2
b_1 a a b a \$ a_3

Sequence maps to that range

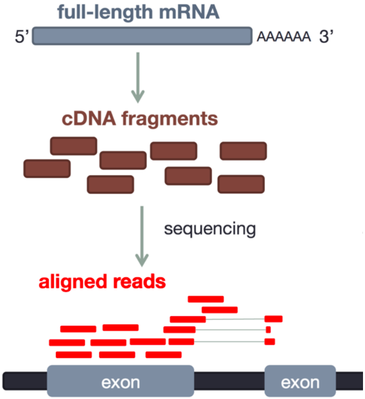
BWA-MEM

BWA MEM [Li, 2013] is the next generation in the BWA family, and is one of the few that works well for both 70bp reads and long sequences up to a few megabases.

1. allows long gaps
2. the allowable error rate adjusts with sequence length
3. can report multiple non-overlapping local hits

- As for BWA, uses a canonical seed-and-extend paradigm, grouping seeds that are colinear and close to each other as a chain.
- Each seed is extended using a banded affine-gap-penalty dynamic programming, stopping when the difference between the best and the current extension score is above some threshold, avoiding extension through poorly aligned regions
- Keep track of the best extension score reaching the end of the query sequence. If the difference between the best score reaching the end and the best local alignment score is below a threshold, the local alignment will be rejected even if it has a higher score.

Mapping to the transcriptome

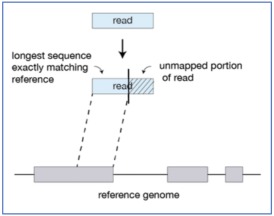


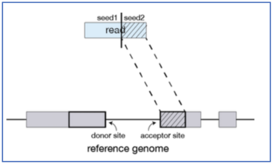
1. Alignment of exon-exon spanning reads
2. Multiple isoforms
3. Identification of novel splice junctions

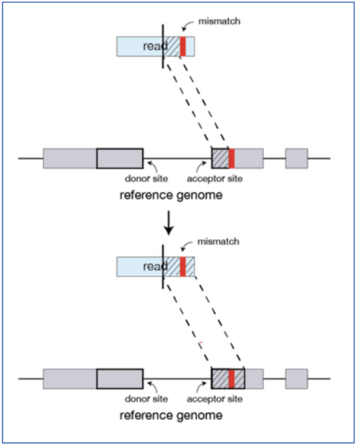
STAR uses an indexed suffix array [generated using both the genomic sequence, and the sequence spanning known exon-exon boundaries (transcriptome)], to find MMPs (longest possible perfect matches), identifies "anchor alignments", and stitches them together.

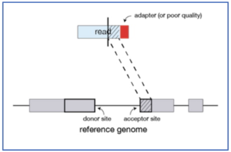
STAR can also identify novel junctions, if it finds enough reads as support. Users can define how many reads must span a novel junction, and how many bases must be covered on either side of the junction.

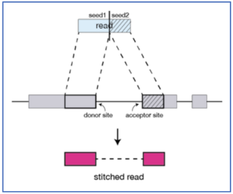
Splice-aware aligner: STAR [Spliced Transcripts Alignment to a Reference]



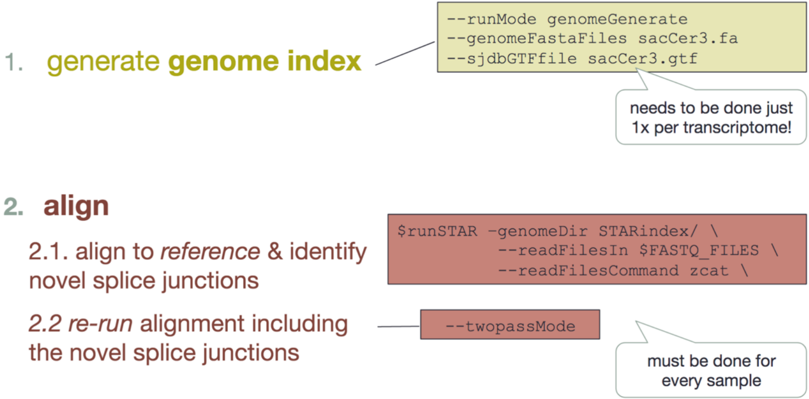








Running STAR



1. generate **genome index**

```
--runMode genomeGenerate
--genomeFastaFiles sacCer3.fa
--sjdbGTFfile sacCer3.gtf
```

needs to be done just 1x per transcriptome!

2. **align**

2.1. align to *reference* & identify novel splice junctions

```
$runSTAR -genomeDir STARindex/ \
         --readFilesIn $FASTQ_FILES \
         --readFilesCommand zcat \
```

2.2 re-run alignment including the novel splice junctions

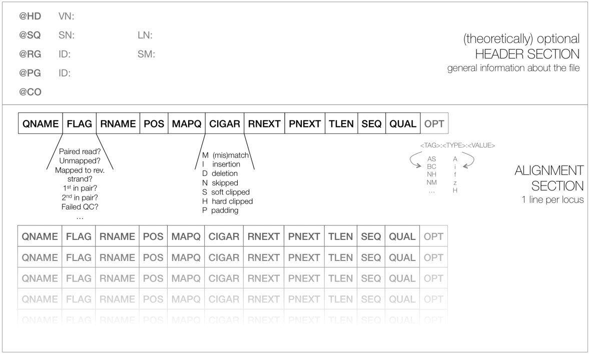
```
--twopassMode
```

must be done for every sample

STAR has many parameters (familiarize yourself with the manual)! See [Ballouz et al., 2018] for a discussion of how parameter selection affects mapping (e.g., handling of multi-mapped reads, intron sizes).

Output files

SAM files



@HD VN:
@SQ SN: LN:
@RG ID: SM:
@PG ID:
@CO

(theoretically) optional
HEADER SECTION
general information about the file

QNAME	FLAG	RNAME	POS	MAPQ	CIGAR	RNEXT	PNEXT	TLEN	SEQ	QUAL	OPT
QNAME	FLAG	RNAME	POS	MAPQ	CIGAR	RNEXT	PNEXT	TLEN	SEQ	QUAL	OPT
QNAME	FLAG	RNAME	POS	MAPQ	CIGAR	RNEXT	PNEXT	TLEN	SEQ	QUAL	OPT
QNAME	FLAG	RNAME	POS	MAPQ	CIGAR	RNEXT	PNEXT	TLEN	SEQ	QUAL	OPT
QNAME	FLAG	RNAME	POS	MAPQ	CIGAR	RNEXT	PNEXT	TLEN	SEQ	QUAL	OPT
QNAME	FLAG	RNAME	POS	MAPQ	CIGAR	RNEXT	PNEXT	TLEN	SEQ	QUAL	OPT

Paired read?
Unmapped?
Mapped to rev. strand?
1st in pair?
2nd in pair?
Failed QC?
...

M (mis)match
I insertion
D deletion
N skipped
S soft clipped
H hard clipped
P padding

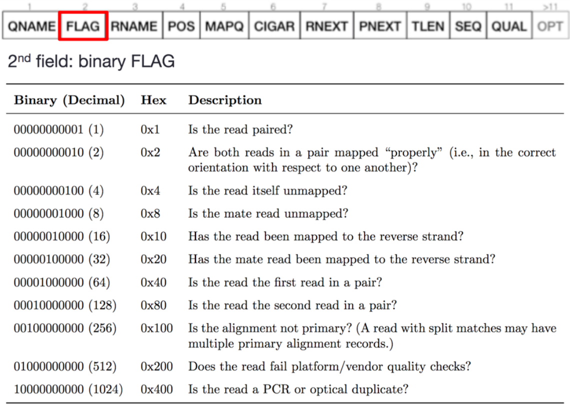
<TAG>:<TYPE>:<VALUE>

AS	A
BC	i
NH	f
NM	z
...	H

ALIGNMENT
SECTION
1 line per locus

Each line of the optional header section starts with `@`, and includes information such as chromosomes names (SN) and their lengths (LN). The vast majority of lines within a SAM file are compact representations of the read alignments where each read is described by the 11 mandatory entries and a variable number of optional fields [Li et al., 2009].

SAM FLAG field



1	2	3	4	5	6	7	8	9	10	11	>11
QNAME	FLAG	RNAME	POS	MAPQ	CIGAR	RNEXT	PNEXT	TLEN	SEQ	QUAL	OPT

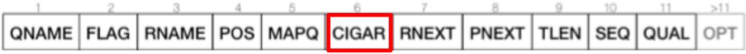
2nd field: binary FLAG

Binary (Decimal)	Hex	Description
00000000001 (1)	0x1	Is the read paired?
00000000010 (2)	0x2	Are both reads in a pair mapped "properly" (i.e., in the correct orientation with respect to one another)?
00000000100 (4)	0x4	Is the read itself unmapped?
00000001000 (8)	0x8	Is the mate read unmapped?
00000010000 (16)	0x10	Has the read been mapped to the reverse strand?
00000100000 (32)	0x20	Has the mate read been mapped to the reverse strand?
00001000000 (64)	0x40	Is the read the first read in a pair?
00010000000 (128)	0x80	Is the read the second read in a pair?
00100000000 (256)	0x100	Is the alignment not primary? (A read with split matches may have multiple primary alignment records.)
01000000000 (512)	0x200	Does the read fail platform/vendor quality checks?
10000000000 (1024)	0x400	Is the read a PCR or optical duplicate?

The FLAG field includes information about the mapping of the individual read. It is a bitwise flag, compactly storing answers to multiple binary Yes/No questions as a short series of bits where each of the single bits can be addressed separately.

See https://broadinstitute.github.io/picard/explain-flags.html to interpret bit flag values.

CIGAR [Concise Idiosyncratic Gapped Alignment Report string]



1	2	3	4	5	6	7	8	9	10	11	>11
QNAME	FLAG	RNAME	POS	MAPQ	CIGAR	RNEXT	PNEXT	TLEN	SEQ	QUAL	OPT

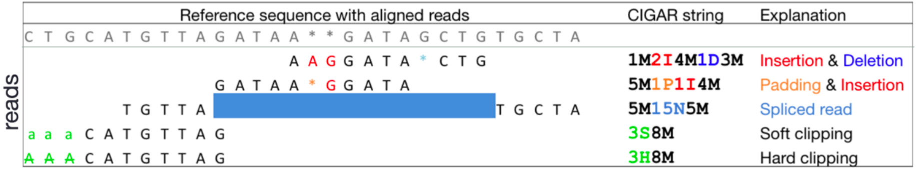
6th field: CIGAR string – which hoops did the aligner have to jump through to align the read to the genome locus that it thought was the best fit?

M alignment (match or **mis**match!!)

I (N) insertion (large insertions) ← spliced out introns = sequences are missing in the read, i.e., they need to be inserted in order to align the read to the genome

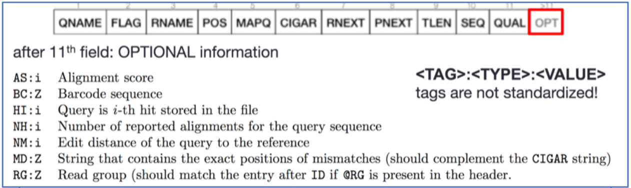
D deletion

S/H clipping



Reference sequence with aligned reads	CIGAR string	Explanation
CTGCATGTTAGATAA**GATAGCTGTGCTA		
AAGGATA*CTG	1M2I4M1D3M	Insertion & Deletion
GATAA*GGATA	5M1P1I4M	Padding & Insertion
TGTTA TGCTA	5M15N5M	Spliced read
aaaCATGTTAG	3S8M	Soft clipping
AAACATGTTAG	3H8M	Hard clipping

SAM OPT field

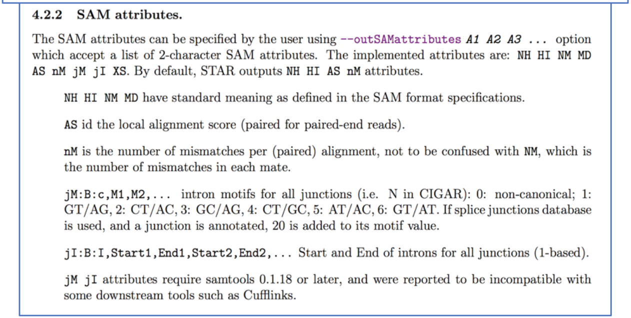


1	2	3	4	5	6	7	8	9	10	11	>11
QNAME	FLAG	RNAME	POS	MAPQ	CIGAR	RNEXT	PNEXT	TLEN	SEQ	QUAL	OPT

after 11th field: OPTIONAL information

<TAG>:<TYPE>:<VALUE>
tags are not standardized!

AS:i	Alignment score
BC:Z	Barcode sequence
HI:i	Query is *i*-th hit stored in the file
NH:i	Number of reported alignments for the query sequence
NM:i	Edit distance of the query to the reference
MD:Z	String that contains the exact positions of mismatches (should complement the **CIGAR** string)
RG:Z	Read group (should match the entry after ID if **@RG** is present in the header.



4.2.2 SAM attributes.

The SAM attributes can be specified by the user using `--outSAMattributes` *A1 A2 A3* ... option which accept a list of 2-character SAM attributes. The implemented attributes are: NH HI NM MD AS nM jM jI XS. By default, STAR outputs NH HI AS nM attributes.

NH HI NM MD have standard meaning as defined in the SAM format specifications.

AS id the local alignment score (paired for paired-end reads).

nM is the number of mismatches per (paired) alignment, not to be confused with NM, which is the number of mismatches in each mate.

jM:B:c,M1,M2,... intron motifs for all junctions (i.e. N in CIGAR): 0: non-canonical; 1: GT/AG, 2: CT/AC, 3: GC/AG, 4: CT/GC, 5: AT/AC, 6: GT/AT. If splice junctions database is used, and a junction is annotated, 20 is added to its motif value.

jI:B:I,Start1,End1,Start2,End2,... Start and End of introns for all junctions (1-based).

jM jI attributes require samtools 0.1.18 or later, and were reported to be incompatible with some downstream tools such as Cufflinks.

The number of optional SAM/BAM fields, their value types and the information stored within them depends on the alignment program and can vary substantially.

Exploring SAM/BAM files

The most widely used tool to explore and manipulate SAM/BAM files is `samtools`.

There are many options to subset reads based on SAM fields such as chromosomal location, or FLAG value, or mapping quality.

```
samtools view <in.bam>
```

Use `egrep` to subset reads based on the optional tags.

Most downstream applications also require the BAM file to be indexed by reference sequence position, to allow the efficient retrieval of all reads aligning to a locus.

```
samtools index <in.bam>
```

References

Sara Ballouz, Alexander Dobin, Thomas R Gingeras, and Jesse Gillis. The fractured landscape of RNA-seq alignment: the default in our STARs. *Nucleic Acids Research*, 46(10):5125–5138, 05 2018. doi: 10.1093/nar/gky325. URL https://dx.doi.org/10.1093/nar/gky325.

Sara Ballouz, Alexander Dobin, and Jesse Gillis. Is it time to change the reference genome? *bioRxiv*, 2019. doi: 10.1101/533166. URL https://www.biorxiv.org/content/early/2019/01/29/533166.

Alexander Dobin, Carrie A. Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R. Gingeras. STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics*, 29 (1):15–21, 2013. doi: 10.1093/bioinformatics/bts635.

Paolo Ferragina and Giovanni Manzini. Opportunistic Data Structures with Applications. Technical report, 2010.

Jürgen Jänes, Fengyuan Hu, Alexandra Lewin, and Ernest Turro. A comparative study of RNA-seq analysis strategies. *Briefings in Bioinformatics*, (January):1–9, 2015. ISSN 1467-5463. doi: 10.1093/bib/bbv007.

Heng Li. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv e-prints*, art. arXiv:1303.3997, March 2013.

Heng Li and Richard Durbin. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14):1754–1760, 05 2009. doi: 10.1093/bioinformatics/btp324.

Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, and Richard Durbin. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–9, August 2009. ISSN 1367-4811. doi: 10.1093/bioinformatics/btp352.

S.P. Pfeifer. From next-generation resequencing reads to a high-quality variant data set. *Heredity*, 118(2):111–124, 2017. doi: 10.1038/hdy.2016.102.

Knut Reinert, Ben Langmead, David Weese, and Dirk J. Evers. Alignment of next-generation sequencing reads. *Annual Review of Genomics and Human Genetics*, 16:133–151, 8 2015. doi: 10.1146/annurev-genom-090413-025358.

Po-Yen Wu, John H. Phan, and May D. Wang. Assessing the impact of human genome annotation choice on RNA-seq expression estimates. *BMC Bioinformatics*, 14(11):S8, Nov 2013. doi: 10.1186/1471-2105-14-S11-S8.

Shanrong Zhao and Baohong Zhang. A comprehensive evaluation of Ensembl, RefSeq, and UCSC annotations in the context of RNA-seq read mapping and gene quantification. *BMC Genomics*, 16(1):97, Feb 2015. doi: 10.1186/s12864-015-1308-8.