

Filetypes for Annotation & Alignment

J Fass | 20 June 2017

Filetypes

- Fasta
- Fastq
- GTF / GFF
- SAM / BAM / CRAM
- BED

See <https://genome.ucsc.edu/FAQ/FAQformat.html>

Fastq

Oct	Dec	Hex	Char
112	74	4A	J

$$74 - 33 = 41$$

Probability of error = $10^{-41/10} \sim 0.0001$

41 is the “phred-scaled Q-value”

Standard FASTQ encodes qualities using “phred + 33” quality characters. See https://en.wikipedia.org/wiki/FASTQ_format for a good graphic about current and older encodings.

Common QC question: “how many reads have average of at least Q30?”

GTF / GFF

Gene Transfer Format / Gene Feature Format ... describes gene locations within the genome. If we align reads to the genome, reads that align to gene regions must have come from mRNA from that gene, yes?

GTF / GFF

- Human/mouse: GENCODE (uses Ensembl IDs) (<http://www.genencodegenes.org/>), but may need some manipulation to work with certain software
- Ensembl genomes (<http://ensemblgenomes.org/>) and Biomart (<http://www.ensembl.org/biomart/martview/>)
- Illumina igenomes (http://support.illumina.com/sequencing/sequencing_software/igenome.html) provides indexes for some software, and files with extra info for Tophat/cufflinks.
- NCBI genomes (<http://www.ncbi.nlm.nih.gov/genome/>)
- Many specialized databases (Phytozome, Patric, VectorBase, FlyBase, WormBase)
- “Do it yourself” genome assembly and gene-finding (don’t forget functional annotation)

GTF / GFF

chr12	unknown exon	4382902	4383401 .	+	.
chr12	unknown CDS	4383207	4383401 .	+	.
chr12	unknown start_codon	4383207	4383209 .	+	.
chr12	unknown CDS	4385171	4385386 .	+	.
chr12	unknown exon	4385171	4385386 .	+	.
chr12	unknown CDS	4387926	4388085 .	+	.
chr12	unknown exon	4387926	4388085 .	+	.
chr12	unknown CDS	4398008	4398156 .	+	.
chr12	unknown exon	4398008	4398156 .	+	.
chr12	unknown CDS	4409026	4409172 .	+	.
chr12	unknown exon	4409026	4414522 .	+	.
chr12	unknown stop_codon	4409173	4409175 .	+	.

The left columns list source, feature type, and genomic coordinates

gene_id "CCND2"; gene_name "CCND2"; p_id "P6197"; transcript_id "NM_001759"; tss_id "TSS231";
gene_id "CCND2"; gene_name "CCND2"; p_id "P6197"; transcript_id "NM_001759"; tss_id "TSS231";
gene_id "CCND2"; gene_name "CCND2"; p_id "P6197"; transcript_id "NM_001759"; tss_id "TSS231";
gene_id "CCND2"; gene_name "CCND2"; p_id "P6197"; transcript_id "NM_001759"; tss_id "TSS231";
gene_id "CCND2"; gene_name "CCND2"; p_id "P6197"; transcript_id "NM_001759"; tss_id "TSS231";
gene_id "CCND2"; gene_name "CCND2"; p_id "P6197"; transcript_id "NM_001759"; tss_id "TSS231";
gene_id "CCND2"; gene_name "CCND2"; p_id "P6197"; transcript_id "NM_001759"; tss_id "TSS231";
gene_id "CCND2"; gene_name "CCND2"; p_id "P6197"; transcript_id "NM_001759"; tss_id "TSS231";
gene_id "CCND2"; gene_name "CCND2"; p_id "P6197"; transcript_id "NM_001759"; tss_id "TSS231";
gene_id "CCND2"; gene_name "CCND2"; p_id "P6197"; transcript_id "NM_001759"; tss_id "TSS231";
gene_id "CCND2"; gene_name "CCND2"; p_id "P6197"; transcript_id "NM_001759"; tss_id "TSS231";
gene_id "CCND2"; gene_name "CCND2"; p_id "P6197"; transcript_id "NM_001759"; tss_id "TSS231";

The right column includes attributes, including gene ID, etc.

GTF

Sequence Name (i.e., chromosome, scaffold, etc.)	chr12
Source (program that generated the gtf file or feature)	unknown
Feature (i.e., gene, exon, CDS, start codon, stop codon)	CDS
Start (starting location on sequence)	3677872
End (end position on sequence)	3678014
Score	.
Strand (+ or -)	+
Frame (0, 1, or 2: which is first base in codon, zero-based)	2
Attribute (“;”-delimited list of tags with additional info)	gene_id "PRMT8"; gene_name "PRMT8"; p_id "P10933"; transcript_id "NM_019854"; tss_id "TSS4368";

GTF file with questionable attributes ...

```
gene_id "AAEL005599";  
gene_id "AAEL005599"; transcript_id "AAEL005599-RA";  
gene_id "AAEL005599"; transcript_id "AAEL005599-RA"; exon_number "1 of 4";  
gene_id "AAEL005599"; transcript_id "AAEL005599-RA"; exon_number "2 of 4";  
gene_id "AAEL005599"; transcript_id "AAEL005599-RA"; exon_number "3 of 4";  
gene_id "AAEL005599"; transcript_id "AAEL005599-RA"; exon_number "4 of 4";  
transcript_id "AAEL005599-RA";  
gene_id "AAEL005599"; transcript_id "AAEL005599-RA";  
gene_id "AAEL005599"; transcript_id "AAEL005599-RA"; exon_number "1 of 4";  
gene_id "AAEL005599"; transcript_id "AAEL005599-RA"; exon_number "2 of 4";  
gene_id "AAEL005599"; transcript_id "AAEL005599-RA"; exon_number "3 of 4";  
gene_id "AAEL005599"; transcript_id "AAEL005599-RA"; exon_number "4 of 4";  
gene_id "AAEL005599"; transcript_id "AAEL005599-RA";  
gene_id "AAEL016379"; transcript_id "AAEL005599-RA"; exon_number "5 of 4";  
gene_id "AAEL016380"; transcript_id "AAEL005599-RA"; exon_number "6 of 4";
```

SAM / BAM / CRAM!

<http://www.htslib.org/>

See also samtools man page: <http://samtools.sourceforge.net/>

SAM spec grew out of 1000 Genomes Project (see Li et al. 2009 *Bioinformatics* 25:2078)

SAM is plain text; BAM is binary, compressed version of SAM; CRAM is further compressed but not widely used / recognizable by many tools.

SAM

QNAME: Query name

Read IDs are truncated at first whitespace (spaces / tabs), which can make them *non-unique*. Illumina reads with older IDs have trailing “/1” and “/2” stripped (this information is recorded in the next field). Illumina reads with newer IDs have second block stripped (read number is recorded in the next field).

@FCC6889ACXX:5:1101:8446:45501#CGATGTATC/1 ⇒ @FCC6889ACXX:5:1101:8446:45501

@HISEQ:153:H8ED7ADXX:1:1101:1368:2069 1:N:0:ATCACG ⇒ @HISEQ:153:H8ED7ADXX:1:1101:1368:2069

HISEQ:153:H8ED7ADXX:1:1104:8193:69947

99

chr1

4773690

50

101M

=

4773721

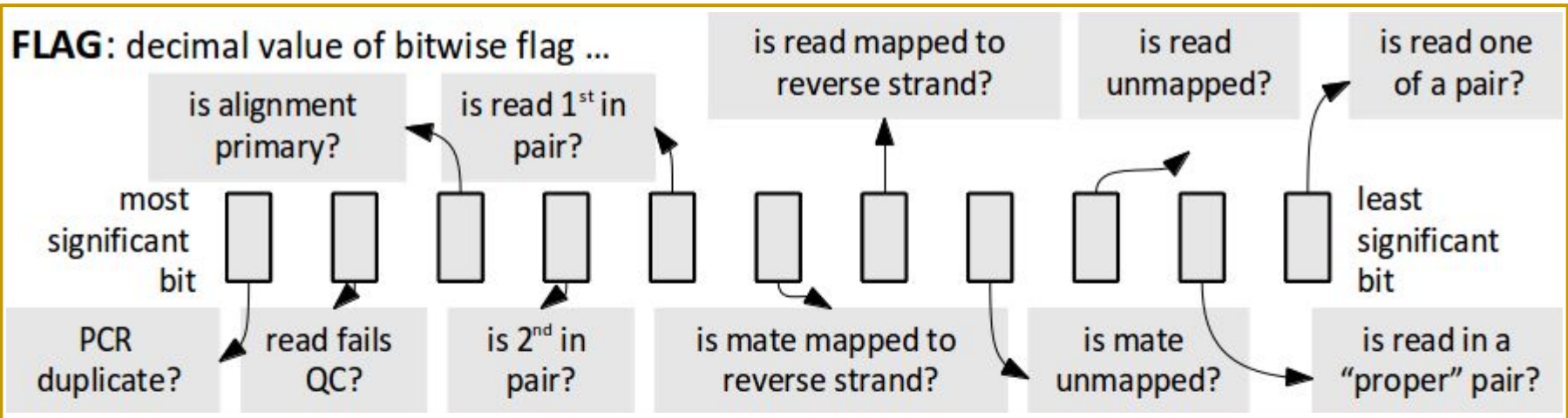
132

GTGCCATCTGTGGGCTGGTGATC[...]AGCAGCATGCTCCATGGTCTCTACATG

BBBFFFFFFBFFFFFFBFFBFFBB[...]FFBFFBBBBBBBBBBBBBBBBBBBBB<

AS:i:0 XN:i:0 XM:i:0 XO:i:0 XG:i:0 NM:i:0 MD:Z:101 YT:Z:UU XS:A:- NH:i:1

FLAG: decimal value of bitwise flag ...



HISEQ:153:H8ED7ADXX:1:1104:8193:69947

99

chr1
4773690

50
101M

=
4773721

132

GTGCCATCTGTGGGCTGGTGATC[...]AGCAGCATGCTCCATGGTCTCTACATG

BBBFFFFFFBFFFFFFBFFBFFBB[...]FFBFFBBBBBBBBBBBBBBBBBBBBB<

AS:i:0 XN:i:0 XM:i:0 XO:i:0 XG:i:0 NM:i:0 MD:Z:101 YT:Z:UU XS:A:- NH:i:1

99 (decimal) = 00001100011 (binary) (0 / NO .. 1 / YES)

... so, (from right to left): read is in a pair; the pair is proper; read *is* mapped (double neg); mate *is* mapped (double neg); read is mapped to forward strand (double neg); mate is mapped to reverse strand; read is 1st in pair ... *remaining bits not used*

SAM

FLAG: still confused?

<https://broadinstitute.github.io/picard/explain-flags.html>

Common flags for SR (single reads): 0, 4, 16, sometimes 20 (hmm..)

Common flags for PE (paired ends): 99/147, 83/163, 77/141, 65/129, 81/161 ...

```
HISEQ:153:H8ED7ADXX:1:1104:8193:69947
```

```
99
```

```
chr1
```

```
4773690
```

```
50
```

```
101M
```

```
=
```

```
4773721
```

```
132
```

```
GTGCCATCTGTGGGCTGGTGATC[...]AGCAGCATGCTCCATGGTCTCTACATG
```

```
BBBFFFFFFBFFFFFFBFFBFFBB[...]FFBFFBBBBBBBBBBBBBBBBBBBBB<
```

```
AS:i:0 XN:i:0 XM:i:0 XO:i:0 XG:i:0 NM:i:0 MD:Z:101 YT:Z:UU XS:A:- NH:i:1
```

SAM

RNAME: reference sequence name

Reference sequence ID (from fasta header), *possibly truncated at first whitespace (still unique??)*

```
>chromosome 1
... becomes ...
chromosome
... (!)
```

HISEQ:153:H8ED7ADXX:1:1104:8193:69947

99

chr1

4773690

50

101M

=

4773721

132

GTGCCATCTGTGGGCTGGTGATC[...]AGCAGCATGCTCCATGGTCTCTACATG

BBBFFFFFFBFFFFFFBFFBFFBB[...]FFBFFBBBBBBBBBBBBBBBBBBBBB<

AS:i:0 XN:i:0 XM:i:0 XO:i:0 XG:i:0 NM:i:0 MD:Z:101 YT:Z:UU XS:A:- NH:i:1

SAM

POS: 1-based *leftmost* position of (post-clipping) aligned read

... 4,773,680 4,773,690 4,773,700 4,773,710 ...

REF: ... TACCCAATGGGGATGACATAAGGTGCCATCTGTGGGCTGGTGATTCCATAGTAGAC ...

READ: GGTGCCATCTGTGGGCTGGTGATCCCATAGTAGAC ...

HISEQ:153:H8ED7ADXX:1:1104:8193:69947

99

chr1

4773690

50

101M

=

4773721

132

GTGCCATCTGTGGGCTGGTGATC[...]AGCAGCATGCTCCATGGTCTCTACATG

BBBFFFFFFBFFFFFFBFFBFFBB[...]FFBFFBBBBBBBBBBBBBBBBBBBBB<

AS:i:0 XN:i:0 XM:i:0 XO:i:0 XG:i:0 NM:i:0 MD:Z:101 YT:Z:UU XS:A:- NH:i:1

SAM

POS: 1-based *leftmost* position of (post-clipping) aligned read

... 4,773,680 4,773,690 4,773,700 4,773,710 ...

REF: ... CCAATGGGGATGACATAAGTGCCATCTGTGGGCTGGTGATCAGTAGAC ...

READ: GTGCCATCTGTGGGCTGGTGATCAGTAGAC ...

HISEQ:153:H8ED7ADXX:1:1104:8193:69947

99

chr1

4773690

50

49H101M

=

4773721

132

GTGCCATCTGTGGGCTGGTGATC[...]AGCAGCATGCTCCATGGTCTCTACATG

BBBFFFFFFBFFFFFFBFFBFFBB[...]FFBFFBBBBBBBBBBBBBBBBBBBBB<

AS:i:0 XN:i:0 XM:i:0 XO:i:0 XG:i:0 NM:i:0 MD:Z:101 YT:Z:UU XS:A:- NH:i:1

SAM

POS: 1-based *leftmost* position of (post-clipping) aligned read

... 4,773,680 4,773,690 4,773,700 4,773,710 ...

REF: ... CCAATGGGGATGACATAAGTGCCATCTGTGGGCTGGTGATCAGTAGAC ...

READ: GCCGTGCCATCTGTGGGCTGGTGATCAGTAGAC ...

HISEQ:153:H8ED7ADXX:1:1104:8193:69947

99

chr1

4773690

50

3S101M

=

4773721

132

GCCGTGCCATCTGTGGGCTGGTGATC[...]AGCAGCATGCTCCATGGTCTCTACATG

BBBBBBFFFFFFBFFFFFFBFFB[...]FFBFFBBBBBBBBBBBBBBBBBBBBB<

AS:i:0 XN:i:0 XM:i:0 XO:i:0 XG:i:0 NM:i:0 MD:Z:GCC101(?) YT:Z:UU XS:A:- NH:i:1

SAM

MAPQ: mapping quality (phred scaled)

Mapping quality is used by some aligners, in different ways. It's generally a function of the edit distance (mismatches, indels), and the uniqueness of the alignment. Multiple equivalent best alignments yield a mapping quality of zero; alignments with an edit distance close to the best alignment lower the mapping quality.

```
HISEQ:153:H8ED7ADXX:1:1104:8193:69947
```

```
99
```

```
chr1
```

```
4773690
```

```
50
```

```
101M
```

```
=
```

```
4773721
```

```
132
```

```
GTGCCATCTGTGGGCTGGTGATC[...]AGCAGCATGCTCCATGGTCTCTACATG
```

```
BBBFFFFFFBFFFFFFBFFBFFBB[...]FFBFFBBBBBBBBBBBBBBBBBBBBB<
```

```
AS:i:0 XN:i:0 XM:i:0 XO:i:0 XG:i:0 NM:i:0 MD:Z:101 YT:Z:UU XS:A:- NH:i:1
```

SAM

CIGAR: extended CIGAR string (Compact Idiosyncratic Gapped Alignment Report)

Format: [0-9][MIDNSHP][0-9][MIDNSHP]...

M = match / mismatch (!), I/D = insertion / deletion, N = skipped bases on reference, S/H = soft / hard clip (hard clipped bases no longer appear in the sequence field), P = padding

... e.g. "101M" means that all bases in the read align to bases in the reference, starting with position (4,773,690), always in the order of the reference.

```
HISEQ:153:H8ED7ADXX:1:1104:8193:69947
```

```
99
```

```
chr1
```

```
4773690
```

```
50
```

```
101M
```

```
=
```

```
4773721
```

```
132
```

```
GTGCCATCTGTGGGCTGGTGATC[...]AGCAGCATGCTCCATGGTCTCTACATG
```

```
BBBFFFFFFBFFFFFFBFFBFFBB[...]FFBFFBBBBBBBBBBBBBBBBBBBBB<
```

```
AS:i:0 XN:i:0 XM:i:0 XO:i:0 XG:i:0 NM:i:0 MD:Z:101 YT:Z:UU XS:A:- NH:i:1
```


SAM

MRNM: reference sequence to which the *mate* of this read is aligned

“=” ... mate is aligned to the same reference sequence as this read

“*” ... this is a single read; no mate exists

```
HISEQ:153:H8ED7ADXX:1:1104:8193:69947
```

```
99
```

```
chr1
```

```
4773690
```

```
50
```

```
101M
```

```
=
```

```
4773721
```

```
132
```

```
GTGCCATCTGTGGGCTGGTGATC[...]AGCAGCATGCTCCATGGTCTCTACATG
```

```
BBBFFFFFFBFFFFFFBFFBFFBB[...]FFBFFBBBBBBBBBBBBBBBBBBBBB<
```

```
AS:i:0 XN:i:0 XM:i:0 XO:i:0 XG:i:0 NM:i:0 MD:Z:101 YT:Z:UU XS:A:- NH:i:1
```

SAM

MPOS: 1-based, left-most position of 1st (post-clipping) nucleotide of mate read

“0” ... no mate exists

HISEQ:153:H8ED7ADXX:1:1104:8193:69947

99

chr1

4773690

50

101M

=

4773721

132

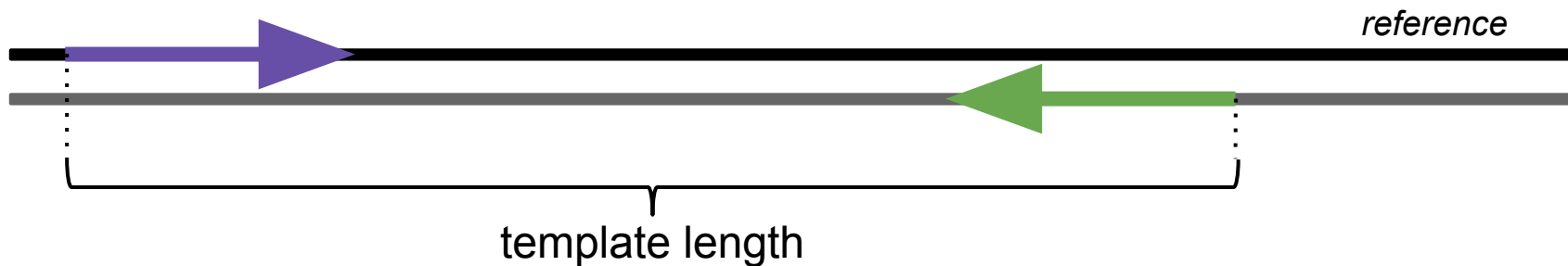
GTGCCATCTGTGGGCTGGTGATC[...]AGCAGCATGCTCCATGGTCTCTACATG

BBBFFFFFFBFFFFFFBFFBFFBB[...]FFBFFBBBBBBBBBBBBBBBBBBBBB<

AS:i:0 XN:i:0 XM:i:0 XO:i:0 XG:i:0 NM:i:0 MD:Z:101 YT:Z:UU XS:A:- NH:i:1

SAM

TLEN: inferred insert size / template length ... "0" if no mate ... "-#" if second read(?)



```
HISEQ:153:H8ED7ADXX:1:1104:8193:69947
```

```
99
```

```
chr1
```

```
4773690
```

```
50
```

```
101M
```

```
=
```

```
4773721
```

```
132
```

```
GTGCCATCTGTGGGCTGGTGATC[...]AGCAGCATGCTCCATGGTCTCTACATG
```

```
BBBFFFFFFBFFFFFFBFFBFFBB[...]FFBFFBBBBBBBBBBBBBBBBBBBBB<
```

```
AS:i:0 XN:i:0 XM:i:0 XO:i:0 XG:i:0 NM:i:0 MD:Z:101 YT:Z:UU XS:A:- NH:i:1
```

SAM

SEQ and **QUAL**: read's nucleotides and base qualities, *always in the order of the reference (forward, top) strand!* ... includes any insertions, deletions, etc. present in the read.

Reads aligned to reverse strand appear in reverse, with reversed base qualities.

```
HISEQ:153:H8ED7ADXX:1:1104:8193:69947
```

```
99
```

```
chr1
```

```
4773690
```

```
50
```

```
101M
```

```
=
```

```
4773721
```

```
132
```

```
GTGCCATCTGTGGGCTGGTGATC[... ]AGCAGCATGCTCCATGGTCTCTACATG
```

```
BBBFFFFFFBFFFFFFBFFBFFBB[... ]FFBFFBBBBBBBBBBBBBBBBBBBBBBB<
```

```
AS:i:0 XN:i:0 XM:i:0 XO:i:0 XG:i:0 NM:i:0 MD:Z:101 YT:Z:UU XS:A:- NH:i:1
```

SAM

OPT: various pre-defined and user-defined tags in the format TAG:VTYPE:VALUE ...
VTYPE is one of [A (printable character); i (signed integer); f (floating point); z (printable string); H (hex string)].

e.g.: NM:i:0 means zero mismatches in this alignment

e.g.: XS:A:- was set by TopHat, RNA that was read was coded by the reverse strand

e.g.: NH:i:1 means that the number of hits for this read was 1 (would be more for repeat)

HISEQ:153:H8ED7ADXX:1:1104:8193:69947

99

chr1

4773690

50

101M

=

4773721

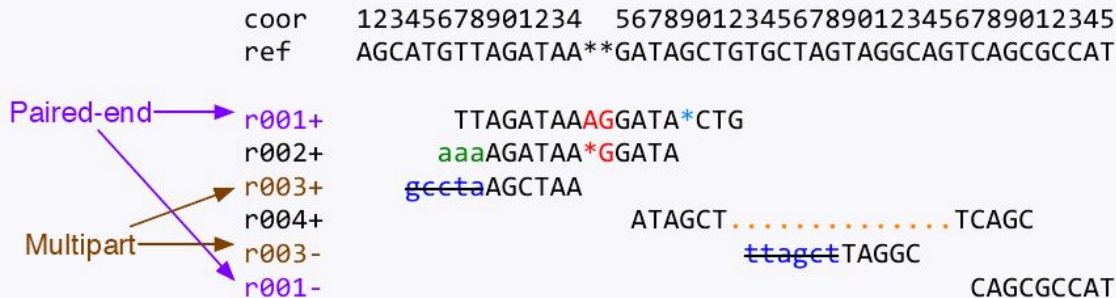
132

GTGCCATCTGTGGGCTGGTGATC[...]AGCAGCATGCTCCATGGTCTCTACATG

BBBFFFFFFBFFFFFFBFFBFFBB[...]FFBFFBBBBBBBBBBBBBBBBBBBBB<

AS:i:0 XN:i:0 XM:i:0 XO:i:0 XG:i:0 NM:i:0 MD:Z:101 YT:Z:UU XS:A:- NH:i:1

SAM - quick summary



Ins & padding
 Soft clipping

Splicing
 Hard clipping

```

@SQ SN:ref LN:45
r001 163 ref 7 30 8M2I4M1D3M = 37 39 TTAGATAAAGGATACTA *
r002 0 ref 9 30 3S6M1P1I4M * 0 0 AAAAGATAAGGATA *
r003 0 ref 9 30 5H6M * 0 0 AGCTAA * NM:i:1
r004 0 ref 16 30 6M14N5M * 0 0 ATAGCTTCAGC *
r003 16 ref 29 30 6H5M * 0 0 TAGGC * NM:i:0
r001 83 ref 37 30 9M = 7 -39 CAGCGCCAT *
  
```

```

ref 7 T 1 . | ref 12 T 3 ... | ref 17 T 3 ...
ref 8 T 1 . | ref 13 A 3 ... | ref 18 A 3 -.1G..
ref 9 A 3 ... | ref 14 A 2 .+2AG.+1G. | ref 19 G 2 *.
ref 10 G 3 ... | ref 15 G 2 .. | ref 20 C 2 ..
ref 11 A 3 ..C | ref 16 A 3 ... | ...
  
```

google "Heng Li slides" - Challenges and Solutions in the Analysis of Next Generation Sequencing Data (2010)

BAM

BAMs are compressed SAMs (so, binary, not human-readable text ... don't look directly at them!). They can be indexed to allow rapid extraction of information, so alignment viewers do not need to uncompress the whole BAM file in order to look at information for a particular read or coordinate range, somewhere in the file.

Indexing your BAM file, `myCoolBamFile.bam`, will create an index file, `myCoolBamFile.bam.bai`, which is needed (in addition to the BAM file) by viewers and other downstream tools. An occasional downstream tool will require an index called `myCoolBamFile.bai` (notice that the “.bai” replaces the “.bam”, instead of being appended after it).

CRAM

Available as of SAMtools 1.0, and is a binary format like BAM. Uses data-specific compression tools (i.e. compressing letters is different than compressing numbers), *specifically* reference-based compression (e.g. for aligned reads, only *mis-matching* bases need to be stored). Also can employ *lossy* compression of base qualities, which appears to have a negligible effect on, say, variant calling (see Illumina [white paper](#)).

Indexing your CRAM file, myCoolBamFile.cram, will create an index file, myCoolBamFile.cram.crai, which is needed (in addition to the CRAM file) by viewers and other downstream tools.

This is still a ***relatively recent development***, so it may be a while before many tools are CRAM-capable.

?’s ...?