Filetypes for Annotation & Alignment

J Fass | 22 August 2017

Filetypes

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

See *https://genome.ucsc.edu/FAQ/FAQformat.html*

Fasta

```
>sequenceName | plus other junk | few maintain a standard here
AGTGGAGCAAGCACAGAGAAGAAACTGCAGTCAGGACATAAAGTAAAGTA
ATTAATCTAAAAAATAGTCTGAGCAGTCTTCTCTGCTGAANNNNNNNNNN
NNNNNNNNNNNNNNNNNNNAATTTCCTTACTAGGAGGTCTTTAGTACAGA
TTCCTGATATGTAATTAATCACTAAATGTCTTTAATGGGATCTCTTTCTA
TTGAGATATTTGTAAACTTTCTTCATGTGATTGGTTTACAGATATTCAGG
TTTCTGCAAATGGGTGCTGTCTATATTATAGAATTTTTAGTTGAAATTTT
CAAAATACTCTTTGagtattctcttgtaattatattactttacaaggttt
gtggggcatctctttcatttgtgattacatggttgcagtattctttttgt
tcttagtcagactgtataattgtctgtgaagtccagtaaacttttgaaag
```

Assembly gap?

Soft-masked repetitive sequence? Low confidence assembly?

Fastq

```
@K00188:264:HG3WJBBXX:1:1101:6289:1595 1:N:0:TAGCTT
GGACTGCCTTTCAGCCCGTCGCAGAGGGAATGGGAGCCTCTGGAGCGGGTGCAGAGGCTCAGCAG
+
AAFFFJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJAJJJJJJ

@header
<sequence>
+(sometimes header?)
<base qualities>
```

Fastq

```
@K00188:264:HG3WJBBXX:1:1101:6289:1595 1:N:0:TAGCTT
GGACTGCCTTTCAGCCCGTCGCAGAGGGAATGGGAGCCTCTGGAGCGGGTGCAGAGGCTCAGCAG
+
AAFFFJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ
```

Oct	Dec	Hex	Char	Oct	Dec	Hex	Char
000	0	00	NUL '\0' (null character)	100	64	40	@
001	1	01	SOH (start of heading)	101	65	41	A
002	2	02	STX (start of text)	102	66	42	B
003	3	03	ETX (end of text)	103	67	43	C
004	4	04	EOT (end of transmission)	104	68	44	D
005	5	05	ENQ (enquiry)	105	69	45	E
006	6	06	ACK (acknowledge)	106	70	46	F
007	7	07	BEL '\a' (bell)	107	71	47	G
010	8	08	BS '\b' (backspace)	110	72	48	H
011	9	09	HT '\t' (horizontal tab)	111	73	49	I
012	10	0A	LF '\n' (new line)	112	74	4A	J
013	11	0B	VT '\v' (vertical tab)	113	75	4B	K

Fastq

Oct	Dec	Hex	Char
112	74	4A	J

74 - 33 = 41

Probability of error = 10 ^ (-41 / 10) ~ 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. Critical for interpreting mutations' effects on protein sequence or possibly intronic and regulatory regions.

GTF / GFF - where do I get ‘em?

- Human/mouse: GENCODE (uses Ensembl IDs) (http://www.gencodegenes.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	.	+	.	gene_id "CCND2"; gene_name "CCND2"; p_id "P6197"; transcript_id "NM_001759"; tss_id "TSS231";
chr12	unknown	CDS	4383207	4383401	.	+	.	gene_id "CCND2"; gene_name "CCND2"; p_id "P6197"; transcript_id "NM_001759"; tss_id "TSS231";
chr12	unknown	start_codon	4383207	4383209	.	+	.	gene_id "CCND2"; gene_name "CCND2"; p_id "P6197"; transcript_id "NM_001759"; tss_id "TSS231";
chr12	unknown	CDS	4385171	4385386	.	+	.	gene_id "CCND2"; gene_name "CCND2"; p_id "P6197"; transcript_id "NM_001759"; tss_id "TSS231";
chr12	unknown	exon	4385171	4385386	.	+	.	gene_id "CCND2"; gene_name "CCND2"; p_id "P6197"; transcript_id "NM_001759"; tss_id "TSS231";
chr12	unknown	CDS	4387926	4388085	.	+	.	gene_id "CCND2"; gene_name "CCND2"; p_id "P6197"; transcript_id "NM_001759"; tss_id "TSS231";
chr12	unknown	exon	4387926	4388085	.	+	.	gene_id "CCND2"; gene_name "CCND2"; p_id "P6197"; transcript_id "NM_001759"; tss_id "TSS231";
chr12	unknown	CDS	4398008	4398156	.	+	.	gene_id "CCND2"; gene_name "CCND2"; p_id "P6197"; transcript_id "NM_001759"; tss_id "TSS231";
chr12	unknown	exon	4398008	4398156	.	+	.	gene_id "CCND2"; gene_name "CCND2"; p_id "P6197"; transcript_id "NM_001759"; tss_id "TSS231";
chr12	unknown	CDS	4409026	4409172	.	+	.	gene_id "CCND2"; gene_name "CCND2"; p_id "P6197"; transcript_id "NM_001759"; tss_id "TSS231";
chr12	unknown	exon	4409026	4414522	.	+	.	gene_id "CCND2"; gene_name "CCND2"; p_id "P6197"; transcript_id "NM_001759"; tss_id "TSS231";
chr12	unknown	stop_codon	4409173	4409175	.	+	.	gene_id "CCND2"; gene_name "CCND2"; p_id "P6197"; transcript_id "NM_001759"; tss_id "TSS231";

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

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

```
chr12	unknown	CDS	3677872	3678014	.	+	2	gene_id "PRMT8"; gene_name "PRMT8"; p_id "P10933"; transcript_id "NM_019854"; tss_id "TSS4368";
```

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!

Sequence **A**lignment / **M**apping

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

Header lines (start with “@”)

Alignment line (one line per alignment)

```
[...]
@SQ	SN:ctg103993	LN:217
@SQ	SN:ctg103994	LN:222
@SQ	SN:ctg103995	LN:205
@SQ	SN:ctg103996	LN:210
@PG	ID:bwa	PN:bwa	VN:0.7.13-r1126	CL:bwa mem -t 4 -M ../../01_Reference/Transcriptome-Contigs-Build2.fna ../../02-Cleaned/3E/3E_SE.fastq
@PG	ID:bwa-7BC92A6F	PN:bwa	VN:0.7.13-r1126	CL:bwa mem -t 4 -M ../../01_Reference/Transcriptome-Contigs-Build2.fna ../../02-Cleaned/3E/3E_R1.fastq ../../02-Cleaned/3E/3E_R2.fastq
K00188:264:HG3WJBBXX:1:1116:14692:35180#0	121	ctg2	128	58	101M	=	128	0	AAGTCTCGACCAAGTGGTTCAGATGGTGACACAGATGTTAGCCCCATCCACCATTCAGTTGCCGTTTTGATAGCTGGAAATCCTGTAAACACAATGCTGAG	FJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJFFFAA	NM:i:10
K00188:264:HG3WJBBXX:1:1116:14692:35180#0	181	ctg2	128	0	*	=	128	0	TTTAGTTTTAATTTTTGACTTTGAATAGCGGGAGTCCAGATCGTGTGAACACAGCAGACTGAGCACTCCATTGACAGCCTTCTTCTGTACTTTAGCTATCC	FJFJJFAAJF7F7JJJAFFFAF<7<AFFJJJFJJJJJJJJJJJJJJJJJJJJJJFJJJJJJFAJJJJJJJFFFJJJJJJJJJJFFJJJJJJFFFAA	AS:i:0	XS:i:0
K00188:264:HG3WJBBXX:1:1202:11028:9596#0	121	ctg5	45	60	101M	=	45	0	TTCTTTTTTCTACAGTTCATTGTCTGTATAAAGTATGCATCAGGAACAATCTGACTAGGAAGGTAAATAATGTAAAACAGATGATTATTGTATGAAAGTTG	JJJJJJJJJJJJJJJJJJJJJJJJFJJJJJJJJJJJJJJJJJJJJJJJJJFJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJFFFAA	NM:i:8
K00188:264:HG3WJBBXX:1:1202:11028:9596#0	181	ctg5	45	0	*	=	45	0	TCAGCTGTATTAGTAATTTAGTAGAAAAGGTCTTGAGAGAATTATGTTTTTTAAAAATCCACATCACTTCAAACAAAAAGCCCCATTAGAATGGAGGGCCA	FJFJJJJJJFJJJJJJFFJJJJFJAJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJFF-JFJJJJJJJJJJJJJJJJJJJJJJJJJJJJJFFFAA	AS:i:0
[...]
```

SAM

```
[...]
@SQ     SN:ctg103993     LN:217
@SQ     SN:ctg103994     LN:222
@SQ     SN:ctg103995     LN:205
@SQ     SN:ctg103996     LN:210
@PG     ID:bwa  PN:bwa  VN:0.7.13-r1126 CL:bwa mem -t 4 -M ../../01_Reference/Transcriptome-Contigs-Build2.fna
../../02-Cleaned/3E/3E_SE.fastq
@PG     ID:bwa-7BC92A6F PN:bwa  VN:0.7.13-r1126 CL:bwa mem -t 4 -M ../../01_Reference/Transcriptome-Contigs-Build2.fna
../../02-Cleaned/3E/3E_R1.fastq ../../02-Cleaned/3E/3E_R2.fastq
K00188:264:HG3WJBBXX:1:1116:14692:35180#0      121     ctg2      128     58      101M =       128     0
AAGTCTCGACCAAGTGGTTCAGATGGTGACACAGATGTTAGCCCATCCACCATTCAGTTGCCGTTTTGATAGCTGGAAATCCTGTAAACACAAT
GCTGAG   FJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJFFFAA   NM:i:10
K00188:264:HG3WJBBXX:1:1116:14692:35180#0      181     ctg2      128     0       *              =       128     0
TTTAGTTTTAATTTTTGACTTTGAATAGCGGGAGTCCAGATCGTGTGAACACAGCAGACTGAGCACTCCATTGACAGCCTTCTTCTGTACTTTAGC
TATCC   FJFJJFAAJF7F7JJJJAFFFAF<7<AFFJJJFJJJJJJJJJJJJJJJJJJJJJFJJJJJJJFAJJJJJJJJFFFJJJJJJJJJJJFFJJJJJJJFFFAA   AS:i:0  XS:i:0
K00188:264:HG3WJBBXX:1:1202:11028:9596#0       121     ctg5      45      60      101M   =       45      0
TTCTTTTTTCTACAGTTCATTGTCTGTATAAAGTATGCATCAGGAACAATCTGACTAGGAAGGTAAATAATGTAAAACAGATGATTATTGTATGAAA
GTTG  JJJJJJJJJJJJJJJJJJJJJJJJJFJJJJJJJJJJJJJJJJJJJJJJJJJJJFJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJFFFAA   NM:i:8
K00188:264:HG3WJBBXX:1:1202:11028:9596#0       181     ctg5      45      0       *              =       45      0
TCAGCTGTATTAGTAATTTAGTAGAAAAGGTCTTGAGAGAATTATGTTTTTTAAAAATCCACATCACTTCAAACAAAAAGCCCCATTAGAATGGAGG
GCCA   FJFJJJJJJFJJJJJJFFJJJJFJAJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJFF-JFJJJJJJJJJJJJJJJJJJJJJJJJJJJJFFFAA   AS:i:0
[...]
```

Read Pair (identical headers)

SAM

```
K00188:264:HG3WJBBXX:1:1116:14692:35180#0       121     ctg2      128     58      101M      =       128     0
AAGTCTCGACCAAGTGGTTCAGATGGTGACACAGATGTTAGCCCCATCCACCATTCAGTTGCCGTTTTGATAGCTGGAAATCCTGTAAACACAATGCTGAG
FJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJFFFAA   NM:i:10
```

<table>
<tr><th>Col</th><th>Field</th><th>Type</th><th>Regexp/Range</th><th>Brief description</th></tr>
<tr><td>1</td><td>QNAME</td><td>String</td><td>[!-?A-~]{1,254}</td><td>Query template NAME</td></tr>
<tr><td>2</td><td>FLAG</td><td>Int</td><td>[0,2^{16}-1]</td><td>bitwise FLAG</td></tr>
<tr><td>3</td><td>RNAME</td><td>String</td><td>*|[!-()+-<>-~][!-~]*</td><td>Reference sequence NAME</td></tr>
<tr><td>4</td><td>POS</td><td>Int</td><td>[0,2^{31}-1]</td><td>1-based leftmost mapping POSition</td></tr>
<tr><td>5</td><td>MAPQ</td><td>Int</td><td>[0,2^{8}-1]</td><td>MAPping Quality</td></tr>
<tr><td>6</td><td>CIGAR</td><td>String</td><td>*|([0-9]+[MIDNSHPX=])+</td><td>CIGAR string</td></tr>
<tr><td>7</td><td>RNEXT</td><td>String</td><td>*|=|[!-()+-<>-~][!-~]*</td><td>Ref. name of the mate/next read</td></tr>
<tr><td>8</td><td>PNEXT</td><td>Int</td><td>[0,2^{31}-1]</td><td>Position of the mate/next read</td></tr>
<tr><td>9</td><td>TLEN</td><td>Int</td><td>[-2^{31}+1,2^{31}-1]</td><td>observed Template LENgth</td></tr>
<tr><td>10</td><td>SEQ</td><td>String</td><td>*|[A-Za-z=.]+</td><td>segment SEQuence</td></tr>
<tr><td>11</td><td>QUAL</td><td>String</td><td>[!-~]+</td><td>ASCII of Phred-scaled base QUALity+33</td></tr>
</table>

SAM

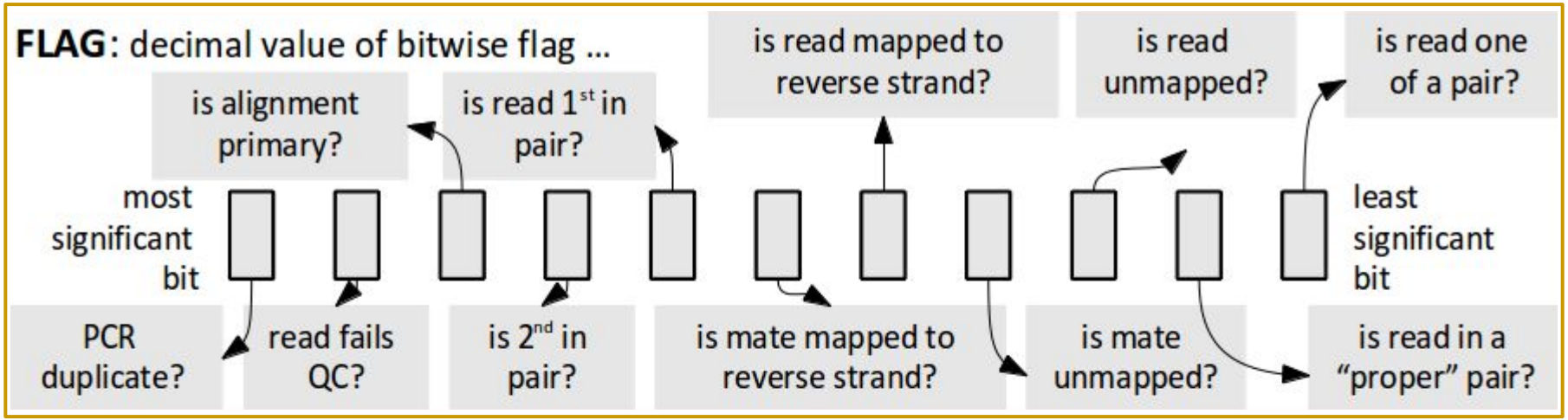
QNAME: Query name

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
BBBFFFFFFBFFFFBFFFBFFBB[...]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
```




```
HISEQ:153:H8ED7ADXX:1:1104:8193:69947
99
chr1
4773690
50
101M
=
4773721
132
GTGCCATCTGTGGGCTGGTGATC[...]AGCAGCATGCTCCATGGTCTCTACATG
BBBFFFFFFBFFFFBFFFBFFBB[...]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
BBBFFFFFFBFFFFBFFFBFFBB[...]FFBFFBBBBBBBBBBBBBBBBBBBB<
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
BBBFFFFFFBFFFFBFFFBFFBB[...]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
101M
=
4773721
132
GTGCCATCTGTGGGCTGGTGATC[...]AGCAGCATGCTCCATGGTCTCTACATG
BBBFFFFFFBFFFFBFFFBFFBB[...]FFBFFBBBBBBBBBBBBBBBBBBBB<
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
BBBFFFFFFBFFFFBFFFBFFBB[...]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
BBBBBBFFFFFFBFFFFBFFFBFFBB[...]FFBFFBBBBBBBBBBBBBBBBBBBB<
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 as a measure of confidence in 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, because they increase the possibility that the read actually arose from a position other that the reported one, but was misplaced due to errors in the read.

```
HISEQ:153:H8ED7ADXX:1:1104:8193:69947
99
chr1
4773690
50
101M
=
4773721
132
GTGCCATCTGTGGGCTGGTGATC[...]AGCAGCATGCTCCATGGTCTCTACATG
BBBFFFFFFBFFFFBFFFBFFBB[...]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
BBBFFFFFFBFFFFBFFFBFFBB[...]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 is 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
BBBFFFFFFBFFFFBFFFBFFBB[...]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

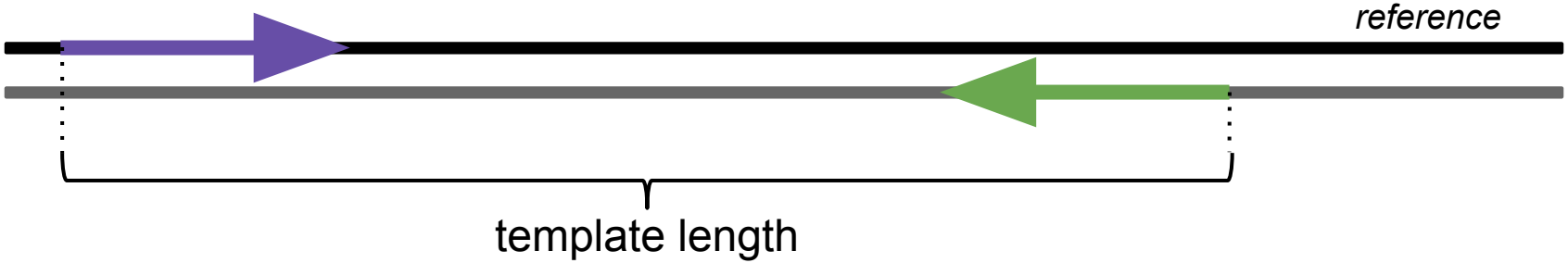
MPOS: 1-based, left-most position of 1st (post-clipping) nucleotide of mate read

"0" … no mate exists

```
HISEQ:153:H8ED7ADXX:1:1104:8193:69947
99
chr1
4773690
50
101M
=
4773721
132
GTGCCATCTGTGGGCTGGTGATC[...]AGCAGCATGCTCCATGGTCTCTACATG
BBBFFFFFFBFFFFBFFFBFFBB[...]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
BBBFFFFFFBFFFFBFFFBFFBB[...]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
BBBFFFFFFBFFFFBFFFBFFBB[...]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
```

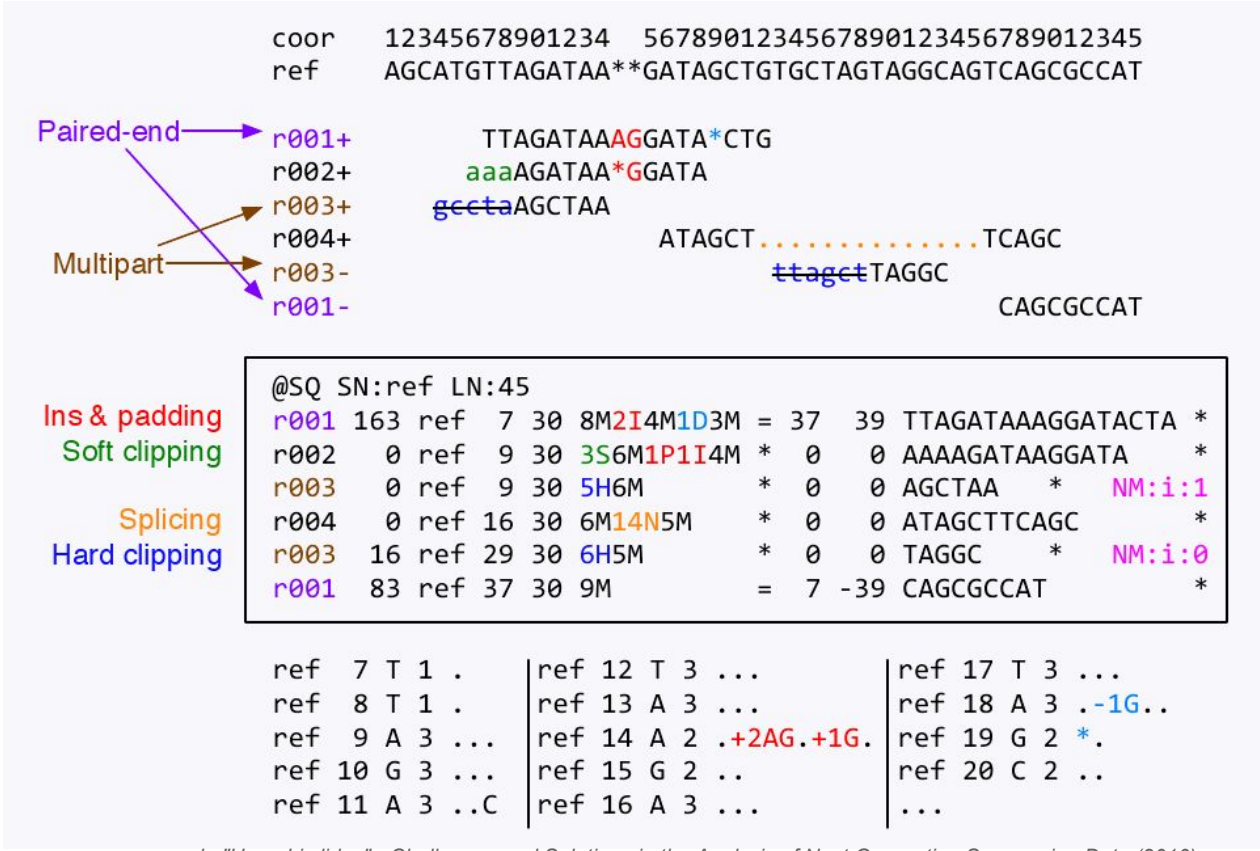
SAM

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
BBBFFFFFFBFFFFBFFFBFFBB[...]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



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

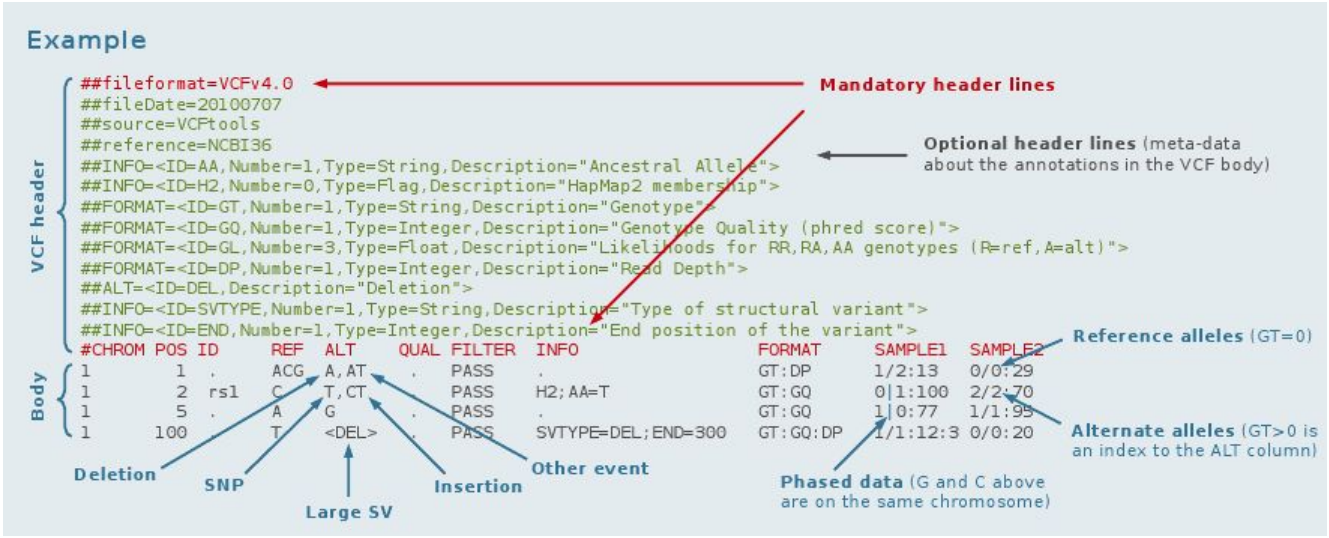
Available as of SAMtools 1.0, and is a binary format like BAM. Uses data-specific compression tools (i.e. compressing letters is different than compressing numbers), *specifically* reference-based compression (e.g. for aligned reads, only *mis-matching* bases need to be stored). Also can employ *lossy* 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.

Variant Calling - VCF format

VCF (variant call format) is now the standard format for variant reporting.



http://vcftools.sourceforge.net/specs.html ... VCF poster

Variant Call Format

```
##fileformat=VCFv4.1
##fileDate=20130825
##source=freeBayes v9.9.2-9-gfbf46fc-dirty
##reference=../results/8/8.fa
##phasing=none
##commandline="../tools/freebayes/bin/freebayes -f ../results/8/8.fa --min-alternate-fraction 0.03 --min-mapping-quality 20 --min-base-quality 20
--ploidy 1 --pooled-continuous --use-best-n-alleles 4 --use-mapping-quality --min-alternate-fraction 0.04 --min-alternate-count 1
../results/8/8.bam"
##INFO=<ID=RO,Number=1,Type=Integer,Description="Reference allele observation count, with partial observations recorded fractionally">
##INFO=<ID=AO,Number=A,Type=Integer,Description="Alternate allele observations, with partial observations recorded fractionally">
##INFO=<ID=TYPE,Number=A,Type=String,Description="The type of allele, either snp, mnp, ins, del, or complex.">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Float,Description="Genotype Quality, the Phred-scaled marginal (or unconditional) probability of the called
genotype">
##FORMAT=<ID=GL,Number=G,Type=Float,Description="Genotype Likelihood, log10-scaled likelihoods of the data given the called genotype for each
possible genotype generated from the reference and alternate alleles given the sample ploidy">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=RO,Number=1,Type=Integer,Description="Reference allele observation count">
##FORMAT=<ID=QR,Number=1,Type=Integer,Description="Sum of quality of the reference observations">
##FORMAT=<ID=AO,Number=A,Type=Integer,Description="Alternate allele observation count">
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO    FORMAT  8
8_PB1   20      .       TGTTACGCG       GCTTTGC,TGTTCTAC        27.2619 .       AO=1,2;RO=6;TYPE=complex,complex        GT:DP:RO:QR:AO:QA:GL
2:3:0:0:1,2:31,70:-4.46,-1.65,0
8_PB1   38      .       TCA     ACG,TA,AGA      0.0495692       .       AO=1,1,1;RO=3;TYPE=complex,del,mnp
        GT:DP:RO:QR:AO:QA:GL    2:6:3:101:1,1,1:31,37,34:0,-4.556,-4.004,-4.28
8_PB1   42      .       G       A       3.94171e-14     .       AO=8;RO=128;TYPE=snp    GT:DP:RO:QR:AO:QA:GL
```

Variant Call Format

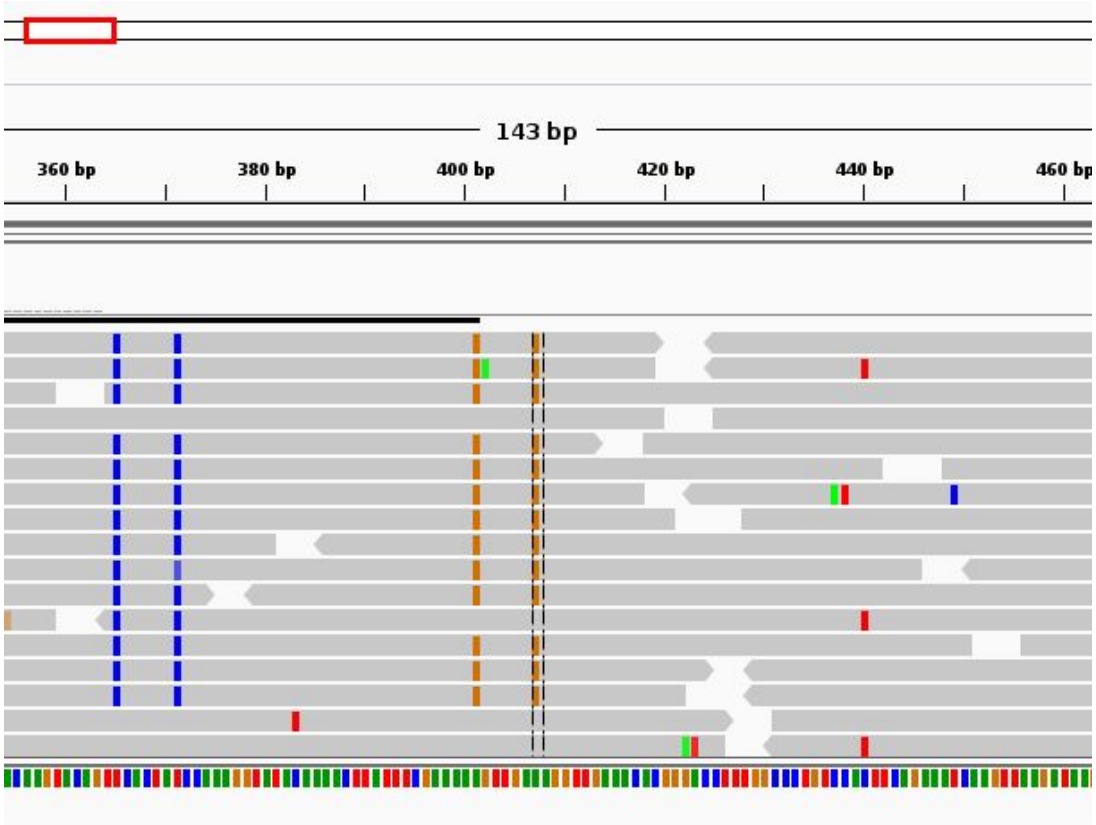
```
##fileformat=VCFv4.1
##fileDate=20130825
##source=freeBayes v9.9.2-9-gfbf46fc-dirty
##reference=../results/8/8.fa
##phasing=none
##commandline="../tools/freebayes/bin/freebayes -f ../results/8/8.fa --min-alternate-fraction 0.03 --min-mapping-quality 20 --min-base-quality 20
--ploidy 1 --pooled-continuous --use-best-n-alleles 4 --use-mapping-quality --min-alternate-fraction 0.04 --min-alternate-count 1
../results/8/8.bam"
##INFO=<ID=RO,Number=1,Type=Integer,Description="Reference allele observation count, with partial observations recorded fractionally">
##INFO=<ID=AO,Number=A,Type=Integer,Description="Alternate allele observations, with partial observations recorded fractionally">
##INFO=<ID=TYPE,Number=A,Type=String,Description="The type of allele, either snp, mnp, ins, del, or complex.">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Float,Description="Genotype Quality, the Phred-scaled marginal (or unconditional) probability of the called
genotype">
##FORMAT=<ID=GL,Number=G,Type=Float,Description="Genotype Likelihood, log10-scaled likelihoods of the data given the called genotype for each
possible genotype generated from the reference and alternate alleles given the sample ploidy">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=RO,Number=1,Type=Integer,Description="Reference allele observation count">
##FORMAT=<ID=QR,Number=1,Type=Integer,Description="Sum of quality of the reference observations">
##FORMAT=<ID=AO,Number=A,Type=Integer,Description="Alternate allele observation count">
##FORMAT=<ID=QA,Number=A,Type=Integer,Description="Sum of quality of the alternate observations">
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO    FORMAT  8
8_PB1   26      .       TGTTACGCG       GCTTTTGC,TGTTTCTAC      27.2619 .       AO=1,2;RO=0;TYPE=complex,complex        GT:DP:RO:QR:AO:QA:GL
2:3:0:0:1,2:31,70:-4.46,-1.65,0
8_PB1   38      .       TCA     ACG,TA,AGA      0.0495692       .       AO=1,1,1;RO=3;TYPE=complex,del,mnp
    GT:DP:RO:QR:AO:QA:GL    2:6:3:101:1,1,1:31,37,34:0,-4.556,-4.004,-4.28
8_PB1   42      .       G       A       3.94171e-14     .       AO=8;RO=128;TYPE=snp    GT:DP:RO:QR:AO:QA:GL
```

Variant Call Format

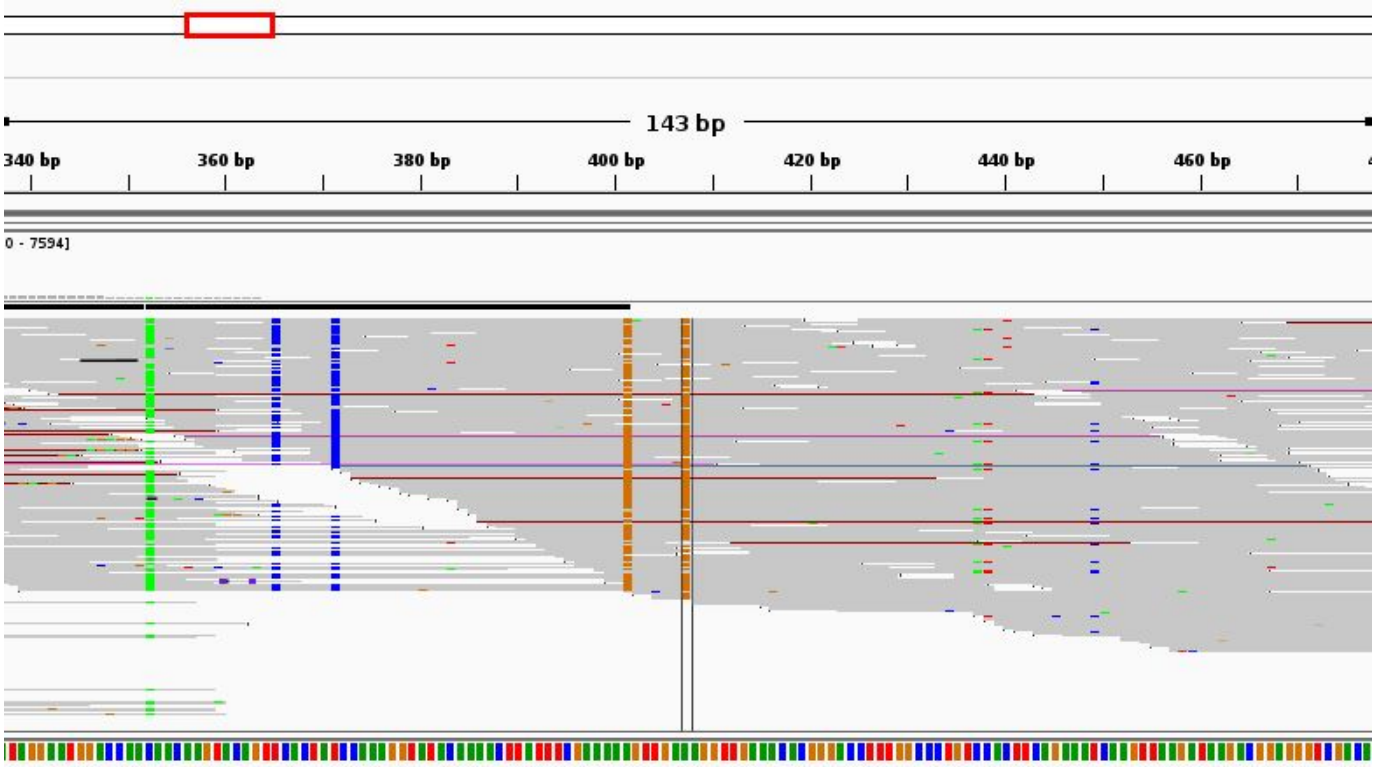
```
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO    FORMAT  8
8_PB2   407     .       A       G       3935.83 .       AO=149;RO=21;TYPE=snp   GT:DP:RO:QR:AO:QA:GL
1:170:21:788:149:5579:-5,0


CHROM  = 8_PB2

POS    = 407

ID     = .

REF    = A

ALT    = G

QUAL   = 3935.83

FILTER = .

INFO   = AO=149;RO=21;TYPE=snp

FORMAT = GT:DP:RO:QR:AO:QA:GL

8      = 1:170:21:788:149:5579:-5,0
```

Variant Call Format

Variant Call Format



Variant Call Format

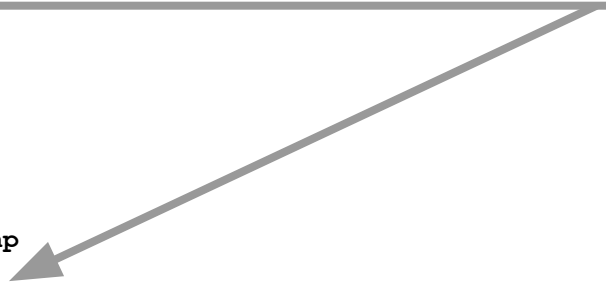


Variant Call Format

```
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO    FORMAT  8
8_PB2   407     .       A       G       3935.83 .       AO=149;RO=21;TYPE=snp   GT:DP:RO:QR:AO:QA:GL
1:170:21:788:149:5579:-5,0
```

```
CHROM  = 8_PB2
POS    = 407
ID     = .
REF    = A
ALT    = G
QUAL   = 3935.83
FILTER = .
INFO   = AO=149;RO=21;TYPE=snp
FORMAT = GT:DP:RO:QR:AO:QA:GL
8      = 1:170:21:788:149:5579:-5,0
```

```
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
```



Variant Call Format

```
##INFO=<ID=RO,Number=1,Type=Integer,Description="Reference allele
observation count, with partial observations recorded
fractionally">

##INFO=<ID=AO,Number=A,Type=Integer,Description="Alternate allele
observations, with partial observations recorded fractionally">

##INFO=<ID=TYPE,Number=A,Type=String,Description="The type of
allele, either snp, mnp, ins, del, or complex.">
```

Variant Call Format

```
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">

##FORMAT=<ID=GQ,Number=1,Type=Float,Description="Genotype
Quality, the Phred-scaled marginal (or unconditional) probability
of the called genotype">

##FORMAT=<ID=GL,Number=G,Type=Float,Description="Genotype
Likelihood, log10-scaled likelihoods of the data given the called
genotype for each possible genotype generated from the reference
and alternate alleles given the sample ploidy">

##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
```

Variant Call Format

```
##FORMAT=<ID=RO,Number=1,Type=Integer,Description="Reference
allele observation count">

##FORMAT=<ID=QR,Number=1,Type=Integer,Description="Sum of quality
of the reference observations">

##FORMAT=<ID=AO,Number=A,Type=Integer,Description="Alternate
allele observation count">

##FORMAT=<ID=QA,Number=A,Type=Integer,Description="Sum of quality
of the alternate observations">
```

Variant Effect Prediction

- snpEff
- Variant Effect Predictor (EMBL)
- SIFT

VCF after Effect Prediction

```
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO    FORMAT  SL134   SL134-2
BRD905
NC_003427.1     831     .       T       C       2105.82 test
AB=0;ABP=0;AC=6;AF=1;AN=6;ANN=C|synonymous_variant|LOW|thrA|STM0002|transcript|STM0002|protei
n_coding|1/1|c.495T>C|p.Val165Val|495/2463|495/2463|165/820||,C|upstream_gene_variant|MODIFIE
R|thrB|STM0003|transcript|STM0003|protein_coding||c.-1970T>C|||||1970|,C|upstream_gene_varian
t|MODIFIER|thrC|STM0004|transcript|STM0004|protein_coding||c.-2903T>C|||||2903|,C|downstream_
gene_variant|MODIFIER|thrL|STM0001|transcript|STM0001|protein_coding||c.*576T>C|||||576|,C|do
wnstream_gene_variant|MODIFIER|STM0005|STM0005|transcript|STM0005|protein_coding||c.*4283A>G|
||||4283|;AO=180;CIGAR=1X;DP=196;DPB=196;DPRA=0;EPP=3.44459;EPPR=4.09604;GTI=0;LEN=1;MEANALT=
1.66667;MQM=59.6;MQMR=60;NS=3;NUMALT=1;ODDS=51.9791;PAIRED=0;PAIREDR=0;PAO=0;PQA=0;PQR=0;PRO=
0;QA=2592;QR=82;RO=8;RPL=37;RPP=138.558;RPPR=7.35324;RPR=143;RUN=1;SAF=90;SAP=3.0103;SAR=90;S
RF=3;SRP=4.09604;SRR=5;TYPE=snp;technology.PacBio=1         GT:AO:DP:PL:QA:QR:RO
1/1:22:22:478,66,0:528:0:0    1/1:23:23:499,69,0:552:0:0    1/1:135:151:1287,231,0:1512:82:8
```

VCF after Effect Prediction

```
ANN=C|synonymous_variant|LOW|thrA|STM0002|transcript|STM0002|protein_coding|1/1|c.495T>C|p
.Val165Val|495/2463|495/2463|165/820||
```

C	ALT allele being annotated
synonymous_variant	Sequence Ontology term describing variant
LOW	putative impact
thrA	Gene Name (affected by variant)
STM0002	Gene ID
transcript	Feature type
STM0002	Feature ID (transcript ID in this case)
protein_coding	transcript biotype (ENSEMBL) (e.g. coding, non coding)
1/1	exon 1 of 1 total in this transcript (can alternatively rank introns)
c.495T>C	HGVS DNA-level variant description
p.Val165Val	HGVS protein-level variant description
495/2463	cDNA position (/ total length)
495/2463	CDS position (/ total length)
165/820	amino acid position (/ total length)
Empty	distance to feature
Empty	errors and warnings

?’s …?