NGS File Formats

FASTA format

- It is the most commonly used sequence file format (text).
- Originally developed for the Fast Align program.
- Output from most first generation platforms
- File can have single or multiple sequences
- Each sequence entry has two parts
  - Header – A single line always starting with '>'
  - Sequence – Starts the next line. Can be one or multiple lines

```
>D86068.1 Human immunodeficiency virus1
TGGAAGGGCTAATTCACTCCCAACGAGGACAAGATATCCTTGATCTGTGGATCTACCACACACAAGGCTACTTCCCTGATTGGCAGAACTACACAC
CAGGACCAGGGATCAGATATCCACTGACCTTTGGATGGTGCTACAAGCTAGTACCAGTTGAGCCAGAGAAGTTAGAAGAAGCCAACAAAGGAGA
GAACACCAGCTTATTACACC
>D86069.1 Human immunodeficiency virus2
TGGAAGGGCTAATTCACTCCCAACGAAGACAAGATATCCTTGATCTGTGGATCTACCACACACAAGGCTACTTCCCTGATTGGCAGAACTACACAC
CAGGACCAGGGATCAGATATCCACTGACCTTTGGATGGTGCTACAAGCTAGTACCAGTTGAGCCAGAGAAGTTAGAAGAAGCCAACAAAGGAGA
GAACACCAGCTTGTTACACC
```

FASTQ format

- Universally used for Next Generation Sequencing data.
- Each file usually contains millions of reads.
- Output from second generation platforms.
- Each entry has four rows
  - 1 – Name: starts with '@'
  - 2 – Sequence: ATGC
  - 3 – Comment field: starts with '+'
  - 4 – Quality Scores
- Line 2 and 4 are equal in length.
- Paired end reads are stored in two FastQ files.

FASTQ format

Line 1 - Name contains additional information about the read

Eg : @EAS139:136:FC706VJ:2:2104:15343:197393 1:Y:18:ATCACG

EAS139	the unique instrument name
136	the run id
FC706VJ	the flowcell id
2	flowcell lane
2104	tile number within the flowcell lane
15343	'x'-coordinate of the cluster within the tile
197393	'y'-coordinate of the cluster within the tile
1	the member of a pair, 1 or 2 *(paired-end or mate-pair reads only)*
Y	Y if the read is filtered, N otherwise
18	0 when none of the control bits are on, otherwise it is an even number
ATCACG	index sequence

FASTQ format

@Read_1
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCC
+
!''**+)%%%++)(%%%1+*55CCF>>>>>>CCCCCCC65
@Read_2
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCC
+
BCCFFFFFFFFCCFFFFCFFFFFFFCBBBBBBBBBBBBBBBB

Phred Score

- Originally developed for Phred base calling in the Human Genome project
- Each nucleotide has a corresponding score
- It calculates the probability of an erroneous base call
- $Q = -10\log_{10}P$
- In FASTQ files the Phred score is converted using Q+33 ASCII
- Q30 is 1:1000 erroneous base calls

Phred Score

ASCII_BASE=33 Illumina, Ion Torrent, PacBio and Sanger

Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII
0	1.00000	33 !	11	0.07943	44 ,	22	0.00631	55 7	33	0.00050	66 B
1	0.79433	34 "	12	0.06310	45 -	23	0.00501	56 8	34	0.00040	67 C
2	0.63096	35 #	13	0.05012	46 .	24	0.00398	57 9	35	0.00032	68 D
3	0.50119	36 $	14	0.03981	47 /	25	0.00316	58 :	36	0.00025	69 E
4	0.39811	37 %	15	0.03162	48 0	26	0.00251	59 ;	37	0.00020	70 F
5	0.31623	38 &	16	0.02512	49 1	27	0.00200	60 <	38	0.00016	71 G
6	0.25119	39 '	17	0.01995	50 2	28	0.00158	61 =	39	0.00013	72 H
7	0.19953	40 (	18	0.01585	51 3	29	0.00126	62 >	40	0.00010	73 I
8	0.15849	41)	19	0.01259	52 4	30	0.00100	63 ?	41	0.00008	74 J
9	0.12589	42 *	20	0.01000	53 5	31	0.00079	64 @	42	0.00006	75 K
10	0.10000	43 +	21	0.00794	54 6	32	0.00063	65 A			

ASCII_BASE=64 Old Illumina

Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII
0	1.00000	64 @	11	0.07943	75 K	22	0.00631	86 V	33	0.00050	97 a
1	0.79433	65 A	12	0.06310	76 L	23	0.00501	87 W	34	0.00040	98 b
2	0.63096	66 B	13	0.05012	77 M	24	0.00398	88 X	35	0.00032	99 c
3	0.50119	67 C	14	0.03981	78 N	25	0.00316	89 Y	36	0.00025	100 d
4	0.39811	68 D	15	0.03162	79 O	26	0.00251	90 Z	37	0.00020	101 e
5	0.31623	69 E	16	0.02512	80 P	27	0.00200	91 [	38	0.00016	102 f
6	0.25119	70 F	17	0.01995	81 Q	28	0.00158	92 \	39	0.00013	103 g
7	0.19953	71 G	18	0.01585	82 R	29	0.00126	93]	40	0.00010	104 h
8	0.15849	72 H	19	0.01259	83 S	30	0.00100	94 ^	41	0.00008	105 i
9	0.12589	73 I	20	0.01000	84 T	31	0.00079	95 _	42	0.00006	106 j
10	0.10000	74 J	21	0.00794	85 U	32	0.00063	96 `			

SAM format

- Sequence Alignment Map (SAM) format.
- Alignment format universal for all reads produced across platforms.
- It is usually in text format
- Has two sections: Header and Alignments
- Compressed SAM files are called BAM files (smaller size)
- They are binary format but machine readable with specific tools.

SAM format

```
Coor      12345678901234  5678901234567890123456789012345
ref       AGCATGTTAGATAA**GATAGCTGTGCTAGTAGGCAGTCAGCGCCAT

+r001/1        TTAGATAAAGGATA*CTG
+r002         aaaAGATAA*GGATA
+r003       gcctaAGCTAA
+r004                       ATAGCT..............TCAGC
-r003                              ttagctTAGGC
-r001/2                                         CAGCGGCAT
```

```
@HD VN:1.6 SO:coordinate
@SQ SN:ref LN:45
r001   99 ref  7 30 8M2I4M1D3M = 37  39 TTAGATAAAGGATACTG *
r002    0 ref  9 30 3S6M1P1I4M *  0   0 AAAAGATAAGGATA    *
r003    0 ref  9 30 5S6M       *  0   0 GCCTAAGCTAA       * SA:Z:ref,29,-,6H5M,17,0;
r004    0 ref 16 30 6M14N5M    *  0   0 ATAGCTTCAGC       *
r003 2064 ref 29 17 6H5M       *  0   0 TAGGC             * SA:Z:ref,9,+,5S6M,30,1;
r001  147 ref 37 30 9M         =  7 -39 CAGCGGCAT         * NM:i:1
```

SAM format - Header

```
@HD    VN:1.0  SO:coordinate
@SQ    SN:Ref.C.IN.95.95IN21068.AF067155      LN:9002
@RG    ID:e065f70d-ebf4-4bd0-8ada-de71d7463f6d PI:145  SM:PG16-BW002087_read1  PL:ILLUMINA
@PG    ID:0    VN:11.0 PN:clcgenomicswb
@PG    ID:samtools    PN:samtools    PP:0    VN:1.9-213-g706c7b4    CL:samtools view -H PG16-BW002087.bam
```

- @HD : The first header line contains version, sorting orders
- @SQ : Reference sequence details
- @RG : Read Information
- @PG : Program information
- Each header line is tab-delimited

SAM format - Alignment

- Should start with any character except '@'
- Contains 11 mandatory fields and optional fields
- Fields are tab delimited

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

SAM format - Flags

Integer	Binary	Description (Paired Read Interpretation)
1	000000000001	template having multiple templates in sequencing (read is paired)
2	000000000010	each segment properly aligned according to the aligner (read mapped in proper pair)
4	000000000100	segment unmapped (read1 unmapped)
8	000000001000	next segment in the template unmapped (read2 unmapped)
16	000000010000	SEQ being reverse complemented (read1 reverse complemented)
32	000000100000	SEQ of the next segment in the template being reverse complemented (read2 reverse complemented)
64	000001000000	the first segment in the template (is read1)
128	000010000000	the last segment in the template (is read2)
256	000100000000	not primary alignment
512	001000000000	alignment fails quality checks
1024	010000000000	PCR or optical duplicate
2048	100000000000	supplementary alignment (e.g. aligner specific, could be a portion of a split read or a tied region)

The FLAG attributes are summed to get the final value, e.g. a SAM row resulting from an Illumina paired-end FASTQ record having the FLAG value 2145 would indicate:

Flag Value	Meaning	Flag Sum
1	read is paired	1
32	read2 was reverse complemented	33
64	read1	97
2048	Supplementary alignment	2145

SAM format – CIGAR scores

Op	BAM	Description
M	0	alignment match (can be a sequence match or mismatch)
I	1	insertion to the reference
D	2	deletion from the reference
N	3	skipped region from the reference
S	4	soft clipping (clipped sequences present in SEQ)
H	5	hard clipping (clipped sequences NOT present in SEQ)
P	6	padding (silent deletion from padded reference)
=	7	sequence match
X	8	sequence mismatch

- H can only be present as the first and/or last operation.
- S may only have H operations between them and the ends of the CIGAR string.
- For mRNA-to-genome alignment, an N operation represents an intron. For other types of alignments, the interpretation of N is not defined.
- Sum of lengths of the M/I/S/=/X operations shall equal the length of SEQ.

```
RefPos:    1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19
Reference: C  C  A  T  A  C  T  G  A  A  C  T  G  A  C  T  A  A  C
Read: ACTAGAATGGCT
```

```
RefPos:    1  2  3  4  5  6  7     8  9 10 11 12 13 14 15 16 17 18 19
Reference: C  C  A  T  A  C  T     G  A  A  C  T  G  A  C  T  A  A  C
Read:                  A  C  T  A  G  A  A     T  A  A  C  T
```

```
POS: 5
CIGAR: 3M1I3M1D1M1X3M
```

SAM format - Alignment

```
1:497:R:-272+13M17D24M 113 1 497 37 37M 15 100338662 0
CGGGTCTGACCTGAGGAGAACTGTGCTCCGCCTTCAG 0;==-==9;>>>>>=>>>>>>>>>>>=>>>>>>>>>>
```

Field	Alignment
QNAME	1:497:R:-272+13M17D24M
FLAG	113
RNAME	1
POS	497
MAPQ	37
CIGAR	37M
MRNM/RNEXT	15
MPOS/PNEXT	100338662
ISIZE/TLEN	0
SEQ	CGGGTCTGACCTGAGGAGAACTGTGCTCCGCCTTCAG
QUAL	0;==-==9;>>>>>=>>>>>>>>>>>=>>>>>>>>>>

Pileup Format

- Intermediary format between a SAM file and a variant call file.
- Contains six tab delimited columns

Sequence	Position	Reference Base	Read Count	Read Results	Quality
seq1	272	T	24	,.$....,,.,.,...,,,.,..^+.	<<<+;<<<<<<<<<<=<;<;7<&
seq1	273	T	23	,.....,,.,.,...,,,.,..A	<<<;<<<<<<<<<3<=<<<;<<+
seq1	274	T	23	,.$....,,.,.,...,,,.,...	7<7;<;<<<<<<<<<=<;<;<<6
seq1	275	A	23	,$....,,.,.,...,,,.,...^l.	<+;9*<<<<<<<<<=<<:;<<<<
seq1	276	G	22	...T,,.,.,...,,,.,....	33;+<<7=7<<7<&<<1;<<6<
seq1	277	T	22	....,,.,.,.C.,,,.,..G.	+7<;<<<<<<<&<=<<:;<<&<
seq1	278	G	23	....,,.,.,...,,,.,....^k.	%38*<<;<7<<7<=<<<;<<<<<
seq1	279	C	23	A..T,,.,.,...,,,.,.....	75&<<<<<<<<<=<<<9<<:<<<

VCF Format

- Variant call file is the standard output from different variant calling tools.
- The file has a header section that starts with '#' contains the meta information
- There are 8 mandatory columns per variant record.
- It is stored in a compressed format and indexed for faster data recovery.

VCF Format

```
##fileformat=VCFv4.3
##fileDate=20090805
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
#CHROM      POS         ID          REF         ALT         QUAL        FILTER      INFO         FORMAT
 20         14370       rs6054257   G           A           29          PASS        NS=3;DP=14   GT:GQ:DP:HQ
 20         17330       .           T           A           3           q10         NS=3;DP=11   GT:GQ:DP:HQ
 20         1110696     rs6040355   A           G,T         67          PASS        NS=2;DP=10   GT:GQ:DP:HQ
 20         1230237     .           T           .           47          PASS        NS=3;DP=13   GT:GQ:DP:HQ
 20         1234567     microsat1   GTC         G,GTCT      50          PASS        NS=3;DP=9    GT:GQ:DP
```

1	CHROM	The name of the sequence (typically a chromosome) or 'the reference sequence', i.e. the sequence against which the given sample varies.
2	POS	The 1-based position of the variation on the given sequence.
3	ID	The identifier of the variation, e.g. a dbSNP rs identifier, or if unknown a ".".
4	REF	The reference base (or bases in the case of an indel) at the given position on the given reference sequence.
5	ALT	The list of alternative alleles at this position.
6	QUAL	A quality score associated with the inference of the given alleles.
7	FILTER	A flag indicating which of a given set of filters the variation has passed.
8	INFO	An extensible list of key-value pairs (fields) describing the variation.
9	FORMAT	An (optional) extensible list of fields for describing the samples.