The BAM/SAM format

The BAM Format is a binary format for storing sequence data. SAM is the corresponding, uncompressed, text file format.

The SAM format includes two type of lines:
- header lines
- alignment lines

Each header line start with an '@' followed by a two-letter code and contain comments in free-form text:

```
@HD     VN:1.0  SO:coordinate
@SQ     SN:1    LN:249250621    AS:NCBI37       UR:file:/data/local/ref/GATK/human_g1k_v37.fasta        M5:1b22b98cdeb4a9304cb5d48026a85128
@SQ     SN:2    LN:243199373    AS:NCBI37       UR:file:/data/local/ref/GATK/human_g1k_v37.fasta        M5:a0d9851da00400dec1098a9255ac712e
@SQ     SN:3    LN:198022430    AS:NCBI37       UR:file:/data/local/ref/GATK/human_g1k_v37.fasta        M5:fdfd811849cc2fadebc929bb925902e5
@RG     ID:UM0098:1     PL:ILLUMINA     PU:HWUSI-EAS1707-615LHAAXX-L001 LB:80   DT:2010-05-05T20:00:00-0400     SM:SD37743      CN:UMCORE
@RG     ID:UM0098:2     PL:ILLUMINA     PU:HWUSI-EAS1707-615LHAAXX-L002 LB:80   DT:2010-05-05T20:00:00-0400     SM:SD37743      CN:UMCORE
@PG     ID:bwa  VN:0.5.4
```

Each alignment line is a tab-delimited text line, with the following fields:

1) QNAME: ID of the read (“query”)

2) FLAG: alignment flags

3) RNAME: ID of the reference (typically: chromosome name)

4) POS: Position in reference (1-based, left side)

5) MAPQ: Mapping quality (as Phred score)

6) CIGAR: Alignment description (gaps etc.) in CIGAR format

7) MRNM: Mate reference sequence name [for paired end data]

8) MPOS: Mate position [for paired end data]

9) SIZE: inferred insert size [for paired end data]

10) SEQ: sequence of the read

11) QUAL: quality string of the read

N) EXTRA fields

The SAM Tools provide various utilities for manipulating alignments in the BAM/SAM format.
http://samtools.sourceforge.net/

The Flag field (field #2 in the SAM/BAM format)

Some of the values that the Flag field can assume:

1 → The read is one of a pair
2 → The alignment is one end of a proper paired-end alignment
4 → The read has no reported alignments
8 → The read is one of a pair and has no reported alignments
16 → The alignment is to the reverse reference strand
32 → The other mate in the paired-end alignment is aligned to the reverse reference strand
64 → The read is the first (#1) mate in a pair
128 → The read is the second (#2) mate in a pair

Note!!! Flag values can sum up!!!
For instance, a Flag value of 83 (= 64 + 16 + 2 + 1) will come from the sum of the following:
1 → The read is one of a pair
2 → The alignment is one end of a proper paired-end alignment
16 → The alignment is to the reverse reference strand
64 → The read is the first (#1) mate in a pair

Here is a useful tool that explain the Flag field:
http://broadinstitute.github.io/picard/explain-flags.html

The CIGAR field (field #6 in the SAM format)

```
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  G  G  C  T
```

```
POS: 5
CIGAR: 3M1I3M1D5M
```

The BED format

The BED (Browser Extensible Data) format is a tab-delimited text format where each line specifies a genomic interval.

The first three BED fields (mandatory) are:

chrom - The name of the chromosome (e.g. chr3, chrY, chr2_random).
chromStart - The starting position of the feature in the chromosome (0-based).
chromEnd - The ending position of the feature in the chromosome. The chromEnd base is not included in the display of the feature. For example, the first 100 bases of a chromosome are defined as chromStart=0, chromEnd=100, and span the bases numbered 0-99.

The above can be followed by up to 9 additional (optional) BED fields, of which the most commonly used are:

name - Defines the name of the BED line.
score - A score between 0 and 1000.
strand - Defines the strand - either '+' or '-'.

Finally, a BED file can also include an header line describing its visual settings in the UCSC Genome Browser

Example of BED format (including header line):

```
track name='my track' description='my track description' color=255,0,0 visibility=2
chr8 128867449 128867455 geneA 99 +
chr8 128902915 128902921 geneB 1000 +
chr8 129001512 129001518 geneC 0 +
```

More on BED format at: http://genome.ucsc.edu/FAQ/FAQformat.html#format1

Il formato GTF

Il formato GTF (Gene Transfer Format) è un formato di testo delimitato da tabulazioni pensato per descrivere annotazioni geniche e altre caratteristiche associate con il DNA.
E' un sottotipo specifico del più generale GFF (General Feature Format).

Contiene 9 campi obbligatori (di cui i primi 8 mutuati dal formato GTF, mentre il 9° è specifico):
<seqname> = nome del cromosoma (o contig) della sequenza di riferimento (es: chr1)
<source> = fonte dell'annotazione
<feature> = tipo di annotazione (es: "CDS", "start_codon", "stop_codon", "exon")
<start> <end> = coordinate di inizio e fine dell'annotazione. Nota: <start> deve essere minore di <end> e la numerazione inizia a 1 (a differenza dei file BED).
<score> = valore numerico opzionale
<frame> = 0 (se l'annotazione è in-frame), 1 o 2 (se l'annotazione è preceduta da 1 o 2 nucleotide/i extra). [Nota: si riferisce alla fase di lettura, vale per i codoni.]
[attributes] = attributi dell'annotazione. Possono essere multipli, ma i principali sono:
gene_id value;
transcript_id value;

Esempio di annotazioni genomiche in formato GTF (con prima riga di intestazione):

```
AB000381 Twinscan  CDS          380   401   .   +   0  gene_id "001"; transcript_id "001.1";
AB000381 Twinscan  CDS          501   650   .   +   2  gene_id "001"; transcript_id "001.1";
AB000381 Twinscan  CDS          700   707   .   +   2  gene_id "001"; transcript_id "001.1";
AB000381 Twinscan  start_codon  380   382   .   +   0  gene_id "001"; transcript_id "001.1";
AB000381 Twinscan  stop_codon   708   710   .   +   0  gene_id "001"; transcript_id "001.1";
```

Ulteriori dettagli sul formato GTF: http://mblab.wustl.edu/GTF2.html

The GTF/GFF formats

The GTF (Gene Transfer Format) format is a tab-delimited text format that consists of one line per feature, each containing 9 columns of data, plus optional track definition lines.
It is a variant of the GFF (General Feature Format) used to hold information about gene structure.

Fields:

seqname - name of the chromosome or scaffold; chromosome names can be given with or without the 'chr' prefix.
source - name of the program that generated this feature, or the data source (database or project name)
feature - feature type name, e.g. Gene, Variation, Similarity
start - Start position of the feature, with sequence numbering starting at 1.
end - End position of the feature, with sequence numbering starting at 1.
score - A floating point value.
strand - defined as + (forward) or - (reverse).
frame - One of '0', '1' or '2'. '0' indicates that the first base of the feature is the first base of a codon, '1' that the second base is the first base of a codon, and so on.
attribute - A semicolon-separated list of tag-value pairs, providing additional information about each feature.

Example of GTF format:

```
AB000381 Twinscan  CDS          380   401   .   +   0  gene_id "001"; transcript_id "001.1";
AB000381 Twinscan  CDS          501   650   .   +   2  gene_id "001"; transcript_id "001.1";
AB000381 Twinscan  CDS          700   707   .   +   2  gene_id "001"; transcript_id "001.1";
AB000381 Twinscan  start_codon  380   382   .   +   0  gene_id "001"; transcript_id "001.1";
AB000381 Twinscan  stop_codon   708   710   .   +   0  gene_id "001"; transcript_id "001.1";
```

More on the GTF format at: http://mblab.wustl.edu/GTF2.html