SAM and BAM format

SAM/BAM Format

- SAM (Sequence Alignment Map) format is a text based format that stores alignment data.
- BAM (Binary Alignment Map) format is the binary version of SAM.
- This is the generally accepted file format for aligned sequence data.
- SAMTools is a utility that can be used to convert between SAM and BAM format.
- http://samtools.sourceforge.net/

SAM/BAM File

- Made of two sections:
  - Header section
    - Optional
    - Lines start with @
    - Contains information such as reference sequence, commands used to generate it, library information
  - Alignment section
    - 11 mandatory columns for essential alignment information
    - Contains information such as alignment position, mapping quality, mate's mapping position

Critical Pieces of Information in SAM

- What did we align the data to?
  - @SQ SN:NC_010473.1 LN:486137

- What sample did we align?
  - @RG ID:SRR959254 LB:library1 PL:ILLUMINA PU:NONE

- What tool did we align with?
  - @PG ID:bwa PN:bwa VN:0.6.2-r126

SAM/BAM File Columns

- The first 11 columns after the header are absolutely required:
    1. Read ID
    2. The SAM flag
    3. Chromosome/contig read aligned to
    4. Position which read aligned to
    5. Mapping quality score
    6. Cigar string
    7. Chromosome/contig which read pair aligned to
    8. Position which read pair aligned to
    9. Insert Size
    10. Sequence in bases
    11. Quality score for each base

Random read from a SAM file

1. ERR001268.25
2. 147
3. chr22
4. 44549174
5. 60
6. 36M
7. =
8. 44548985
9. -225
10. GGTTGGATGTGTATTTTTCTAGATATTTTCTTTTTC
11.)(1)+.-5+<.@9A%<;=0IIIHCII?III;IIIII

SAM Flag

Bit : Description
0x1 : 1: template having multiple segments in sequencing
0x2 : 2 : each segment properly aligned according to the aligner
0x4 : 4 : segment unmapped
0x8 : 8: next segment in the template unmapped
0x10 : 16 : SEQ being reverse complemented
0x20 : 32: SEQ of the next segment in the template being reversed
0x40 : 64 :the first segment in the template
0x80 : 128 :the last segment in the template
0x100 : 256 : secondary alignment
0x200 : 512 : not passing quality controls
0x400 : 1024 : PCR or optical duplicate
0x800 : 2048 : supplementary alignment

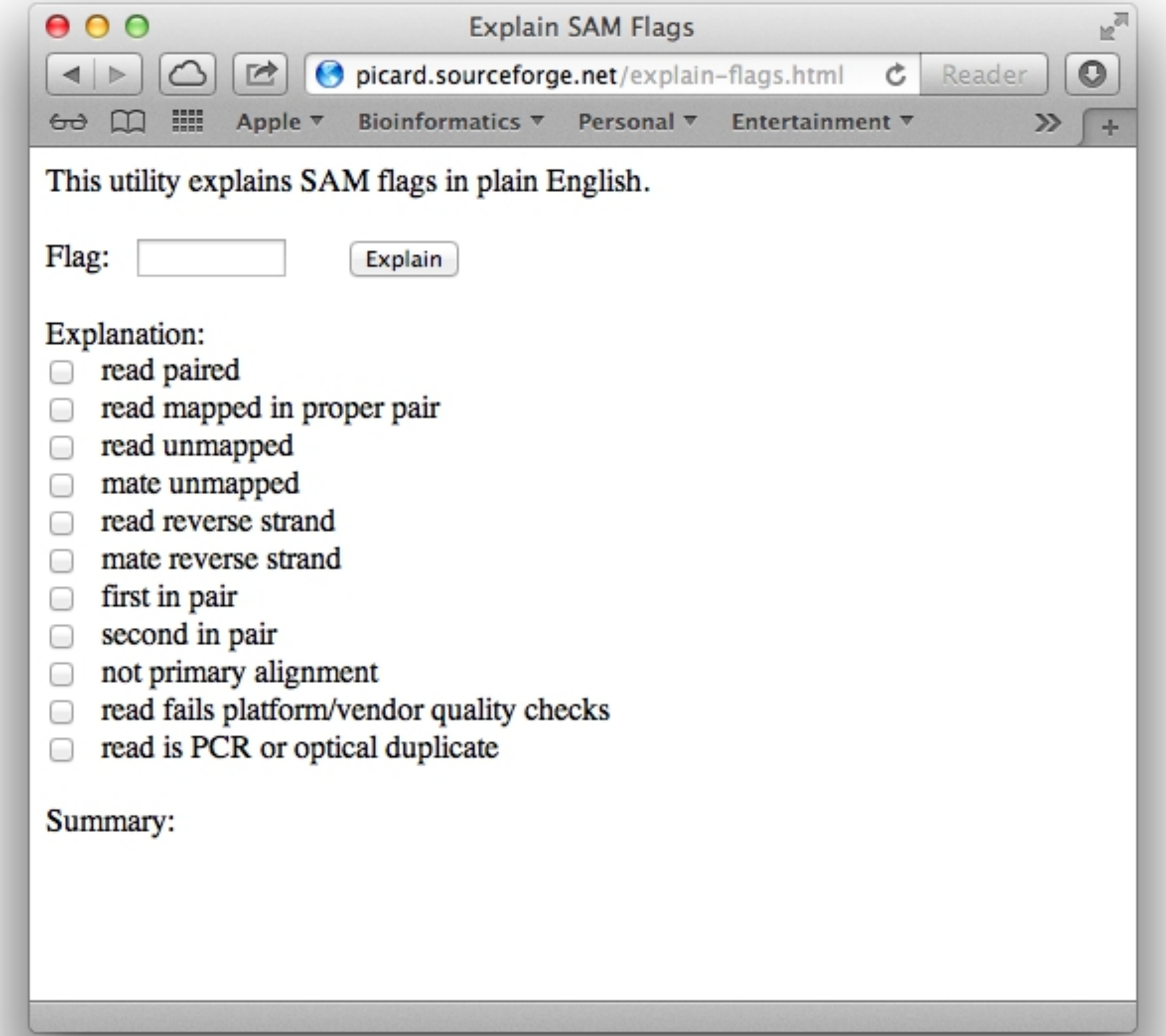
77 = 64 + 8 + 4 + 1 : paired, unmapped, mate unmapped, first in pair
163 = 128 + 32 + 2 + 1: paired, pair is mapped, mate is reversed, second in pair

SAM Flag

- The flag tries to give you a lot of information in a single number

- Instead of trying to memorize all the codes, I go to the following website to find out what they mean:

`http://picard.sourceforge.net/explain-flags.html`



This utility explains SAM flags in plain English.

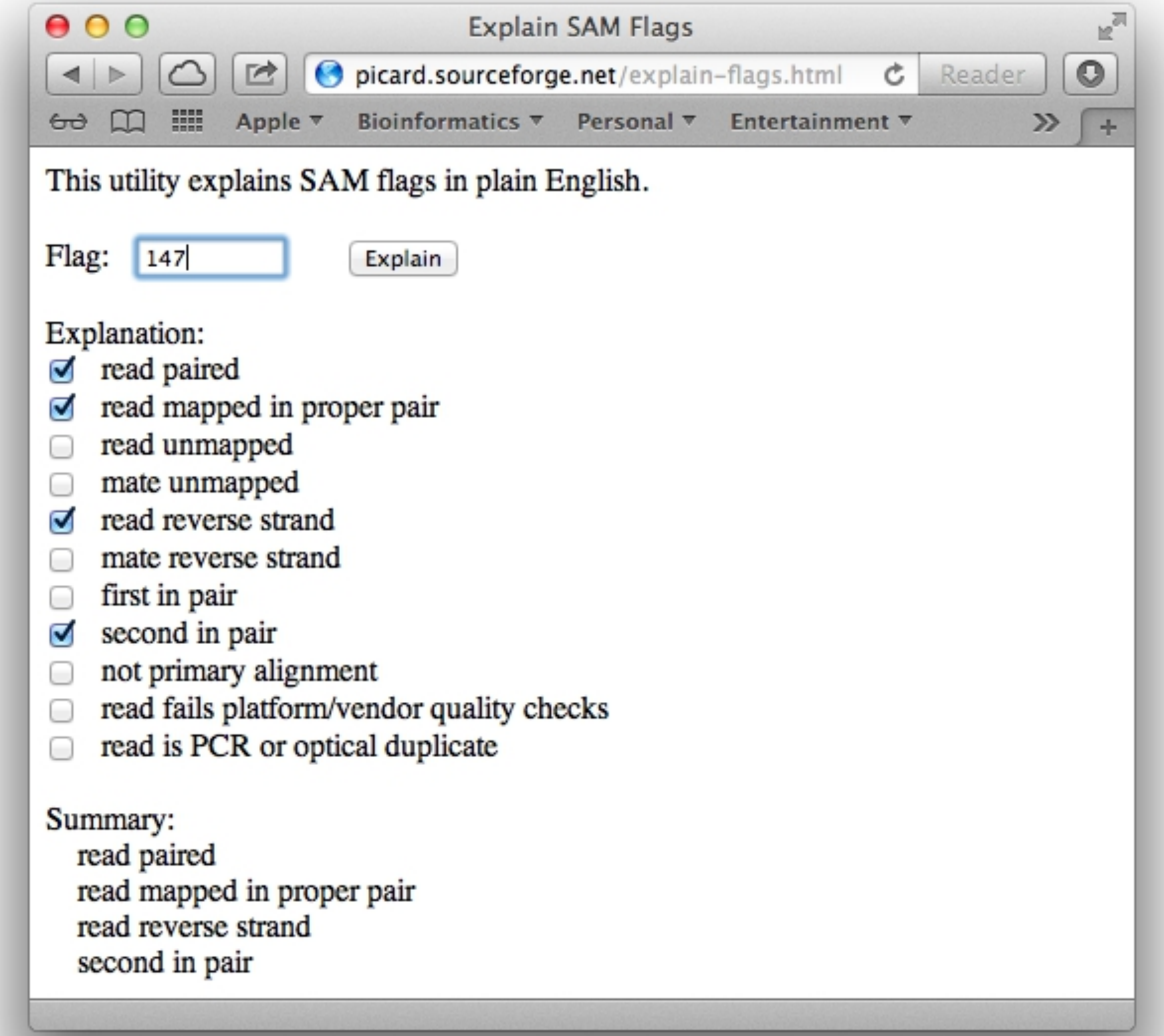
Flag: □ Explain

Explanation:

- □ read paired
- □ read mapped in proper pair
- □ read unmapped
- □ mate unmapped
- □ read reverse strand
- □ mate reverse strand
- □ first in pair
- □ second in pair
- □ not primary alignment
- □ read fails platform/vendor quality checks
- □ read is PCR or optical duplicate

Summary:

http://picard.sourceforge.net/explain-flags.html



This utility explains SAM flags in plain English.

Flag: 147

Explanation:

☑ read paired
☑ read mapped in proper pair
☐ read unmapped
☐ mate unmapped
☑ read reverse strand
☐ mate reverse strand
☐ first in pair
☑ second in pair
☐ not primary alignment
☐ read fails platform/vendor quality checks
☐ read is PCR or optical duplicate

Summary:
read paired
read mapped in proper pair
read reverse strand
second in pair

http://picard.sourceforge.net/explain-flags.html

Mapping Quality Score
Phred base score, estimates a probability that the mapping is correct
Determined by aligner

That exampe of two aligners using different scores and downstream effects.

Same data, aligned with two different aligners.
Reads were aligned at the same position
Mapping quality of the reads that aligned at these positions were essentially the same, except at the lowe end

    one aligner assigned values of 0-5 for low quality
    one aligner assigned all of these as 0

Downtream variant caller interpreted this differently, calling the variant from one aligner, but NOT from the other

- CIGAR string

  Compact **I**diosyncratic **G**apped **A**lignment **R**eport

op	Description
M	Alignment match (can be a sequence match or mismatch
I	Insertion to the reference
D	Deletion from the reference
N	Skipped region from the reference
S	Soft clip on the read (clipped sequence present in <seq>)
H	Hard clip on the read (clipped sequence NOT present in <seq>)
P	Padding (silent deletion from the padded reference sequence)

First 100 reads in a bam file, cigar strings

```
 40 *
      7 100M1S
      7 101M
      1 10S3M1D88M
      3 1S2M1D6M1I91M
      1 1S2M1I97M
      6 2M1I98M
     22 2S3M1D96M
      1 3S2M1I95M
      2 3S3M1D95M
```

Optional Fields
Tag:Type:Value

Tag : 2 letter identifier

Some examples

Z? reserved for end users

SM:i:37 - Mapping quality if the read is mapped as a single read rather than as a read pair

AM:i:37 - Smaller single-end mapping quality of the two reads in a pair

RG:Z:SRR035022 - Read group. Value matches the header RG-ID tag if @RG is present in the header

NM:i:2 - Number of nucleotide differences (i.e. edit distance to the reference sequence)

MD:Z:4A95 - String for mismatching positions

Type :

A	Printable character
i	Signed 32-bin integer
f	Single-precision float number
Z	Printable string
H	Hex string (high nybble first)