SAM格式详细说明

SAM格式，是一种用于存储测序序列比对结果的通用格式，以TAB为分割符。SAM格式的设计理念是：灵活存储所有比对信息。SAM一般为11列以上，其中前11列为已定义列，12列及以后为可选列，算附加信息。
由于测序分两种：单端测序(Single-read)和双端测序(Paired-end和Mate-pair)。下文SAM格式中read表示Single-read或Paired-end Read1序列，mate表示Paired-end Read2序列。

1. 第一列：read name，read的名字通常包括测序平台等信息；
2. 第二列：sum of flags，比对flag数字之和，比对flag用数字表示，分别为：

- 1（1）该read是成对的paired reads中的一个
- 2（10）paired reads中每个都正确比对到参考序列上
- 4（100）该read没比对到参考序列上
- 8（1000）与该read成对的matepair read没有比对到参考序列上
- 16（10000）该read其反向互补序列能够比对到参考序列
- 32（100000）与该read成对的matepair read其反向互补序列能够比对到参考序列
- 64（1000000）在paired reads中，该read是与参考序列比对的第一条
- 128（10000000）在paired reads中，该read是与参考序列比对的第二条
- 256（100000000）该read是次优的比对结果
- 512（1000000000）该read没有通过质量控制
- 1024（10000000000）由于PCR或测序错误产生的重复reads
- 2048（100000000000）补充匹配的read
  通过这个和可以直接推断出匹配的情况。假如说标记不是以上列举出的数字，比如说83=（64+16+2+1），就是这几种情况值和。

3. 第三列：RNAM，reference sequence name，实际上就是比对到参考序列上的染色体号。若是无法比对，则是*；
4. 第四列：position，read比对到参考序列上，第一个碱基所在的位置。若是无法比对，则是0；
5. 第五列：Mapping quality，比对的质量分数，越高说明该read比对到参考基因组上的位置越唯一；
6. 第六列：CIGAR值，read比对的具体情况，

- “M”表示 match或 mismatch；
- “I”表示 insert；
- “D”表示 deletion；
- “N”表示 skipped（跳过这段区域）；
- “S”表示 soft clipping（被剪切的序列存在于序列中）；
- “H”表示 hard clipping（被剪切的序列不存在于序列中）；
- “P”表示 padding；
- “=”表示 match；
- “X”表示 mismatch（错配，位置是一一对应的）；

7. 第七列：MRNM(chr)，mate的reference sequence name，实际上就是mate比对到的染色体号，若是没有mate，则是*；
8. 第八列：mate position，mate比对到参考序列上的第一个碱基位置，若无mate,则为0；
9. 第九列：ISIZE，Inferred fragment size.详见Illumina中paired end sequencing 和 mate pair sequencing，是负数，推测应该是两条read之间的间隔(待查证)，若无mate则为0；
10. 第十列：Sequence，就是read的碱基序列，如果是比对到互补链上则是reverse completed
    eg. CGTTTCTGTGGGTGATGGGCCTGAGGGGCGTTCTCN
11. 第十一列：ASCII，read质量的ASCII编码。
12. 第十二列之后：Optional fields，可选的自定义区域

- AS:i 匹配的得分
- XS:i 第二好的匹配的得分
- YS:i mate 序列匹配的得分
- XN:i 在参考序列上模糊碱基的个数
- XM:i 错配的个数
- XO:i gap open的个数
- XG:i gap 延伸的个数
- NM:i 经过编辑的序列
- YF:i 说明为什么这个序列被过滤的字符串
- MD:Z 代表序列和参考序列错配的字符串(例如MD:Z:45A2C3 失配碱基的位点，45,45+2两个位点失配)
- XT:A:U read只有一个完整比对，U unique
- XT:A:R read有一个以上位置完整比对，R repeat
- NM:i:2 read有2个碱基mismatch
- X0:i:2 read有2个位置完整比对(与XT:A有对应关系)
- X1:i:2 read有2个位置以1个碱基失配比对
- XA:Z:Chr3,+1530, 50M,0;Chr4,-7568,50M,1 有0/1个碱基失配的详细比对情况(与XT:A、X0:i、X1:i有对应关系)

Tag	Type	Description
X?	?	Reserved fields for end users (together with Y? and Z?)
AM	i	The smallest template-independent mapping quality of fragments in the rest
AS	i	Alignment score generated by aligner
BQ	Z	Offset to base alignment quality (BAQ), of the same length as the read sequence. At the i-th read base, $\mathrm{BAQ}_i = Q_i - (\mathrm{BQ}_i - 64)$ where Q_i is the i-th base quality.
CM	i	Edit distance between the color sequence and the color reference (see also NM)
CQ	Z	Color read quality on the original strand of the read. Same encoding as QUAL; same length as CS.
CS	Z	Color read sequence on the original strand of the read. The primer base must be included.
E2	Z	The 2nd most likely base calls. Same encoding and same length as QUAL.
FI	i	The index of fragment in the template.
FS	Z	Fragment suffix.
LB	Z	Library. Value to be consistent with the header RG-LB tag if @RG is present.
H0	i	Number of perfect hits
H1	i	Number of 1-difference hits (see also NM)
H2	i	Number of 2-difference hits
HI	i	Query hit index, indicating the alignment record is the i-th one stored in SAM
IH	i	Number of stored alignments in SAM that contains the query in the current record
MD	Z	String for mismatching positions. *Regex*: [0-9]+(([ACGTN]|\^[ACGTN]+)[0-9]+)* [1]
MQ	i	Mapping quality of the mate/next fragment
NH	i	Number of reported alignments that contains the query in the current record
NM	i	Edit distance to the reference, including ambiguous bases but excluding clipping
OQ	Z	Original base quality (usually before recalibration). Same encoding as QUAL.
OP	i	Original mapping position (usually before realignment)
OC	Z	Original CIGAR (usually before realignment)
PG	Z	Program. Value matches the header PG-ID tag if @PG is present.
PQ	i	Phred likelihood of the template, conditional on both the mapping being correct
PU	Z	Platform unit. Value to be consistent with the header RG-PU tag if @RG is present.
Q2	Z	Phred quality of the mate/next fragment. Same encoding as QUAL.
R2	Z	Sequence of the mate/next fragment in the template.
RG	Z	Read group. Value matches the header RG-ID tag if @RG is present in the header.
SM	i	Template-independent mapping quality
TC	i	The number of fragments in the template.
U2	Z	Phred probility of the 2nd call being wrong conditional on the best being wrong. The same encoding as QUAL.
UQ	i	Phred likelihood of the fragment, conditional on the mapping being correct