Mansoura University
Faculty of Computers and Information
Department of Computer Science
Second Semester: 2020-2021

[MED-145] Genomics: Genome Indexing & Reads Mapping

Introduction to SAM/BAM formats

Grade: Third Year (Medical Informatics Program)

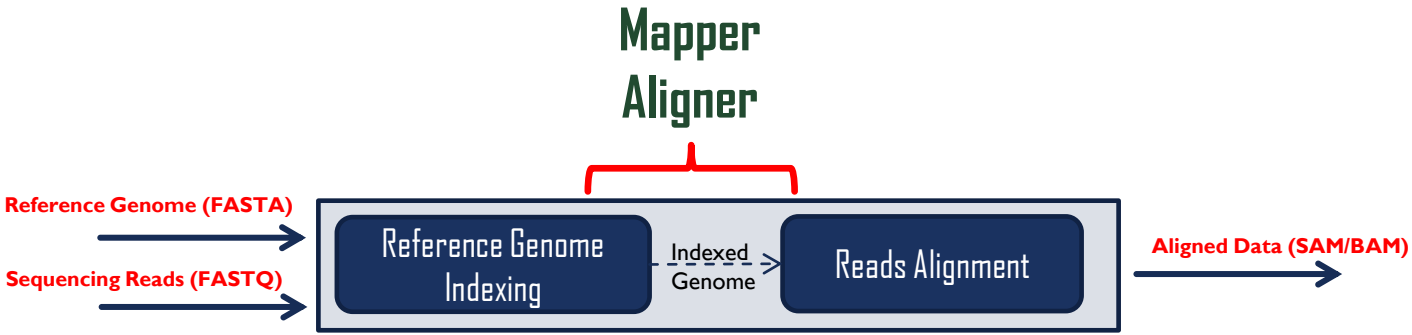
Sara El-Metwally, Ph.D.

Faculty of Computers and Information,

Mansoura University,

Egypt.

TYPICAL MAPPING/ALIGNMENT WORKFLOW

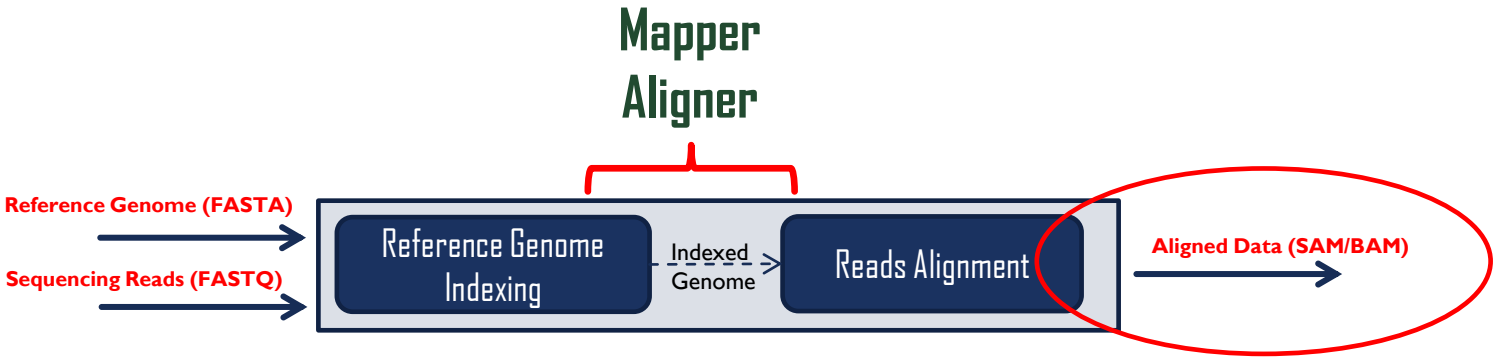


Genome Indexing and Mapping Approaches

hash-based

Burrows-Wheeler

TYPICAL MAPPING/ALIGNMENT WORKFLOW



Genome Indexing and Mapping Approaches

hash-based

Burrows-Wheeler

Sequence alignment software

Aligner	Approach	Applications	Availability
BWA-mem	Burrows-Wheeler	DNA, SE, PE, SV	open-source
Bowtie2	Burrows-Wheeler	DNA, SE, PE, SV	open-source
Novoalign	hash-based	DNA, SE, PE	free for academic use
TopHat	Burrows-Wheeler	RNA-seq	open-source
STAR	hash-based (reads)	RNA-seq	open-source
GSNAP	hash-based (reads)	RNA-seq	open-source

Slides curiosity from Aaron Quinlan: https://github.com/quinlan-lab/applied-computational-genomics

BWA-MEM: never "published" ; widely used.

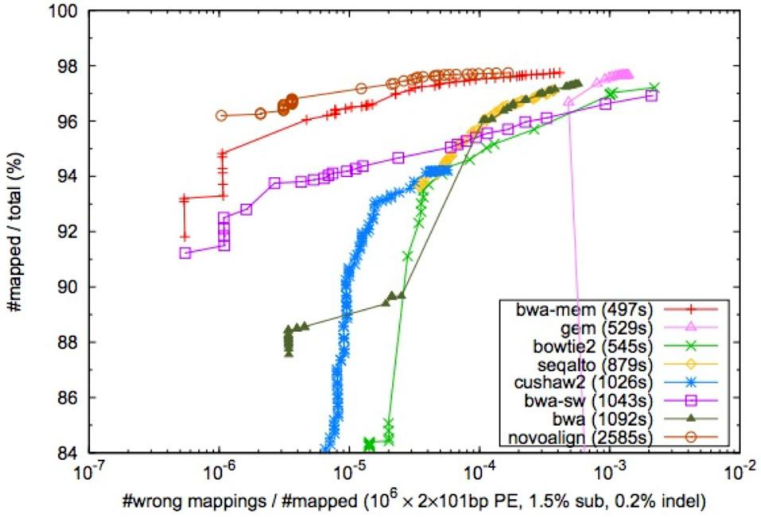


Fig. 1. Percent mapped reads as a function of the false alignment rate under different mapping quality cutoff. Alignments with mapping quality 3 or lower are excluded. An alignment is *wrong* if after correcting clipping, its start position is within 20bp from the simulated position. 10^6 pairs of 101bp reads are simulated from the human reference genome using wgsim (http://bit.ly/wgsim2) with 1.5% substitution errors and 0.2% indel variants. The insert size follows a normal distribution $N(500, 50^2)$. The reads are aligned back to the genome either as single end (SE; top panel) or as paired end (PE; bottom panel). GEM is configured to allow up to 5 gaps and to output suboptimal alignments (option '–e5 –m5 –s1' for SE and '–e5 –m5 –s1 –pb' for PE). GEM does not compute mapping quality. Its mapping quality is estimated with a BWA-like algorithm with suboptimal alignments available. Other mappers are run with the default setting except for specifying the insert size distribution. The run time in seconds on a single CPU core is shown in the parentheses.

Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM

Heng Li

Broad Institute of Harvard and MIT, 7 Cambridge Center, Cambridge, MA 02142, USA

https://arxiv.org/pdf/1303.3997v2.pdf

Slides curiosity from Aaron Quinlan: https://github.com/quinlan-lab/applied-computational-genomics

BWA-MEM

Unaligned Sample Data In FASTQ (SE or PE)

```
@seq1
ATTCGAAACA...
+
DDED88(999...
@seq2
CCCCGTTTCA...
+
AAC887BBAC...
```

Reference genome (FASTA)

```
>chr1
TACCTCCAGGGGGCATCCTCCCCCCCAATTCG
AAACACAATCGTAGCCCCTGGCACTACCTATG
TGTGTCAATTCGGAGAGAGAGAGATTCACGAA
AAAAAAGTCTGGACTCAACTAGGATACACACA
TTCGGCTACAGATACCAAAAAAAAAAAAAAAA
AAATTTTCACCATTGAGGCACCACCTTCTCGT
CGCTGCGTCGCTCTGCTCGCTTCGGCTAAAAA
TTCGCGCAATACATTCGGCTACAGATACCAAA
```

BWA MEM

Aligned Sample Data in SAM format

```
seq1
99              1
3666901         60
149M            =
3666935         185
ATTCGAAACA...DDED88(999
MC:Z:151M       MD:Z:149
RG:Z:15-0017315_1
NM:i:0          MQ:i:60
AS:i:149        XS:i:44
seq2
147             1
3666935         60
151M            =
3666901         -185
CCCCGTTTCA...AAC887BBAC...
MC:Z:149M       MD:Z:151
RG:Z:15-0017315_1
NM:i:0          MQ:i:60
AS:i:151        XS:i:59
```

Slides curiosity from Aaron Quinlan: https://github.com/quinlan-lab/applied-computational-genomics

BWA-MEM workflow

This takes a long time, but you do it once

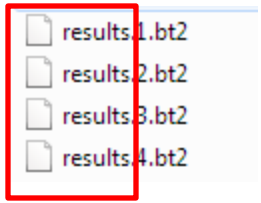
Create BWT of reference genome.

```
$ bwa index grch38.fa
```

Output is in SAM format.
Use multiple threads if you have a computer with multiple CPUs.

Align paired-end FASTQ to BWT index.

```
$ bwa mem -t 16 grch38.fa 1.fq 2.fq > sample.sam
```



Same prefix: grch38.fa

Slides adapted from Aaron Quinlan: https://github.com/quinlan-lab/applied-computational-genomics

SAM format: a **text**-based **standard(!)** for representing sequence alignments

BIOINFORMATICS **APPLICATIONS NOTE** *Vol. 25 no. 16 2009, pages 2078–2079*
doi:10.1093/bioinformatics/btp352

Sequence analysis

The Sequence Alignment/Map format and SAMtools

Heng Li[1,†], Bob Handsaker[2,†], Alec Wysoker[2], Tim Fennell[2], Jue Ruan[3], Nils Homer[4], Gabor Marth[5], Goncalo Abecasis[6], Richard Durbin[1,*] and 1000 Genome Project Data Processing Subgroup[7]

[1]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Cambridge, CB10 1SA, UK, [2]Broad Institute of MIT and Harvard, Cambridge, MA 02141, USA, [3]Beijing Institute of Genomics, Chinese Academy of Science, Beijing 100029, China, [4]Department of Computer Science, University of California Los Angeles, Los Angeles, CA 90095, [5]Department of Biology, Boston College, Chestnut Hill, MA 02467, [6]Center for Statistical Genetics, Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, USA and [7]http://1000genomes.org

Received on April 28, 2009; revised on May 28, 2009; accepted on May 30, 2009

Advance Access publication June 8, 2009

Associate Editor: Alfonso Valencia

Table 1. Mandatory fields in the SAM format

No.	Name	Description
1	`QNAME`	Query NAME of the read or the read pair
2	`FLAG`	Bitwise FLAG (pairing, strand, mate strand, etc.)
3	`RNAME`	Reference sequence NAME
4	`POS`	1-Based leftmost POSition of clipped alignment
5	`MAPQ`	MAPping Quality (Phred-scaled)
6	`CIGAR`	Extended CIGAR string (operations: `MIDNSHP`)
7	`MRNM`	Mate Reference NaMe ('=' if same as `RNAME`)
8	`MPOS`	1-Based leftmost Mate POSition
9	`ISIZE`	Inferred Insert SIZE
10	`SEQ`	Query SEQuence on the same strand as the reference
11	`QUAL`	Query QUALity (ASCII-33=Phred base quality)

Slides curiosity from Aaron Quinlan: https://github.com/quinlan-lab/applied-computational-genomics

SAM format overview

- **In the dark ages,** sequence aligners used disparate output formats. **Pain.**
- 1000 Genomes Project sought to <u>standardize</u>. **Standards are good.**
- The result is imperfect, but it's a **huge** improvement.
- **Strengths of the SAM and BAM formats**
    - Compressed: less disk hungry
    - Indexed: fast viewing, slicing, etc.
    - Single-end and paired-end
    - Relatively simple to produce
    - ***Good toolkits available***

Slides curiosity from Aaron Quinlan: https://github.com/quinlan-lab/applied-computational-genomics

SAM/BAM/CRAM

- SAM: Sequence Alignment Map format is a tab delineated text file describing mapping information for short read data.

- With the millions of reads generated from NGS machines, the SAM files become really big.

- BAM files contain the same information as SAM files but are encoded in condensed computer readable binary format to save disk space.

- Use samtools to convert SAM to BAM.

SAM/BAM

- There are two types of BAM files: unsorted or sorted.
- Sorted BAM file: the alignments are sorted left-to-right along the reference genome.
- Most alignment tools output SAM (not BAM), and the alignments come out in an arbitrary order -- not sorted.
- An authoritative and complete document describing the SAM and BAM formats is the SAM specification.

SAM/BAM

The header contains general information about the alignment, such as the reference name, reference length, the program used for the alignment, etc.

```
@HD     VN:1.0  SO:coordinate
@SQ     SN:chr20        LN:64444167
@PG     ID:TopHat       VN:2.0.14       CL:/srv/dna_tools/tophat/tophat -N 3 --read-edit-dist 5 --read-rea
lign-edit-dist 2 -i 50 -I 5000 --max-coverage-intron 5000 -M -o out /data/user446/mapping_tophat/index/chr
20 /data/user446/mapping_tophat/L6_18_GTGAAA_L007_R1_001.fastq
HWI-ST1145:74:C101DACXX:7:1102:4284:73714       16      chr20   190930  3       100M    *       0       0
       CCGTGTTTAAAGGTGGATGCGGTCACCTTCCCAGCTAGGCTTAGGGATTCTTAGTTGGCCTAGGAAATCCAGCTAGTCCTGTCTCTCAGTCCCCCCTCT
C    BBDCCDDCCDDDDCDDDDDDCDCCCDBC?DDDDDDDDDDDDDDCCDCDDDDDDDDDDCCCCEDDDC?DDDDDDDDDDDDDDDDDDDDDBDHFFFFDC@@
    AS:i:-15        XM:i:3  XO:i:0  XG:i:0  MD:Z:55C20C13A9 NM:i:3  NH:i:2  CC:Z:=  CP:i:55352714   HI:i:0
HWI-ST1145:74:C101DACXX:7:1114:2759:41961       16      chr20   193953  50      100M    *       0       0
       TGCTGGATCATCTGGTTAGTGGCTTCTGACTCAGAGGACCTTCGTCCCCTGGGGCAGTGGACCTTCCAGTGATTCCCCTGACATAAGGGGCATGGACGA
G    DCDDDDEDDDDDDDCDDDDDDDCCCDDDCDDDDDEEC>DFFFEJJJJJIGJJJJIHGBHHGJIJJJJJJGJJJIJJJJJIHJJJJJJHHHHHFFFFFCCC
    AS:i:-16        XM:i:3  XO:i:0  XG:i:0  MD:Z:60G16T18T3 NM:i:3  NH:i:1
HWI-ST1145:74:C101DACXX:7:1204:14760:4030       16      chr20   270877  50      100M    *       0       0
       GGCTTTATTGGTAAAAAAGGAATAGCAGATTTAATCAGAAATTCCCACCTGGCCCAGCAGCACCAACCAGAAAGAAGGGAAGAAGACAGGAAAAAACCA
C    DDDDDDDDDCCDDDDDDDDDDEEEEEEEFFFEFFEGHHHHFGDJJIHJJIJIJJJIIIIGGFJJIHIIIIJJJJJJIGHHFAHGFHJHFGGHFFFDD@BB
    AS:i:-11        XM:i:2  XO:i:0  XG:i:0  MD:Z:0A85G13    NM:i:2  NH:i:1
HWI-ST1145:74:C101DACXX:7:1210:11167:8699       0       chr20   271218  50      50M4700N50M     *       0
       0       GTGGCTCTTCCACAGGAATGTTGAGGATGACATCCATGTCTGGGGTGCACTTGGGTCTCCGAAGCAGAACATCCTCAAATATGACCTCTCG
```

Sequence Alignment/Map Format Specification

The SAM/BAM Format Specification Working Group

7 Jan 2021

The master version of this document can be found at `https://github.com/samtools/hts-specs`.
This printing is version 981fc0f from that repository, last modified on the date shown above.

@HD

1 The SAM Format Specification

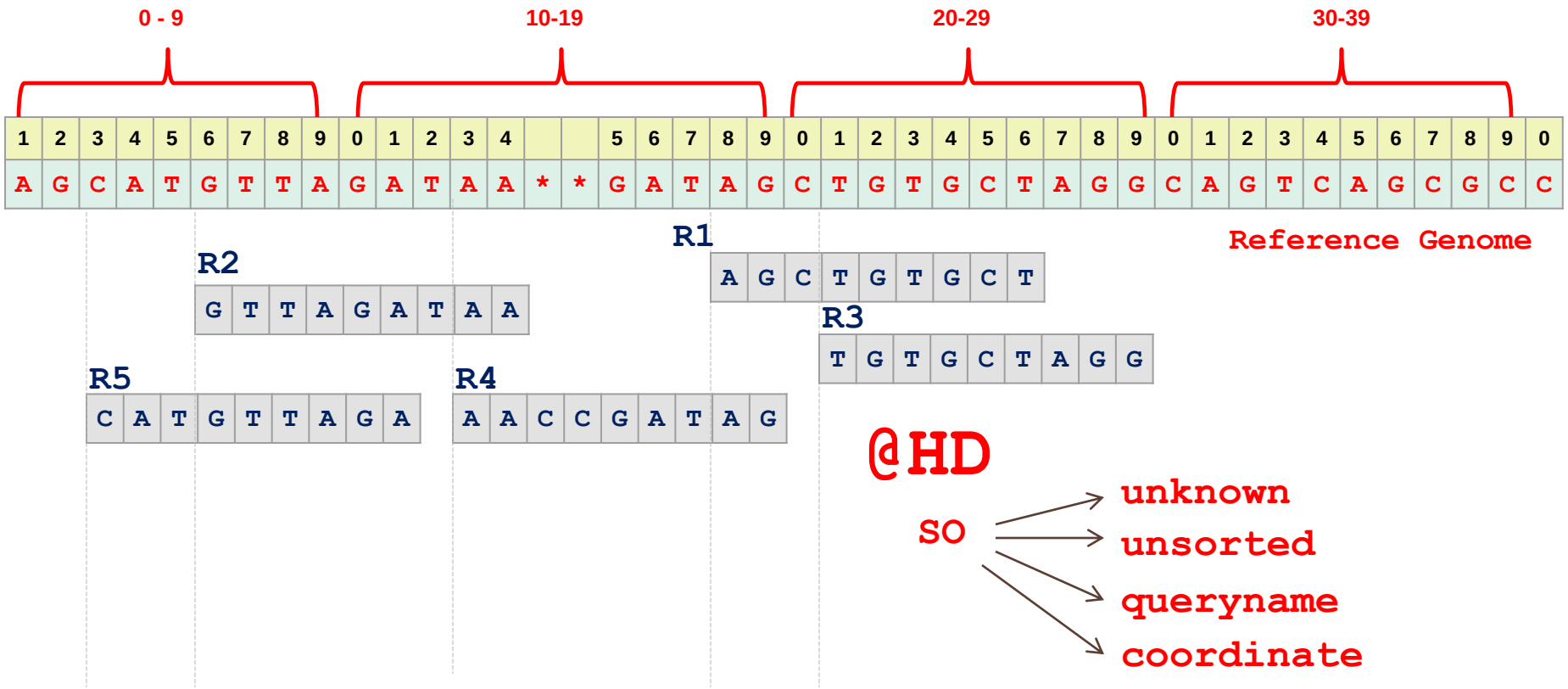
SAM stands for Sequence Alignment/Map format. It is a TAB-delimited text format consisting of a header section, which is optional, and an alignment section. If present, the header must be prior to the alignments. Header lines start with '@', while alignment lines do not. Each alignment line has 11 mandatory fields for essential alignment information such as mapping position, and variable number of optional fields for flexible or aligner specific information.

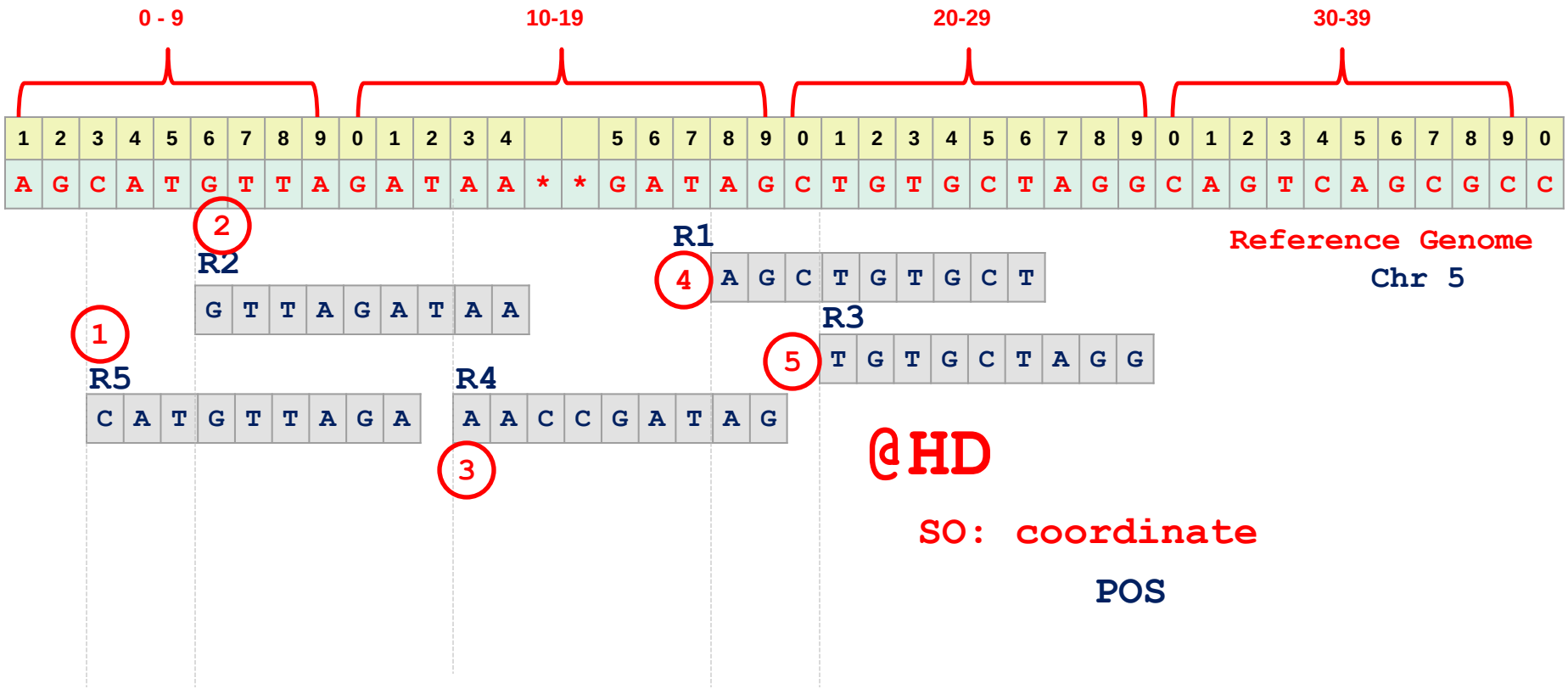
VN

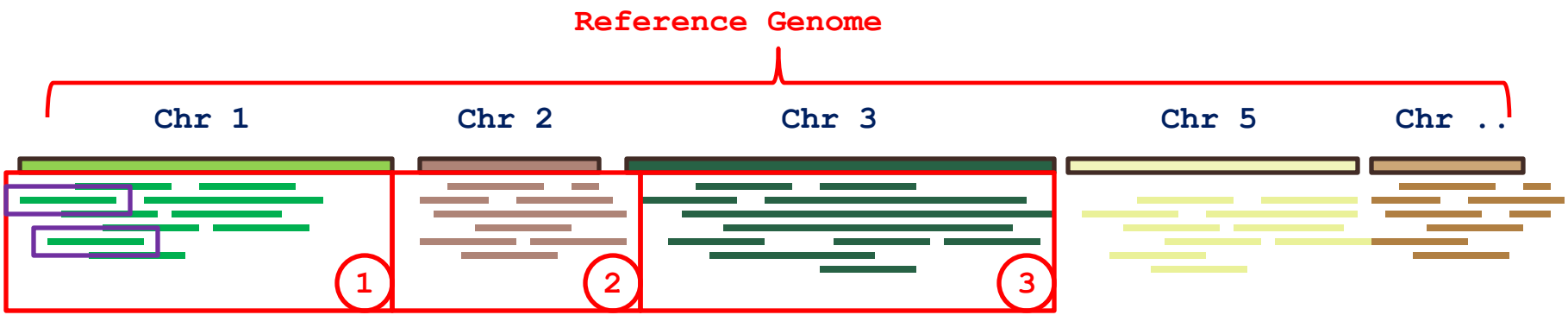
This specification is for version 1.6 of the SAM and BAM formats. Each SAM and BAM file may optionally specify the version being used via the `@HD VN` tag. For full version history see Appendix B.

Unless explicitly specified elsewhere, all fields are encoded using 7-bit US-ASCII [1] in using the POSIX / C locale. Regular expressions listed use the POSIX / IEEE Std 1003.1 extended syntax.

Image credit: https://samtools.github.io/hts-specs/SAMv1.pdf


0 - 9
10-19
20-29
30-39
1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0
A G C A T G T T A G A T A A * * G A T A G C T G T G C T A G G C A G T C A G C G C C
Reference Genome
R1
A G C T G T G C T
R2
G T T A G A T A A
R3
T G T G C T A G G
R4
A A C C G A T A G
R5
C A T G T T A G A
@HD
SO
unknown
unsorted
queryname
coordinate


0 - 9
10-19
20-29
30-39
1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0
A G C A T G T T A G A T A A * * G A T A G C T G T G C T A G G C A G T C A G C G C C
Reference Genome
Chr 5
R1
4
A G C T G T G C T
R2
2
G T T A G A T A A
R3
5
T G T G C T A G G
R5
1
C A T G T T A G A
R4
A A C C G A T A G
3
@HD
SO: coordinate
POS



@HD

SO: coordinate

RNAME

POS

BWA-MEM

@SQ

Reference genome (FASTA)

```
>chr1   SN
TACCTCCAGGGGGCATCCTCCCCCCCAATTCG
AAACACAATCGTAGCCCCTGGCACTACCTATG
TGTGTCAATTCGGAGAGAGAGAGATTCACGAA
AAAAAAGTCTGGACTCAACTAGGATACACACA
TTCGGCTACAGATACCAAAAAAAAAAAAAAAA
AAATTTTCACCATTGAGGCACCACCTTCTCGT
CGCTGCGTCGCTCTGCTCGCTTCGGCTAAAAA
TTCGCGCAATACATTCGGCTACAGATACCAAA
```

LN

Aligned Sample Data in SAM format

Unaligned Sample Data In FASTQ (SE or PE)

```
@seq1
ATTCGAAACA...
+
DDED88(999...
```

@PG

BWA MEM

ID,VN

```
seq1        99          1
            3666901     60
            149M        =
            3666935     185
            ATTCGAAACA...DDED88(999
            MC:Z:151M   MD:Z:149
            RG:Z:15-0017315_1
            NM:i:0      MQ:i:60
            AS:i:149    XS:i:44
```

```
$ bwa mem -t 16 grch38.fa 1.fq 2.fq > sample.sam
```

CL

Slides curiosity from Aaron Quinlan: https://github.com/quinlan-lab/applied-computational-genomics

SAM/BAM

```
@HD     VN:1.0  SO:coordinate
@SQ     SN:chr20        LN:64444167
@PG     ID:TopHat       VN:2.0.14       CL:/srv/dna_tools/tophat/tophat -N 3 --read-edit-dist 5 --read-rea
lign-edit-dist 2 -i 50 -I 5000 --max-coverage-intron 5000 -M -o out /data/user446/mapping_tophat/index/chr
20 /data/user446/mapping_tophat/L6_18_GTGAAA_L007_R1_001.fastq
HWI-ST1145:74:C101DACXX:7:1102:4284:73714       16      chr20   190930  3       100M    *       0       0
       CCGTGTTTAAAGGTGGATGCGGTCACCTTCCCAGCTAGGCTTAGGGATTCTTAGTTGGCCTAGGAAATCCAGCTAGTCCTGTCTCTCAGTCCCCCCTCT
C    BBDCCDDCCDDDDCDDDDDDCDCCCDBC?DDDDDDDDDDDDDDCCDCDDDDDDDDDDCCCCEDDDC?DDDDDDDDDDDDDDDDDDDDDBDHFFFFDC@@
    AS:i:-15        XM:i:3  XO:i:0  XG:i:0  MD:Z:55C20C13A9 NM:i:3  NH:i:2  CC:Z:=  CP:i:55352714   HI:i:0
HWI-ST1145:74:C101DACXX:7:1114:2759:41961       16      chr20   193953  50      100M    *       0       0
       TGCTGGATCATCTGGTTAGTGGCTTCTGACTCAGAGGACCTTCGTCCCCTGGGGCAGTGGACCTTCCAGTGATTCCCCTGACATAAGGGGCATGGACGA
G    DCDDDDEDDDDDDDCDDDDDDDCCCDDDCDDDDDEEC>DFFFEJJJJJIGJJJJIHGBHHGJIJJJJJJGJJJIJJJJJIHJJJJJJHHHHHFFFFFCCC
    AS:i:-16        XM:i:3  XO:i:0  XG:i:0  MD:Z:60G16T18T3 NM:i:3  NH:i:1
HWI-ST1145:74:C101DACXX:7:1204:14760:4030       16      chr20   270877  50      100M    *       0       0
       GGCTTTATTGGTAAAAAAGGAATAGCAGATTTAATCAGAAATTCCCACCTGGCCCAGCAGCACCAACCAGAAAGAAGGGAAGAAGACAGGAAAAAACCA
C    DDDDDDDDDCCDDDDDDDDDDEEEEEEEFFFEFFEGHHHHFGDJJIHJJIJIJJJIIIIGGFJJIHIIIIJJJJJJIGHHFAHGFHJHFGGHFFFDD@BB
    AS:i:-11        XM:i:2  XO:i:0  XG:i:0  MD:Z:0A85G13    NM:i:2  NH:i:1
HWI-ST1145:74:C101DACXX:7:1210:11167:8699       0       chr20   271218  50      50M4700N50M     *       0
       0       GTGGCTCTTCCACAGGAATGTTGAGGATGACATCCATGTCTGGGGTGCACTTGGGTCTCCGAAGCAGAACATCCTCAAATATGACCTCTCG
```

Every line after the header represents a single read, with 11 mandatory tab separated fields of information.

What critical information do we need for sequence alignments?

Slides curiosity from Aaron Quinlan: https://github.com/quinlan-lab/applied-computational-genomics

SAM format overview

Col #	Name	Meaning	Example
1	QNAME	Read or Pair name	HWI:ST156_1:278:1:1058:4544:0
2	FLAG	Bitwise FLAG	***Much more soon!***
3	RNAME	Reference sequence name	chr1
4	POS	1-based alignment start coordinate	8,724,005
5	MAPQ	Mapping quality	60
6	CIGAR	Extended CIGAR string	***Much more soon!***
7	MRNM	If paired, the mate's reference seq.	chr1
8	MPOS	If paired, the mate's alignment start	8,724,505
9	ISIZE	If paired, the insert size	562
10	SEQ	The sequence of the query/mate	ACAAATTCAG...
11	QUAL	The quality string for the query/mate	HHH$^^%$$$...
12	OPT	Optional Tags	XA:i:2, MD:Z:0T34G15

http://samtools.sourceforge.net/samtools.shtml

Slides curiosity from Aaron Quinlan: https://github.com/quinlan-lab/applied-computational-genomics

Col #	Name	Meaning	Example
1	QNAME	Read or Pair name	HWI:ST156_1:278:1:1058:4544:0
2	FLAG	Bitwise FLAG	***Much more soon!***
3	RNAME	Reference sequence name	chr1
4	POS	1-based alignment start coordinate	8,724,005
5	MAPQ	Mapping quality	60
6	CIGAR	Extended CIGAR string	***Much more soon!***
7	MRNM	If paired, the mate's reference seq.	chr1
8	MPOS	If paired, the mate's alignment start	8,724,505
9	ISIZE	If paired, the insert size	562
10	SEQ	The sequence of the query/mate	ACAAATTCAG...
11	QUAL	The quality string for the query/mate	HHH$^^%$$$...
12	OPT	Optional Tags	XA:i:2, MD:Z:0T34G15

arq5x@beast:~/beast_data/Pat — ssh — 101×8

arq5x@beas.../Pat — bash | arq5x@beas.../Pat — bash | arq5x@beas...a/Pat — ssh | java

```
[arq5x@beast Pat]$ samtools view 1094PC0005.possrt.conc.bam | head -1
HWI-ST156_1:278:66:2461:8880:0 99     chr10 50133   0       51M     =       50214   132     TAATT
GACGCGCTGTTCAGCCCTTTGAGTTCGGTTGAGTTTTGGGTTGGAG  GFDGDFFFDBEEEEE:DDDDFDFEFGGEEGFAFDFGFFB?BBABB@EBDA? X
T:A:R   NM:i:0  SM:i:0  AM:i:0  X0:i:2  X1:i:0  XM:i:0  XO:i:0  XG:i:0  MD:Z:51 XA:Z:chr10,+50133,51M
,0;
```

1 2 3 4

Slides curiosity from Aaron Quinlan: https://github.com/quinlan-lab/applied-computational-genomics

```
ST-E00223:32:H5J57CCXX:6:2123:15189:52872       97      1       10001   0       4S15M1I54M2I50M25S      4       699063  0
ACCCTAACCCTAACCCTAACCCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAAACCCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCAACCCTAACCC
CCCCCCACCCAACCCCACCCCCCAC
AAFFFKKKKKKKKKKKKKKKKKKKFKKFKKKKKKFKKAKKKKFFKKKKKFKF<FF7FFFFK7FK,AA,FKFFKF,FKKK7<KKK,,7FFF7F,AAFFK,AKA,,FFF<,A7FKK,,,7A,,,7,,
,,,((,(,,(,,,,,,,(,,A(,,( MC:Z:80S11M1D58M        MD:Z:119        RG:Z:15-0017315_1       NM:i:3  MQ:i:47 AS:i:104
XS:i:103
ST-E00223:46:HG7V5CCXX:2:1116:12601:22862       1123    1       10006   0       81M70S  =       10106   143
CTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACACTCACCCTAACCCTACCCCTAACCCTAACCCTATCATTCACTCGAACCCTAACACTACCGCTAGCGCTAACTCCAGCC
CGCACACTATCGCTAACCCTCACGC
AAFFFKKKKKKKKKKKKKKKFKKA,,A,A,,77<,,A,,7FFK7,,,,AF,,,A7,,,77,,,7AA,,7,FFK,<A,,,,7<<,,,,AA,7,AA,,7,,,7,,,,,,,,,,A7,,,,,,7,7,7(
,(,((,,,7,A,F,7,,,<,,AA,, MC:Z:108S43M    MD:Z:47C2A12A17 RG:Z:15-0017315_1       NM:i:3  MQ:i:2  AS:i:66 XS:i:71
ST-E00223:32:H5J57CCXX:5:2208:10074:43308       99      1       10008   36      101M1I41M7S     =       10107   137
AACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTACCCCTAACCCTAACCCTAACCCTAACCCTCACCCTCACCCTCACCCTAGCCCAAACCCTAACCCTAACCCTAACCCT
AACCCTAACCCTAACCCTACCCCG
AAFFFKFKKKKKKKKKKKKKKKKFKKKKKKKFKKK7<KA7<F,AF7F,7,A,FA7<KKKKKFKKKK<,,FKAKKAFFA,A7,7,,,7,7,,,,,7F<,,7,,,,7,,A,,,<F7,,,7FA,7,,<F
,,,<A,7<AFKK<,<,,7,,,,,(  MC:Z:112S38M    MD:Z:49A28A5A5A6A44     RG:Z:15-0017315_1       NM:i:6  MQ:i:36 AS:i:110
XS:i:113
ST-E00223:46:HG7V5CCXX:5:2119:12936:64896       99      1       10013   0       90M61S  =       10176   211
TAACCCTAAGCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCATAACCCAAACCCTAACCCTAACCCTAACCCGAACCGTAAGCCAAAACATAACCACAACCATAACAA
TAACCAAAACCTTAACGTTAAACAT
A<AFFKKKK,AAAFKK<FKKKKAK,<A,AFKKKKAFFFKKKKAFKKFAFFFKKA7,FFA,F,F,7F,AFF77AFFAFFKAFKKA,,FFKF,AA,F,,7F,A,,7A,,,7A,,,,,,AFK,,,AA,,
,7FKA,A,AFF,<,,,,,,,77<AA MC:Z:99S48M4S   MD:Z:9C49C6T23  RG:Z:15-0017315_1       NM:i:3  MQ:i:0  AS:i:75 XS:i:75
ST-E00223:32:H5J57CCXX:1:1205:17290:54577       99      1       10019   1       92M59S  =       10354   414
TAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCAAACCCCCACCCCAAACCCAAGCCCAACCCTACCCC
CTAACCCCTAACCCCAACCCTGACC
AAFFFKKKKFKKKKAKKKKKKKKKKKKKKKF7KKKKKKKKKKKKKFKKKKFKKKKKKKAKKKFFKKFFKKKK,<,7<FA,FFFK,,7FFF,,,,,A,((,(((,7,7A,,,,A,,,777A,AAKK
<<,7FFA7,A,,AA,<,,AFA,7AF MC:Z:72S79M     MD:Z:92 RG:Z:15-0017315_1       NM:i:0  MQ:i:20 AS:i:92 XS:i:97
```

Slides curiosity from Aaron Quinlan: https://github.com/quinlan-lab/applied-computational-genomics

SAM/BAM

qname: sequence name.

```
HWI-ST1145:74:C101DACXX:7:1102:4284:73714        16      chr20   190930  3       100M    *       0       0
        CCGTGTTTAAAGGTGGATGCGGTCACCTTCCCAGCTAGGCTTAGGGATTCTTAGTTGGCCTAGGAAATCCAGCTAGTCCTGTCTCTCAGTCCCCCCTCT
C    BBDCCDDCCDDDDCDDDDDDCDCCCDBC?DDDDDDDDDDDDDDDCCDCDDDDDDDDDDCCCCEDDDC?DDDDDDDDDDDDDDDDDDDDDDBDHFFFFDC@@
    AS:i:-15        XM:i:3  XO:i:0  XG:i:0  MD:Z:55C20C13A9 NM:i:3  NH:i:2  CC:Z:=  CP:i:55352714   HI:i:0
```

Field 1, qname is the name of the read. Read names often contain information about:

1.The scientific study for which the read was sequenced.

2.The sequencing instrument, and the exact part of the sequencing instrument, where the DNA was sequenced.

SAM/BAM

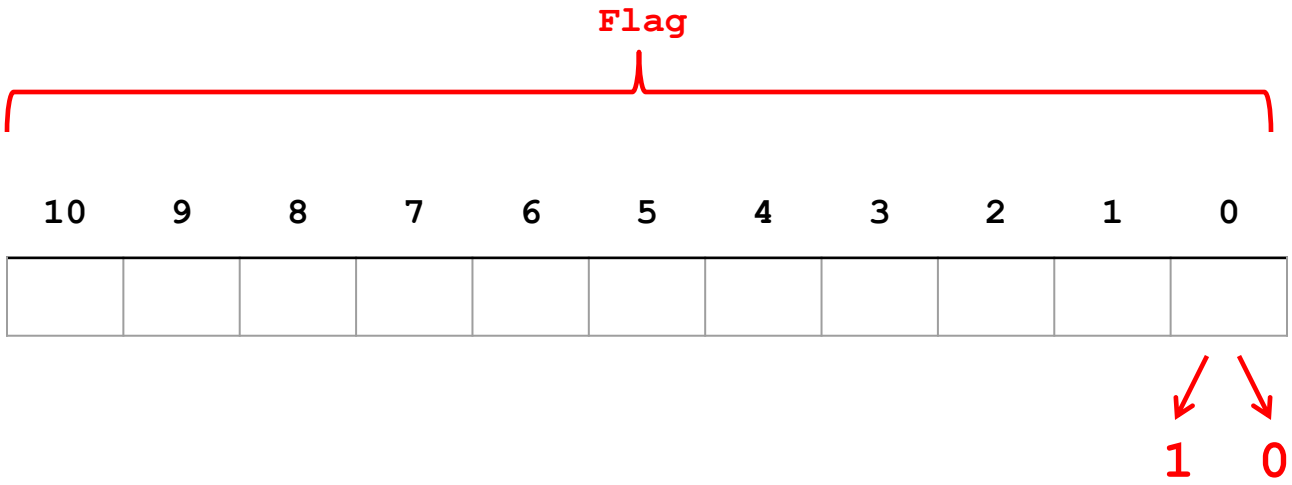
flag is a bit field encoding some yes/no pieces of information about whether and how the read aligned

flag

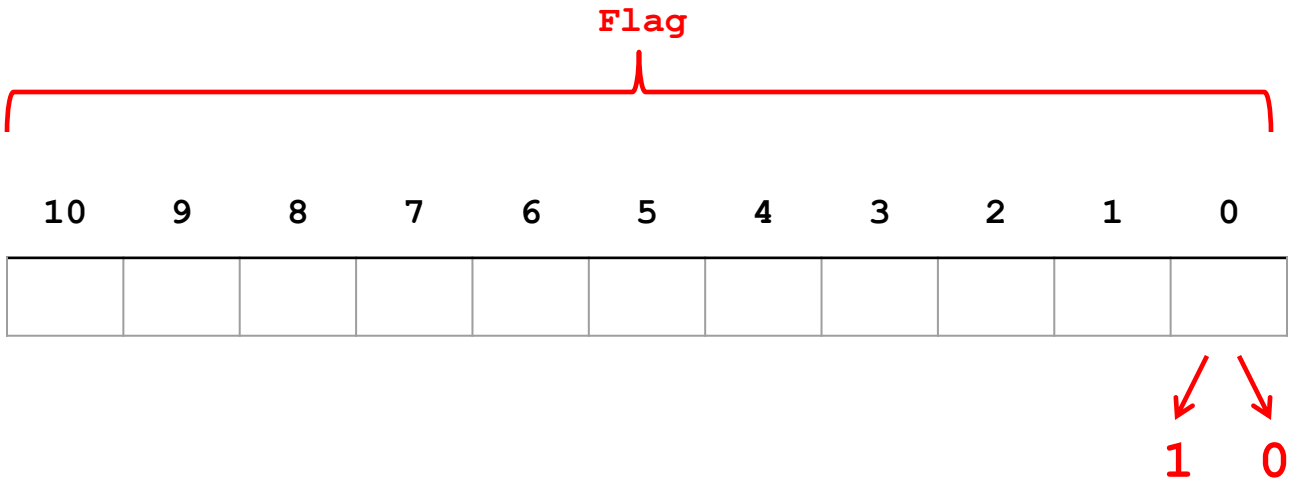
```
HWI-ST1145:74:C101DACXX:7:1102:4284:73714      16      chr20    190930  3       100M    *       0       0
        CCGTGTTTAAAGGTGGATGCGGTCACCTTCCCAGCTAGGCTTAGGGATTCTTAGTTGGCCTAGGAAATCCAGCTAGTCCTGTCTCTCAGTCCCCCCTCT
C     BBDCCDDCCDDDDCDDDDDDCDCCCDBC?DDDDDDDDDDDDDDDCCDCDDDDDDDDDDCCCCEDDDC?DDDDDDDDDDDDDDDDDDDDDBDHFFFFDC@@
    AS:i:-15        XM:i:3  XO:i:0  XG:i:0  MD:Z:55C20C13A9 NM:i:3  NH:i:2  CC:Z:=  CP:i:55352714   HI:i:0
```

base2	base10	base16	Meaning	Applies to:
00000000001	1	0x0001	The read originated from a paired sequencing molecule	Both
00000000010	2	0x0002	The read is mapped in a **proper** pair	Pairs only
00000000100	4	0x0004	The query sequence itself is unmapped	Both
00000001000	8	0x0008	The query's mate is unmapped	Pairs only
00000010000	16	0x0010	Strand of the query (0 for forward; 1 for reverse strand)	Both
00000100000	32	0x0020	Strand of the query's mate	Pairs only
00001000000	64	0x0040	The query is the first read in the pair	Pairs only
00010000000	128	0x0080	The read is the second read in the pair	Pairs only
00100000000	256	0x0100	The alignment is not primary	Both
01000000000	512	0x0200	The read fails platform/vendor quality checks	Both
10000000000	1024	0x0400	The read is either a PCR duplicate or an optical duplicate	Both

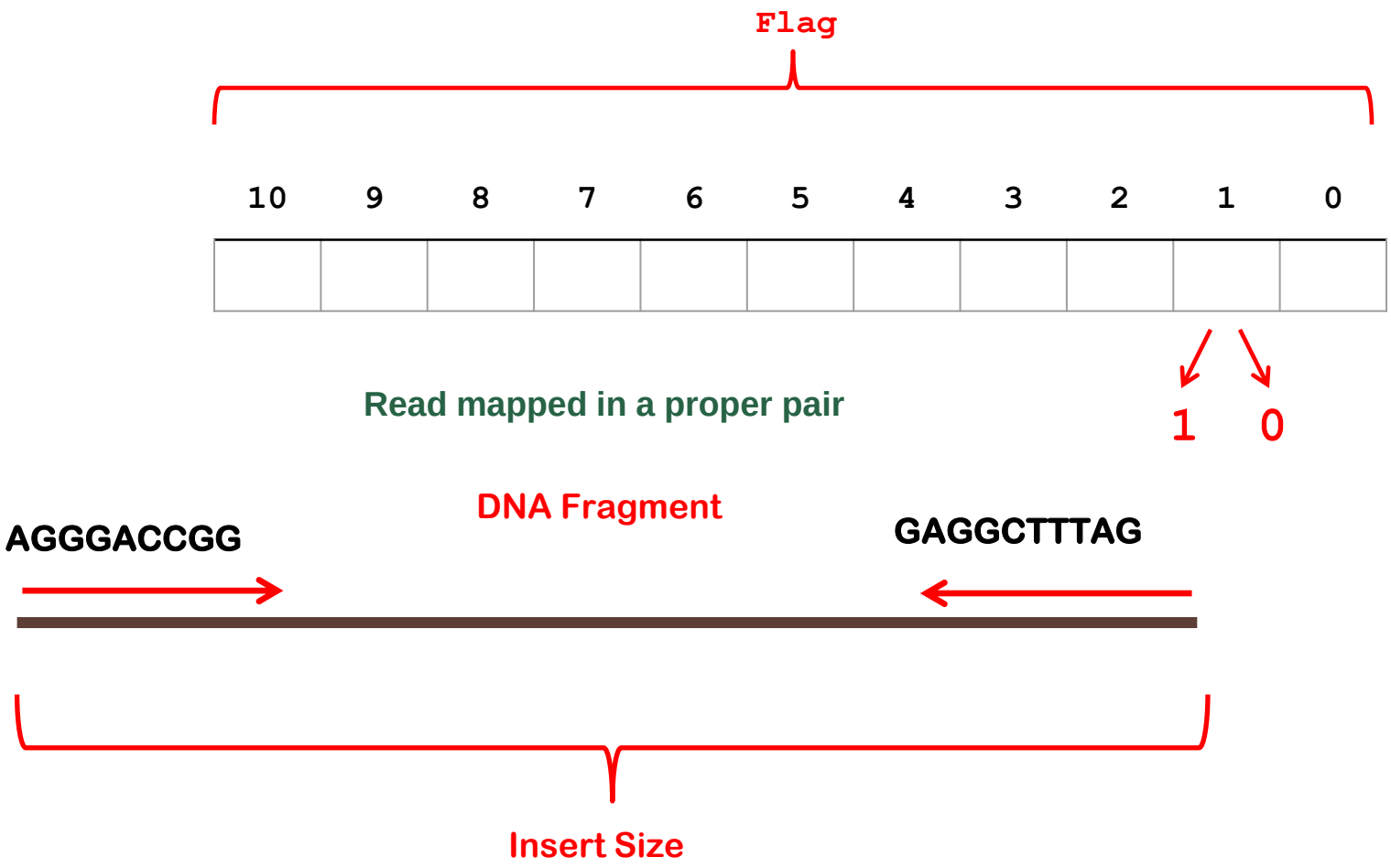
Slides adapted from Aaron Quinlan: https://github.com/quinlan-lab/applied-computational-genomics

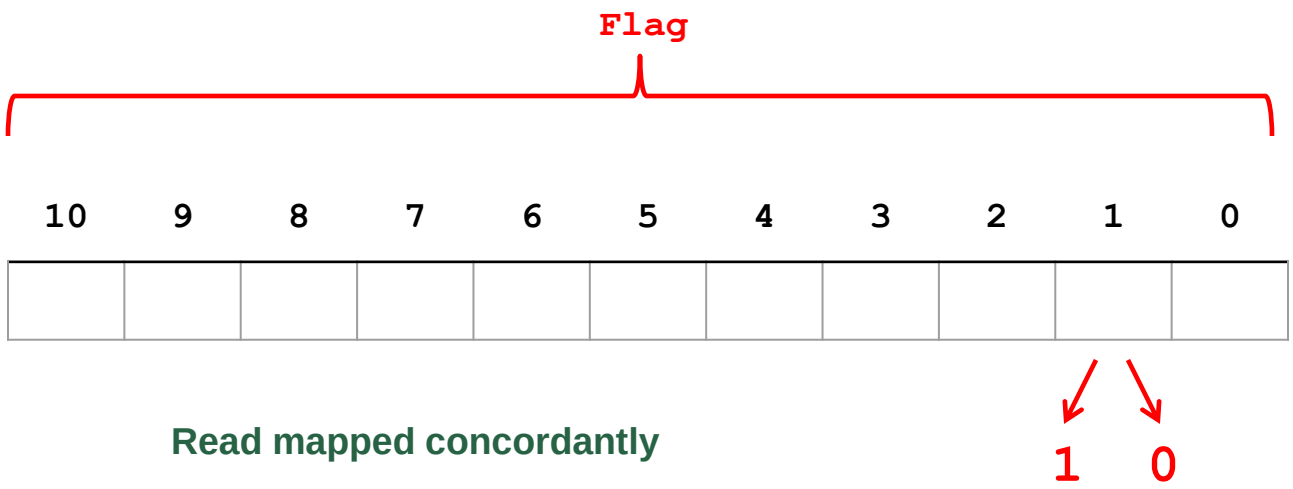


Single-end sequencing read



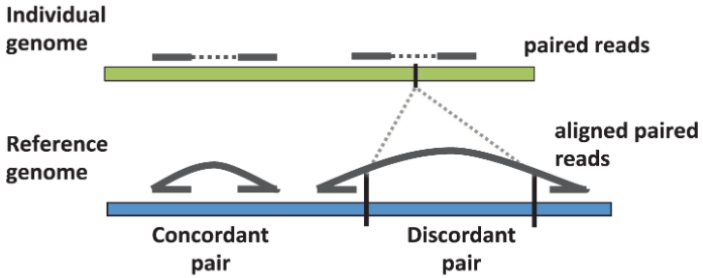
paired-end sequencing read


Flag
10
9
8
7
6
5
4
3
2
1
0
1
0
Read mapped in a proper pair
DNA Fragment
AGGGACCGG
GAGGCTTTAG
Insert Size

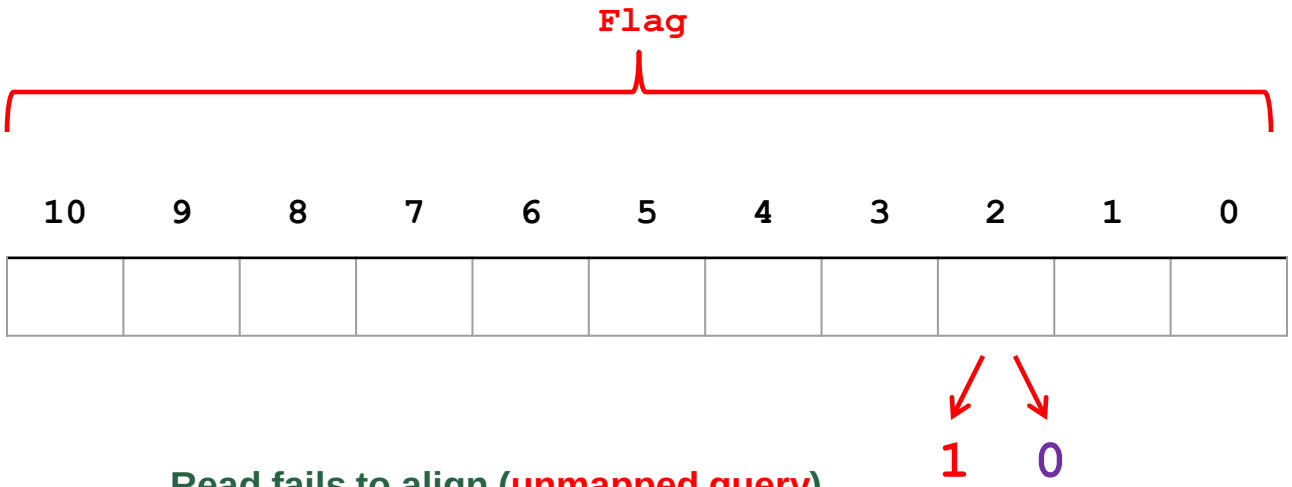


Read mapped concordantly

```
AGGGTTTGGTTCGTTTTAGGGTTTGGTTCGTTGAGGCTTTAG
|||| |||||||||||           |||||||||||  ||||
AGGGATTGGTTCGTT ___________ GGTTCGTTGAAGCTT
```

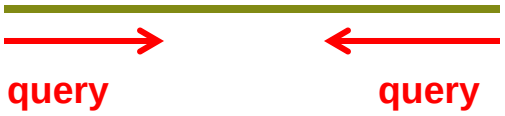


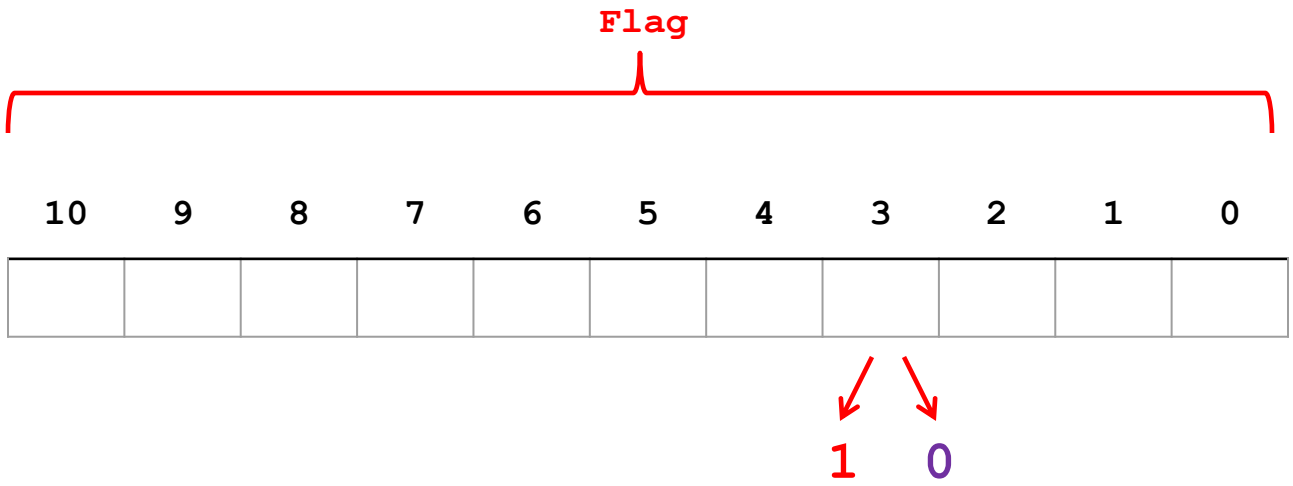
doi: https://doi.org/10.1371/journal.pcbi.1002821.g005



Read fails to align (unmapped query)

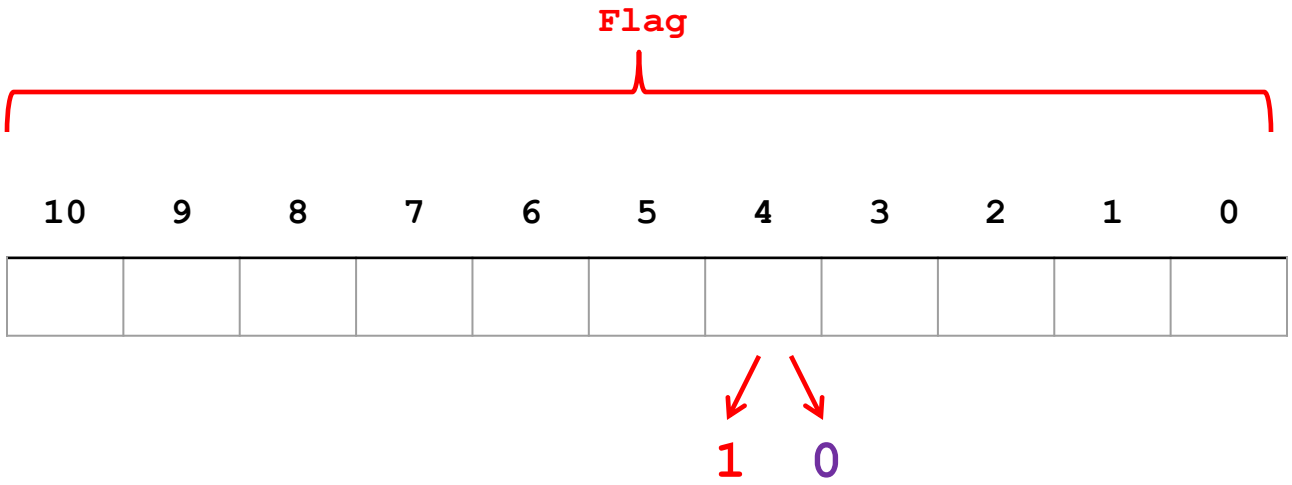
Read aligned successfully (mapped query)





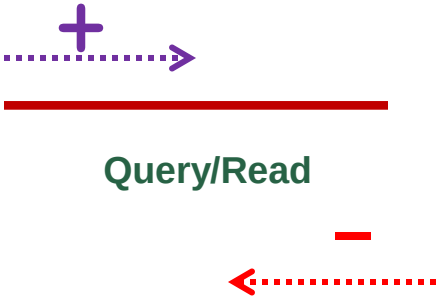
Read mate fails to align (unmapped query mate)

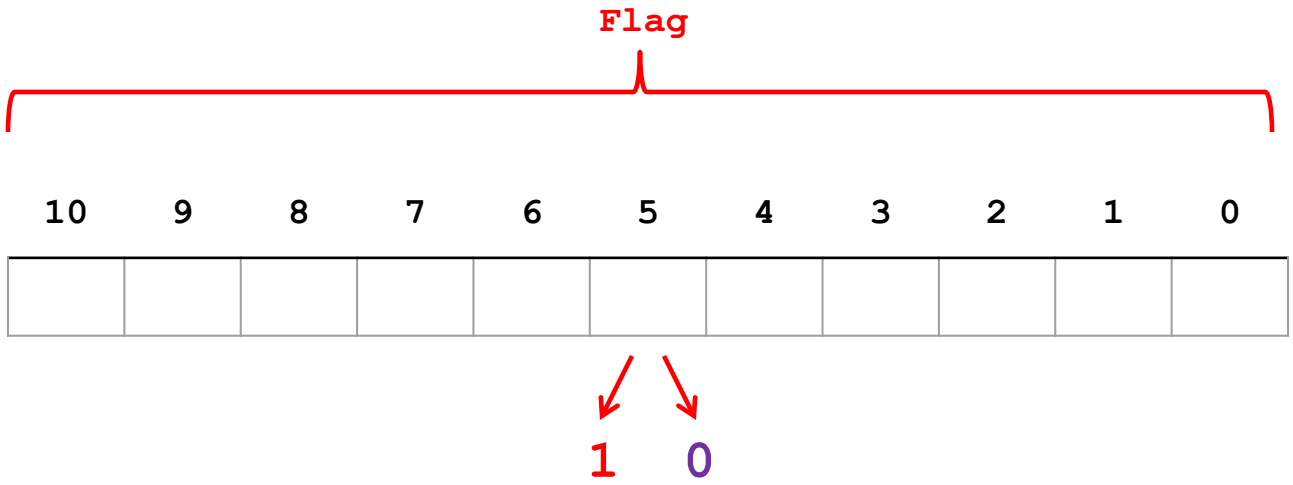
Read mate aligned successfully (mapped query mate)



Strand of the query (0: forward direction,+, Watson strand)

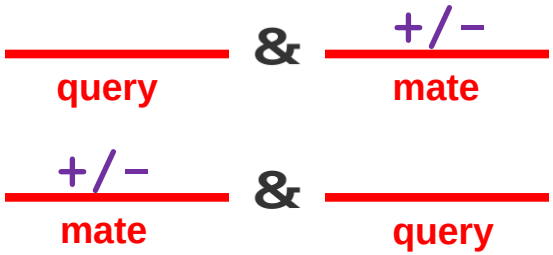
Strand of the query (1: reverse complement,-, Crick strand)

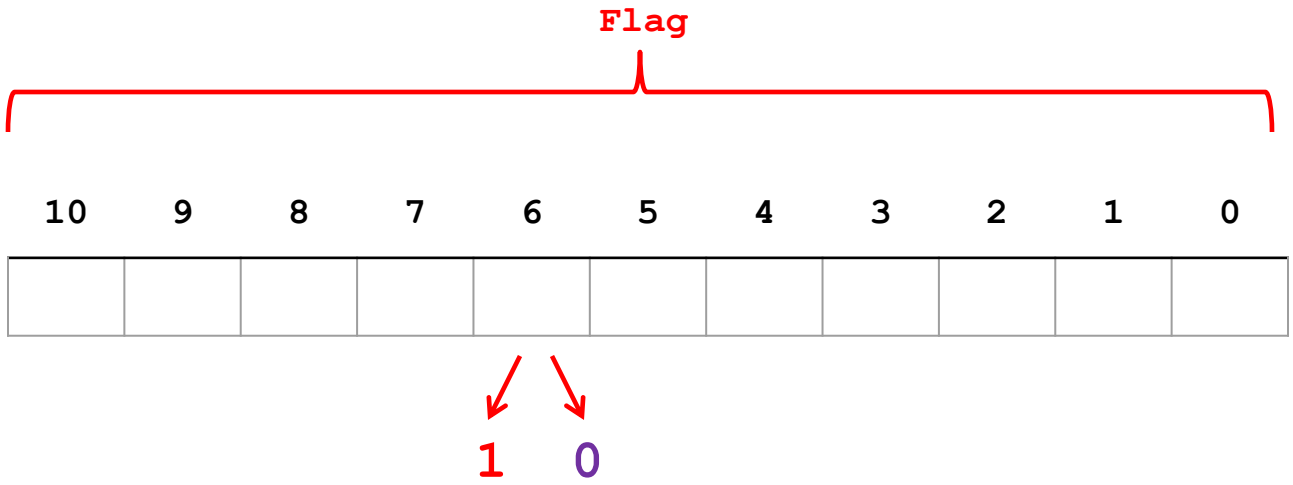




Strand of the query's mate (0: forward direction,+, Watson strand)

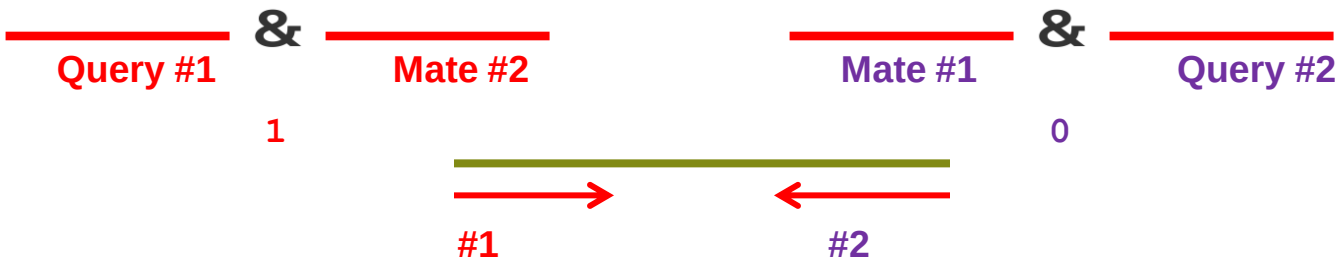
Strand of the query's mate (1: reverse complement,-, Crick strand)

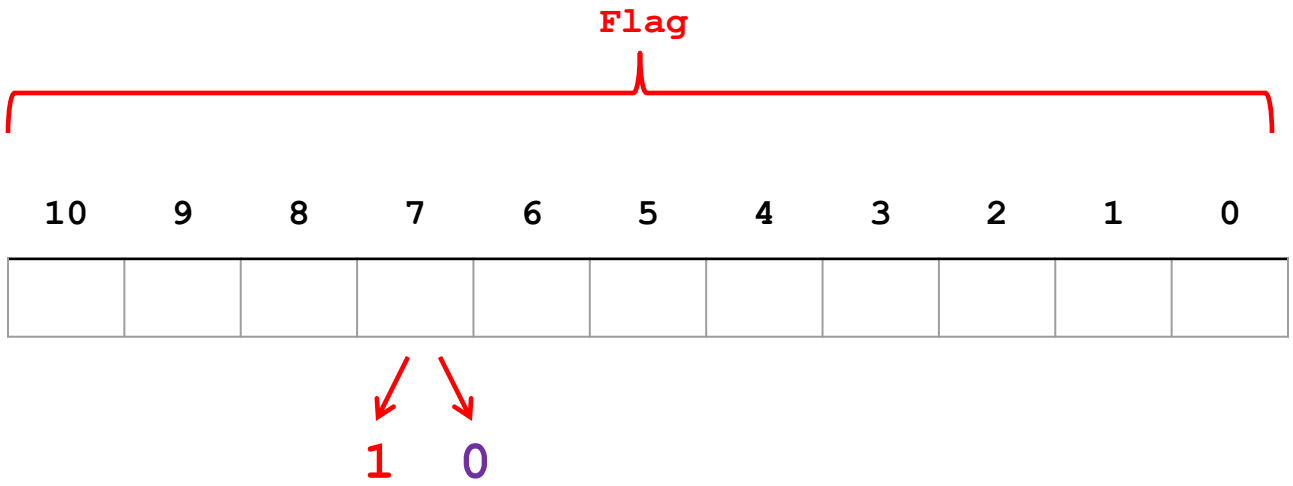




the query is the second read in the pair (0: #2)

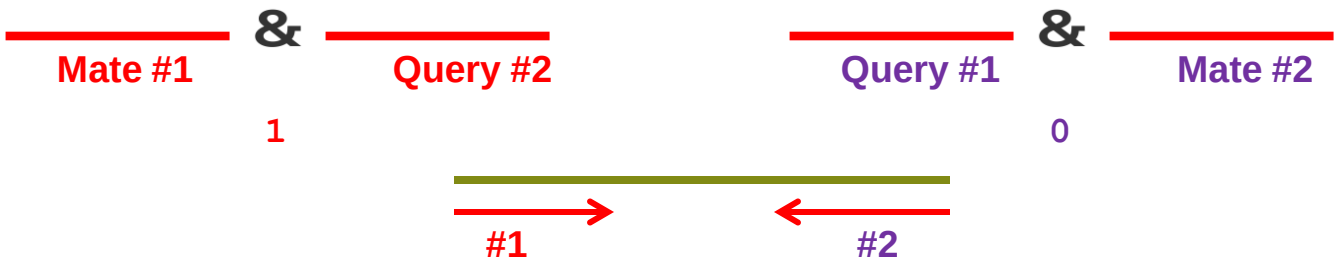
the query is the first read in the pair (1: #1)





the query is the first read in the pair (0: #1)

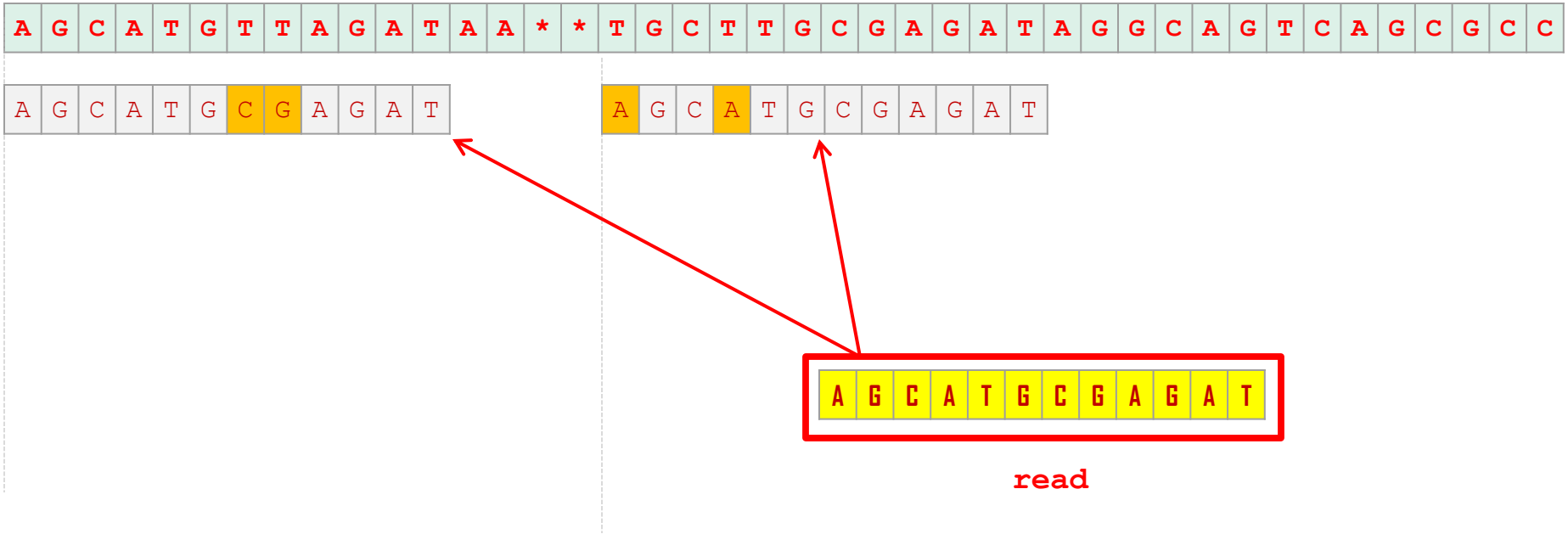
the query is the second read in the pair (1: #2)

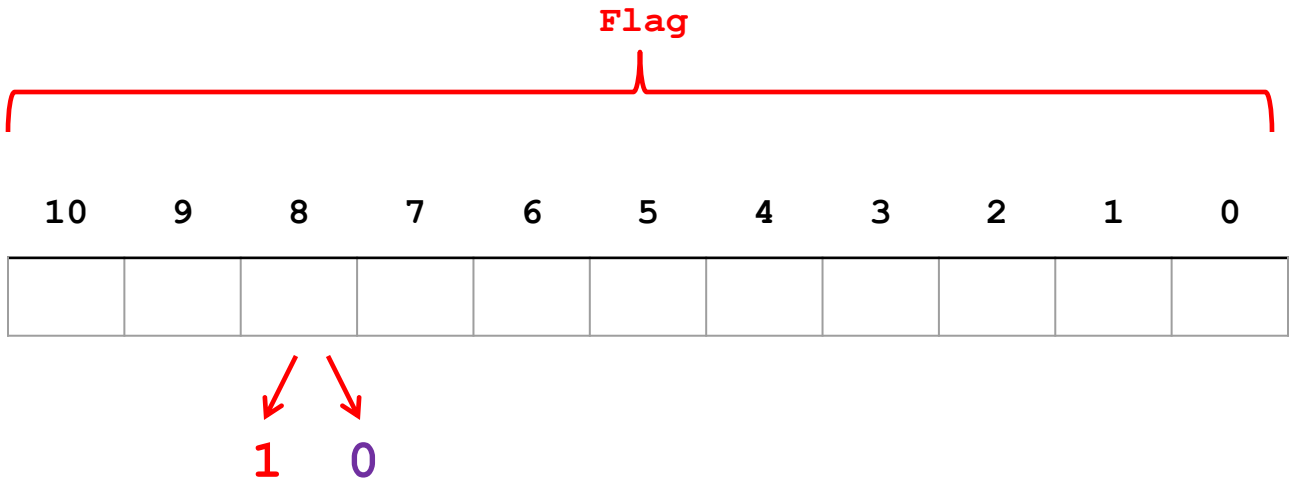


Note

Multiple mapping positions

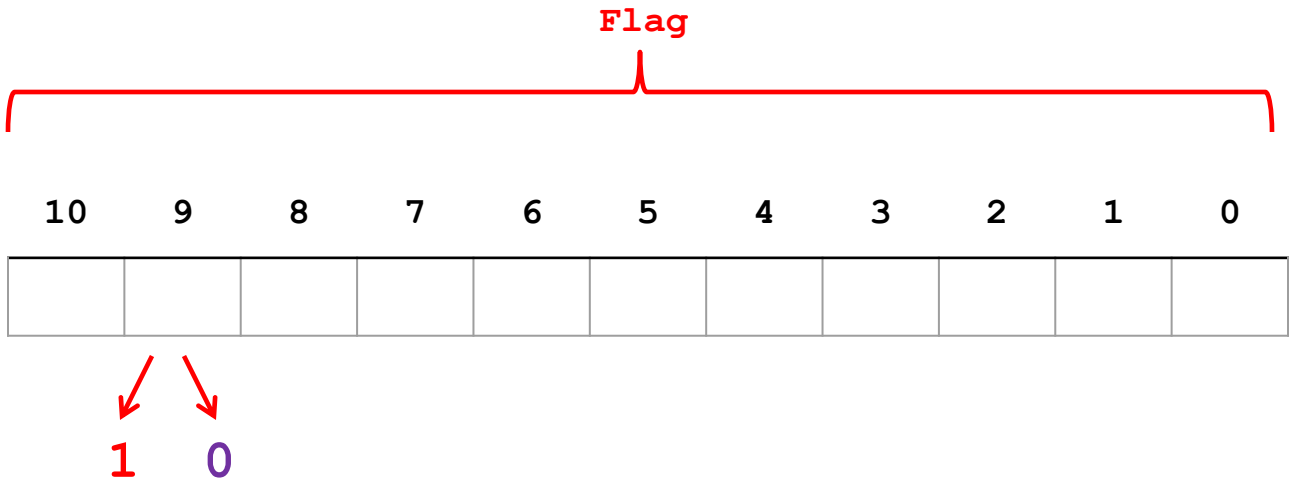
Which one is Primary ? Secondary?





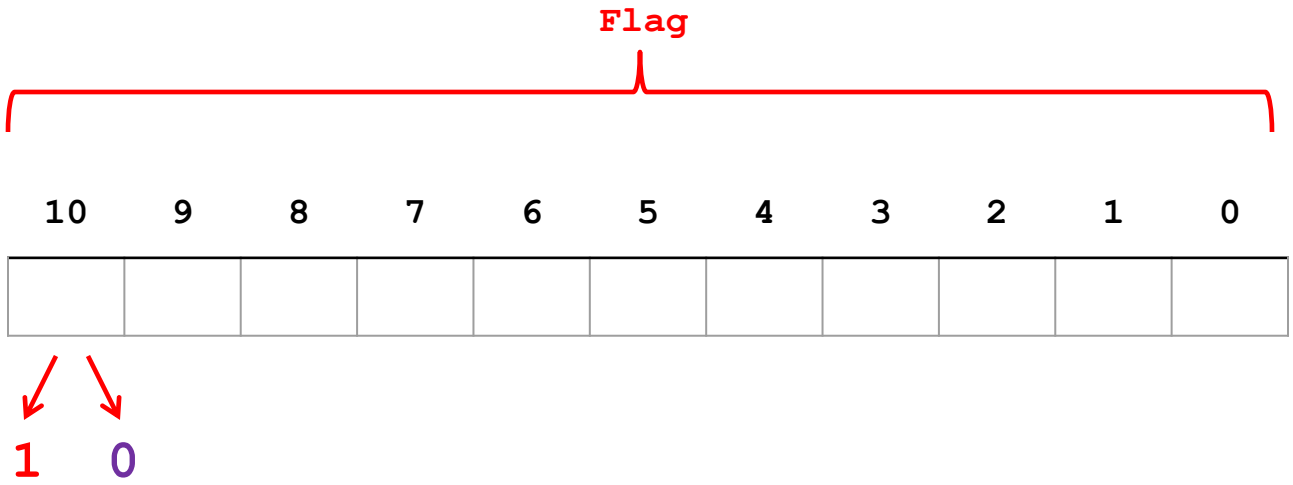
The alignment is primary(0: Primary Alignment)

The alignment is not primary (1: Secondary Alignment)



The read passing filters, such as platform/vendor quality controls (0)

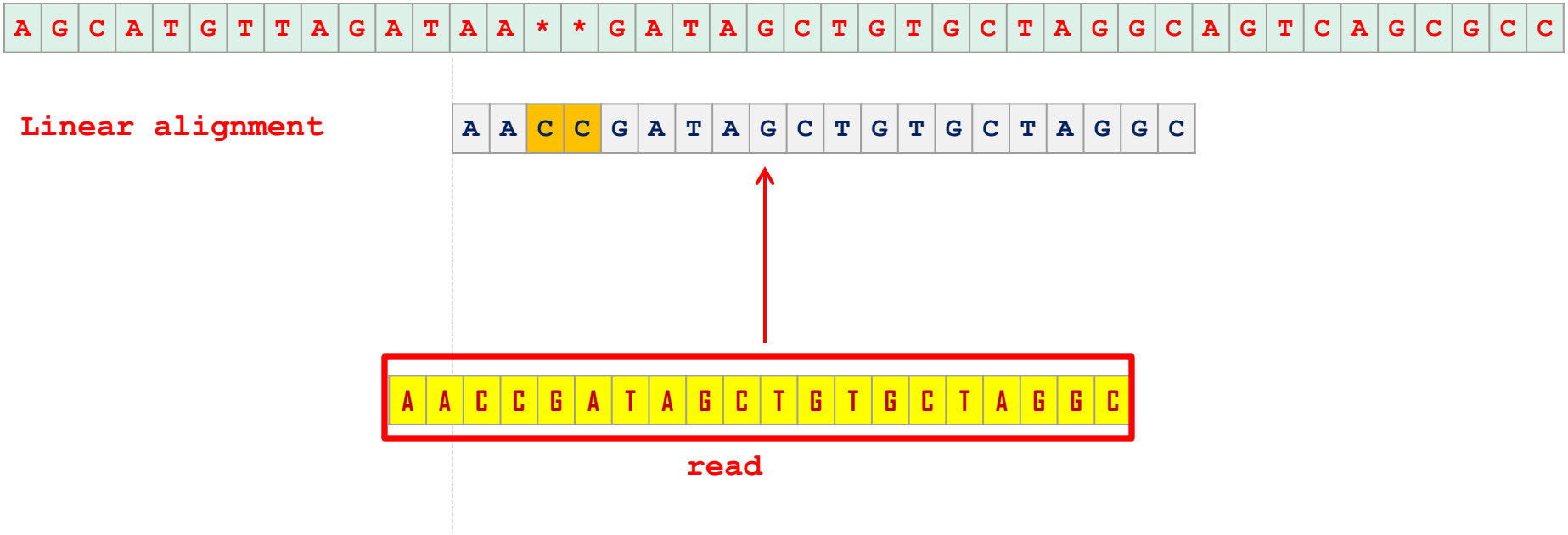
The read not passing filters, such as platform/vendor quality controls (1)



Read is not PCR or optical duplicate(0)

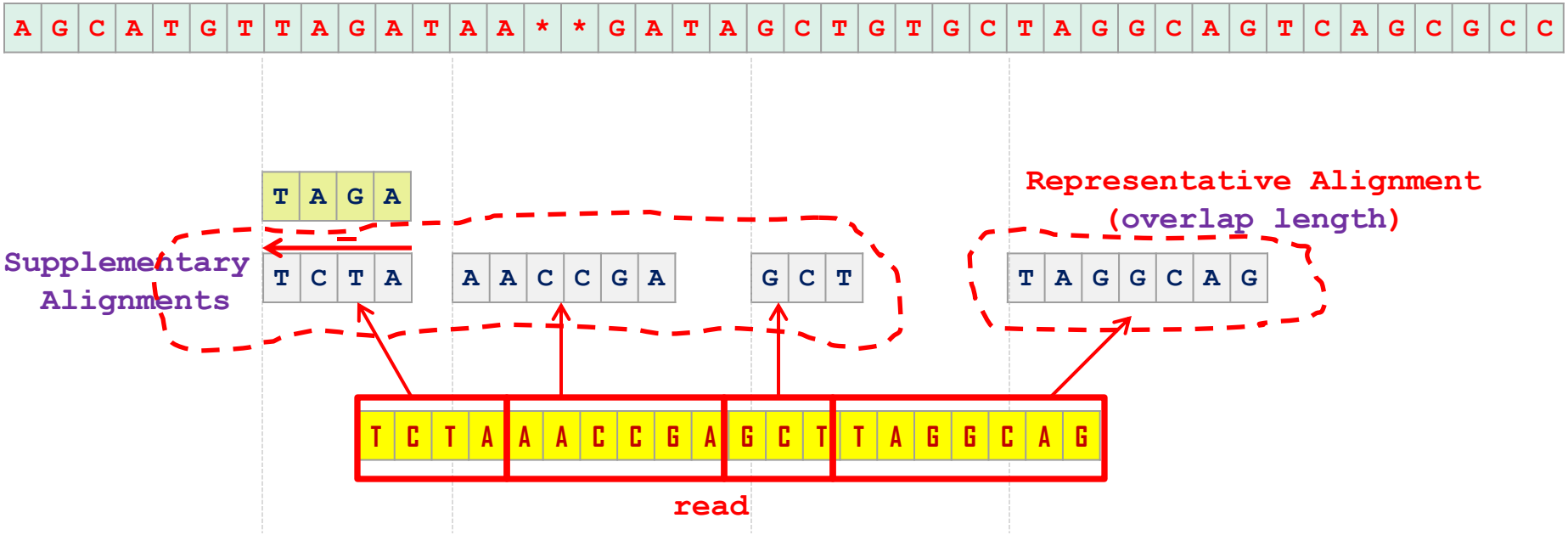
Read is PCR or optical duplicate (1)

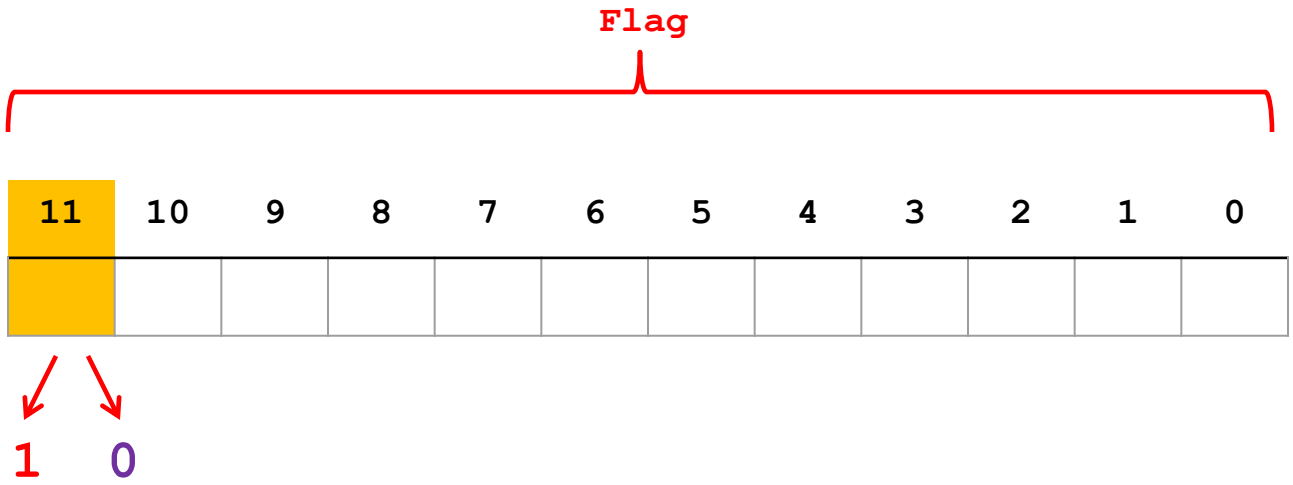
Note


A G C A T G T T A G A T A A * * G A T A G C T G T G C T A G G C A G T C A G C G C C
Linear alignment
A A C C G A T A G C T G T G C T A G G C
A A C C G A T A G C T G T G C T A G G C
read

Note

Chimeric/Non-linear alignment





Not supplementary alignment (0)

Supplementary alignment (1)

Flag

10	9	8	7	6	5	4	3	2	1	0
1024	512	256	128	64	32	16	8	4	2	1

Flag

ST-E00223:32:H5J57CCXX:4:1220:14651:8868 99 1 10086

Flag

10	9	8	7	6	5	4	3	2	1	0
1024	512	256	128	64	32	16	8	4	2	1
0	0	0	0	1	1	0	0	0	1	1

Read is aligned to the Watson strand, +

Flag

ST-E00223:32:H5J57CCXX:4:1220:14651:8868 99 1 10086

- ❑ **Read is a paired-end.**
- ❑ **Pairs are aligned concordantly (proper pairs).**
- ❑ **Read's mate is aligned to the Crick strand, - .**
- ❑ **This is the first end (read #1).**

Read

SAM/BAM

Flags

The `flags` field is a bitfield. Individual bits correspond to certain yes/no properties of the alignment. Here are the most relevant ones:

- Bit 0 (least significant): 1 if read is paired-end, 0 otherwise
- Bit 1: for paired-end reads only: 1 if the pair aligns concordantly, 0 otherwise
- Bit 2: 1 if read failed to align, 0 otherwise
- Bit 3: for paried-end reads only: 1 if the other end failed to align, 0 otherwise
- Bit 4: 1 if read aligned to Crick strand, 0 if Watson strand
- Bit 5: for paired-end reads only: 1 if the other end aligned to Crick strand, 0 if Watson strand
- Bit 6: for paired-end reads only: 1 if this is the first (#1) end, 0 if this is the second (#2) end
- Bit 7: for paired-end reads only: 0 if this is the first (#1) end, 1 if this is the second (#2) end

There are a few more that are used less often; see the [SAM specification] for details.

SAM/BAM

POS

```
HWI-ST1145:74:C101DACXX:7:1102:4284:73714       16      chr20   190930  3       100M    *       0       0
        CCGTGTTTAAAGGTGGATGCGGTCACCTTCCCAGCTAGGCTTAGGGATTCTTAGTTGGCCTAGGAAATCCAGCTAGTCCTGTCTCTCAGTCCCCCCTCT
C     BBDCCDDCCDDDDCDDDDDDDCDCCCDBC?DDDDDDDDDDDDDDDCCDCDDDDDDDDDDCCCCEDDDC?DDDDDDDDDDDDDDDDDDDDDDBDHFFFFDC@@
    AS:i:-15        XM:i:3  XO:i:0  XG:i:0  MD:Z:55C20C13A9 NM:i:3  NH:i:2  CC:Z:=  CP:i:55352714   HI:i:0
```

POS is the 1-based offset into the reference sequence where the read aligned

SAM/BAM

MAPQ

```
HWI-ST1145:74:C101DACXX:7:1102:4284:73714        16      chr20    190930  3       100M     *       0       0
        CCGTGTTTAAAGGTGGATGCGGTCACCTTCCCAGCTAGGCTTAGGGATTCTTAGTTGGCCTAGGAAATCCAGCTAGTCCTGTCTCTCAGTCCCCCCTCT
C     BBDCCDDCCDDDDCDDDDDDDCDCCCDBC?DDDDDDDDDDDDDDDCCDCDDDDDDDDDDCCCCEDDDC?DDDDDDDDDDDDDDDDDDDDDDBDHFFFFDC@@
    AS:i:-15        XM:i:3  XO:i:0  XG:i:0  MD:Z:55C20C13A9 NM:i:3  NH:i:2  CC:Z:=  CP:i:55352714   HI:i:0
```

MAPQ: For an aligned read, this is a confidence value; high when we're very confident we've found the correct alignment, low when we're not confident.

$$MAPQ = -10*\log_{10}(P_{map\_loc\_wrong})$$

$(P_{map\_loc\_wrong})$	$\log_{10}(P_{map\_loc\_wrong})$	MAPQ
1	0	0
0.1	-1	10
0.01	-2	20
0.001	-3	30
0.0001	-4	40

SAM/BAM

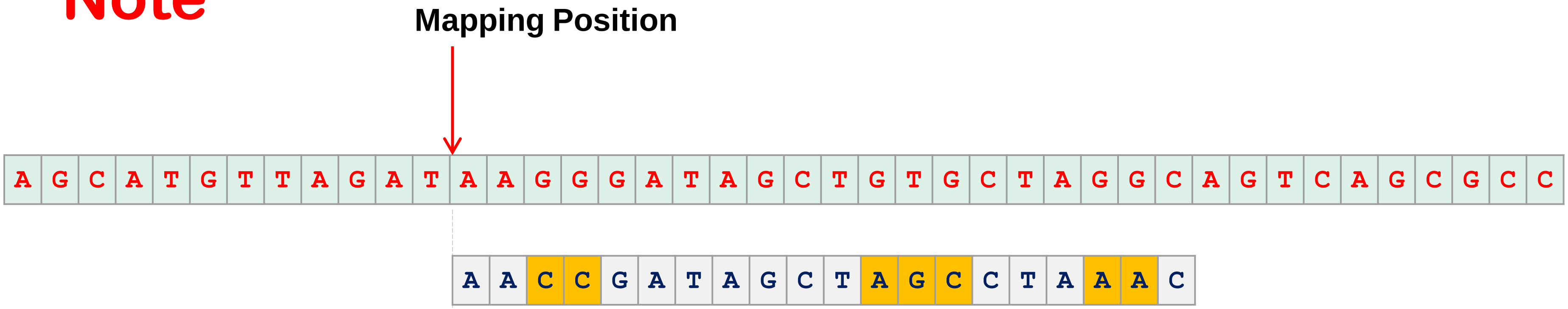
cigar

```
HWI-ST1145:74:C101DACXX:7:1102:4284:73714       16      chr20   190930  3       100M    *       0       0
      CCGTGTTTAAAGGTGGATGCGGTCACCTTCCCAGCTAGGCTTAGGGATTCTTAGTTGGCCTAGGAAATCCAGCTAGTCCTGTCTCTCAGTCCCCCCTCT
C    BBDCCDDCCDDDDCDDDDDDCDCCCDBC?DDDDDDDDDDDDDDCCDCDDDDDDDDDDCCCCEDDDC?DDDDDDDDDDDDDDDDDDDDDDBDHFFFFDC@@
    AS:i:-15        XM:i:3  XO:i:0  XG:i:0  MD:Z:55C20C13A9 NM:i:3  NH:i:2  CC:Z:=  CP:i:55352714   HI:i:0
```

The cigar string : **encode the details of the alignment.**

Slides adapted from Aaron Quinlan: https://github.com/quinlan-lab/applied-computational-genomics

Note



CIGAR String	20M
Extended CIGAR String	2=2X7=3X3=2X1=

Note

Mapping Position

A	G	C	A	T	G	T	T	A	G	A	T	A	A	-	-	G	A	T	A	G	C	T	G	T	G	C	T	A	G	G	C	A	G	T	C	A	G	C	G	C	C

A	A	C	C	G	A	T	A	G	C	T	A	G	C	C	T	A	A	A	C

CIGAR String	2M2I16M

Extended CIGAR String	2=2I7=3X3=2X1=

Note

Mapping Position

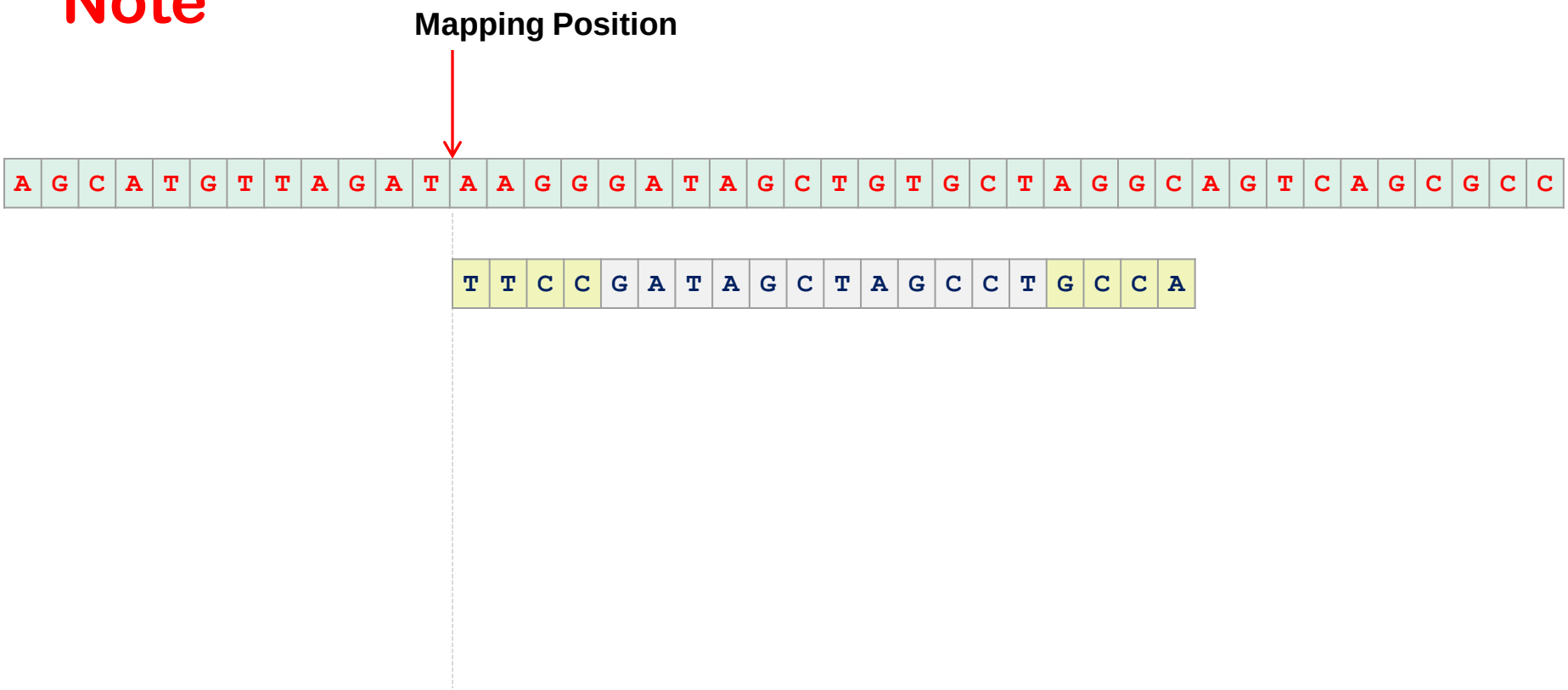
A	G	C	A	T	G	T	T	A	G	A	T	A	A	-	-	G	A	T	A	G	C	T	G	T	G	C	T	A	G	G	C	A	G	T	C	A	G	C	G	C	C

A	A	C	C	G	A	T	A	G	C	T	A	-	-	C	T	A	A	A	C

CIGAR String	2M2I8M2D6M

Extended CIGAR String	2=2I7=1X2D3=2X1=

Note


Mapping Position
A G C A T G T T A G A T A A G G G A T A G C T G T G C T A G G C A G T C A G C G C C
T T C C G A T A G C T A G C C T G C C A

Note

Start Position

End Position

A	G	C	A	T	G	T	T	A	G	A	T	A	A	G	G	G	A	T	A	G	C	T	G	T	G	C	T	A	G	G	C	A	G	T	C	A	G	C	G	C	C

T	T	C	C	G	A	T	A	G	C	T	A	G	C	C	T	G	C	C	A

CIGAR String	4S13M3S

SEQ	TTCCGATAGCTAGCCTGCCA

Note

Start Position

End Position

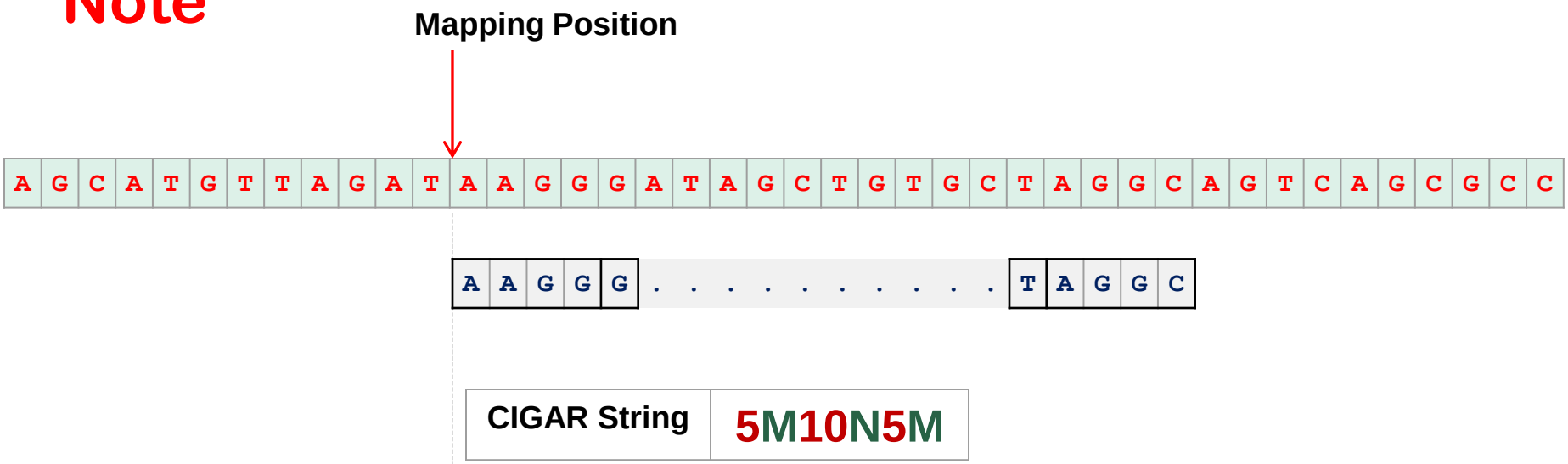
A	G	C	A	T	G	T	T	A	G	A	T	A	A	G	G	G	A	T	A	G	C	T	G	T	G	C	T	A	G	G	C	A	G	T	C	A	G	C	G	C	C

T	T	C	C	G	A	T	A	G	C	T	A	G	C	C	T	G	C	C	A

CIGAR String	4H13M3H

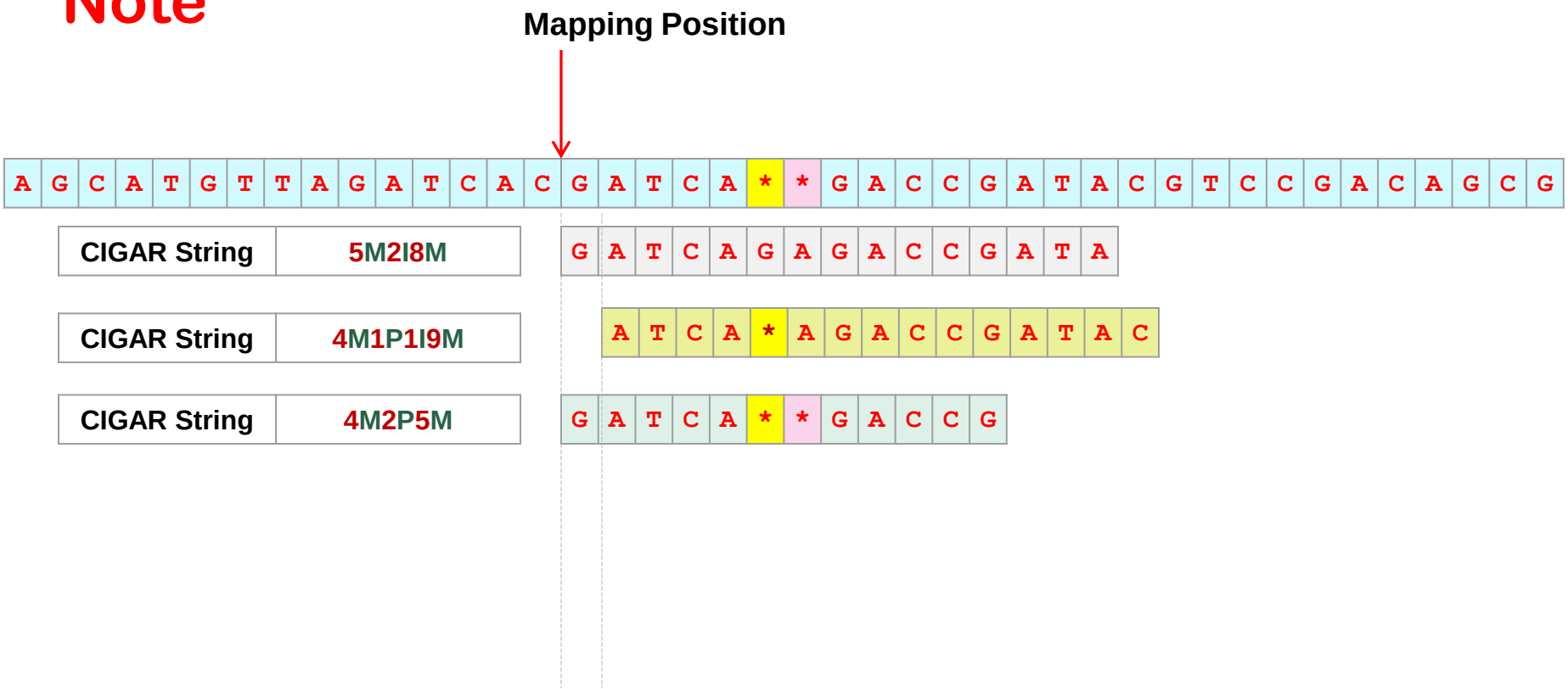
SEQ	GATAGCTAGCCTG

Note



CIGAR String	5M10N5M

Note


Mapping Position
A G C A T G T T A G A T C A C G A T C A * * G A C C G A T A C G T C C G A C A G C G
CIGAR String
5M2I8M
G A T C A G A G A C C G A T A
CIGAR String
4M1P1I9M
A T C A * A G A C C G A T A C
CIGAR String
4M2P5M
G A T C A * * G A C C G

SAM/BAM

cigar

```
HWI-ST1145:74:C101DACXX:7:1102:4284:73714        16       chr20    190930  3       100M     *        0        0
        CCGTGTTTAAAGGTGGATGCGGTCACCTTCCCAGCTAGGCTTAGGGATTCTTAGTTGGCCTAGGAAATCCAGCTAGTCCTGTCTCTCAGTCCCCCCTCT
C     BBDCCDDCCDDDDCDDDDDDCDCCCDBC?DDDDDDDDDDDDDDCCDCDDDDDDDDDDCCCCEDDDC?DDDDDDDDDDDDDDDDDDDDDDBDHFFFFDC@@
    AS:i:-15         XM:i:3  XO:i:0  XG:i:0  MD:Z:55C20C13A9 NM:i:3  NH:i:2  CC:Z:=  CP:i:55352714   HI:i:0
```

The cigar string : **encode the details of the alignment.**

Operation	Meaning
M	Match*
D	Deletion w.r.t. reference
I	Insertion w.r.t. reference
N	Split or spliced alignment
S	Soft-clipping
H	Hard-clipping
P	Padding

```
Reference:      ACCTGTC--TACCTTACG
Experimental:   ACCT-TCCATACTTTATC
                4M  1D 2M 2I    7M    2S
```

CIGAR string: 4M1D2M2I7M2S

LENGTH/OPERATION

Slides adapted from Aaron Quinlan: https://github.com/quinlan-lab/applied-computational-genomics

SAM/BAM

```
REF:AGCTAGCATCGTGTCGCCCGTCTAGCATACGCATGATCGACTGTCAGCTAGTCAGACTAGTC
Read:          GTGTAACCC.............................TCAGAATA
```

Operation	Meaning
=	Exact match
X	Mismatch
D	Deletion w.r.t. reference
I	Insertion w.r.t. reference
N	Split or spliced alignment
S	Soft-clipping
H	Hard-clipping
P	Padding

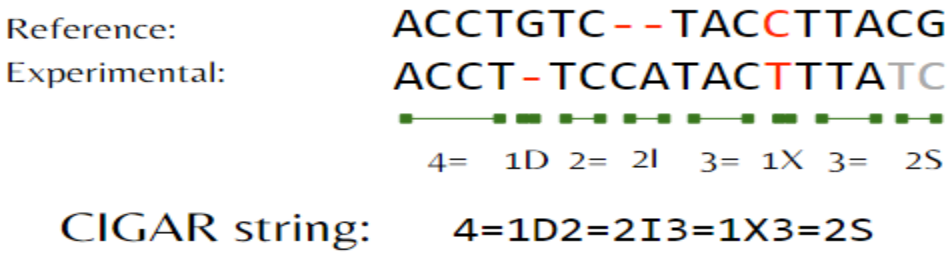
The CIGAR for this alignment is : 9M32N8M.

Slides adapted from Aaron Quinlan: https://github.com/quinlan-lab/applied-computational-genomics

SAM/BAM

The extended CIGAR string: M become = and X

Operation	Meaning
=	Exact match
X	Mismatch
D	Deletion w.r.t. reference
I	Insertion w.r.t. reference
N	Split or spliced alignment
S	Soft-clipping
H	Hard-clipping
P	Padding



Slides adapted from Aaron Quinlan: https://github.com/quinlan-lab/applied-computational-genomics

SAM/BAM

```
HWI-ST1145:74:C101DACXX:7:1102:4284:73714        16      chr20   190930  3       100M    *       0       0
        CCGTGTTTAAAGGTGGATGCGGTCACCTTCCCAGCTAGGCTTAGGGATTCTTAGTTGGCCTAGGAAATCCAGCTAGTCCTGTCTCTCAGTCCCCCCTCT
C     BBDCCDDCCDDDDCDDDDDDDCDCCCDBC?DDDDDDDDDDDDDDDCCDCDDDDDDDDDDCCCCEDDDC?DDDDDDDDDDDDDDDDDDDDDBDHEFFFDC@@
    AS:i:-15        XM:i:3  XO:i:0  XG:i:0  MD:Z:55C20C13A9 NM:i:3  NH:i:2  CC:Z:=  CP:i:55352714   HI:i:0
```

RNEXT PNEXT tlen

✓**rnext** only relevant for paired-end reads; name of the reference sequence where other end aligned.
✓**pnext** only relevant for paired-end reads; 1-based offset into the reference sequence where other end aligned.
✓**Tlen/Isize** only relevant for paired-end reads; insert length inferred from alignment.

SAM/BAM

```
HWI-ST1145:74:C101DACXX:7:1102:4284:73714        16      chr20   190930  3       100M    *       0       0
        CCGTGTTTAAAGGTGGATGCGGTCACCTTCCCAGCTAGGCTTAGGGATTCTTAGTTGGCCTAGGAAATCCAGCTAGTCCTGTCTCTCAGTCCCCCCTCT
C     BBDCCDDCCDDDDCDDDDDDCDCCCDBC?DDDDDDDDDDDDDDDCCDCDDDDDDDDDDCCCCEDDDC?DDDDDDDDDDDDDDDDDDDDBDHEFFFDC@@
    AS:i:-15        XM:i:3  XO:i:0  XG:i:0  MD:Z:55C20C13A9 NM:i:3  NH:i:2  CC:Z:=  CP:i:55352714   HI:i:0
```

RNEXT PNEXT tlen

= : RNAME equals RNEXT

*: information is not available

```
my_read 99  chr2:172936693-172938111  129 60  100M  =  429  400  CTAACTAGCCTGGGAAA
my_read 147 chr2:172936693-172938111  429 60  100M  =  129  -400  TCGAGCTCTGCATTC
```

SAM/BAM

```
HWI-ST1145:74:C101DACXX:7:1102:4284:73714        16      chr20   190930  3       100M    *       0       0
        CCCTCTTTAAAGCTGCATCCGCTCACCTTCCCACCTAGGCTTAGCGATTCTTACTTGGCCTAGGAAATCCAGCTACTCCTCTCTCTCACTCCCCCCTCT
C       BBDCCDDCCDDDDCDDDDDDDCDCCCDBC?DDDDDDDDDDDDDDDDCCDCDDDDDDDDDDDCCCCEDDDC?DDDDDDDDDDDDDDDDDDDDDBDHEEEEDC@@
        AS:i:-13        XM:i:3  X0:i:0  XG:i:0  MD:Z:53C20C13A9 NM:i:3  NH:i:2  CC:Z:=  CP:i:53352714   HI:i:0
```

seq

qual

extras

tab-separated "extra" fields, usually optional and aligner-specific but often very important!

SAM/BAM

```
HWI-ST1145:74:C101DACXX:7:1102:4284:73714        16       chr20    190930  3        100M     *        0        0
        CCGTGTTTAAAGGTGGATGCGGTCACCTTCCCAGCTAGGCTTAGGGATTCTTAGTTGGCCTAGGAAATCCAGCTAGTCCTGTCTCTCAGTCCCCCCTCT
C    BBBCCDDCCDDDDCDDDDDDDCDCCCDBC?DDDDDDDDDDDDDDDCCDCDDDDDDDDDDDCCCCEDDDC?DDDDDDDDDDDDDDDDDDDDDDBDHFFFFDC@@
    AS:i:-15        XM:i:3   XO:i:0   XG:i:0   MD:Z:55C20C13A9 NM:i:3   NH:i:2   CC:Z:=   CP:i:55352714    HI:i:0
```

Alignment Score

TAG:TYPE:VALUE

X?

reserved fields for end users

String for mismatching positions.

http://chagall.med.cornell.edu/galaxy/references/SAM_BAM_Specification.pdf

https://samtools.github.io/hts-specs/SAMtags.pdf

https://github.com/vsbuffalo/devnotes/wiki/The-MD-Tag-in-BAM-Files

SAM/BAM

- MD: String for mismatching positions.
- The MD field aims to achieve SNP/indel calling without looking at the reference.
- The MD field ought to match the CIGAR string.

MD: Z: 10A5^AC6

Thank you!