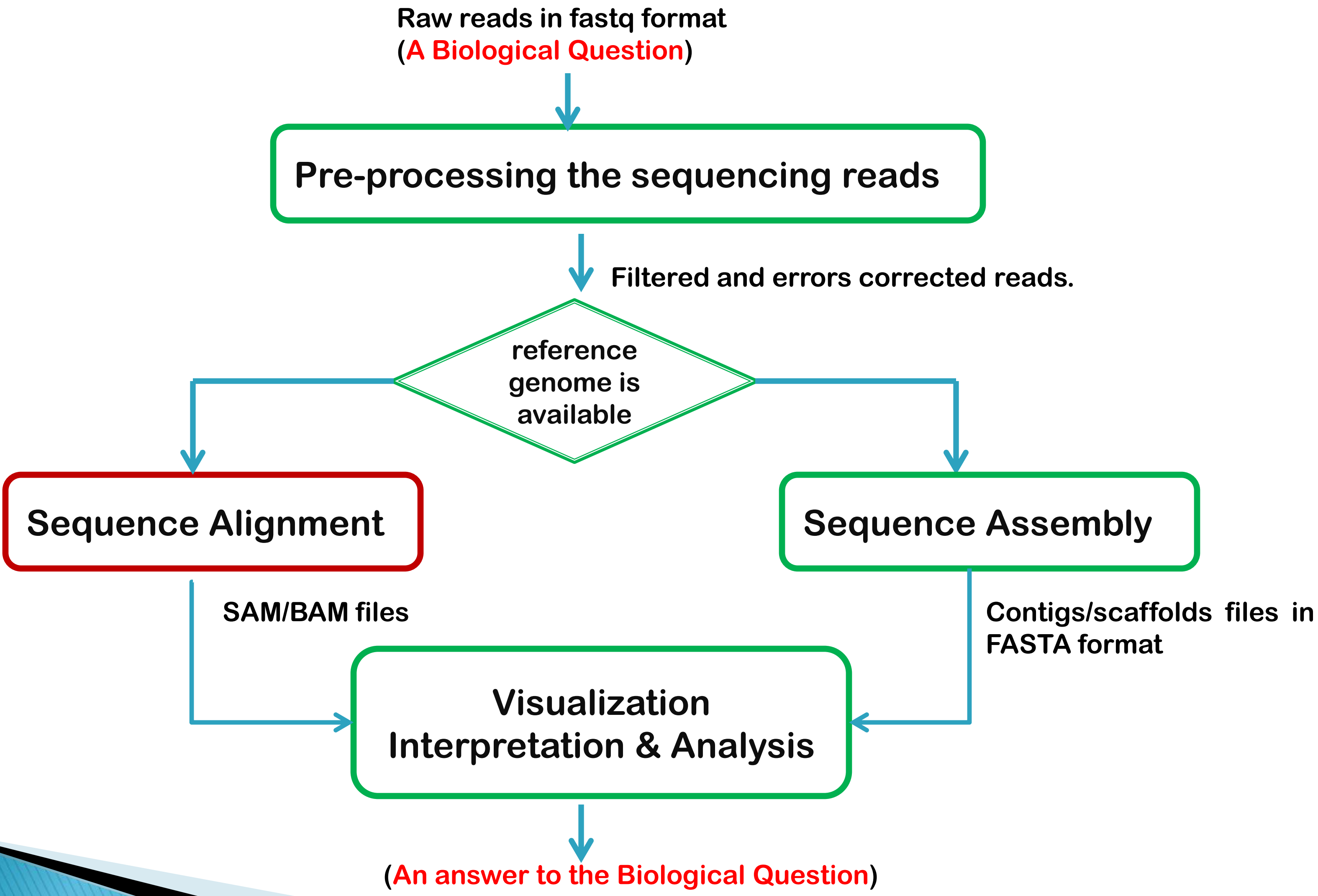
Informatics on High-throughput Sequencing Data

(Summer Course 2020)

Day 14

A typical Data Analysis pipeline



Raw reads in fastq format
(A Biological Question)

Pre-processing the sequencing reads

Filtered and errors corrected reads.

reference genome is available

Sequence Alignment

Sequence Assembly

SAM/BAM files

Contigs/scaffolds files in FASTA format

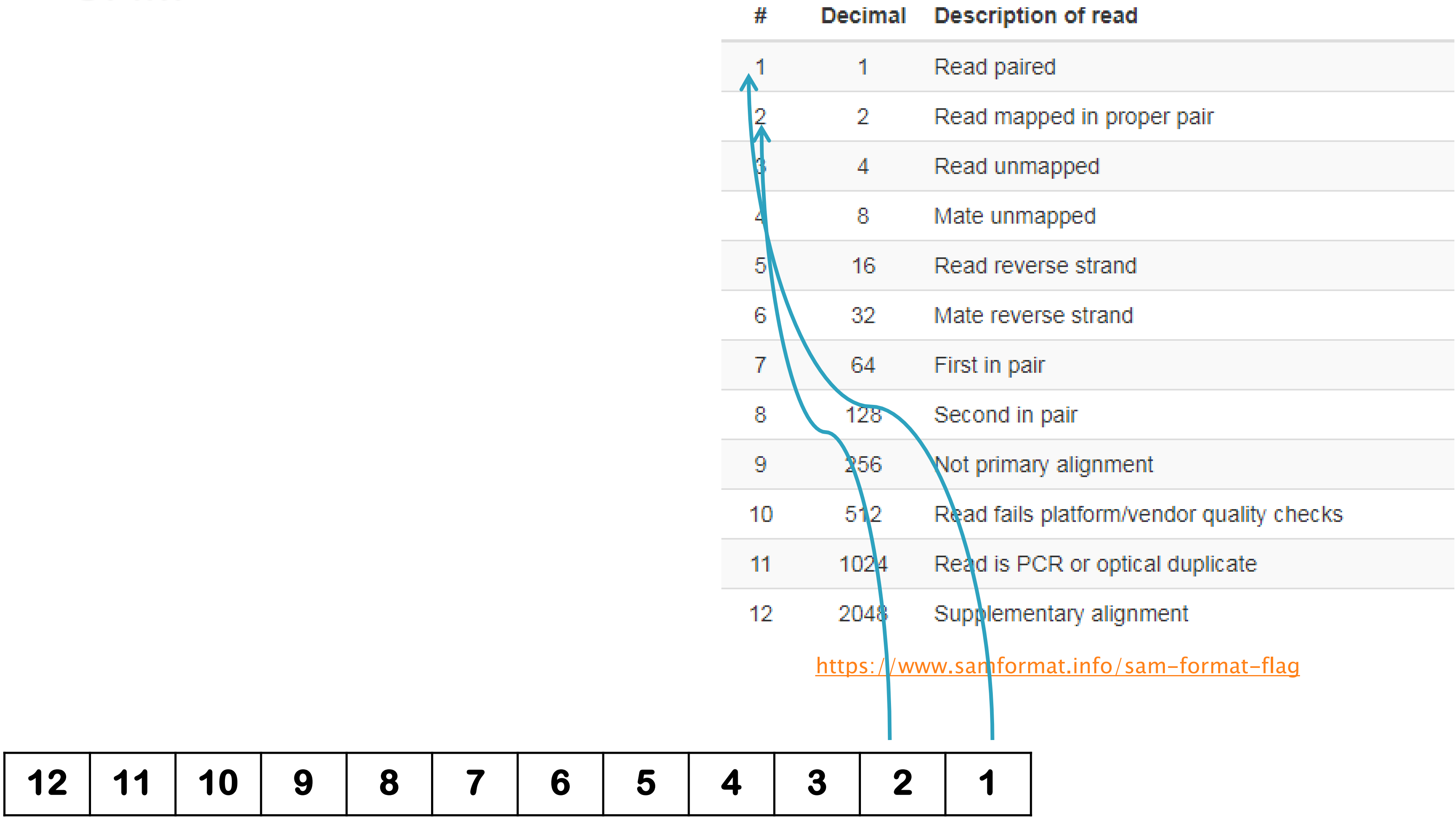
Visualization
Interpretation & Analysis

(An answer to the Biological Question)

SAM

Field	Meaning
GAII05_0002:1:2:12086:1654	Read ID
16	Flag
Chr2	Chr
1694072	start
0	MAPQ
51M	CIGAR
*	Mate Chr
0	Mate start
0	Mate dis
CCTTGTAAAATCATTATTAATGTTTTTAAACCCCTTTTAAAAATCCTTGTA	read
CCCCCCCCCCCCCCCCCCCBBCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC	qual
NM:i:1 MD:Z:20C30 AS:i:46 XS:i:46	Tag-Type-Value

SAM

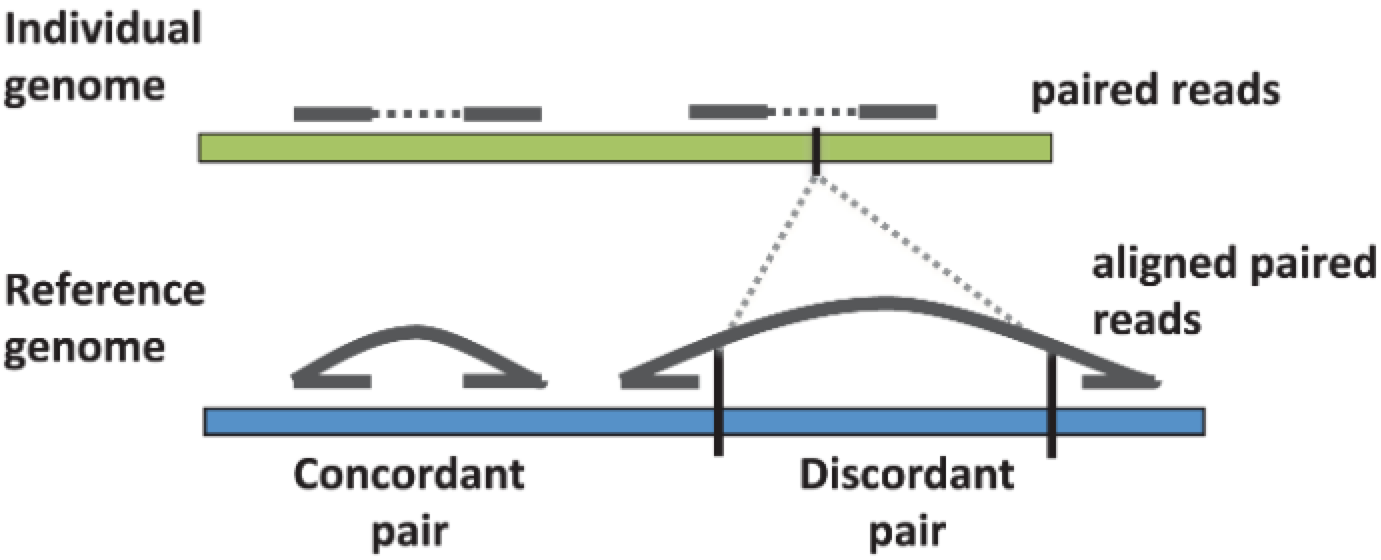


#	Decimal	Description of read
1	1	Read paired
2	2	Read mapped in proper pair
3	4	Read unmapped
4	8	Mate unmapped
5	16	Read reverse strand
6	32	Mate reverse strand
7	64	First in pair
8	128	Second in pair
9	256	Not primary alignment
10	512	Read fails platform/vendor quality checks
11	1024	Read is PCR or optical duplicate
12	2048	Supplementary alignment

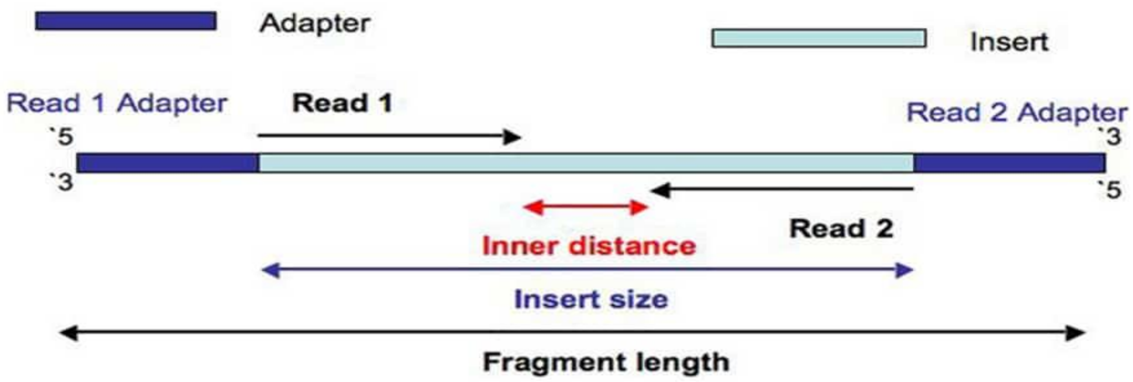
https://www.samformat.info/sam-format-flag

12	11	10	9	8	7	6	5	4	3	2	1

SAM

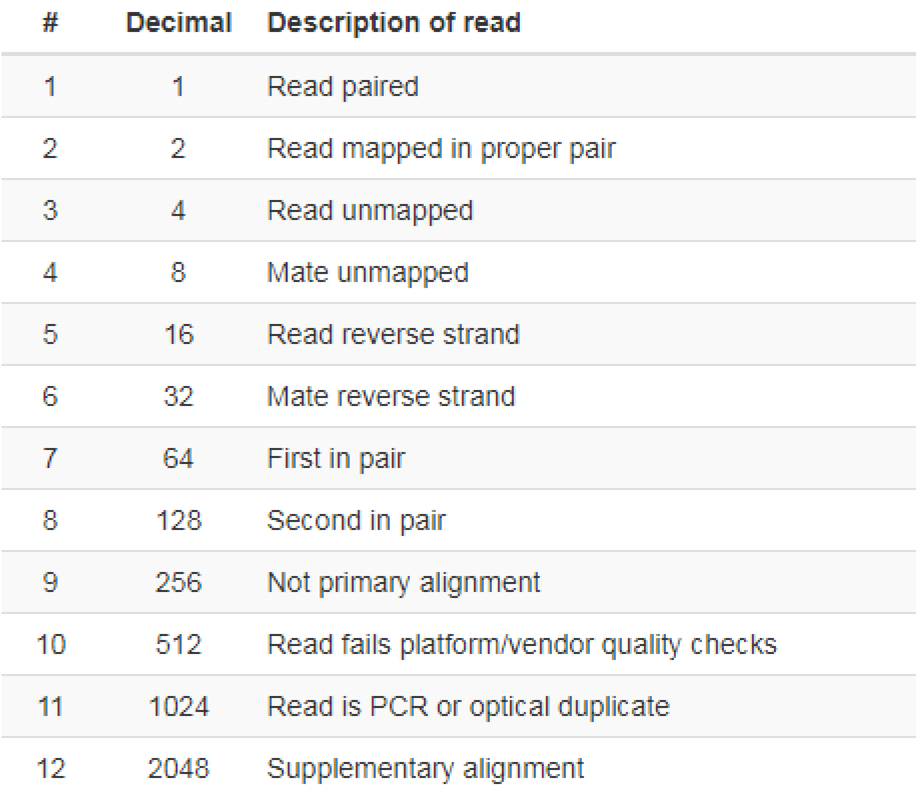


doi: https://doi.org/10.1371/journal.pcbi.1002821.g005



https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1002821

SAM



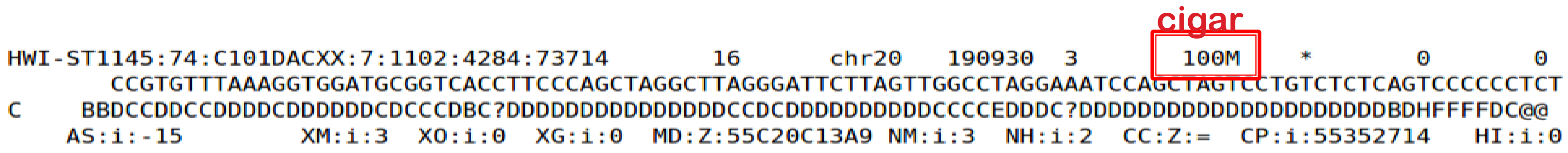
#	Decimal	Description of read
1	1	Read paired
2	2	Read mapped in proper pair
3	4	Read unmapped
4	8	Mate unmapped
5	16	Read reverse strand
6	32	Mate reverse strand
7	64	First in pair
8	128	Second in pair
9	256	Not primary alignment
10	512	Read fails platform/vendor quality checks
11	1024	Read is PCR or optical duplicate
12	2048	Supplementary alignment

https://www.samformat.info/sam-format-flag

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

0	**0**	**0**	**0**	**0**	**1**	**1**	**0**	**0**	**0**	**1**	**1**

SAM

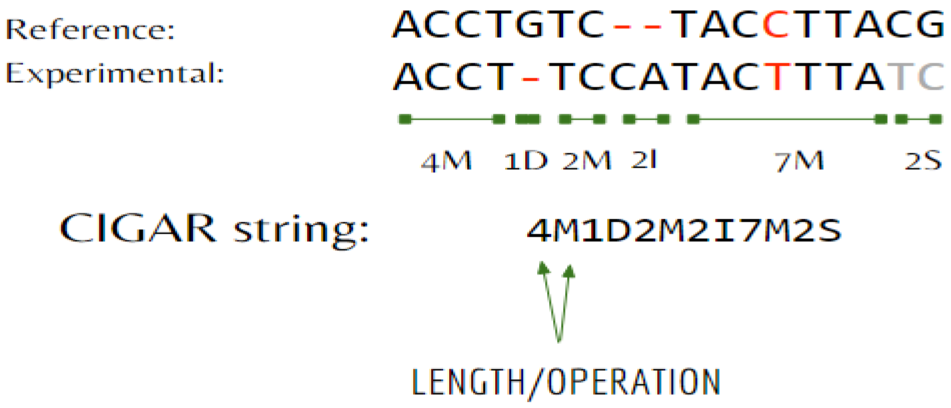


cigar

```
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
```

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

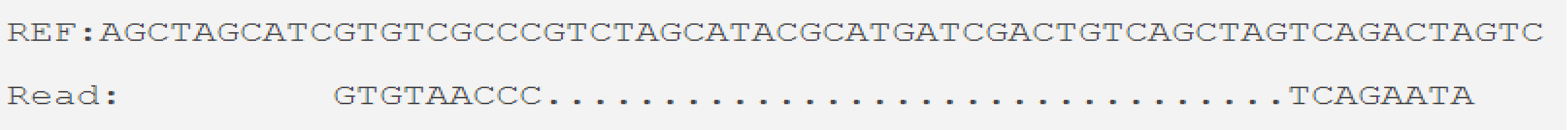
4M 1D 2M 2I 7M 2S

CIGAR string: 4M1D2M2I7M2S

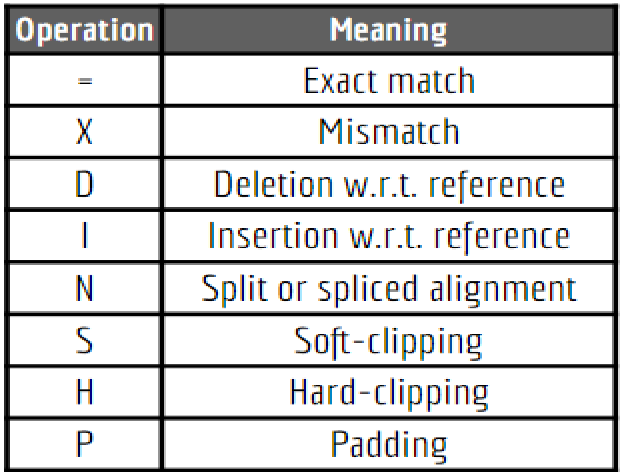
LENGTH/OPERATION

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

SAM



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



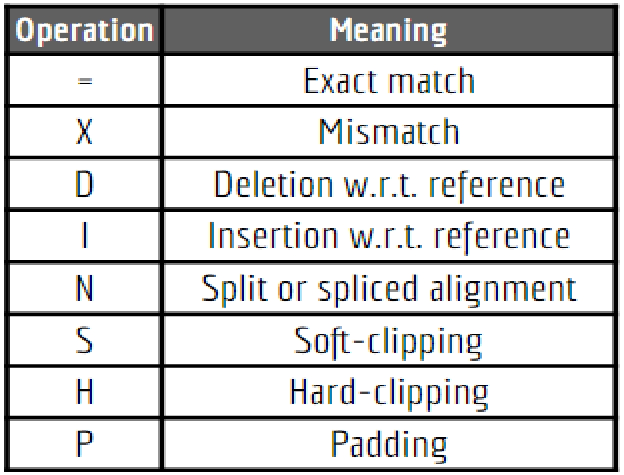
Operation	Meaning
=	Exact match
X	Mismatch
D	Deletion w.r.t. reference
I	Insertion w.r.t. reference
N	Split or spliced alignment
S	Soft-clipping
H	Hard-clipping
P	Padding

The CIGAR for this alignment is : 9M32N8M.

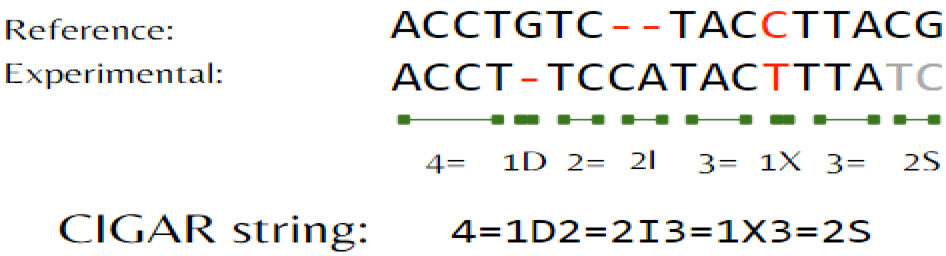
Slides adapted from Aaron Quinlan: https://github.com/quinlan-lab/applied-computational-genomics

SAM

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



```
Reference:      ACCTGTC--TACCTTACG
Experimental:   ACCT-TCCATACTTTATC
```

CIGAR string: 4=1D2=2I3=1X3=2S

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

SAM

- 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.

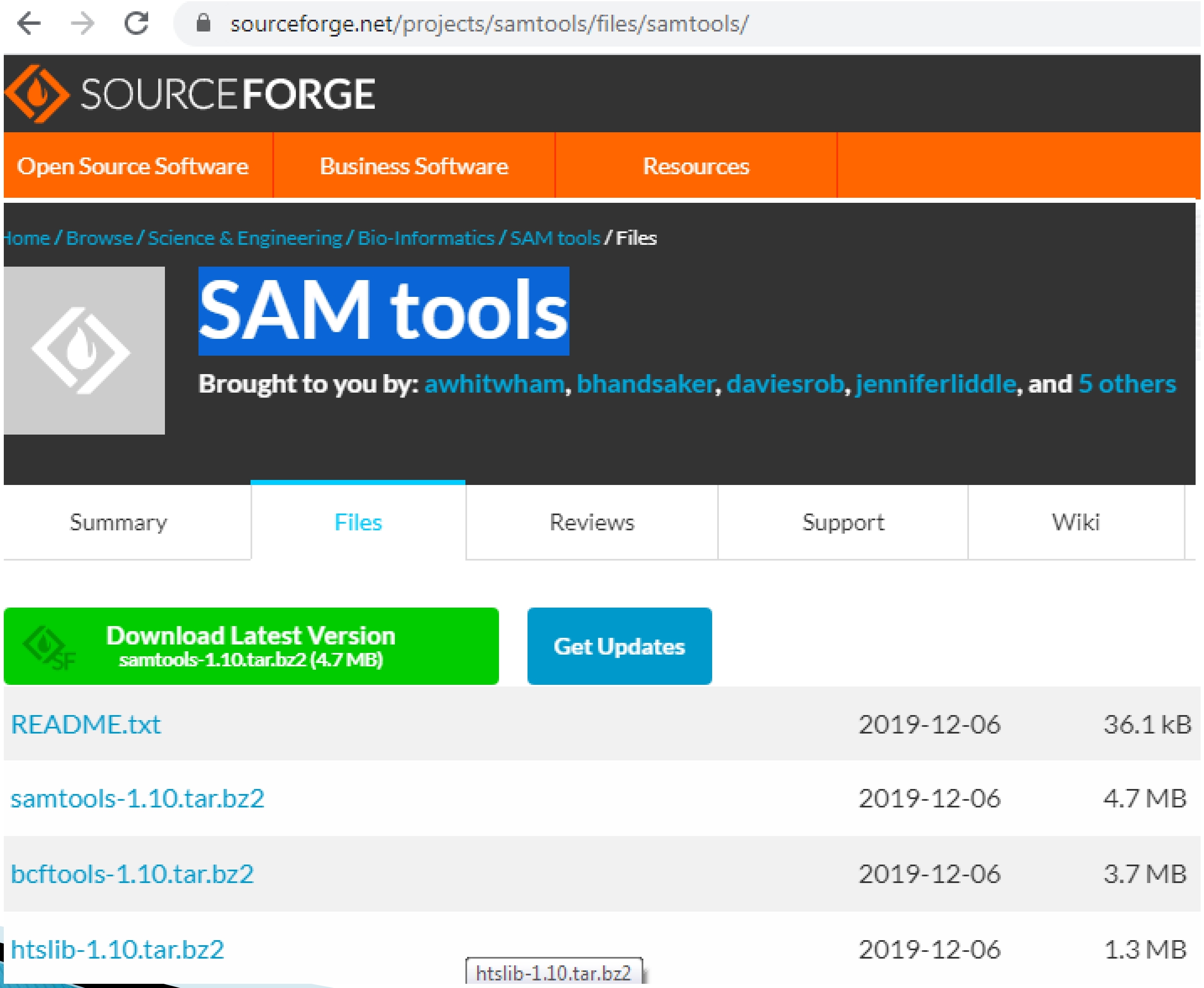
MD: Z: 10A5^AC6

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 tools



sourceforge.net/projects/samtools/files/samtools/

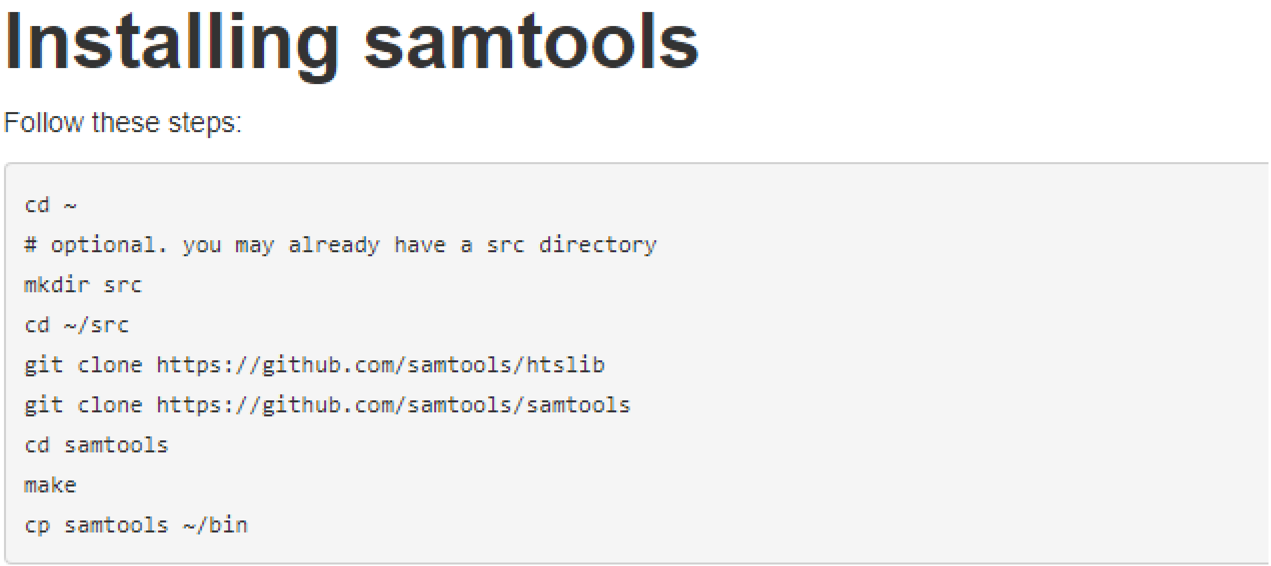
Home / Browse / Science & Engineering / Bio-Informatics / SAM tools / Files

SAM tools

Brought to you by: awhitwham, bhandsaker, daviesrob, jenniferliddle, and 5 others

README.txt	2019-12-06	36.1 kB
samtools-1.10.tar.bz2	2019-12-06	4.7 MB
bcftools-1.10.tar.bz2	2019-12-06	3.7 MB
htslib-1.10.tar.bz2	2019-12-06	1.3 MB

SAM tools



Installing samtools

Follow these steps:

```
cd ~
# optional. you may already have a src directory
mkdir src
cd ~/src
git clone https://github.com/samtools/htslib
git clone https://github.com/samtools/samtools
cd samtools
make
cp samtools ~/bin
```

http://quinlanlab.org/tutorials/samtools/samtools.html

SAM tools

- `./samtools view -S -b sample.sam > sample.bam`
- `./samtools view sample.bam | head` (i.e. without header info)
- `./samtools view -h sample.bam > sample.sam`
- `./samtools flagstat sample.bam`
- `./samtools sort sample.bam -o sample.sorted.bam`
- `./samtools view sample.sorted.bam | head`
- `./samtools index sample.sorted.bam` (i.e. sample.sorted.bam.bai) (IGV viewer and easy to access alignment regions)
- `./samtools view sample.sorted.bam Chr5:1000000-1900000 | wc -l` (i.e. files must be sorted then indexed)
- `./samtools view -L example.bed sample.sorted.bam` (i.e. combine with head shows the lower range / tail for upper range)

http://genome.ccbr.utoronto.ca/FAQ/FAQformat.html#format1

http://quinlanlab.org/tutorials/samtools/samtools.html

Thanks!

// | ?