Samtools and the SAM format

Alessandro Romanel
CIBIO

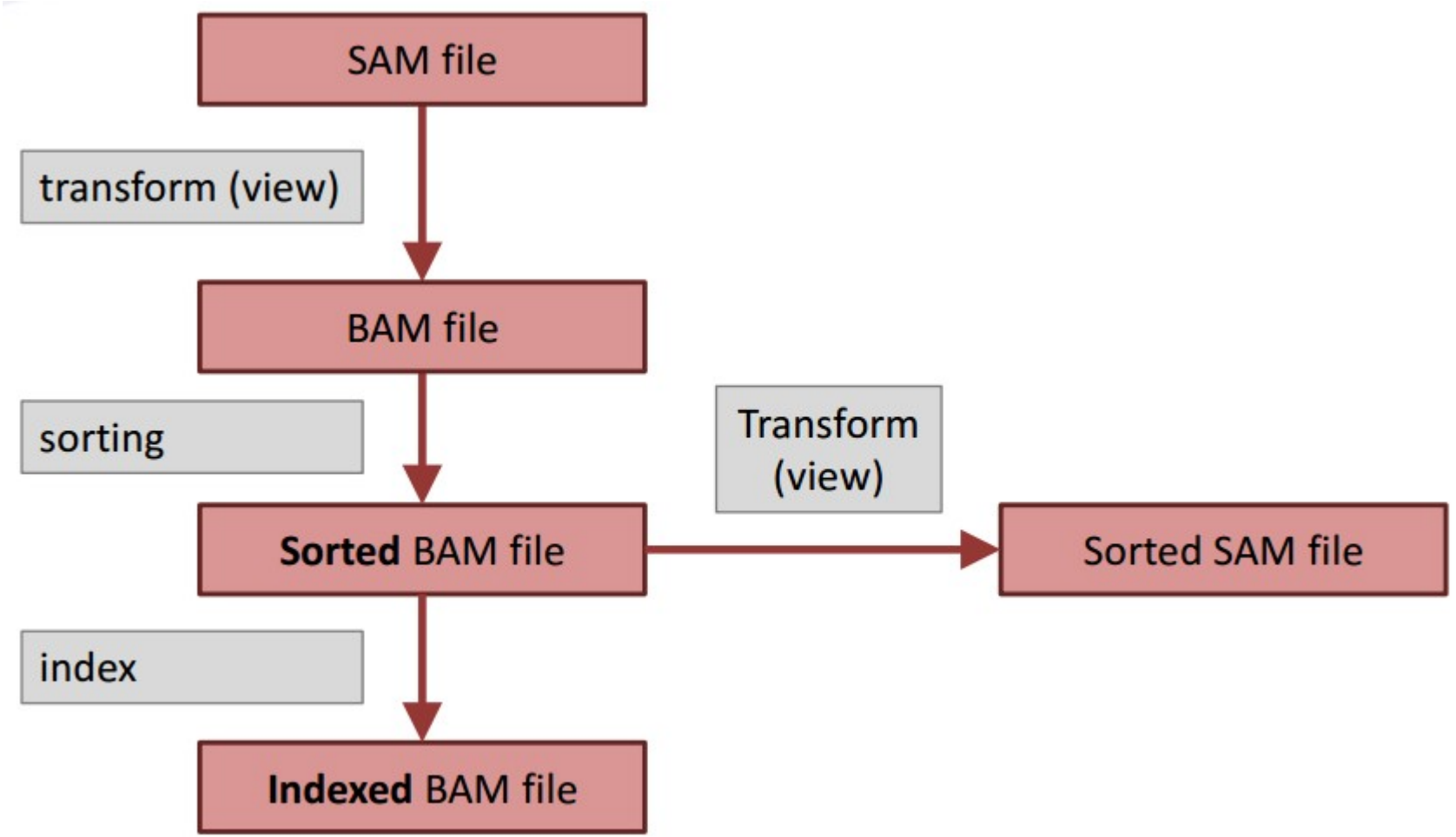
Files

```
mkdir ~/Genomics
cd ~/Genomics

wget https://github.com/aromanel/EthSEQ_Data/raw/master/Genomics/Normal.sam.gz
wget https://github.com/aromanel/EthSEQ_Data/raw/master/Genomics/Tumor.sam.gz
wget https://github.com/aromanel/EthSEQ_Data/raw/master/Genomics/samtools-1.4.tar.gz
wget https://github.com/aromanel/EthSEQ_Data/raw/master/Genomics/samtools-1.4.tar.gz
wget https://github.com/aromanel/EthSEQ_Data/raw/master/Genomics/CG100.bed

gunzip Normal.sam.gz
gunzip Tumor.sam.gz
tar -xvzf samtools-1.4.tar.gz

cd samtools-1.4
./configure
make

export PATH="~/Genomics/samtools-1.4/:$PATH"
```

SAM and BAM files

- **SAM** file
  - Information on the alignment of each read
  - optimized for readability and sequential access
- **BAM** (binary **SAM**):
  - Compression saves space (optimized for size)
  - May be *sorted* + *indexed* at location query (optimized for random access)
  - The file is not readable by eye
- Your default format should be BAM - only turn it into SAM when viewing the file

SAM/BAM hierarchy



Some tools have certain requirements of what type of SAM/BAM they take.
Your default data format should be a sorted, indexed BAM file!

```
romanel@silk:~/Desktop/samtools-1.4$ ./samtools

Program: samtools (Tools for alignments in the SAM format)
Version: 1.4 (using htslib 1.4)

Usage:   samtools <command> [options]

Commands:
  -- Indexing
     dict           create a sequence dictionary file
     faidx          index/extract FASTA
     index          index alignment

  -- Editing
     calmd          recalculate MD/NM tags and '=' bases
     fixmate        fix mate information
     reheader       replace BAM header
     rmdup          remove PCR duplicates
     targetcut      cut fosmid regions (for fosmid pool only)
     addreplacerg   adds or replaces RG tags

  -- File operations
     collate        shuffle and group alignments by name
     cat            concatenate BAMs
     merge          merge sorted alignments
     mpileup        multi-way pileup
     sort           sort alignment file
     split          splits a file by read group
     quickcheck     quickly check if SAM/BAM/CRAM file appears intact
     fastq          converts a BAM to a FASTQ
     fasta          converts a BAM to a FASTA

  -- Statistics
     bedcov         read depth per BED region
     depth          compute the depth
     flagstat       simple stats
     idxstats       BAM index stats
     phase          phase heterozygotes
     stats          generate stats (former bamcheck)

  -- Viewing
     flags          explain BAM flags
     tview          text alignment viewer
     view           SAM<->BAM<->CRAM conversion
     depad          convert padded BAM to unpadded BAM
```

```
romanel@silk:~/Desktop/samtools-1.4$ ./samtools view

Usage: samtools view [options] <in.bam>|<in.sam>|<in.cram> [region ...]

Options:
  -b       output BAM
  -C       output CRAM (requires -T)
  -1       use fast BAM compression (implies -b)
  -u       uncompressed BAM output (implies -b)
  -h       include header in SAM output
  -H       print SAM header only (no alignments)
  -c       print only the count of matching records
  -o FILE  output file name [stdout]
  -U FILE  output reads not selected by filters to FILE [null]
  -t FILE  FILE listing reference names and lengths (see long help) [null]
  -L FILE  only include reads overlapping this BED FILE [null]
  -r STR   only include reads in read group STR [null]
  -R FILE  only include reads with read group listed in FILE [null]
  -q INT   only include reads with mapping quality >= INT [0]
  -l STR   only include reads in library STR [null]
  -m INT   only include reads with number of CIGAR operations consuming
           query sequence >= INT [0]
  -f INT   only include reads with all bits set in INT set in FLAG [0]
  -F INT   only include reads with none of the bits set in INT set in FLAG [0]
  -s FLOAT subsample reads (given INT.FRAC option value, 0.FRAC is the
           fraction of templates/read pairs to keep; INT part sets seed)
  -x STR   read tag to strip (repeatable) [null]
  -B       collapse the backward CIGAR operation
  -?       print long help, including note about region specification
  -S       ignored (input format is auto-detected)
      --input-fmt-option OPT[=VAL]
               Specify a single input file format option in the form
               of OPTION or OPTION=VALUE
  -O, --output-fmt FORMAT[,OPT[=VAL]]...
               Specify output format (SAM, BAM, CRAM)
      --output-fmt-option OPT[=VAL]
               Specify a single output file format option in the form
               of OPTION or OPTION=VALUE
  -T, --reference FILE
               Reference sequence FASTA FILE [null]
  -@, --threads INT
               Number of additional threads to use [0]
```

SAM format

- Currently version 1.4
- Structure
  - Header
    - version, sort order, reference sequences, read groups, program/processing history
  - Alignments records

SAM header

COMMAND
samtools view -H file.bam

RESULT

@HD VN:1.4 GO:none SO:coordinate
@SQ SN:1 LN:249250621
@SQ SN:2 LN:243199373
@SQ SN:3 LN:198022430
@SQ SN:4 LN:191154276
@SQ SN:5 LN:180915260
@SQ SN:6 LN:171115067
@SQ SN:7 LN:159138663
@SQ SN:8 LN:146364022
...
@RG ID:PM207 PL:Illumina LB:GA LNID:L001 FCID:H9CF5ADXX DT:2014-05-27T00:00:00-0400 BCID:ACAAGCTA SM:PM207_EBC5_1_Ctrl_HALO
...
@PG ID:bwa PN:bwa VN:0.6.2-r126
@PG ID:GATK PrintReads VN:2.5-2-gf57256b CL:readGroup=null platform=null number=-1 downsample_coverage=1.0 sample_file=[] sample_name=[] simplify=false no_pg_tag=false

Sorting

Reference sequences
Names with lengths

Read groups with platform
Library and sample information

Programs (analysis) history

Alignment record

```
HWI-D00163:119:H9CF5ADXX:1:1101:18401:36465     163     1       13314   16      83M     =       13383
   152     GGATCTGAGCCCTGGTGGAGGTCAAAGCCACCTTTGGTTCTGCCATTGCTGCTGTGTGGAAGTTCACTCCTGCCTTTTCCTTT     AAA
ADEGEFECBCFBDFBCDBDEECCEECDDCDCCFCDCEDFECDDCGEDFFDFEGDGDECECCEEEEGDEGGFGFEEFDBBC     X0:i:1  X1:i:5  B
D:Z:NNOQSRSSSTROQRRQNQOPQPPPNFMPPPNPNNFMOONNNOOOPPNNPQOOPONMMNPOONQOOQNPOPPQQQPPIIQPOOG        MD:Z:83
 RG:Z:PM207      XG:i:0  BI:Z:PPQQTTSSQTQOSSRSPPQONQQPOJNROONOPOHMQQNOPPPNPONNPQPPQPOOPPPQPNPOOPNONPRQ
POQPJJQPRPJ        AM:i:0  NM:i:0  SM:i:16 XM:i:0  XO:i:0  XT:A:U
HWI-D00163:119:H9CF5ADXX:1:1102:11031:63853     163     1       13314   16      83M     =       13383
   152     GGATCTGAGCCCTGGTGGAGGTCAAAGCCACCTTTGGTTCTGCCATTGCTGCTGTGTGGAAGTTCACTCCTGCCTTTTCCTTT     AAA
ADEGEFECBCFBDGDDDBDFECCEECDDCDCCFCDCEDFECDDCFEDFEDFDGDGDECEDDFEEFGDEGGFGFEEFDBBC     X0:i:1  X1:i:5  B
D:Z:NNOQSRSSSTROQRRQNQOPQPPPNFMPPPNPNNFMOONNNOOOPPNNPQOOPONMMNPOONQOOQNPOPPQQQPPIIQPOOG        MD:Z:83
 RG:Z:PM207      XG:i:0  BI:Z:PPQQTTSSQTQOSSRSPPQONQQPOJNROONOPOHMQQNOPPPNPONNPQPPQPOOPPPQPNPOOPNONPRQ
POQPJJQPRPJ        AM:i:0  NM:i:0  SM:i:16 XM:i:0  XO:i:0  XT:A:U
HWI-D00163:119:H9CF5ADXX:1:1102:11554:26103     163     1       13314   16      83M     =       13383
   152     GGATCTGAGCCCTGGTGGAGGTCAAAGCCACCTTTGGTTCTGCCATTGCTGCTGTGTGGAAGTTCACTCCTGCCTTTTCCTTT     AAA
ADEGEFECBEEBCEBEFBDEDCCEECDDCDCCFCDCEDFECDDCGEDFEDFDFCGDFCGDCEEEEFDEGGFGFEBEDBBC     X0:i:1  X1:i:5  B
D:Z:NNOQSRSSSTROQRRQNQOPQPPPNFMPPPNPNNFMOONNNOOOPPNNPQOOPONMMNPOONQOOQNPOPPQQQPPIIQPOOG        MD:Z:83
 RG:Z:PM207      XG:i:0  BI:Z:PPQQTTSSQTQOSSRSPPQONQQPOJNROONOPOHMQQNOPPPNPONNPQPPQPOOPPPQPNPOOPNONPRQ
POQPJJQPRPJ        AM:i:0  NM:i:0  SM:i:16 XM:i:0  XO:i:0  XT:A:U
HWI-D00163:119:H9CF5ADXX:1:1105:13972:48206     163     1       13314   16      83M     =       13383
   152     GGATCTGAGCCCTGGTGGAGGTCAAAGCCACCTTTGGTTCTGCCATTGCTGCTGTGTGGAAGTTCACTCCTGCCTTTTCCTTT     AAA
ADEFEFECDEFDBFEDFDCEDDDCFDFDCDCCFCDCEEFECDDCGFEFEFGDGDGDECGEDFEEFGCFHGFGCEEFDBBC     X0:i:1  X1:i:5  B
D:Z:NNOQSRSSSTROQRRQNQOPQPPPNFMPPPNPNNFMOONNNOOOPPNNPQOOPONMMNPOONQOOQNPOPPQQQPPIIQPOOG        MD:Z:83
 RG:Z:PM207      XG:i:0  BI:Z:PPQQTTSSQTQOSSRSPPQONQQPOJNROONOPOHMQQNOPPPNPONNPQPPQPOOPPPQPNPOOPNONPRQ
POQPJJQPRPJ        AM:i:0  NM:i:0  SM:i:16 XM:i:0  XO:i:0  XT:A:U
HWI-D00163:119:H9CF5ADXX:1:1106:19105:75254     163     1       13314   16      83M     =       13383
:
```

Alignment record (essential fields)

QNAME: *Query template NAME*
HWI-D00163:119:H9CF5ADXX:1:1101:18401:36465
FLAG: *bitwise FLAG*
163
RNAME: *Reference sequence NAME*
1
POS: *1-based leftmost mapping POSition*
13314
MAPQ: *MAPping Quality*
16
CIGAR: *CIGAR string*
83M
RNEXT: *Ref. name of the mate/next read*
=
PNEXT: *Position of the mate/next read*
13383
TLEN: *observed Template LENgth*
152
SEQ: *segment SEQuence*
GGATCTGAGCCCTGGTGGAGGTCAAAGCCACCTTTGGTTCTGCCATTGCTGCTGTGTGGAAGTTCA
CTCCTGCCTTTTCCTTT
QUAL: *ASCII of Phred-scaled base QUALity+33*
AAAADEGEFECBCFBDFBCDBDEECCEECDDCDCCFCDCEDFECDDCGEDFFDFEGDGDECECCEE
EEGDEGGFGFEEFDBBC

FLAG

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

163 = 128+32+2+1

FLAG

2048 = 100000000000
1024 = 010000000000
512 = 001000000000
256 = 000100000000
128 = 000010000000
64 = 000001000000
32 = 000000100000
16 = 000000010000
8 = 000000001000
4 = 000000000100
2 = 000000000010
1 = 000000000001

128 = 000010000000
32 = 000000100000
2 = 000000000010
1 = 000000000001

163 = 000010100011

CIGAR

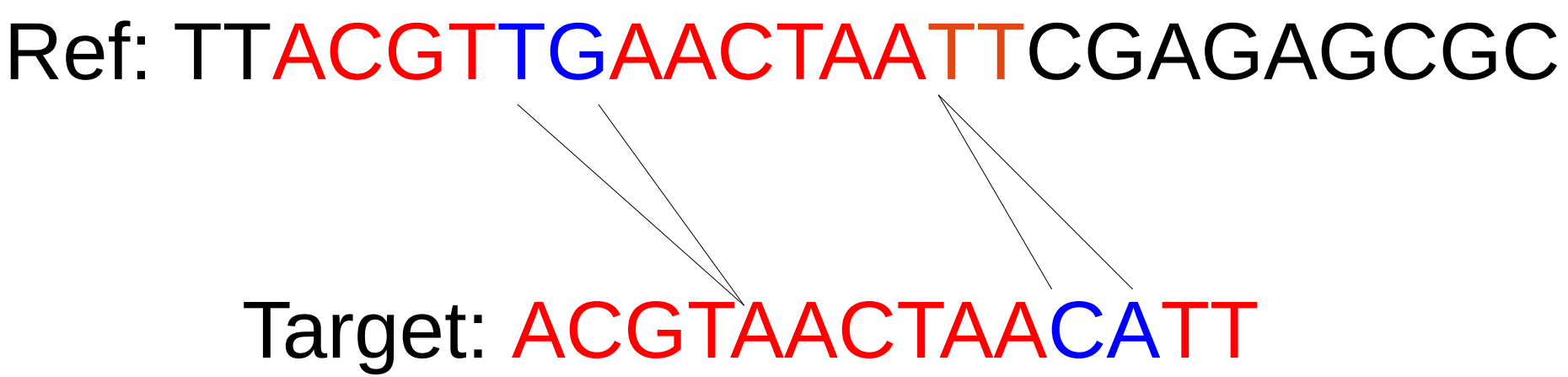
Op	BAM	Description
M	0	alignment match (can be a sequence match or mismatch)
I	1	insertion to the reference
D	2	deletion from the reference
N	3	skipped region from the reference
S	4	soft clipping (clipped sequences present in SEQ)
H	5	hard clipping (clipped sequences NOT present in SEQ)
P	6	padding (silent deletion from padded reference)
=	7	sequence match
X	8	sequence mismatch

CIGAR

Ref: TTACGTTGAACTAATTCGAGAGCGC

Target: ACGTAACTAACATT

CIGAR



CIGAR

Ref: TTACGTTGAACTAATTCGAGAGCGC

Target: ACGTAACTAACATT

Cigar = 4M2D6M2I2M

Alignment record (additional)

X0:i:1
X1:i:5
BD:Z:NNOQSRSSSTROQRRQNQOPQPPPNFMPPPNPNNFMOONNNOOOPPNN
PQOOPONMMNPOONQOOQNPOPPQQQPPIIQPOOG
MD:Z:83
RG:Z:PM207
XG:i:0
BI:Z:PPQQTTSSQTQOSSRSPPQONQQPOJNI
QPPQPOOPPPQPNPOOPNONPRQPOQPJJQ
AM:i:0
NM:i:0
SM:i:16
XM:i:0
XO:i:0
XT:A:U

Tag	Meaning
NM	Edit distance
MD	Mismatching positions/bases
AS	Alignment score
BC	Barcode sequence
X0	Number of best hits
X1	Number of suboptimal hits found by BWA
XN	Number of ambiguous bases in the referenece
XM	Number of mismatches in the alignment
XO	Number of gap opens
XG	Number of gap extentions
XT	Type: Unique/Repeat/N/Mate-sw
XA	Alternative hits; format: (chr,pos,CIGAR,NM;)*
XS	Suboptimal alignment score
XF	Support from forward/reverse alignment
XE	Number of supporting seeds

Default operations

- By default samtools (not all operations) expects a BAM file as input and will produce a SAM file as output
- Alignment results are typically stored as a sorted and indexed BAM file
- Aligners produce SAM files so our first job is usually to convert those to BAM formats.

SAM to BAM

- Convert a SAM into a BAM
  - *samtools view -Sbh Normal.sam > Normal.bam*
- Sort a BAM file
  - *samtools sort Normal.bam > Normal.sorted.bam*
  - *samtools sort Normal.bam Normal.sorted (v0.19)*
- Create index
  - *samtools index Normal.sorted.bam*

Sorting

- Sorted so that read pairs are next to one another (typically the same order as the FastQ file)
- Sorted by alignment position
- Depending on the next analysis method your file has to be sorted a certain way

Sorting

- Compare the two sortings
  - *samtools sort -n Normal.bam > Normal.sorted.rname.bam*
  - *samtools view Normal.sorted.bam* | *less*
  - *samtools view Normal.sorted.rname.bam* | *less*

Filtering

- Required flag (keep if matches)
  - *samtools view -f ...*
- Filtering (remove if matches)
  - *samtools view -F ...*

Filtering

- Count reads in BAM file
  - *samtools view -c Normal.sorted.bam*
- Reads that map to reverse strand
  - *samtools view -c -f 16 Normal.sorted.bam*
- Reads that map to the forward strand
  - *samtools view -c -F 16 Normal.sorted.bam*
- *Reads that have a mapping quality >30*
  - *samtools view -c -q 30 Normal.sorted.bam*

Explore statistics

- General statistics
  - *samtools flagstat Normal.sorted.bam*
- Detailed statistics
  - *samtools stats Normal.sorted.bam > Stats.txt*
  - *less Stats.txt*

Explore coverage statistics

- Single base sum coverage per region
  - *samtools bedcov CG100.bed Normal.sorted.bam > BEDCov.txt*
  - *less BEDCov.txt*

- Single base depth
  - *samtools depth -b CG100.bed Normal.sorted.bam > BEDDepth.txt*
  - *less BEDDepth.txt*

mpileup

- Test mpileup
  - *samtools mpileup Normal.sorted.bam | less*

- *Pileup of a region*
  - *samtools mpileup -r 1:3410684-3410690 Normal.sorted.bam*

- *Control base and mapping quality*
  - *samtools mpileup -r 1:3410684-3410690 -q 60 -Q 60 Normal.sorted.bam*

Tasks

- Repeat previous commands on file Tumor.sam
- Check pileup of following positions
  - 1:196642233
  - 10:114192285
  - 17:7578265
  - 3:128204654

  and comment results
- Test if results change by filtering reads on base and mapping quality
- Use bedcov command to further comment on position 17:7578265