Week 8, Lecture 16

István Albert

Biochemistry and Molecular Biology and Bioinformatics Consulting Center
Penn State

Binary SAM (BAM) files

SAM file:
- information on the alignment of each read
- optimized for readability and sequential access

BAM (binary SAM):
- compression $\rightarrow$ saves space (optimized for size)
- may be sorted + indexed $\rightarrow$ 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

Download and ‘make’ SAMTOOLS

http://samtools.sourceforge.net/
Samtools: is suite of commands

Usage: samtools <command> [options]

Command: view SAM<->BAM conversion
sort sort alignment file
mpileup multi-way pileup
depth compute the depth
faidx index/extract FASTA
tview text alignment viewer
index index alignment
idxstats BAM index stats (r595 or later)
fixmate fix mate information
flagstat simple stats
colv recycle column and '=' bases
merge merge sorted alignments
rmdup remove PCR duplicates
reheader replace BAM header
cat concatenate BAMs
bedcov read depth per BED region
targetcut cut fosmid regions (for fosmid pool only)
phase phase heterocigotes
bamshuf shuffle and group alignments by name

Most actions will provide help on their usage

Usage: samtools view [options] <in.bam> <in.sam> [region1 [...]]

Options: -b output BAM
-h print header for the BAM output
-H print header only (no alignments)
-S input is SAM
-u uncompressed BAM output (force -b)
-1 fast compression (force -b)
-x output FLAG in HEX (samtools-C specific)
-c print only the count of matching records
-b collapse the backward CIGAR operation
-@ INT number of BAM compression threads [0]
-L FILE output alignments overlapping the input BED FILE [null]
-t FILE list of reference names and lengths (force -S) [null]
-T FILE reference sequence file (force -S) [null]
-o FILE output file name [stdout]
-R FILE list of read groups to be outputted [null]
-f INT required flag, @ for unset [0]

Default Operation

- By default samtools expects a BAM file as input and will produce a SAM file as output

- Every alignment result should be stored as a sorted and indexed BAM file

Transform SAM to BAM

transform to bam

samtools view -Sb input.sam > tempfile.bam

sort bam file

samtools sort -f tempfile.bam output.bam

Index bam file

samtools index output.bam
Add the following to the previous week’s shell script

```
# perform the alignments via bwa
~/bin/bwa aln $REF $QUERY > $SAI
~/bin/bwa samse $REF $SAI $QUERY > $SAM

# transform the SAM file to BAM
~/bin/samtools view -Sb $SAM > $TMP

# sort the samfile
~/bin/samtools sort -f $TMP $BAM

# index the BAM file
~/bin/samtools index $BAM

echo "Finished ref=$REF, query=$QUERY, bam=$BAM"
```

Filtering SAM/BAM files

- Required flag (keep if matches)
  - `samtools view -f`

- Filtering flag (remove if matches)
  - `samtools view -F`

Flags are using a bitwise representation

<table>
<thead>
<tr>
<th>Flag</th>
<th>Bitwise Representation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>paired end read</td>
</tr>
<tr>
<td>2</td>
<td>mapped as proper pair</td>
</tr>
<tr>
<td>4</td>
<td>unmappable read</td>
</tr>
<tr>
<td>8</td>
<td>read mate unmapped</td>
</tr>
<tr>
<td>16</td>
<td>read mapped on reverse strand</td>
</tr>
</tbody>
</table>

```
ialbert@porthos ~/work/lec12
$ ~/bin/samtools view -c -f 4 results.bam
1

ialbert@porthos ~/work/lec12
$ ~/bin/samtools view -c -f 4 results.bam
3
```

```
# save on typing
alias samtools="/bin/samtools"

# how many reads in total
samtools view -c results.bam
5

# reads that cannot be mapped
samtools view -c -f 4 results.bam
10

# reads that can be mapped
samtools view -c -f 4 results.bam
40

# reads that map to reverse strand
samtools view -c -f 16 results.bam
50

# reads that map to forward strand
samtools view -c -f 16 results.bam
0

# reads that have a minimum mapping quality of 1
samtools view -c -q 1 results.bam
100
```

*Note: For BWA this also means unique alignment!*
A sorted file will stay sorted during transformation

• Once sorted all output will stay sorted regardless of the output type (SAM, BAM)

• You can creating a second, smaller and filtered file that does not need to be sorted again.

• You do need to index the new file though!

Explore other commands

Flag statistics

```
samtools flagstat data.bam
```

Index stats

```
samtools idxstats data.bam
```

Depth of coverage

```
samtools depth data.bam | head
```

Querying a BAM file name:start-end

Samtools allows querying:

```
samtools view data.bam chrV:1000-2000
```

Homework 16

Generate a sorted and indexed BAM file based on the data

`lect15.fq.gz`

1. Find the number of uniquely mapped reads

2. Find the number of high quality alignments (MAPQ>30) for each strand separately

3. A genomic feature has its start site on the forward strand on chromosome I at position 111,000.
   - How many reads fall within 500b upstream of this location?
   - Print the position of each read (hint: there are not that many)
   - Report the number of reads in this region for each strand separately.