Handling SAM files

Estudios in silico en Biomedicina

(Máster en Bioinformática, Universidad de Valencia)

David Montaner

2014-10-21

SAM Format

Header section: @

Alignment section:

1. QNAME: read name
2. FLAG: bitwise FLAG
3. RNAME: reference sequence name (chromosome)
4. POS: 1-based leftmost mapping position (where the read starts mapping)
5. MAPQ: mapping quality
6. CIGAR: CIGAR string
7. RNEXT: name of the mate/next read (for paired ends)
8. PNEXT: Position of the mate/next read (BP O)
9. TLEN: observed Template LENgth
10. SEQ: read sequence
11. QUAL: read quality (phred-scale +33)

BAM format

SAM

SAM : Sequence Alignment/Map format.
Is a TAB-delimited **text** format storing the alignment information.

BAM

A **binary** and *compressed* version of the file is easier to handle.
Generally **sorted** by chromosome position.
Supplementary **index** files may be created for certain purposes (quick access)

SAMtools

```
Program: samtools (Tools for alignments in the SAM format)
Version: 0.1.19-96b5f2294a

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
```

SAMtools

```
Program: samtools (Tools for alignments in the SAM format)
Version: 0.1.19-96b5f2294a

Usage:   samtools <command> [options]

Command: calmd       recalculate MD/NM tags 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 or
         phase       phase heterozygotes
         bamshuf     shuffle and group alignments by name
```

SAMtools Commands

```
Usage:   samtools sort [options] <in.bam> <out.prefix>

Options: -n         sort by read name
         -f         use <out.prefix> as full file name inste
         -o         final output to stdout
         -l INT     compression level, from 0 to 9 [-1]
         -@ INT     number of sorting and compression threa
         -m INT     max memory per thread; suffix K/M/G rec
```

SAMtools First Commands

- view SAM- BAM conversion
- sort sort alignment file
- index index alignment

- reheader replace BAM header
- cat concatenate BAMs
- merge merge sorted alignments

- mpileup multi-way pileup
- depth compute the depth

SAMtools more Commands

- faidx index/extract FASTA
- tview text alignment viewer
- idxstats BAM index stats (r595 or later)
- fixmate fix mate information
- flagstat simple stats
- calmd recalculate MD/NM tags and '=' bases
- rmdup remove PCR duplicates
- bedcov read depth per BED region
- targetcut cut fosmid regions (for fosmid pool only)
- phase phase heterozygotes
- bamshuf shuffle and group alignments by name