The Biostar Handbook: A Beginner's Guide to Bioinformatics[6]

#course

SAM

SAM: Sequence Alignment/Map

> "SAM format is a TAB-delimited, line oriented text format"

SAM:
1. tab-delimited
**2. line oriented **
3. txt format
4. header section
5. alignment section

method:
cat
bwa
bowtie

> "A BAM file is a a binary, compressed (and almost always sorted) representation of the SAM information."

BAM:
1. binary
2. compressed
3. sorted (mostly by alignment coordinate; rarely by read name)

method:
samtools view
samtools fastq
bamtools
piscard

> “BAM files are always sorted, usually by the alignment coordinate information and more rarely by the read names.”

> “CRAM files are conceptually similar to BAM files. They represent a compressed version, where some of the compression is driven by compressing the reference genome that the sequence data is aligned to. This makes the data smaller than the BAM compressed formats.”

CRAM:
1. compressing reference genome
**2. smaller than BAM **

method:
samtools sort

> “A CRAM file can only be decoded if the same reference sequence is present at a given location on the target computer.”

读取CRAM格式需要提供参考序列，不然打不开
samtools view -T $REF bwa.cram

> “We will call the format and the data as SAM with the understanding that the actual representation may be BAM or CRAM.”
>
> “SAM files were invented to allow the representation of hundreds of millions of short alignments”

> “the differences are mostly in the accuracy or performance characteristics of the tool”

DIFFERENCE?
we care about the tools which will be used to analyze data

> “We’ll mention here that attempting to transform one “standard” SAM file into another “standard” SAM file that contains all the information that another tool might have reported is a surprisingly challenging (and most likely impossible) task - often the only recourse is to actually use this other tool.”

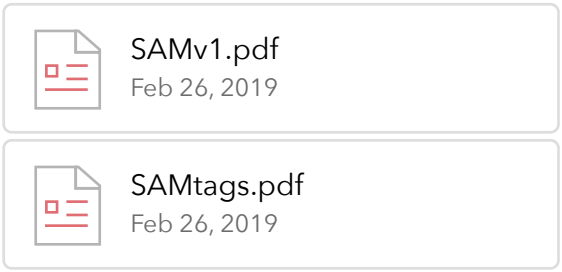
it is imposible to transform one SAM file to other SAM file; the only way is using different tool to get different SAM file

> "not only for representing and storing alignment information, but also for storing raw "unaligned" FASTQ files."

BAM can store alignment information and unaligned FASTQ files

> "Instead of having to deal with 100 paired end samples distributed over 200 files we can just store a single BAM file."

samtools



> "one of the linear alignments in a chimeric alignment is considered the "representative" or "primary" alignment, and the others are called "supplementary"

Chimeric alignment 有多行记录, 第一个是representative/primary，第二个是supplementary
Linear alignment 只有一行记录

Tools for sam:

samtools, bamtools, picard, sambamba, samblaster

```
# get overview of the alignments in bam
samtools flagstat # produces a report on flags
samtools idxstats # produces a report on how many reads align to each
chromosome
bamtools stats # produces a report on flags
```

```
# sam -> bam
samtools sort bwa.sam > bwa_sorted.bam
samtools index bwa_sorted.bam
```

```
# sam -> cram
samtools sort --reference $REF -O cram bwa.sam > bwa.cram
samtools index bwa.cram
```

```
# samtools view
samtools view bwa_sorted.bam
samtools view -T $REF bwa.cram
# choose quality > 10
samtools view -q 10 bwa_sorted.bam -b -o bwa_sorted_mq10.bam
# choose (0x3 3 PARIED,PROPER_PAIR)
samtools view -c -f 3 bwa_sorted.bam
# choose reversely
samtools view -c -F 3 bwa_sorted.bam
```

While using GATK, we need RG in the header of bam file, @RG include three records(ID,LB,SM). There are two ways to add RG to bam file:

```
TAG='@RG\tID:xzg\tSM:Ebola\tLB:patient_100'
# Add the tags during alignment
bwa mem -R $TAG $REF $R1 $R2 | samtools sort > bwa.bam
samtools index bwa.bam
# Add tags with samtools addreplacerg
samtools addreplacerg -r $TAG bwa_sorted.bam -o bwa_sorted_with_rg.bam
```