Coverage Calculation

QiongJia

2024-11-14

Contents

Design

To calculate the average coverage(sequencing depth) for the given sample BAM file, two approach are presented below.

1. The first one is based on the coverage calculation equation:

$$C = \frac{LN}{G}$$

where

- C is coverage.
- G is the haploid genome length.
- L is the read length in the sequencing.
- N is the number of reads.

The ***samtool idxstats*** can be used to get chromosome lengths and number of mapped reads.

2. The second one is more precise. The ***samtool depth*** can be used to calculate the coverage at each genomic position and the average coverage of the given BAM file.

3. At the end, two other tools ***bedtools genomecov*** and ***mosdepth*** are presented briefly in coverage calculation.

Download BAM file

```
cd ~
mkdir TakeHomeFulgent
cd TakeHomeFulgent
wget ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/data/
NA12878/exome_alignment/
NA12878.mapped.illumina.mosaik.CEU.exome.20110411.bam

bam=NA12878.mapped.illumina.mosaik.CEU.exome.20110411.bam
```

Quick estimate

```
samtools idxstats $bam \
| awk -vreadlen=100 '
   {
       len += $2
       nreads += $3
   }
   END {
       print nreads * readlen / len
   }
   '
```

The average coverage of given BAM file is: **5.26501**.

However, we don't have the information of the read length and a arbitrary number is used here so the estimation is not accurate.

Coverage calculation for each position

Step1: calculate the coverage at each genomic position.

```
samtools depth -a $bam > NA12878_coverage.txt
```

Overlook the coverage output.

```
head -n 5 NA12878_coverage.txt
```

```
1       1       0
1       2       0
1       3       0
1       4       0
1       5       0
```

Each line represents a genomic position. Three columns are included int he coverage output:

- Chromosome;
- Position;
- Reads covered this position.

Step2: calculate the average coverage.

```
awk '{sum+=$3} END { print "Average coverage = ",sum/NR}' NA12878_coverage.txt
```

Average coverage = **3.64239**

Alternative calculation:

The total length of the genome can also be calculated as below:

```
## @SQ is the reference sequence dictionary and LN in this line shows the reference sequence length.
## So the $tot here represent the totle length of sample genome
tot=$(samtools view -H $bam | awk -vFS=: '/^@SQ/ {sum+=$3} END {print sum}')
echo $tot
# 3101804739
```

Then the average coverage is calculated as below:

```
sum=$(awk '{sum+=$3} END {print sum}' NA12878_coverage.txt)
echo $sum
# 11297985096
avg=$(echo "$sum/$tot" | bc -l)
echo $avg
# 3.64239
printf "The average coverage is: %.2f\n" "$avg"
```

The average coverage is: **3.64**

Other methods

bedtools

bedtools genomecov -d also reports the genome coverage per base as below:

```
# To use -ibam flag in bedtools genomecov, the bam file is needed to be sorted by position
samtools sort $bam | bedtools genomecov -ibam stdin -d > NA12878_genomecov.txt
```

Then the average coverage would be:

```
awk '{sum+=$3} END { print "Average coverage = ",sum/NR}' NA12878_genomecov.txt
```

Average coverage = **3.99539**

mosdepth

mosdepth can report coverage for both per-base and summary result at the same time.

```
mosdepth NA12878 $bam
```

The file ended with ***.mosdepth.summary.txt*** contain the average coverage result.

```
awk 'NR==1 {print} {last=$0} END {print last}' NA12878.mosdepth.summary.txt
```

```
##    chrom     length       bases mean min  max
## 85 total 3101804739 10082458770 3.25   0 4343
```

```
awk '{last=$4} END {print "Average coverage = ",last}' NA12878.mosdepth.summary.txt
```

Average coverage = **3.25**

It can also report coverage based on the user defined region by using ***–by <bed/window>*** .

Runing time comparison

- *samtools depth* : 1348s;
- *bedtools genomecov* : 21074s;
- *mosdepth* : 433s;

```
## R version 4.2.3 (2023-03-15)
## Platform: x86_64-apple-darwin17.0 (64-bit)
## Running under: macOS Big Sur ... 10.16
##
## Matrix products: default
## BLAS:   /Library/Frameworks/R.framework/Versions/4.2/Resources/lib/libRblas.0.dylib
## LAPACK: /Library/Frameworks/R.framework/Versions/4.2/Resources/lib/libRlapack.dylib
##
## locale:
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## loaded via a namespace (and not attached):
##  [1] compiler_4.2.3    fastmap_1.2.0     cli_3.6.3         tools_4.2.3
##  [5] htmltools_0.5.8.1 rstudioapi_0.15.0 yaml_2.3.10       rmarkdown_2.28
##  [9] knitr_1.48        xfun_0.47         digest_0.6.37     rlang_1.1.4
## [13] evaluate_0.24.0
```