Algorithms for Next Generation Sequencing 1st edition Sung Solutions Manual

SKU: 9781466565500-SOLUTIONS

Chapter 2

NGS file formats

1. Suppose chromosome 2 is the following sequence `ACACGACTAA`....

   - For the genomic region in chromosome 2 containing the DNA sequence `ACGAC`, if we describe it using bed format, what are the chromosome, start, and end?
   - In SAM, what is the alignment position of the DNA sequence `ACGAC`?
   - In BAM, what is the alignment position of the DNA sequence `ACGAC`?

   Solution:

   - The bed format is `chr2 2 7`.
   - SAM is 1-based coordinate format. The alignment position is 3.
   - BAM is 0-based coordinate format. The alignment position is 2.

   ∎

2. Please perform the following conversions.

   (a) Convert the following set of intervals from the 0-based coordinate format to the 1-based coordinate format: 3..100, 0..89 and 1000..2000.

   (b) Convert the following set of intervals from the 1-based coordinate format to the 0-based coordinate format: 3..100, 1..89 and 1000..2000.

   Solution:

   - 4..100, 1..89 and 1001..2000.
   - 2..100, 0..89 and 999..2000.

   ∎

3. Given a BAM file `input.bam`, we want to find all alignments with $maQ > 0$ using samtools. What should be the command?

   Solution: `samtools view -q 1 input.bam`. ∎

Algorithms for Next Generation Sequencing 1st edition Sung Solutions Manual

4. Given two BED files `input1.bed` and `input2.bed`, we want to find all genomic regions in `input1.bed` that overlap with some genomic regions in `input2.bed`. What should be the command?

   Solution: `bedtools intersect -a input1.bed -b input2.bed -wa`. ∎

5. For the following wiggle file, can you compute $coverage(3,8)$, $mean(3,8)$, $minVal(3,8)$, $maxVal(3,8)$ and $stdev(3,8)$?

```
fixedStep chrom=chr1 start=1 step=1 span=1
20
10
15
30
20
25
30
20
10
30
```

   Solution: $coverage(3,8) = 1$, $mean(3,8) = \frac{15+30+20+25+30+20}{6}$, $minVal(3,8) = 15$, $maxVal(3,8) = 30$ and $stdev(3,10) =$. ∎

6. Can you propose a script to convert a BAM file into a bigWig file?

   Solution: The input is a bam file `file.bam` while the output is a bigWig file `file.bw`. In the conversion process, we requires a file containing the sizes of the chromosomes of the human genome, which is `hg.chrom.sizes`. The conversion can be done in 3 steps as follows.

```
bamToBed -i file.bam | awk '{OFS=""; print $1, $2, $3, $4}' |
sort -k 1,1 > file.bed
genomeCoverageBed -i file.bed -g hg.chrom.sizes -bg > file.cov
bedGraphToBigWig file.cov hg.chrom.sizes file.bw
```

∎

Algorithms for Next Generation Sequencing 1st edition Sung Solutions Manual