现代生物学计算环境
Homework02 Solution

姜文渊
学号：1951510
School of Life Science, Tongji University

版本：0.01
更新：March 14, 2021

1 环境准备

同作业 1 的环境。

2 小问 01

问题：*Are there any sequence reads whose centers locate inside gene smad2 (genomic location chr18:45357922-45457515 -)? If yes, how many?*

解：

```
$ cat CTCF.bed | awk ’$1=="chr18"␣&&␣($2+$3)>=45357922*2␣&&␣($2+$3)
   <=45457515*2’ | wc -l
  109
```

3 小问 02

问题： *Change the file from BED format to GFF format. In this study, the 2nd column of GFF file can be set to ‘chipseq’, and the 3rd column set to ‘insulator’. For the 9th column of GFF file, grouping sequence reads by 10,000 bp bins. For example, for a read locates in chr1:123,456-123,481, its group is ‘chr1_120K_130K’.*

解：

```
$ cat CTCF.bed | awk ’{printf("%s\tchipseq\tinsulator\t%s\t%s\t%s\t%s\t
   .\t%s_%d0K_%d0K\n",$1,$2,$3,$5,$6,$1,$2/10000,($3-1)/10000+1)}’ >
   CTCF.gff
```

```
$ head -n1 CTCF.gff
  chr2  chipseq insulator 209779152 209779176 2 + .
     chr2_209770K_209780K
```

4 小问 03

问题：*Based on CTCF.gff generated in step 2), list top 5 bins (containing more reads than other bins), and the read number in each of the 5 bins.*

解：

```
$ cat CTCF.gff | cut -f 9 | sort | uniq -c | sort -k1nr | head -n 5
  5749 chrM_10K_20K
  4375 chrM_00K_10K
  4186 chr8_43090K_43100K
  2282 chr10_42380K_42390K
  2218 chr10_42390K_42400K
```

5 小问 04

问题：*For the following bin "chr17_37070K_37080K", re-group the reads in this bin by 50 bp sub-bins. The reads number of a sub-bin depends on how many read centers in it. After that, display the number of reads in each 50 bp sub-bin by '-'as follows.*

解：

```
$ cat CTCF.gff | grep 'chr17_37070K_37080K$' | awk '{printf("%s\t%d\t%d
   \n",$1,$4/50,($5-1)/50+1)}' | awk '{printf("%s\t%d\t%d\n",$1,$2*50,
   $3*50)}' | sort | uniq -c |  awk '{printf("%s_%d_%d\t",$2,$3,$4);for
   (i=1;i<=$1;i++){printf("-")};printf("\n")}'
...
  chr17_37074550_37074600 --
  chr17_37075050_37075100 -
  chr17_37076200_37076250 -
...
```

6 小结

awk 工具效率低下，但是使用较为方便。如果处理的数据量不是很大，还是可以使用的。但是若处理的数据量较大，例如在数十 GB 级别的文件，则应当避免使用 awk。

此外，作业中有很多细节没有进行进一步的处理，还望阅这份作业的老师斧正。