现代生物学计算环境
Homework01 Solution

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

版本：0.01
更新：March 1, 2021

1 环境准备

笔者使用的软件环境如下所列：

- Ubuntu 18.04.2 LTS (GNU/Linux 4.15.0-132-generic x86_64)
- GNU Wget 1.19.4
- grep (GNU grep)3.1
- wc (GNU coreutils)8.28
- sort (GNU coreutils)8.28
- GNU bash, version 4.4.19(1)-release (x86_64-pc-linux-gnu)

建立工作目录，并下载所需的数据文件：

```
$ mkdir homework01/
$ cd homework01/
$ wget http://10.10.187.253/~course/BioComp2021/data/CTCF.bed.zip
--2021-03-01 14:53:19--  http://10.10.187.253/~course/BioComp2021/data/
   CTCF.bed.zip
Connecting to 10.10.187.253:80... connected.
HTTP request sent, awaiting response... 200 OK
Length: 38360002 (37M) [application/zip]
Saving to: ‘CTCF.bed.zip’
...
2021-03-01 14:53:23 (11.2 MB/s) - ‘CTCF.bed.zip’ saved
   [38360002/38360002]
$ unzip CTCF.bed.zip
Archive:  CTCF.bed.zip
  inflating: CTCF.bed
```

2 小问 01

问题：*Select top 1 million sequence reads and sort them by chromosome and then by start position. Save output in a new BED file.*

分析：利用 `head` 工具选取文件中前 1 Million 个序列，然后利用 `sort` 工具进行排序，排序事所用参数应当格外注意。此外，sort 时如果不能得到预期的效果，则应当将第一列预处理后 sort。

解：

```
$ head -n 1000000 CTCF.bed | tr -d chr | sort -t $'\t' -k 1n,1 -k 2n,2
    | sed 's|^|chr&|g' > CTCF_top1M_sorted.bed
```

效果如下：

```
$ head CTCF_top1M_sorted.bed
chrM  0 24  CTCF.135870 42  +
chrM  0 24  CTCF.597101 42  +
chrM  1 25  CTCF.333481 42  +
chrM  1 25  CTCF.690978 42  -
chrM  2 26  CTCF.312222 42  -
```

3 小问 02

问题：*How many reads locate in each chromosome? Sort output by reads number in descending order.*

分析：利用 `grep` 工具选取同一条染色体上的序列，然后利用 `wc` 工具进行计数，利用 `sort` 工具进行排序后利用 echo 输出。可以写成下面 for 循环的形式。

解：

```
$ for chr in $(grep -o 'chr[a-zA-Z0-9]*' CTCF.bed | sort | uniq);do
    echo $chr $(grep $chr CTCF.bed| wc -l); done | sort -t' ' -k2nr
```

效果如下：

```
chr1 1410122
chr2 413093
chr3 197647
chr6 180673
chr5 175297
chr4 173135
...
```

4 小问 03

问题：*If two or more reads have identical genomic coordinates and strands, they are called duplicated reads. How many genomic locations (considering both coordinates and strands) on chr10 have at least one read? How many locations on chr10 only have one read?*

分析：利用 `grep` 工具选取同一条染色体上的序列，利用 `cut` 对列进行选取后利用 `uniq` 工具进行计数。注意，在使用 `uniq` 工具进行计数前应当先 sort。最后利用 `wc` 工具统计行数。这里使用了 `bc` 做减法。

解：

```
# at least one read
$ grep '^chr10' CTCF.bed | cut -f 2,3,6 | sort | uniq |wc -l
133604

# only have one read
echo $(grep '^chr10' CTCF.bed | cut -f 2,3,6 | sort | uniq |wc -l) - $(
    grep '^chr10' CTCF.bed | cut -f 2,3,6 | sort | uniq -d |wc -l) | bc
121406
```

5 小问 04

问题：*Which genomic location except chrM has the largest number of duplicated reads? What's the number?*

分析：方法类似。

解：

```
$ grep -v '^chrM' CTCF.bed | cut -f 1,2,3,6 | sort | uniq -dc | sort -
    k1nr | head -n1
182 chr16 23650830  23650854  +
```

6 小问 05

问题：*How many genomic locations on autosomes have 5 or more duplicated reads?*

分析：方法类似。这里使用的方法较为笨重（即构造完全符合格式的正则表达式，并且多次使用管道），因为笔者的电脑 RAM 较大，故没有进行更多的优化。

解：

```
$ cut -f 1,2,3,6 CTCF.bed | sort | uniq -dc | sort -k2nr | grep "[[:
    space:]][0-9]*[[:space:]]chr[0-9]*[[:space:]][0-9]*[[:space
    :]][0-9]*[[:space:]][+-]" | grep -v "[[:space:]][0-5][[:space:]]chr
    [0-9]*[[:space:]][0-9]*[[:space:]][0-9]*[[:space:]][+-]" | wc -l
898
```

7 小结

Unix Shell 和 Linux 下的众多小程序，结合管道指令，可以完成大部分数据预处理的工作。当然，想要熟练使用，除了熟悉工具的参数外（个人以为这是最不重要的，毕竟可以 man 一下），还需要在写 shell 命令前构思好工作流（即大问题可以拆分成几个小问题，需要怎么用管道连接不同的指令等）。

这次作业由于输入文件只有 123MB，加上 `grep`，`sort` 等指令效率较高，故而没有考虑对命令执行速度的优化。笔者使用的是一台老旧的台式机（Xeon Gold 5117 @2.00Ghz, 128GB RAM），运行上述命令等待时间在 10s 以内，内存没有溢出，勉强符合要求。

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。