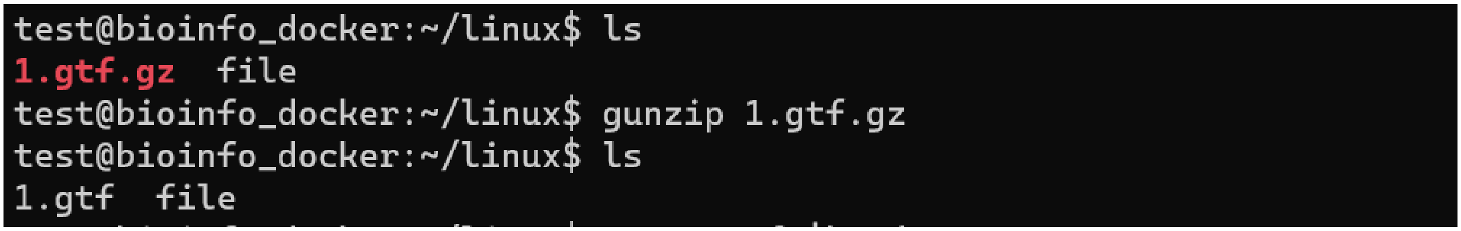
第二周上机作业

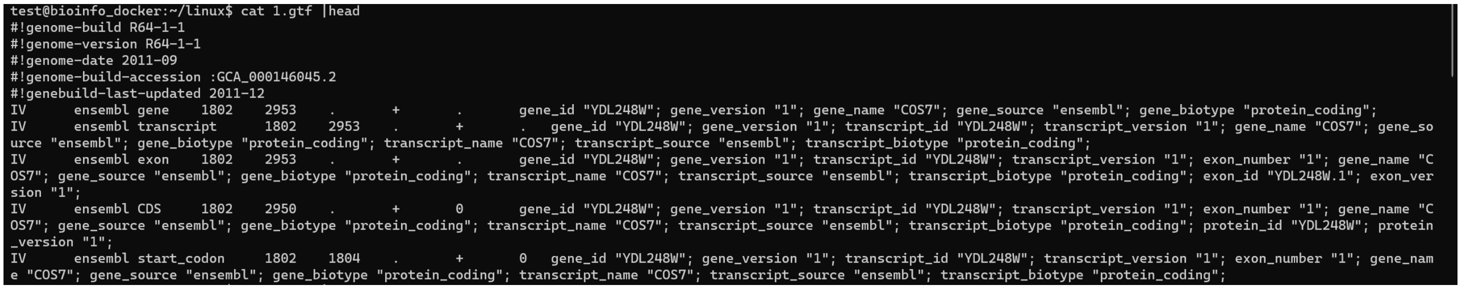
Step 0 解压缩



```
test@bioinfo_docker:~/linux$ ls
1.gtf.gz  file
test@bioinfo_docker:~/linux$ gunzip 1.gtf.gz
test@bioinfo_docker:~/linux$ ls
1.gtf  file
```

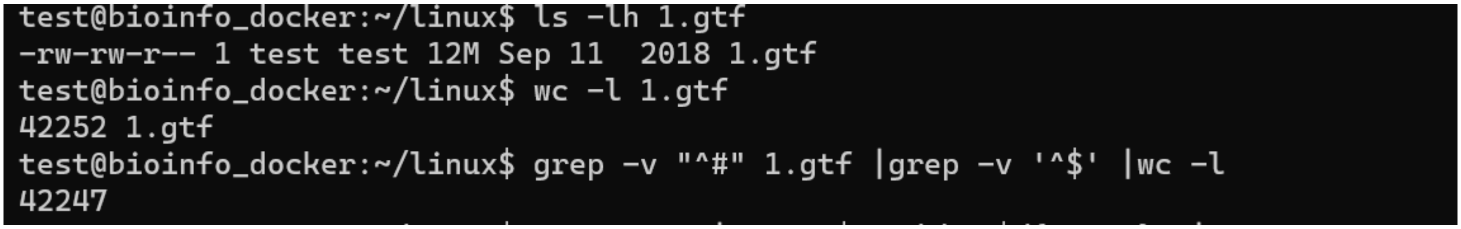
Step 1 查看文件基本信息

- 显示前10行（显示后10行/前15行操作省略）



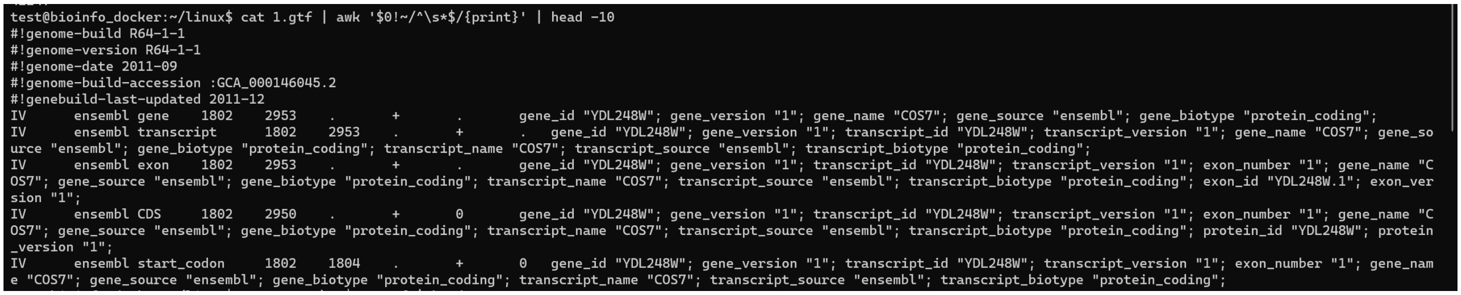
```
test@bioinfo_docker:~/linux$ cat 1.gtf |head
#!genome-build R64-1-1
#!genome-version R64-1-1
#!genome-date 2011-09
#!genome-build-accession :GCA_000146045.2
#!genebuild-last-updated 2011-12
IV      ensembl gene    1802    2953    .       +       .       gene_id "YDL248W"; gene_version "1"; gene_name "COS7"; gene_source "ensembl"; gene_biotype "protein_coding";
IV      ensembl transcript      1802    2953    .       +       .   gene_id "YDL248W"; gene_version "1"; transcript_id "YDL248W"; transcript_version "1"; gene_name "COS7"; gene_so
urce "ensembl"; gene_biotype "protein_coding"; transcript_name "COS7"; transcript_source "ensembl"; transcript_biotype "protein_coding";
IV      ensembl exon    1802    2953    .       +       .       gene_id "YDL248W"; gene_version "1"; transcript_id "YDL248W"; transcript_version "1"; exon_number "1"; gene_name "C
OS7"; gene_source "ensembl"; gene_biotype "protein_coding"; transcript_name "COS7"; transcript_source "ensembl"; transcript_biotype "protein_coding"; exon_id "YDL248W.1"; exon_ver
sion "1";
IV      ensembl CDS     1802    2950    .       +       0       gene_id "YDL248W"; gene_version "1"; transcript_id "YDL248W"; transcript_version "1"; exon_number "1"; gene_name "C
OS7"; gene_source "ensembl"; gene_biotype "protein_coding"; transcript_name "COS7"; transcript_source "ensembl"; transcript_biotype "protein_coding"; protein_id "YDL248W"; protein
_version "1";
IV      ensembl start_codon     1802    1804    .       +       0   gene_id "YDL248W"; gene_version "1"; transcript_id "YDL248W"; transcript_version "1"; exon_number "1"; gene_nam
e "COS7"; gene_source "ensembl"; gene_biotype "protein_coding"; transcript_name "COS7"; transcript_source "ensembl"; transcript_biotype "protein_coding";
```

- 显示文件大小行数



```
test@bioinfo_docker:~/linux$ ls -lh 1.gtf
-rw-rw-r-- 1 test test 12M Sep 11  2018 1.gtf
test@bioinfo_docker:~/linux$ wc -l 1.gtf
42252 1.gtf
test@bioinfo_docker:~/linux$ grep -v "^#" 1.gtf |grep -v '^$' |wc -l
42247
```

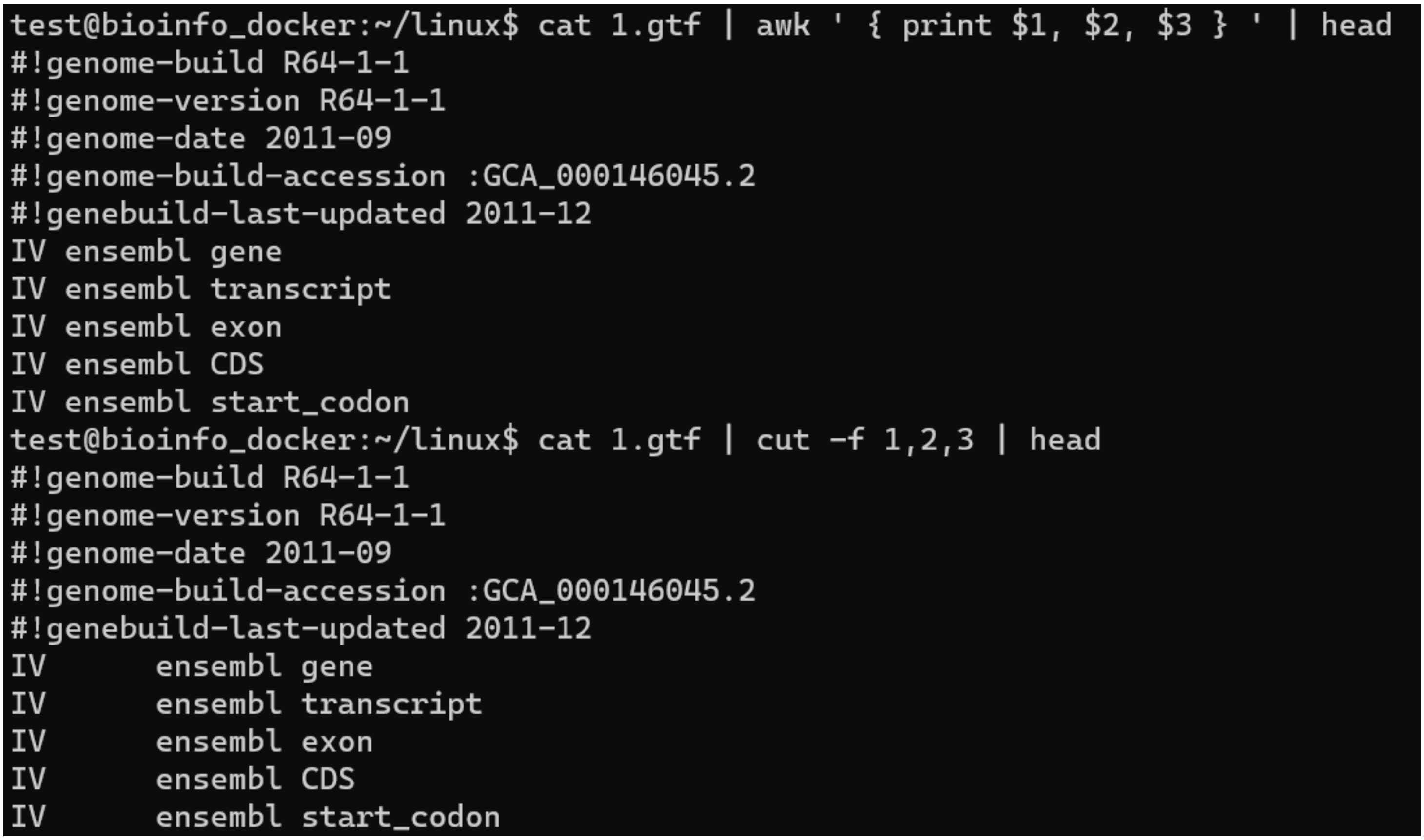
- 过滤操作



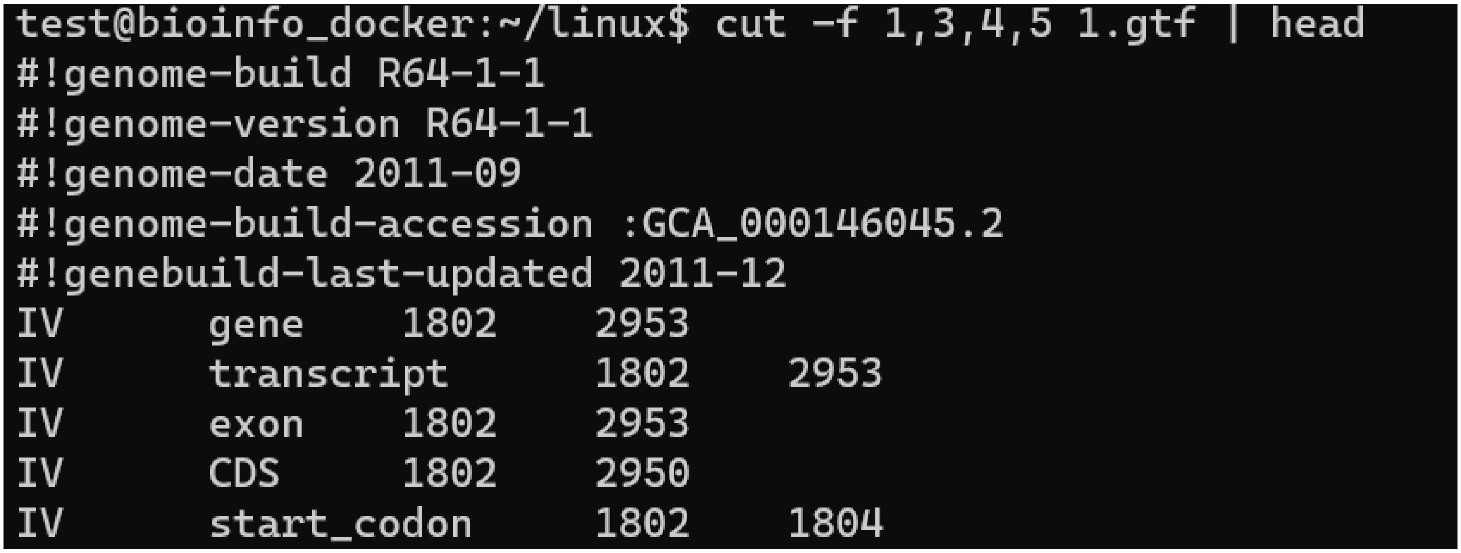
```
test@bioinfo_docker:~/linux$ cat 1.gtf | awk '$0!~/^\s*$/{print}' | head -10
#!genome-build R64-1-1
#!genome-version R64-1-1
#!genome-date 2011-09
#!genome-build-accession :GCA_000146045.2
#!genebuild-last-updated 2011-12
IV      ensembl gene    1802    2953    .       +       .       gene_id "YDL248W"; gene_version "1"; gene_name "COS7"; gene_source "ensembl"; gene_biotype "protein_coding";
IV      ensembl transcript      1802    2953    .       +       .   gene_id "YDL248W"; gene_version "1"; transcript_id "YDL248W"; transcript_version "1"; gene_name "COS7"; gene_so
urce "ensembl"; gene_biotype "protein_coding"; transcript_name "COS7"; transcript_source "ensembl"; transcript_biotype "protein_coding";
IV      ensembl exon    1802    2953    .       +       .       gene_id "YDL248W"; gene_version "1"; transcript_id "YDL248W"; transcript_version "1"; exon_number "1"; gene_name "C
OS7"; gene_source "ensembl"; gene_biotype "protein_coding"; transcript_name "COS7"; transcript_source "ensembl"; transcript_biotype "protein_coding"; exon_id "YDL248W.1"; exon_ver
sion "1";
IV      ensembl CDS     1802    2950    .       +       0       gene_id "YDL248W"; gene_version "1"; transcript_id "YDL248W"; transcript_version "1"; exon_number "1"; gene_name "C
OS7"; gene_source "ensembl"; gene_biotype "protein_coding"; transcript_name "COS7"; transcript_source "ensembl"; transcript_biotype "protein_coding"; protein_id "YDL248W"; protein
_version "1";
IV      ensembl start_codon     1802    1804    .       +       0   gene_id "YDL248W"; gene_version "1"; transcript_id "YDL248W"; transcript_version "1"; exon_number "1"; gene_nam
e "COS7"; gene_source "ensembl"; gene_biotype "protein_coding"; transcript_name "COS7"; transcript_source "ensembl"; transcript_biotype "protein_coding";
```

Step 2 数据提取

- 筛选特定列

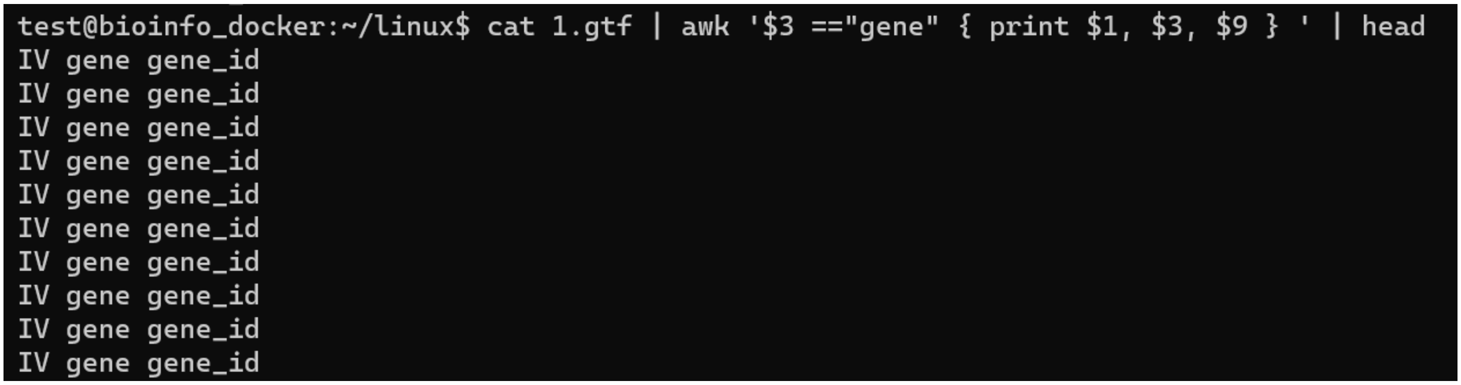


```
test@bioinfo_docker:~/linux$ cat 1.gtf | awk ' { print $1, $2, $3 } ' | head
#!genome-build R64-1-1
#!genome-version R64-1-1
#!genome-date 2011-09
#!genome-build-accession :GCA_000146045.2
#!genebuild-last-updated 2011-12
IV ensembl gene
IV ensembl transcript
IV ensembl exon
IV ensembl CDS
IV ensembl start_codon
test@bioinfo_docker:~/linux$ cat 1.gtf | cut -f 1,2,3 | head
#!genome-build R64-1-1
#!genome-version R64-1-1
#!genome-date 2011-09
#!genome-build-accession :GCA_000146045.2
#!genebuild-last-updated 2011-12
IV      ensembl gene
IV      ensembl transcript
IV      ensembl exon
IV      ensembl CDS
IV      ensembl start_codon
```



```
test@bioinfo_docker:~/linux$ cut -f 1,3,4,5 1.gtf | head
#!genome-build R64-1-1
#!genome-version R64-1-1
#!genome-date 2011-09
#!genome-build-accession :GCA_000146045.2
#!genebuild-last-updated 2011-12
IV      gene    1802    2953
IV      transcript      1802    2953
IV      exon    1802    2953
IV      CDS     1802    2950
IV      start_codon     1802    1804
```

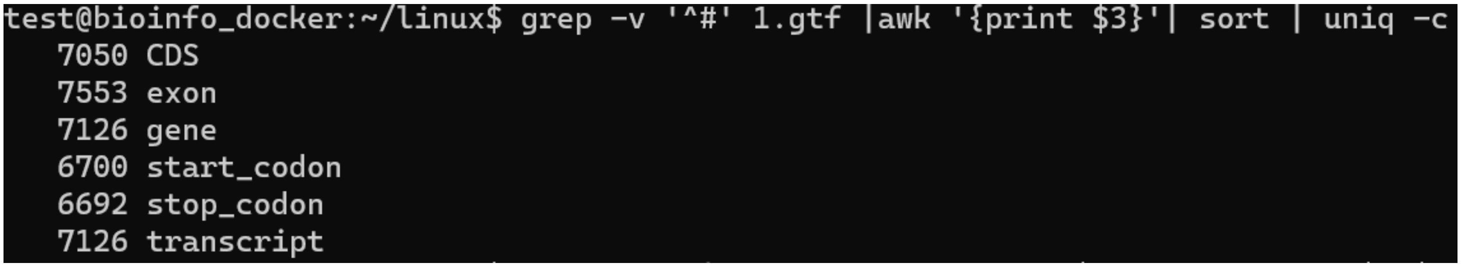
- 筛选特定行



```
test@bioinfo_docker:~/linux$ cat 1.gtf | awk '$3 =="gene" { print $1, $3, $9 } ' | head
IV gene gene_id
IV gene gene_id
IV gene gene_id
IV gene gene_id
IV gene gene_id
IV gene gene_id
IV gene gene_id
IV gene gene_id
IV gene gene_id
IV gene gene_id
```

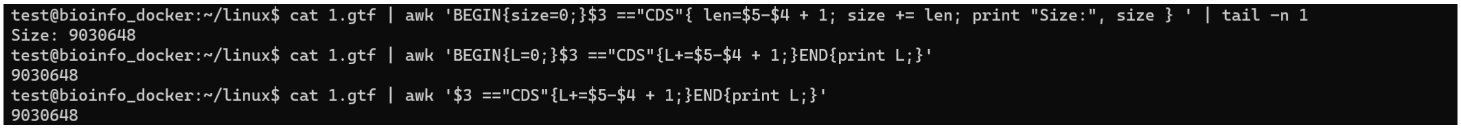
Step 3 提取和计算特定的feature

- 提取并统计featrue类型



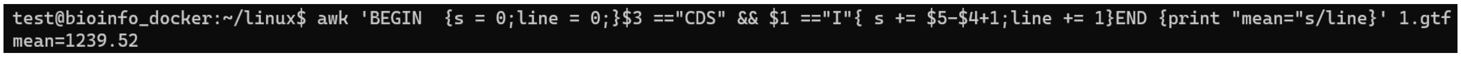
```
test@bioinfo_docker:~/linux$ grep -v '^#' 1.gtf |awk '{print $3}'| sort | uniq -c
   7050 CDS
   7553 exon
   7126 gene
   6700 start_codon
   6692 stop_codon
   7126 transcript
```

- 计算所有CDS



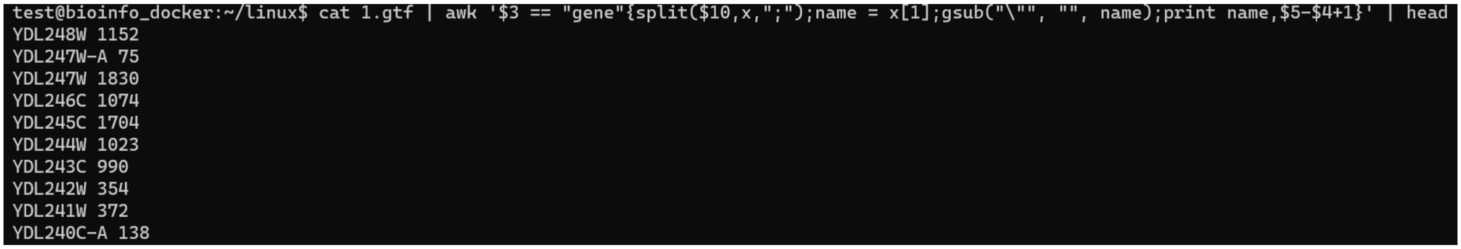
```
test@bioinfo_docker:~/linux$ cat 1.gtf | awk 'BEGIN{size=0;}$3 =="CDS"{ len=$5-$4 + 1; size += len; print "Size:", size } ' | tail -n 1
Size: 9030648
test@bioinfo_docker:~/linux$ cat 1.gtf | awk 'BEGIN{L=0;}$3 =="CDS"{L+=$5-$4 + 1;}END{print L;}'
9030648
test@bioinfo_docker:~/linux$ cat 1.gtf | awk '$3 =="CDS"{L+=$5-$4 + 1;}END{print L;}'
9030648
```

- 计算1号染色体cds的平均长度



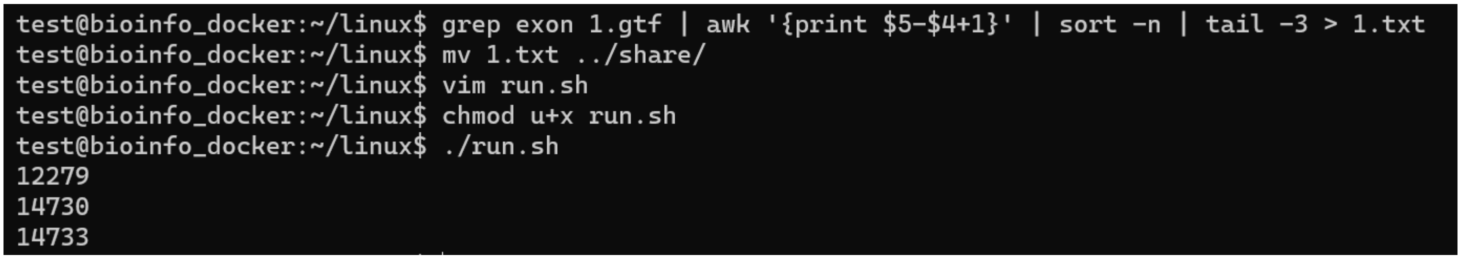
```
test@bioinfo_docker:~/linux$ awk 'BEGIN  {s = 0;line = 0;}$3 =="CDS" && $1 =="I"{ s += $5-$4+1;line += 1}END {print "mean="s/line}' 1.gtf
mean=1239.52
```

- 分离并提取基因名字



```
test@bioinfo_docker:~/linux$ cat 1.gtf | awk '$3 == "gene"{split($10,x,";");name = x[1];gsub("\"", "", name);print name,$5-$4+1}' | head
YDL248W 1152
YDL247W-A 75
YDL247W 1830
YDL246C 1074
YDL245C 1704
YDL244W 1023
YDL243C 990
YDL242W 354
YDL241W 372
YDL240C-A 138
```

Step 4 提取数据并存入新文件



```
test@bioinfo_docker:~/linux$ grep exon 1.gtf | awk '{print $5-$4+1}' | sort -n | tail -3 > 1.txt
test@bioinfo_docker:~/linux$ mv 1.txt ../share/
test@bioinfo_docker:~/linux$ vim run.sh
test@bioinfo_docker:~/linux$ chmod u+x run.sh
test@bioinfo_docker:~/linux$ ./run.sh
12279
14730
14733
```