bash_bioinfo

November 13, 2020

1 Bash for Bioinformatics

A scripting language is a programmable language that supports scripts, namely programs written for a special run-time environment that automate the execution of tasks that could alternatively be executed one-by-one by a human operator. **Bash** is the most common Unix textual shell that also provide a bash programming language. In this course, the student will learn how to **automate extensive computational tasks** (i.e. running programs, dealing with their outputs and making pipelines), and the concept of **batch processing**, namely the execution of a series of jobs in a program on a computer without manual intervention Bash scripts will be used in order to build (bioinformatics) pipelines that transform raw data, execute programs, and present results. The focus is given to real applications in the fields of bioinformatics for what concern the analysis do the input data and the performance valuation of computational tools.

2 Exercises

2.1 Exercise 1

1. Create a script to download the gff3 file of *mycoplasma genitalium* from ftp://ftp.ensemblgenomes.org/pub/bacteria/release-45/gff3/bacteria_13_collection/mycoplasma_genitalium_g37 such that no multiple copies of the file are made.

- the gff3 file must be saved as `myco.gff3`
- Suggested tools: `wget`, `if [ -f file ]`, `gunzip`

[1]:

```
filename="Mycoplasma_genitalium_g37.ASM2732v1.37.gff3"

if [ -f "${filename}.gz" ]; then
    rm ${filename}.gz
fi

wget ftp://ftp.ensemblgenomes.org/pub/bacteria/release-45/gff3/
 ↪bacteria_13_collection/mycoplasma_genitalium_g37/${filename}.gz
gunzip ${filename}.gz
mv $filename myco.gff3
```

```
--2019-10-24 14:17:58--  ftp://ftp.ensemblgenomes.org/pub/bacteria/release-45/gf
f3/bacteria_13_collection/mycoplasma_genitalium_g37/Mycoplasma_genitalium_g37.AS
M2732v1.37.gff3.gz
           => ‘Mycoplasma_genitalium_g37.ASM2732v1.37.gff3.gz’
Resolving ftp.ensemblgenomes.org (ftp.ensemblgenomes.org)… 193.62.197.94
Connecting to ftp.ensemblgenomes.org
(ftp.ensemblgenomes.org)|193.62.197.94|:21… connected.
Logging in as anonymous … Logged in!
==> SYST … done.    ==> PWD … done.
==> TYPE I … done.  ==> CWD (1)
/pub/bacteria/release-45/gff3/bacteria_13_collection/mycoplasma_genitalium_g37
… done.
==> SIZE Mycoplasma_genitalium_g37.ASM2732v1.37.gff3.gz … 39544
==> PASV … done.    ==> RETR Mycoplasma_genitalium_g37.ASM2732v1.37.gff3.gz
… done.
Length: 39544 (39K) (unauthoritative)

Mycoplasma_genitali 100%[===================>]  38.62K  --.-KB/s    in 0.07s

2019-10-24 14:17:59 (537 KB/s) -
‘Mycoplasma_genitalium_g37.ASM2732v1.37.gff3.gz’ saved [39544]
```

[4]:

```
head -n 20 myco.gff3
```

```
##gff-version 3
##sequence-region   Chromosome 1 580076
#!genome-build European Nucleotide Archive ASM2732v1
#!genome-version ASM2732v1
#!genome-date 2014-05
#!genome-build-accession GCA_000027325.1
#!genebuild-last-updated 2014-05
Chromosome      European Nucleotide Archive     chromosome      1       580076
.       .       .       ID=chromosome:Chromosome;Alias=L43967.2;Is_circular=true
###
Chromosome      ena_misc_feature        biological_region       1       580076
.       +       .       external_name=The isolate originally sequenced%2C while
still G37%2C came from the laboratory of P.C. Hu at the University of North
Carolina. Dr. Hu has retired and the sequenced stock is no longer available. The
stock used for re-sequencing came directly from ATCC%2C
and…;logic_name=ena_misc_feature
Chromosome      ena     gene    686     1828    .       +       .
ID=gene:MG_001;Name=dnaN;biotype=protein_coding;description=DNA polymerase
III%2C beta subunit;gene_id=MG_001;logic_name=ena
Chromosome      ena     mRNA    686     1828    .       +       .       ID=trans
cript:AAC71217;Parent=gene:MG_001;Name=dnaN-1;biotype=protein_coding;transcript_
id=AAC71217
Chromosome      ena     exon    686     1828    .       +       .       Parent=t
```

```
ranscript:AAC71217;Name=AAC71217-1;constitutive=1;ensembl_end_phase=0;ensembl_ph
ase=0;exon_id=AAC71217-1;rank=1
Chromosome      ena     CDS     686     1828    .       +       0
ID=CDS:AAC71217;Parent=transcript:AAC71217;protein_id=AAC71217
###
Chromosome      ena_variation   biological_region       728     728     .
+       .       external_name=MG_001%3B
L43967.2:variation:728;logic_name=ena_variation
Chromosome      ena     gene    1828    2760    .       +       .
ID=gene:MG_002;biotype=protein_coding;description=DnaJ domain
protein;gene_id=MG_002;logic_name=ena
Chromosome      ena     mRNA    1828    2760    .       +       .       ID=trans
cript:AAC71218;Parent=gene:MG_002;biotype=protein_coding;transcript_id=AAC71218
Chromosome      ena     exon    1828    2760    .       +       .       Parent=t
ranscript:AAC71218;Name=AAC71218-1;constitutive=1;ensembl_end_phase=0;ensembl_ph
ase=0;exon_id=AAC71218-1;rank=1
Chromosome      ena     CDS     1828    2760    .       +       0
ID=CDS:AAC71218;Parent=transcript:AAC71218;protein_id=AAC71218
```

2.2 Excericise 2

2. From the file `myco.gff3`:
3. count how many chromosomes are reported inside the file
4. count how many genes are in the list
5. give the non redundant list of sources of the reported features
6. which feature types are reported in the file, and for each of them how many feature are reported
7. list gene symbols (Name) without duplicates
8. count how many genes do not have a name

The queries must be done in one single line, when it is possible. Suggested tools: `grep`, `wc`, `sort`, `uniq`, `sed`

[1]:
```
grep -v -P "^#" myco.gff3 | cut -f1 | sort | uniq | wc -l
```

```
1
```

[2]:
```
grep -v -P "^#" myco.gff3 | grep -P "\tgene\t" | wc -l
```

```
477
```

[3]:
```
grep -v -P "^#" myco.gff3 | cut -f2 | sort | uniq
```

```
ena
ena_gene
ena_misc_feature
ena_misc_rna
ena_variation
```

```
European Nucleotide Archive
Rfam
```

[4]:
```
grep -v -P "^#" myco.gff3 | cut -f3 | sort | uniq -c
```

```
     47 biological_region
    476 CDS
      1 chromosome
    559 exon
    477 gene
    476 mRNA
      2 ncRNA_gene
      9 rRNA
      9 rRNA_gene
     74 transcript
     71 tRNA_gene
```

[46]:
```
fields=`grep -v -P "^#" myco.gff3 | grep -P "\tgene\t" | cut -f9`
echo "$fields" | sed s/\;/\\\n/g | grep -P "^Name=" | sed s/Name=//g | sort |␣
 ↪uniq | wc -l
```

```
213
```

[48]:
```
grep -v -P "^#" myco.gff3 | grep -P "\tgene\t" | grep -v "Name=" | wc -l
```

```
264
```

2.3 Excercies 3

Download the human annotation file at this link: ftp://ftp.ensembl.org/pub/release-98/gff3/homo_sapiens/Homo_sapiens.GRCh38.98.gff3.gz and rename it as grch38.gff3. Answer to the same questions of excercise 2, then make the same statistics but only for features on the chromsome 1. Suggested tools: `grep`, `wc -l`, `mktemp`, `sort`, `uniq`, `sed`

[51]:
```
rm Homo_sapiens.GRCh38.98.gff3.gz
wget ftp://ftp.ensembl.org/pub/release-98/gff3/homo_sapiens/Homo_sapiens.GRCh38.
 ↪98.gff3.gz
gunzip Homo_sapiens.GRCh38.98.gff3.gz
mv Homo_sapiens.GRCh38.98.gff3 grch38.gff3
```

```
rm: cannot remove 'Homo_sapiens.GRCh38.98.gff3.gz': No such file or directory
--2019-10-24 15:21:02--  ftp://ftp.ensembl.org/pub/release-98/gff3/homo_sapiens/
Homo_sapiens.GRCh38.98.gff3.gz
           => 'Homo_sapiens.GRCh38.98.gff3.gz'
Resolving ftp.ensembl.org (ftp.ensembl.org)… 193.62.193.8
Connecting to ftp.ensembl.org (ftp.ensembl.org)|193.62.193.8|:21… connected.
Logging in as anonymous … Logged in!
==> SYST … done.    ==> PWD … done.
==> TYPE I … done.  ==> CWD (1) /pub/release-98/gff3/homo_sapiens … done.
==> SIZE Homo_sapiens.GRCh38.98.gff3.gz … 41163349
==> PASV … done.    ==> RETR Homo_sapiens.GRCh38.98.gff3.gz … done.
Length: 41163349 (39M) (unauthoritative)

Homo_sapiens.GRCh38 100%[===================>]  39.26M  5.01MB/s    in 7.6s

2019-10-24 15:21:11 (5.19 MB/s) - ‘Homo_sapiens.GRCh38.98.gff3.gz’ saved
[41163349]
```

[53]:
```
head -n 10 grch38.gff3
```

```
##gff-version 3
##sequence-region   1 1 248956422
##sequence-region   10 1 133797422
##sequence-region   11 1 135086622
##sequence-region   12 1 133275309
##sequence-region   13 1 114364328
##sequence-region   14 1 107043718
##sequence-region   15 1 101991189
##sequence-region   16 1 90338345
##sequence-region   17 1 83257441
```

[57]:
```
grep -v "#" grch38.gff3 | head -n 25
```

```
1       Ensembl chromosome      1       248956422       .       .       .
ID=chromosome:1;Alias=CM000663.2,chr1,NC_000001.11
1       .       biological_region       10469   11240   1.3e+03 .       .
external_name=oe %3D 0.79;logic_name=cpg
1       .       biological_region       10650   10657   0.999   +       .
logic_name=eponine
1       .       biological_region       10655   10657   0.999   -       .
logic_name=eponine
1       .       biological_region       10678   10687   0.999   +       .
logic_name=eponine
1       .       biological_region       10681   10688   0.999   -       .
logic_name=eponine
1       .       biological_region       10707   10716   0.999   +       .
logic_name=eponine
1       .       biological_region       10708   10718   0.999   -       .
logic_name=eponine
1       .       biological_region       10735   10747   0.999   -       .
logic_name=eponine
1       .       biological_region       10737   10744   0.999   +       .
logic_name=eponine
1       .       biological_region       10766   10773   0.999   +       .
logic_name=eponine
```

```
1       .       biological_region       10770   10779   0.999   -       .
logic_name=eponine
1       .       biological_region       10796   10801   0.999   +       .
logic_name=eponine
1       .       biological_region       10810   10819   0.999   -       .
logic_name=eponine
1       .       biological_region       10870   10872   0.999   +       .
logic_name=eponine
1       .       biological_region       10889   10893   0.999   -       .
logic_name=eponine
1       havana  pseudogene      11869   14409   .       +       .       ID=gene:
ENSG00000223972;Name=DDX11L1;biotype=transcribed_unprocessed_pseudogene;descript
ion=DEAD/H-box helicase 11 like 1 [Source:HGNC Symbol%3BAcc:HGNC:37102];gene_id=
ENSG00000223972;logic_name=havana_homo_sapiens;version=5
1       havana  lnc_RNA 11869   14409   .       +       .       ID=transcript:EN
ST00000456328;Parent=gene:ENSG00000223972;Name=DDX11L1-202;biotype=lncRNA;tag=ba
sic;transcript_id=ENST00000456328;transcript_support_level=1;version=2
1       havana  exon    11869   12227   .       +       .       Parent=transcrip
t:ENST00000456328;Name=ENSE00002234944;constitutive=0;ensembl_end_phase=-1;ensem
bl_phase=-1;exon_id=ENSE00002234944;rank=1;version=1
1       havana  exon    12613   12721   .       +       .       Parent=transcrip
t:ENST00000456328;Name=ENSE00003582793;constitutive=0;ensembl_end_phase=-1;ensem
bl_phase=-1;exon_id=ENSE00003582793;rank=2;version=1
1       havana  exon    13221   14409   .       +       .       Parent=transcrip
t:ENST00000456328;Name=ENSE00002312635;constitutive=0;ensembl_end_phase=-1;ensem
bl_phase=-1;exon_id=ENSE00002312635;rank=3;version=1
1       havana  pseudogenic_transcript  12010   13670   .       +       .
ID=transcript:ENST00000450305;Parent=gene:ENSG00000223972;Name=DDX11L1-201;bioty
pe=transcribed_unprocessed_pseudogene;tag=basic;transcript_id=ENST00000450305;tr
anscript_support_level=NA;version=2
1       havana  exon    12010   12057   .       +       .       Parent=transcrip
t:ENST00000450305;Name=ENSE00001948541;constitutive=0;ensembl_end_phase=-1;ensem
bl_phase=-1;exon_id=ENSE00001948541;rank=1;version=1
1       havana  exon    12179   12227   .       +       .       Parent=transcrip
t:ENST00000450305;Name=ENSE00001671638;constitutive=0;ensembl_end_phase=-1;ensem
bl_phase=-1;exon_id=ENSE00001671638;rank=2;version=2
1       havana  exon    12613   12697   .       +       .       Parent=transcrip
t:ENST00000450305;Name=ENSE00001758273;constitutive=0;ensembl_end_phase=-1;ensem
bl_phase=-1;exon_id=ENSE00001758273;rank=3;version=2
grep: write error: Broken pipe
```

[3]:

```
function stats {
    ifile="$1"

    echo -n "number of chomosomes: "
    grep -v -P "^#" $ifile | grep -P "\tchromosome\t" | wc -l
    grep -v -P "^#" $ifile | grep -P "\tchromosome\t"
    echo ""
    echo -n "number of genes: "
    grep -v -P "^#" $ifile | grep -P "\tgene\t" | wc -l

    echo ""
    echo "sources:"
    grep -v -P "^#" $ifile | cut -f2 | sort | uniq

    echo ""
    echo "feature types:"
    grep -v -P "^#" $ifile | cut -f3 | sort | uniq -c

    echo ""
    echo -n "number of gene names: "
    fields=`grep -v -P "^#" $ifile | grep -P "\tgene\t" | cut -f9`
    echo "$fields" | sed s/\;/\\\n/g | grep -P "^Name=" | sed s/Name=//g | sort␣
 ↪| uniq | wc -l

    echo ""
    echo -n "number of genes without names: "
    grep -v -P "^#" $ifile | grep -P "\tgene\t" | grep -v "Name=" | wc -l
}

#stats "grch38.gff3"
```

[4]:

```
stats "grch38.gff3"
```

```
number of chomosomes: 25
1       Ensembl      chromosome      1 248956422       .
.       .       ID=chromosome:1;Alias=CM000663.2,chr1,NC_000001.11
10      Ensembl      chromosome      1 133797422       .
.       .       ID=chromosome:10;Alias=CM000672.2,chr10,NC_000010.11
11      Ensembl      chromosome      1 135086622       .
.       .       ID=chromosome:11;Alias=CM000673.2,chr11,NC_000011.10
12      Ensembl      chromosome      1 133275309       .
.       .       ID=chromosome:12;Alias=CM000674.2,chr12,NC_000012.12
13      Ensembl      chromosome      1 114364328       .
.       .       ID=chromosome:13;Alias=CM000675.2,chr13,NC_000013.11
14      Ensembl      chromosome      1 107043718       .
.       .       ID=chromosome:14;Alias=CM000676.2,chr14,NC_000014.9
15      Ensembl      chromosome      1 101991189       .
.       .       ID=chromosome:15;Alias=CM000677.2,chr15,NC_000015.10
16      Ensembl      chromosome      1 90338345        .
.       .       ID=chromosome:16;Alias=CM000678.2,chr16,NC_000016.10
17      Ensembl      chromosome      1 83257441        .
.       .       ID=chromosome:17;Alias=CM000679.2,chr17,NC_000017.11
```

```
18      Ensembl      chromosome      1 80373285        .
.       .       ID=chromosome:18;Alias=CM000680.2,chr18,NC_000018.10
19      Ensembl      chromosome      1 58617616        .
.       .       ID=chromosome:19;Alias=CM000681.2,chr19,NC_000019.10
2       Ensembl      chromosome      1 242193529       .
.       .       ID=chromosome:2;Alias=CM000664.2,chr2,NC_000002.12
20      Ensembl      chromosome      1 64444167        .
.       .       ID=chromosome:20;Alias=CM000682.2,chr20,NC_000020.11
21      Ensembl      chromosome      1 46709983        .
.       .       ID=chromosome:21;Alias=CM000683.2,chr21,NC_000021.9
22      Ensembl      chromosome      1 50818468        .
.       .       ID=chromosome:22;Alias=CM000684.2,chr22,NC_000022.11
3       Ensembl      chromosome      1 198295559       .
.       .       ID=chromosome:3;Alias=CM000665.2,chr3,NC_000003.12
4       Ensembl      chromosome      1 190214555       .
.       .       ID=chromosome:4;Alias=CM000666.2,chr4,NC_000004.12
5       Ensembl      chromosome      1 181538259       .
.       .       ID=chromosome:5;Alias=CM000667.2,chr5,NC_000005.10
6       Ensembl      chromosome      1 170805979       .
.       .       ID=chromosome:6;Alias=CM000668.2,chr6,NC_000006.12
7       Ensembl      chromosome      1 159345973       .
.       .       ID=chromosome:7;Alias=CM000669.2,chr7,NC_000007.14
8       Ensembl      chromosome      1 145138636       .
.       .       ID=chromosome:8;Alias=CM000670.2,chr8,NC_000008.11
9       Ensembl      chromosome      1 138394717       .
.       .       ID=chromosome:9;Alias=CM000671.2,chr9,NC_000009.12
MT      Ensembl      chromosome      1 16569   .       .
.       ID=chromosome:MT;Alias=chrM,J01415.2,NC_012920.1
X       Ensembl      chromosome      1 156040895       .
.       .       ID=chromosome:X;Alias=CM000685.2,chrX,NC_000023.11
Y       Ensembl      chromosome      2781480   56887902
.       .       .       ID=chromosome:Y;Alias=CM000686.2,chrY,NC_000024.10

number of genes: 21487

sources:
.
ensembl
Ensembl
ensembl_havana
ensembl_havana_tagene
havana
havana_tagene
insdc
mirbase

feature types:
 182510 biological_region
 762023 CDS
     29 C_gene_segment
     25 chromosome
     41 D_gene_segment
1371695 exon
 152699 five_prime_UTR
  21487 gene
     97 J_gene_segment
 103513 lnc_RNA
   1879 miRNA
  99916 mRNA
   2235 ncRNA
  23934 ncRNA_gene
  15202 pseudogene
  15251 pseudogenic_transcript
     60 rRNA
    169 scaffold
     50 scRNA
    954 snoRNA
   1915 snRNA
 153974 three_prime_UTR
     22 tRNA
   1155 unconfirmed_transcript
      1 vaultRNA_primary_transcript
    250 V_gene_segment

number of gene names: 21473

number of genes without names: 0
```

[74]:

```
tmp=`mktemp`
echo $tmp
grep -P "^1\t" grch38.gff3 > $tmp
head -n 10 $tmp
stats $tmp

rm $tmp
echo "done"
```

```
/tmp/tmp.qouf2gzepA
1       Ensembl chromosome      1       248956422       .       .       .
ID=chromosome:1;Alias=CM000663.2,chr1,NC_000001.11
1       .       biological_region       10469   11240   1.3e+03 .       .
external_name=oe %3D 0.79;logic_name=cpg
1       .       biological_region       10650   10657   0.999   +       .
logic_name=eponine
1       .       biological_region       10655   10657   0.999   -       .
logic_name=eponine
```

```
1       .       biological_region       10678   10687   0.999   +       .
logic_name=eponine
1       .       biological_region       10681   10688   0.999   -       .
logic_name=eponine
1       .       biological_region       10707   10716   0.999   +       .
logic_name=eponine
1       .       biological_region       10708   10718   0.999   -       .
logic_name=eponine
1       .       biological_region       10735   10747   0.999   -       .
logic_name=eponine
1       .       biological_region       10737   10744   0.999   +       .
logic_name=eponine
number of chomosomes: 1
1       Ensembl         chromosome      1 248956422       .
.       .       ID=chromosome:1;Alias=CM000663.2,chr1,NC_000001.11

number of genes: 2091

sources:
.
ensembl
Ensembl
ensembl_havana
ensembl_havana_tagene
havana
havana_tagene
mirbase

feature types:
  16825 biological_region
  71699 CDS
      1 chromosome
 126208 exon
  12935 five_prime_UTR
   2091 gene
   9208 lnc_RNA
    158 miRNA
   8692 mRNA
    192 ncRNA
   2088 ncRNA_gene
   1293 pseudogene
   1298 pseudogenic_transcript
     22 rRNA
     14 scRNA
     68 snoRNA
    220 snRNA
  13910 three_prime_UTR
     43 unconfirmed_transcript
```

```
number of gene names: 2090

number of genes without names: 0
done
```

2.4 Excercise 4

Construct a series of files,one for each chromosome (by discarding scaffolds) and make them as valid GFF3 files. The files must be saved into a folder named ”GRCH38”.

[89]:

```
mkdir GRCH38

#grep -v -P "^#" grch38.gff3 | cut -f1 | sort | uniq
chrs=`grep -v -P "^#" grch38.gff3 | cut -f1 | sort | uniq | grep -P␣
 ↪"^([1-9]+|X|Y|MT)$"`

for chr in $chrs
do
    echo $chr
    ofile="GRCH38/grch38.${chr}.gff3"
    head -n 1 grch38.gff3 > $ofile
    grep "##sequence-region   ${chr} " grch38.gff3 >> $ofile
    #cat $ofile
    grep -P "^${chr}\t" grch38.gff3 >>$ofile
    #head -n 10 $ofile

done
```

```
mkdir: cannot create directory ‘GRCH38’: File exists
1
11
12
13
14
15
16
17
18
19
2
21
22
3
4
5
6
```

```
7
8
9
MT
X
Y
```

[90]:
```
head -n 10 GRCH38/grch38.1.gff3
```

```
##gff-version 3
##sequence-region   1 1 248956422
1       Ensembl chromosome      1       248956422       .       .       .
ID=chromosome:1;Alias=CM000663.2,chr1,NC_000001.11
1       .       biological_region       10469   11240   1.3e+03 .       .
external_name=oe %3D 0.79;logic_name=cpg
1       .       biological_region       10650   10657   0.999   +       .
logic_name=eponine
1       .       biological_region       10655   10657   0.999   -       .
logic_name=eponine
1       .       biological_region       10678   10687   0.999   +       .
logic_name=eponine
1       .       biological_region       10681   10688   0.999   -       .
logic_name=eponine
1       .       biological_region       10707   10716   0.999   +       .
logic_name=eponine
1       .       biological_region       10708   10718   0.999   -       .
logic_name=eponine
```

2.5 Excercise 5

Make a csv file that reports the statistics of excercise 2 for each chromosome. Every statistic must be reported as number of elements, not as a list of elements.

[6]:
```
tmp=`mktemp`

chrs=`grep -v -P "^#" grch38.gff3 | cut -f1 | sort | uniq | grep -P␣
 ↪"^([1-9]+|X|Y|MT)$"`

echo "# chr file nof_chrs nof_genes nof_sources not_ftypes nof_names␣
 ↪nof_nonames"

for chr in $chrs
do
    ifile="GRCH38/grch38.${chr}.gff3"
    #echo $ifile
    stats $ifile > $tmp
    #cat $tmp
    c1=`grep "number of chomosomes: " $tmp | sed s/number\ of\ chomosomes:\ //g`
    c2=`grep "number of genes: " $tmp | sed s/number\ of\ genes:\ //g`

    start=`grep -n "sources:" $tmp | cut -d":" -f1`
    end=`grep -n "feature types:" $tmp | cut -d":" -f1`
    (( c3 = $end - $start - 2 ))

    start=`grep -n "feature types:" $tmp | cut -d":" -f1`
    end=`grep -n "number of gene names:" $tmp | cut -d":" -f1`
    (( c4 = $end - $start - 2 ))

    c5=`grep "number of gene names: " $tmp | sed s/number\ of\ gene\ names:\ //g`
    c6=`grep "number of genes without names: " $tmp | sed s/number\ of\ genes\ without\ names:\ //g`
    echo "# $chr $ifile $c1 $c2 $c3 $c4 $c5 $c6"
done

rm $tmp
```

```
# chr file nof_chrs nof_genes nof_sources not_ftypes nof_names nof_nonames
# 1 GRCH38/grch38.1.gff3 1 2091 8 19 2090 0
# 11 GRCH38/grch38.11.gff3 1 1393 7 19 1393 0
# 12 GRCH38/grch38.12.gff3 1 1133 8 18 1133 0
# 13 GRCH38/grch38.13.gff3 1 349 8 17 349 0
# 14 GRCH38/grch38.14.gff3 1 839 8 23 839 0
# 15 GRCH38/grch38.15.gff3 1 659 7 20 659 0
# 16 GRCH38/grch38.16.gff3 1 991 8 19 991 0
# 17 GRCH38/grch38.17.gff3 1 1285 8 19 1285 0
# 18 GRCH38/grch38.18.gff3 1 318 7 18 318 0
# 19 GRCH38/grch38.19.gff3 1 1546 7 18 1546 0
# 2 GRCH38/grch38.2.gff3 1 1355 7 22 1354 0
# 21 GRCH38/grch38.21.gff3 1 261 7 19 261 0
# 22 GRCH38/grch38.22.gff3 1 504 7 22 503 0
# 3 GRCH38/grch38.3.gff3 1 1104 8 19 1102 0
# 4 GRCH38/grch38.4.gff3 1 789 8 19 789 0
# 5 GRCH38/grch38.5.gff3 1 957 8 19 955 0
# 6 GRCH38/grch38.6.gff3 1 1082 8 19 1080 0
# 7 GRCH38/grch38.7.gff3 1 1038 7 22 1036 0
# 8 GRCH38/grch38.8.gff3 1 715 7 18 714 0
# 9 GRCH38/grch38.9.gff3 1 800 8 20 800 0
# MT GRCH38/grch38.MT.gff3 1 13 4 9 13 0
# X GRCH38/grch38.X.gff3 1 868 8 19 867 0
# Y GRCH38/grch38.Y.gff3 1 48 6 15 48 0
```

2.6 Excercise 6

The file `families.fa` in the `data` folder contains the aminoacidic sequences of genes belonging to different genomes. Each sequence is composed by a description line, which starts with a character `>` and contains three fields separeted by a tabulation character. The first field is teh name of the genome, the second one is a unique gene identifier, and the thrid field report the family to which the gene belongs. Subsequently, the aminoacidic sequence of the gene is reported by slpitting it into multiple lines that are no longer than 80 characters.

Answer to the following questions: 1. how many genomes are reported in the file? 2. how many genes are reported in the file? 3. how many genes each genome has? 4. how many genes each family has? 5. which is the average length of the aminoacidic sequences? Is there any difference betwen using `bc` and the standard numerical environment? 6. make a distrubution of the gene length. For each length, report how many genes have such a length.

[2]:

```
grep ">" data/families.fa | cut -f1 | sort | uniq | wc -l
```

```
50
```

[3]:

```
grep ">" data/families.fa | cut -f2 | sort | uniq | wc -l
```

```
29479
```

[8]:

```
grep ">" data/families.fa | sed s/\>//g | cut -f1 | sort | uniq -c   | head -n 20
```

```
    596 genome_1051
    591 genome_1108
    590 genome_1122
    601 genome_1132
    604 genome_1163
    594 genome_1208
    577 genome_1229
    595 genome_1252
    576 genome_1308
    581 genome_1309
    593 genome_1328
    575 genome_1366
    592 genome_1371
    600 genome_1374
    594 genome_1375
    582 genome_1407
    579 genome_1435
    578 genome_1474
    588 genome_1569
    580 genome_1602
```

[9]:

```
grep ">" data/families.fa |  cut -f3 | sort | uniq -c   | head -n 20
```

```
     50 sequence_0
     50 sequence_1
     50 sequence_10
     49 sequence_100
      1 sequence_1000
      1 sequence_1001
      1 sequence_1002
      1 sequence_1003
      1 sequence_1004
      1 sequence_1005
      1 sequence_1006
      1 sequence_1007
     50 sequence_101
      1 sequence_10122
      1 sequence_10123
      1 sequence_10124
      1 sequence_10125
      1 sequence_10126
      1 sequence_10127
      3 sequence_1014
uniq: write error: Broken pipe
```

[12]:

```
tmp=`mktemp`


head -n 1000 data/families.fa | sed s/\>.*$/\>/g | tr -d '\n' | sed s/\>/\\n/g >$tmp
#sed s/\>.*$/\>/g data/families.fa | tr -d '\n' | sed s/\>/\\n/g >$tmp

function getlengths {
    while read line
    do
        #echo "@" $line
        echo $line | wc -c
    done < $1
}

lengths=`getlengths $tmp`
n=0
avg=0
avgi=0
for l in $lengths
do
(( avgi =  (($avgi * $n) + $l) / ($n + 1)  ))
avg=`echo "(($avg * $n ) + $l ) / ($n + 1.0)"  | bc -l`
#echo $l $avgi $avg
(( n = n + 1 ))
```

```
done

echo $n $avgi $avg

rm $tmp
```

```
170 310 354.57647058823529411720
```

[11]:

```
tmp=`mktemp`


head -n 1000 data/families.fa | sed s/\>.*$/\>/g | tr -d '\n' | sed s/\>/\\n/g >$tmp
#sed s/\>.*$/\>/g data/families.fa | tr -d '\n' | sed s/\>/\\n/g >$tmp

function getlengths {
    while read line
    do
        #echo "@" $line
        echo $line | wc -c
    done < $1
}

lengths=`getlengths $tmp | sort | uniq -c`
echo "$lengths"

rm $tmp
```

```
      1 1
      1 1026
      2 104
      1 109
      1 1117
      1 116
      1 1191
      1 122
      1 123
      1 124
      1 126
      1 128
      1 129
      1 1290
      2 131
      1 132
      2 134
      2 136
      1 137
```

```
1 138
2 140
3 142
1 146
1 150
1 151
1 153
1 156
1 158
1 161
1 166
1 168
1 176
1 182
1 187
1 188
1 189
1 190
1 195
1 196
1 201
1 204
1 207
1 208
1 213
1 214
1 217
1 219
1 225
1 235
1 237
1 238
1 240
1 241
1 246
1 248
1 250
1 258
2 261
1 263
1 264
1 267
2 271
1 275
1 281
1 285
1 288
1 289
```

```
1 292
1 295
2 299
1 303
1 306
1 313
1 315
2 316
1 325
1 326
1 327
1 333
1 336
1 340
1 341
1 348
1 351
1 352
1 361
1 363
1 370
1 372
1 374
1 375
1 376
1 380
4 390
1 393
1 396
1 401
1 409
1 411
1 421
1 427
1 435
1 452
1 458
1 460
1 470
1 473
1 474
1 492
1 502
2 513
1 522
1 528
2 56
2 561
```

```
1 566
1 572
1 588
1 59
1 598
1 602
1 604
1 617
1 62
1 626
1 640
1 644
1 654
1 664
1 668
1 669
1 713
1 723
1 728
1 74
1 743
1 752
1 758
1 76
1 790
1 796
1 82
1 840
1 870
1 883
2 90
1 903
1 93
1 938
1 946
1 97
1 98
```