

PRÀCTICA FINAL PROGRAMACIÓ EN SCRIPTING

1. OBJECTIU

L'objectiu del present treball és simplificar, mitjançant els diferents scripts, el tractament de les dades, ja sigui per modificar-les i adaptar-les a les nostres necessitats, o per extreure conclusions del conjunt de les dades en brut. És a dir, fer un anàlisi de dades d'expressió gènica per establir correlació entre les diferents variables.

Es treballa amb dos arxius que estan relacionats:

L'arxiu on s'emmagatzemen les dades dels pacients:

"GTEEx_Analysis_v8_Annotations_SubjectPhenotypesDS.txt"

i l'arxiu que emmagatzema les dades de les mostres:

"GTEEx_Analysis_v8_Annotations_SampleAttributesDS.txt"

SCRIPT A

Es crea un script anomenat a.sh que, en el cas de que s'executi sense opcions:

- descarrega els dos arxius, limitant la grandària d'un d'ells amb la comanda "truncate -s 1480k", ja que la seva mida original és molt gran.
- Mostra la URL de descàrrega dels datasets.
- Mostra les respectives mides, nombre de columnes, i nombre de registres.

Les opcions que admet són -v -x -y. En cas contrari, mostra en pantalla les opcions disponibles:

```
GTEEx_Analysis_v8_Anno 100%[=====>] 10,36K --KB/s
2022-01-02 18:15:02 (5,69 MB/s) - 'GTEEx_Analysis_v8_Annotations_Subject
sx.3' saved [10605/10605]

=====
URL del dataset: https://www.gtexportal.org/home/datasets
=====
Mida del fitxer(K):
20 GTEEx_Analysis_v8_Annotations_SubjectPhenotypesDS.txt
1480 GTEEx_Analysis_v8_Annotations_SampleAttributesDS.txt
=====
Nombre de registres:
981 GTEEx_Analysis_v8_Annotations_SubjectPhenotypesDS.txt
2990 GTEEx_Analysis_v8_Annotations_SampleAttributesDS.txt
=====
Nombre de columnes:
4
63
=====
Invalid argument. Options:
- v: Format del fitxer
- x: Data type
- y: Data info
pscrmartinezsequ@pscrmartinezsequ-VirtualBox:~$
```

i executem ./a.sh -v, mostra, a més, el format dels fitxers:

```
=====
Nombre de registres:
981 GTEEx_Analysis_v8_Annotations_SubjectPhenotypesDS.txt
2990 GTEEx_Analysis_v8_Annotations_SampleAttributesDS.txt
=====
Nombre de columnes:
4
63
=====
Format del fitxer:
GTEEx_Analysis_v8_Annotations_SampleAttributesDS.txt: text/plain; charset=us-ascii
GTEEx_Analysis_v8_Annotations_SubjectPhenotypesDS.txt: text/plain; charset=us-ascii
pscrmartinezsequ@pscrmartinezsequ-VirtualBox:~$
```

Si executem `./a.sh -x`, mostra el tipus de dades:

```
2990 GTEx_Analysis_v8_Annotations_SampleAttributesDS.txt
=====
Nombre de columnes:
4
63
=====
Phenotypes file:
Data type integer:
3
Data type decimal:
0
Data type string:
1
SampleAttributes file:
Data type integer:
26
Data type decimal:
23
Data type string:
14
pscrarmartinezsequ@pscrarmartinezsequ-VirtualBox:~$
```

Si executem `./a.sh -y`, mostra informació dels datasets:

```
=====
URL del dataset: https://www.gtexportal.org/home/datasets
=====
Mida del fitxer(K):
20      GTEx_Analysis_v8_Annotations_SubjectPhenotypesDS.txt
1480    GTEx_Analysis_v8_Annotations_SampleAttributesDS.txt
=====
Nombre de registres:
981 GTEx_Analysis_v8_Annotations_SubjectPhenotypesDS.txt
2990 GTEx_Analysis_v8_Annotations_SampleAttributesDS.txt
=====
Nombre de columnes:
4
63
=====
Data available include:
BAM files for RNA-Seq, Whole Exome Seq, and Whole Genome Seq
Genotype Calls (.vcf) for OMNI SNP Arrays, WES, and WGS
OMNI SNP Array Intensity files (.idat and .gtc)
Affymetrix Expression Array Intensity files (.cel)
Allele Specific Expression (ASE) tables
All expression matrices from the Portal, including samples that did not pass the Analysis Freeze QC
Sample Attributes
Subject Phenotypes
pscrarmartinezsequ@pscrarmartinezsequ-VirtualBox:~$
```

SCRIPT B

Aquest script, anomenat `b.sh`, pretén analitzar algunes de les característiques dels datasets. Això ens permetrà establir correlacions entre les variables i obtenir conclusions de les dades en brut.

Quin és el tipus de mort més freqüent per gènere?

El gènere del pacient s'indica a la segona columna de l'arxiu dels pacients i el tipus de mort en la quarta columna. Seleccionem primer el gènere utilitzant un `if` (`$2 == "1"`) i mostrem el nombre de línies que apareix cada tipus de mort amb `| sort | uniq -c`, considerant les dades de la columna 4 `$4` com a numèriques `| sort -n` i ordenades de major a menor `| sort -r`.

Quin és el rang d'edat més freqüent en què es tenen mostres?

El rang d'edat dels pacients el trobem a la tercera columna de l'arxiu de pacients. Seleccionem la columna 3 `$3` de l'arxiu i comptem en quantes línies apareix el mateix rang d'edat amb `sort | uniq -c`

(uniqu -c funciona millor si ordenem abans les línies), i ordenem la sortida amb opció -n (considera la dada com numèric) i l'opció -r (ordre de més a menys).

Podem comprovar que la mort més freqüent per a ambdós gèneres és la 0 = Ventilator Case

```
pvcramartinezsequ@pvcramartinezsequ-VirtualBox:~$ ./b.sh
Quin és el tipus de mort més freqüent per gènere?
Per al gènere home:
  317 0
  189 2
   73 4
   37 3
   23 1
   14
Per al gènere dona:
  194 0
   50 2
   46 4
   20 3
   12 1
    5
Quin és el rang d'edat més freqüent en què es tenen mostres?
  317 60-69
  315 50-59
  153 40-49
   84 20-29
   78 30-39
   33 70-79
```

Quants tipus de regions de cervell (brain) podem trobar?

Per calcular els diferents tipus de regions de cervell hem de consultar la columna 7 de l'arxiu de mostres. En aquest arxiu hi ha camps amb text que té espais. Per estar segur que llegim correctament els camps utilitzem FS = "\t" per assignar un caràcter separador de camps per a la lectura. Seleccionem les línies amb la paraula "Brain" a la columna 7. Ordenem amb sort i comptem les línies úniques `uniq | wc -l`.

El resultat indica 13 tipus de regions de cervell.

Quin és el pacient que més mostres té?

Per calcular quin és el pacient que més mostres té, vam consultar la primera columna de l'arxiu de mostres assignant un caràcter separador de camps per a la lectura FS = "\t", i aprofitant que cada apartat de l'identificador de mostres es separa amb un guió " - ", substituïm el guió per un caràcter tabulació gsub (/ - /, "\t", \$1, i obtenim els apartats de l'identificador com columnes diferents. Les dues primeres columnes resultants seran els identificadors de pacients. Mostrem el nombre de línies que apareix cada pacient amb `| sort | uniq -c`, considerant les dades de la columna 1 de \$1 com numèriques `| sort -n` i ordenades de major a menor `| sort -r`, i només mostrem primer registre amb la comanda `head`.

Quines són les 7 mostres que tenen més "Split Reads"?

El nombre de "Split Reads" es troba a la columna 55 de l'arxiu de mostres. Mostrem les columnes 1 i 55 de l'arxiu assignant un caràcter separador de camps per a la lectura FS = "\t", i ho ordenem per la segona columna, canviem el format del nombre exponencial a nombre enter `printf "%s %d\n", $1, $55` i ordenem pel nombre enter de major a menor `| sort -k2nr | head -7` perquè només mostrem els 7 primers.

També podríem fer servir:

```
gawk 'BEGIN {FS="\t"} {print $1,$55}' GTEX_Analysis_v8_Annotations_SampleAttributesDS.txt
| sort -k2,2gr | head -7
```

utilitzant l'opció -g (g de general) per tractar amb nombres exponencials, no enters.

Quantes dones han mort de manera violenta i tenen mostres de sang?

¿I quina és la mitjana de "mapped unique" d'aquesta selecció?

El gènere dels pacients i el tipus de mort els trobem a les columnes 2 i 4 de l'arxiu de pacients. Obtenim primer els camps 2 i 4 de l'arxiu que compleixen la condició que siguin dones `if ($ 2 == "2") print $1, $4`. Com ara només tenim dues columnes seleccionem el tipus "mort violenta" ara situat a la columna 2 `if ($ 2 == "1")`. Substituïm el guió per un caràcter tabulació a la primera columna `gsub (/ - /, "\ t", $ 1`, i obtenim els apartats de l'identificador de pacient com columnes diferents, ordenem les línies i el resultat el guardem en un altre arxiu anomenat "dones1 .txt ", que el farem servir per fer el join per poder respondre a la següent qüestió:

El tipus de mostra i el nombre de "mapped unique" els trobem en les columnes 6 i 48 de l'arxiu de mostres. Per obtenir les mostres de sang amb el nombre de "mapped unique", vam consultar l'arxiu de mostres i seleccionem les columnes 1, 6 i 48 i les línies que tinguin la paraula "Blood" `grep "Blood"` I descartem les línies que també tinguin la paraula "Vessel" `grep -v "Vessel"`. Substituïm el guió per un caràcter tabulació a la primera columna `gsub (/ - /, "\ t", $1`, i obtenim els apartats de l'identificador de mostra com columnes diferents, ordenem les línies i el resultat el guardem en un altre arxiu anomenat "mostresSang .txt ", que el farem servir per fer el join.

Per calcular quantes dones han mort de manera violenta i tenen mostres de sang fem el join dels dos arxius, descartem les mostres amb l'identificador repetit i comptem les línies.

Per calcular la mitjana dels "mapped unique" descartem les línies que no hi hagi dades a la columna del nombre de "mapped unique" situada a la posició 8 després de fer el join `if ($ 8! = "") Print $8`, sumem tots les dades de "mapped unique" i el dividim pel nombre total de línies `BEGIN {t = 0} {t = t + $ 1} END {print t / NR}`

```
Quants tipus de regions de cervell podem trobar?
13
Brain - Amygdala
Brain - Anterior cingulate cortex (BA24)
Brain - Caudate (basal ganglia)
Brain - Cerebellar Hemisphere
Brain - Cerebellum
Brain - Cortex
Brain - Frontal Cortex (BA9)
Brain - Hippocampus
Brain - Hypothalamus
Brain - Nucleus accumbens (basal ganglia)
Brain - Putamen (basal ganglia)
Brain - Spinal cord (cervical c-1)
Brain - Substantia nigra
Quin és el pacient que més mostres té?
42 GTEX 12WSD
Quines són les 7 mostres que tenen més Split Reads?
GTEX-11TTK-0005-SM-509BX 39129800
GTEX-131XE-0006-SM-5P9F9 38413500
GTEX-11NSD-0526-SM-5A5LT 38210500
GTEX-110NC-0005-SM-509CY 36385400
GTEX-12C56-1626-SM-5FQUO 35898200
GTEX-110F3-0006-SM-509CM 35647500
GTEX-1339X-2726-SM-5PNYU 34955400
Quantes dones han mort de manera violenta i tenen mostres de sang?
1
¿I quina és la mitjana de mapped unique d'aquestes?
88055100
pscrmartinezsequ@pscrmartinezsequ-VirtualBox:~$
```

SCRIPT C

Inclou dos scripts que funcionen usant la invocació següent:

```
./c_bash.sh directori
```

```
gawk -f c_awk.awk GTEX_Analysis_v8_Annotations_SubjectPhenotypesDS.txt
```

```

GTEX-ZVT3      2      60-69  Unexpected
GTEX-ZVT4      2      50-59   Slow
GTEX-ZVTK      1      20-29  Natural
GTEX-ZVZO      1      40-49
GTEX-ZVZP      1      50-59  Violenta
GTEX-ZVZQ      2      60-69  Natural
GTEX-ZWKS      1      30-39  Ventilator_Case
GTEX-ZXES      2      30-39  Violenta
GTEX-ZXGS      1      60-69   Slow
GTEX-ZY6K      1      50-59  Violenta
GTEX-ZYFC      1      50-59  Unexpected
GTEX-ZYFD      1      50-59   Slow
GTEX-ZYFG      2      60-69  Violenta
GTEX-ZYT6      1      30-39  Natural
GTEX-ZYVF      2      50-59  Violenta
GTEX-ZYW4      1      60-69   Slow
GTEX-ZYWO      2      40-49  Violenta
GTEX-ZYY3      2      60-69  Ventilator_Case
GTEX-ZZ64      1      20-29  Violenta
GTEX-ZZPT      1      50-59  Ventilator_Case
GTEX-ZZPU      2      50-59  Violenta
K-562      2      50-59

Death classification based on the 4-point Hardy Scale

1) Violent and fast death Deaths due to accident, blunt force trauma or suicide, terminal phase estimated at < 10 min.
2) Fast death of natural causes Sudden unexpected deaths of people who had been reasonably healthy, after a terminal phase estimated at < 1 hr (with sudden death from a myocardial infarction as a model cause of death for this category).
3) Intermediate death Death after a terminal phase of 1 to 24 hrs (not classifiable as 2 or 4); patients who were ill but death was unexpected.
4) Slow death Death after a long illness, with a terminal phase longer than 1 day (commonly cancer or chronic pulmonary disease); deaths that are not unexpected.
5) Ventilator Case All cases on a ventilator immediately before death.
.....
pscrmartinezsequ@pscrmartinezsequ-VirtualBox:~$

```

En aquesta imatge veiem el resultat de
`gawk -f c_awk.awk GTEEx_Analysis_v8_Annotations_SubjectPhenotypesDS.txt`

L'script `c_awk.awk` renombra la quarta columna de l'arxiu dels pacients, substituint la classificació numèrica per un string. Fem servir dos arrays, un amb els ordinals i un altre amb les categories (string). Amb 2 bucles for associem els dos arrays. També mostra al final la classificació que es fa servir al dataset, i una explicació de cada tipus de mort.

L'objectiu d'aquest script és facilitar la visualització i comprensió de les dades.

```

BEGIN {
    split("0 1 2 3 4", nrsArr)
    split("Violenta Natural Unexpected Slow Ventilator_Case", namesArr)
    for (i in nrsArr) {
        nr2name[nrsArr[i]] = namesArr[i]
    }
}
{
    n = split($NF, nrs, /,/)
    sub(/^[^[:space:]]+$/, "")
    printf "%s", $0
    for (i=1; i<=n; i++) {
        printf "%s%s", nr2name[nrs[i]], (i<n ? ", " : ORS)
    }
}

```

```

pscrmartinezsequ@pscrmartinezsequ-VirtualBox:~$ ./c_bash.sh directori
Iniciant execució:
PID és 6534
Nom del directori:
directori
Fitxers del directori directori:
GTEx_Analysis_v8_Annotations_SampleAttributesDS.txt
GTEx_Analysis_v8_Annotations_SubjectPhenotypesDS.txt
Fitxers del directori directori renombrats:
GTEx_Analysis_v8_Annotations_SampleAttributesDS.csv
GTEx_Analysis_v8_Annotations_SubjectPhenotypesDS.csv
Treballant sobre cada fitxer...
Iniciant execució:
PID és 6534
Nom del fitxer:
directori/GTEx_Analysis_v8_Annotations_SampleAttributesDS.csv
Mida del fitxer:
11M
Nombre de registres:
22952
Iniciant execució:
PID és 6534
Nom del fitxer:
directori/GTEx_Analysis_v8_Annotations_SubjectPhenotypesDS.csv
Mida del fitxer:
20K
Nombre de registres:
981
pscrmartinezsequ@pscrmartinezsequ-VirtualBox:~$ █

```

L'script `c_bash.sh` renombra cadascun dels fitxers del directori amb un bucle `for` : canviem l'extensió `.txt` per `.csv`.

```

for filename in $1;
do
    find . -name "*.txt" | awk -v mvCmd='mv "%s" "%s"\n' \
    '{ old=$0;
      sub(/[/.]txt$/, ".csv");
      printf mvCmd,old,$0;
    }' | sh
done

```

Amb un bucle `while`, opera sobre cadascun d'ells. L'executarem amb el nostre directori “directori” que conté els dos datasets, però es podria fer servir amb altres directoris.

```

ls $1 | while read file;
do
    echo "Iniciant execució:";
    echo "PID és $$";
    echo "Nom del fitxer:";
    echo "$1/$file";
    echo "Mida del fitxer:";
    ls -sh $1/$file | gawk '{print $1}';
    echo "Nombre de registres:";
    wc -l $1/$file | gawk '{print $1}';
done

```

L'objectiu d'aquest script és mecanitzar una feina recursiva, com renombrar varis fitxers o analitzar els fitxers d'un directori un per un.

SCRIPT D

S'elabora un script anomenat d.sh que genera un document en format HTML5. Mostra en una taula totes les columnes:

```
awk '
BEGIN {print "<table border=2 cellspacing=2 cellpadding=2>"}
{
    print "<tr>"
    if (NR==1) {starttag="th scope=\"col\""; endtag="th"}
    else {starttag="td"; endtag="td"}
    for(i=1;i<=NF;i++) {
        (NR>1 && i==4 && substr($i,0,length($i)-1) >85) ? bg=" style=\"background-
color:#ff0000\" : bg=""
        print "<" starttag bg ">" $i "</" endtag ">"
    }
    print "</tr>"
}
END {print "</table>"}'
/home/pscrmartinezsequ/directori/GTEEx_Analysis_v8_Annotations_SubjectPhenotypesDS.csv >
d.html
```

Opció 2:

```
awk 'BEGIN {print "<table>"} print "<tr>"; for(i= 1; i <=NF; i++) print "<td bgcolor=#00FF00>"
$i "</td>"; print "</tr>"}END {print "</table>"}'
/home/pscrmartinezsequ/GTEEx_Analysis_v8_Annotations_SubjectPhenotypesDS.txt >
/home/pscrmartinezsequ/d.html
```

SCRIPT E

És un script principal, anomenat run.sh que executa pas a pas tot el projecte, cridant als scripts anteriors. Funciona usant la invocació: ./run.sh.

```
echo "LLancem a.sh ";
./a.sh;
echo "LLancem b.sh ";
./b.sh;
echo "LLancem c_bash.sh";
./c_bash.sh directori;
echo "LLancem c_awk.awk";
gawk -f c_awk.awk GTEEx_Analysis_v8_Annotations_SubjectPhenotypesDS.txt;
echo "LLancem d.sh";
./d.sh;
```

firefox d.html;

La taula que obtenim no té cap agrupació de dades. Aquest ha sigut un dels entrebancs més grans que he tingut per elaborar el projecte. Queda pendent aprofundir en aquest aspecte.

/home/pscrmartinezsequ/d × +			
← → ↻ file:///home/pscrmartinezsequ/d.html			
SUBJID	SEX	AGE	DTHHRDY
GTEX-1117F	2	60-69	4
GTEX-111CU	1	50-59	0
GTEX-111FC	1	60-69	1
GTEX-111VG	1	60-69	3
GTEX-111YS	1	60-69	0
GTEX-1122O	2	60-69	0
GTEX-1128S	2	60-69	2
GTEX-113IC	1	60-69	
GTEX-113JC	2	50-59	2
GTEX-117XS	1	60-69	2
GTEX-117YW	1	50-59	3
GTEX-117YX	1	50-59	0
GTEX-1192W	1	60-69	2
GTEX-1192X	1	50-59	4
GTEX-11DXW	1	40-49	2
GTEX-11DXX	2	60-69	0
GTEX-11DXY	1	60-69	2
GTEX-11DXZ	1	50-59	0
GTEX-11DYG	1	60-69	2
GTEX-11DZ1	1	50-59	4
GTEX-11EI6	1	60-69	4
GTEX-11EM3	2	20-29	0