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 modificarles 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.

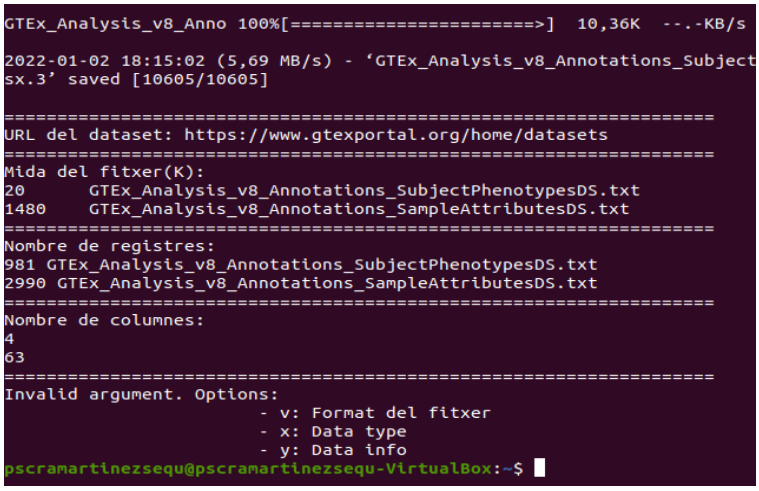
Es treballa amb dos arxius que estan relacionats:
L'arxiu on s'emmagatzemen les dades dels pacients:
"GTEx_Analysis_v8_Annotations_SubjectPhenotypesDS.txt"
i l'arxiu que emmagatzema les dades de les mostres:
"GTEx_Analysis_v8_Annotations_SampleAttributesDS.txt"

SCRIPT A

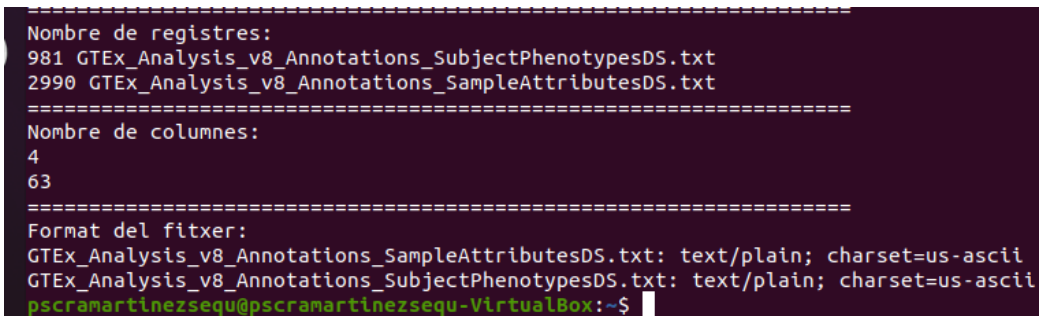
Es crea un script anomenat a.sh que, en el cas de que s'executi sense opcions:

○ 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.

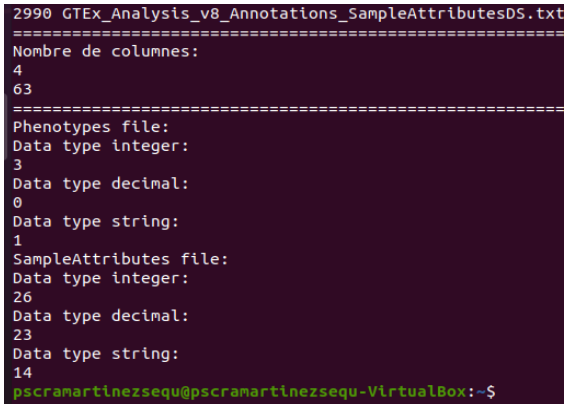
Les opcions que adment són -v -x -y. En cas contrari, mostra en pantalla les opcions disponibles:



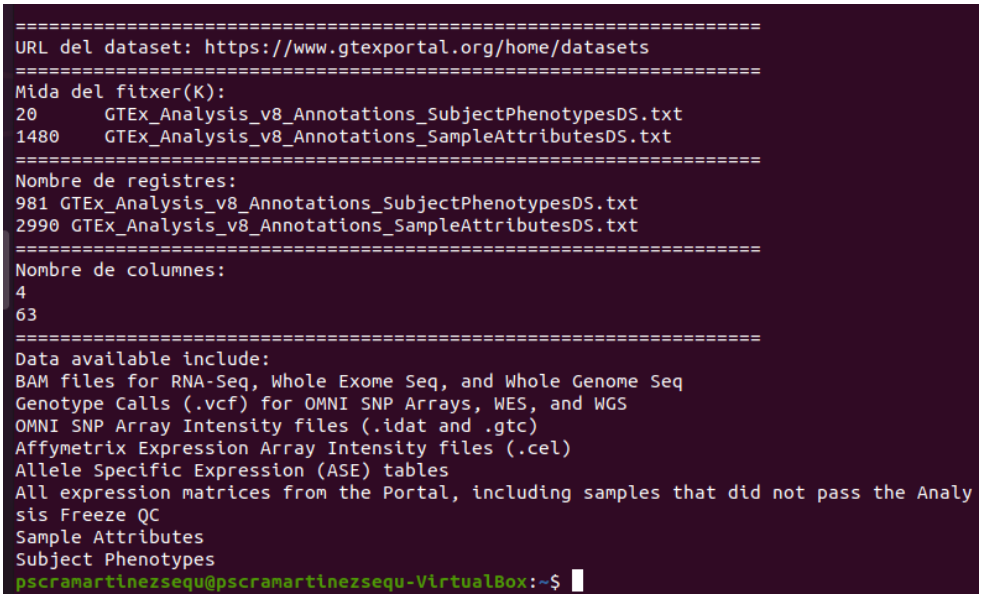
i executem ./a.sh -v, mostra, a més, el format dels fitxers:



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



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



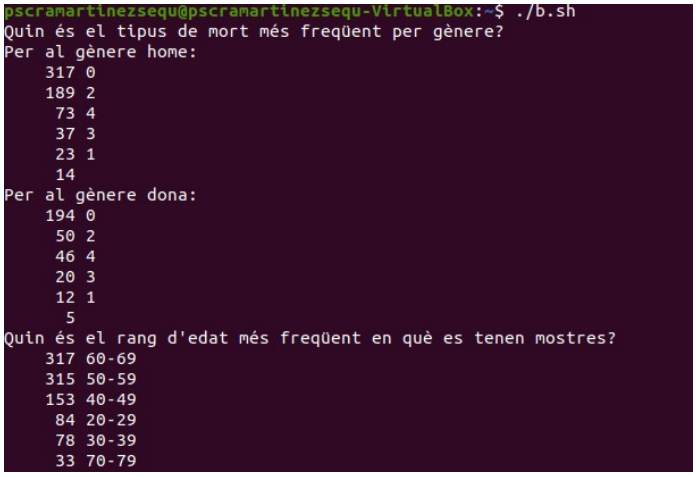
SCRIPT B

Aquest script, anomenat b.sh, pretèn 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 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

(uniq -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



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.

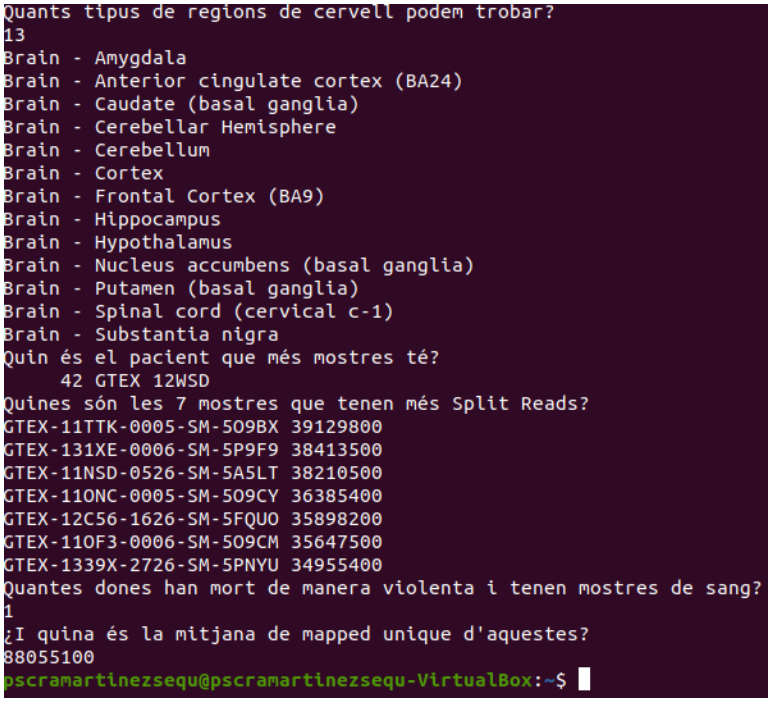
Quines són les 7 mostres que tenen més "Split Reads"?
El nombre de "Split Reads" es troba a la columna 55 de l'arxiu de mostres. Mostrem les columnes 1 i 55 de l'arxiu assignant un caràcter separador de camps per a la lectura FS = "\ t", I ho ordenem per la segona columna, canviem el format del nombre 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 poguer respondre a la següent qüestió:

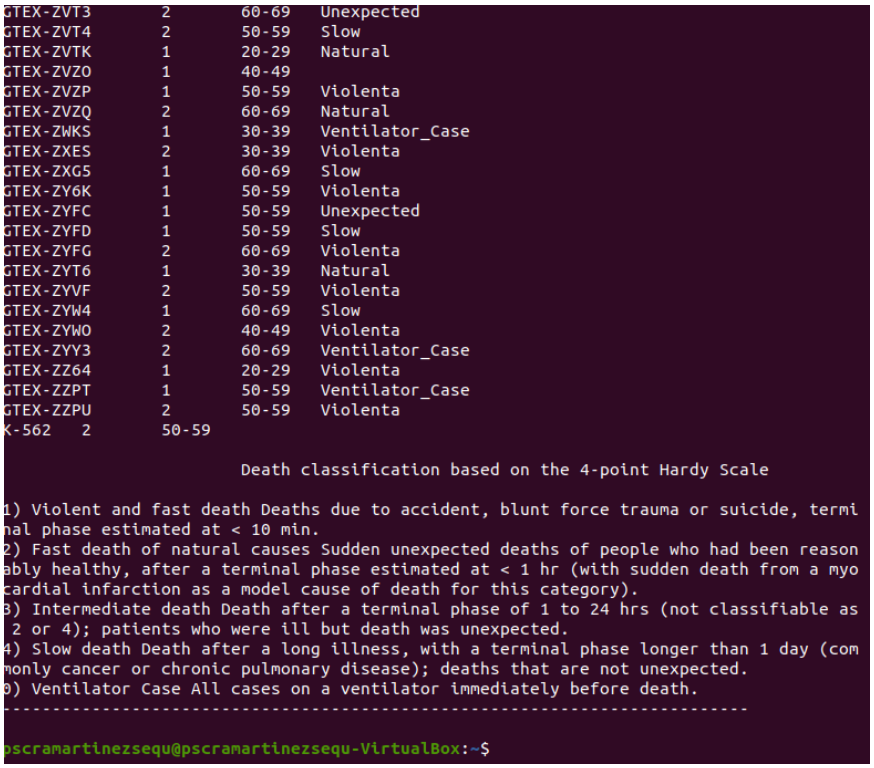
El tipus de mostra i el nombre de "mapped unique" els trobem en les columnes 6 i 48 de l'arxiu de mostres. Per obtenir les mostres de sang amb el nombre de "mapped unique", vam consultar l'arxiu de mostres i seleccionem les columnes 1, 6 i 48 i les línies que tinguin la paraula "Blood" grep "Blood" I 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}



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

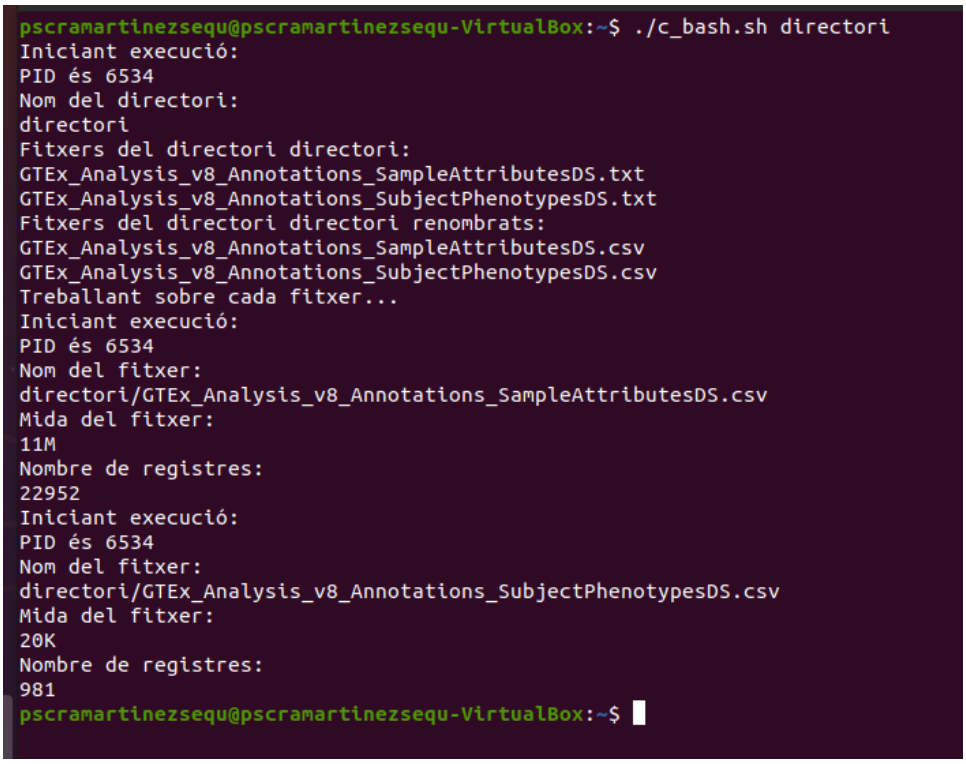


En aquesta imatge veiem el resultat de
gawk -f c_awk.awk GTEx_Analysis_v8_Annotations_SubjectPhenotypesDS.txt

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 lescategories (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)
   }
}
```



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/pscramartinezsequ/directori/GTEx_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/pscramartinezsequ/GTEx_Analysis_v8_Annotations_SubjectPhenotypesDS.txt >
/home/pscramartinezsequ/d.html
```

SCRIPT E

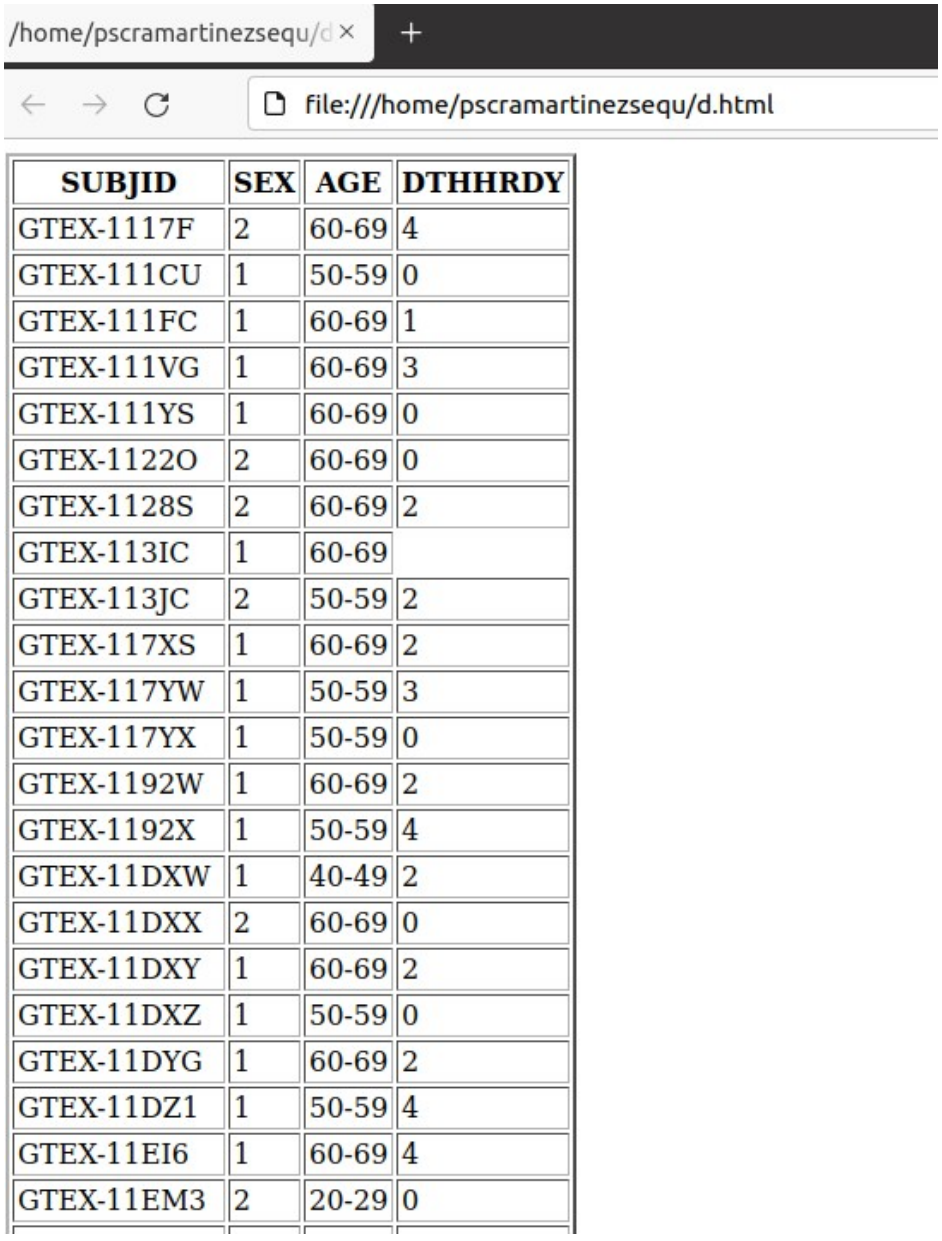
És un script principal, anomenat run.sh que executa pas a pas tot el projecte, cridant als scripts anteriors. Funciona usant la invocació: ./run.sh.

```
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 GTEx_Analysis_v8_Annotations_SubjectPhenotypesDS.txt;
echo "LLancem d.sh";
./d.sh;
```

8

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.



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