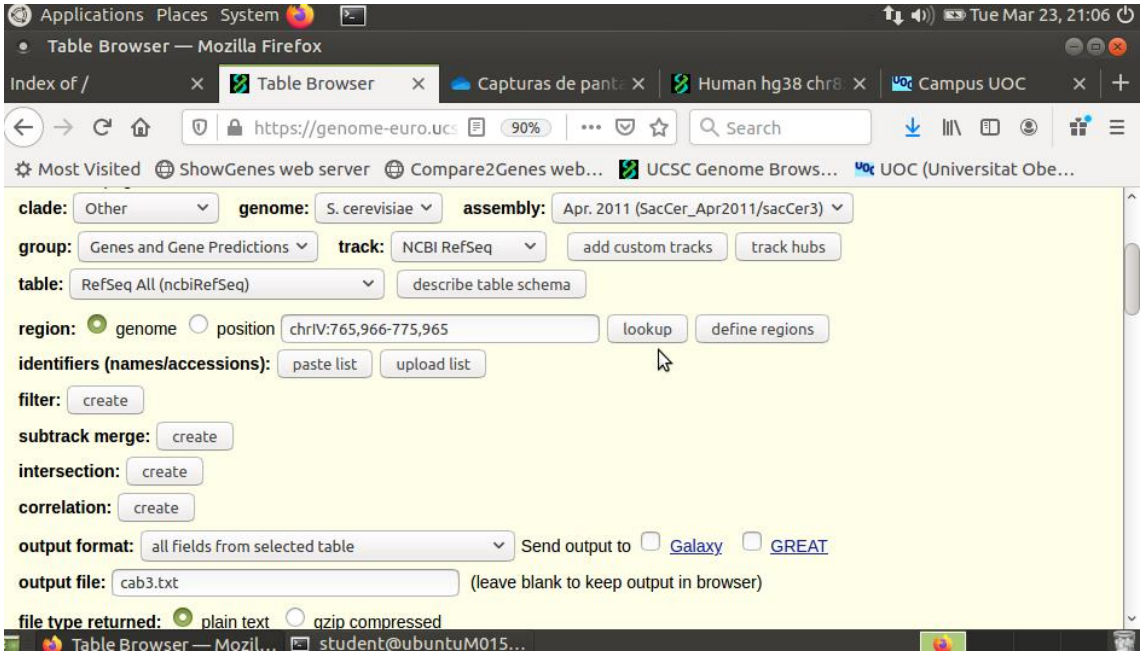
EINES INFORMÀTIQUES PER A LA BIOINFORMÀTICA

Amelia Martínez Sequera

PAC1. ANÀLISI BIOINFORMÀTIC AMB EL TERMINAL

Exercici 1. Descripció dels catàlegs de gens.

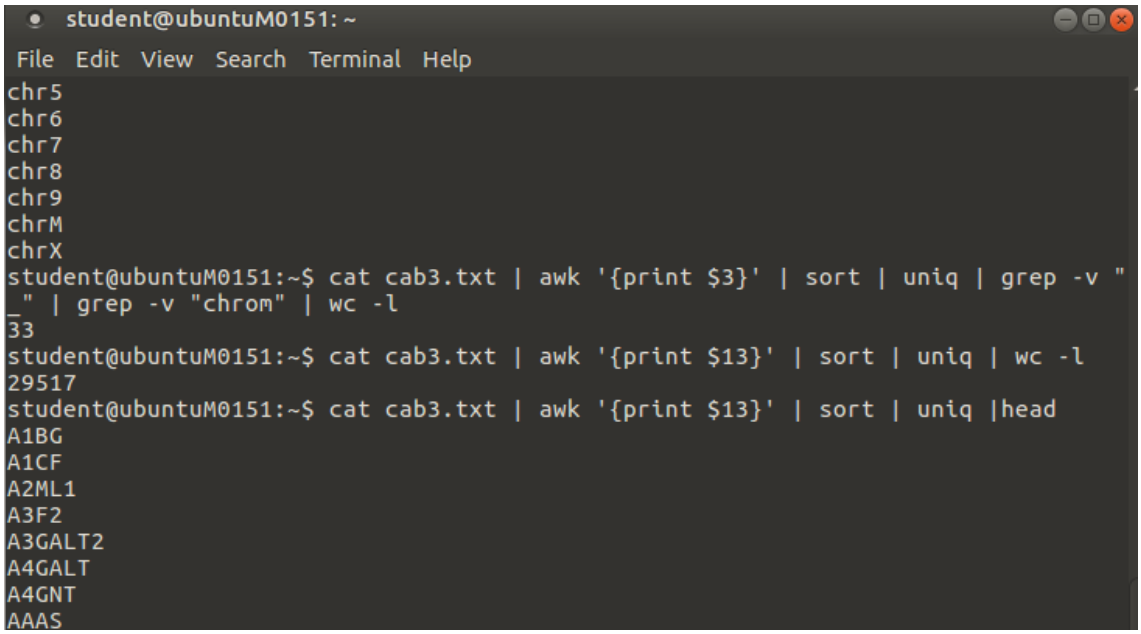
Descarreguem els arxius de USCS Table Browser (ja descomprimits):



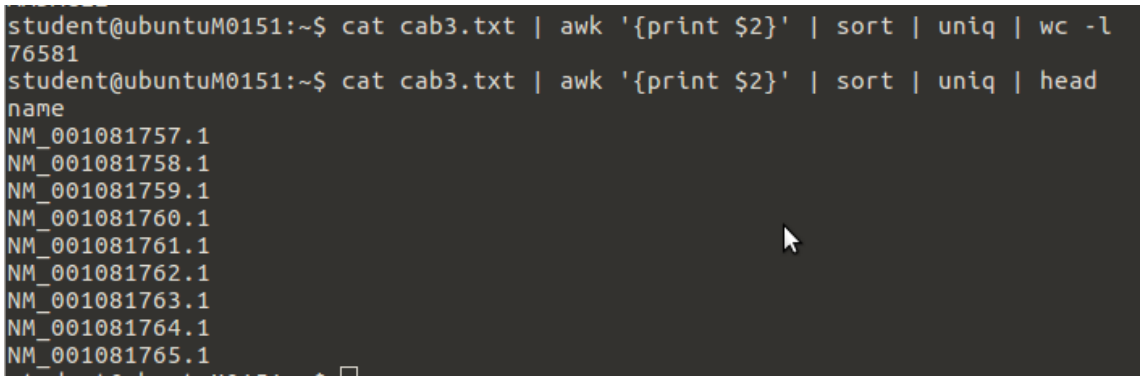
1. **Nombre de cromosomes diferents**: amb wc -l contem la columna 3, evitem les formes "_alt" i "_random", i la forma "chrom" amb grep -v.

Es mostra l´exemple pel cavall, 33 cromosomes.

2. **Nombre de gens diferents**: columna 13. Com en el cas anterior, fem servir uniq per evitar repetits. Normalment, s´obtenen millors resultats si estan ordenats (sort). A l´exemple veiem 29517 pel cavall.



3**. Nombre de trànscrits diferents**: columna 2.



4. **Nombre de trànscrits codificants:** comptem els trànscrits de la columna 2 que contenen NM, amb grep NM.

A l´exemple veiem que el genoma humà hg38 té 163989 trànscrits diferents, del quals 60847 són codificants

```
NM_000017.4
NM_000018.4
NM_000019.4
NM_000020.3
NM_000021.4
NM_000022.4
student@ubuntuM0151:~$ cat hg38.txt | awk '{print $2}' | sort | uniq | wc -l
163989
student@ubuntuM0151:~$ cat hg38.txt | awk '{print $2}' | sort | uniq | grep NM |
head
NM_000014.6
NM_000015.3
NM_000016.6
NM_000017.4
NM_000018.4
NM_000019.4
NM_000020.3
NM_000021.4
NM_000022.4
NM_000023.4
student@ubuntuM0151:~$ cat hg38.txt | awk '{print $2}' | sort | uniq | grep NM |
wc -l
60847
student@ubuntuM0151:~$
```

5. **Nombre de trànscrits no codificants**: comptem els trànscrits NR.

Pel cavall, per exemple, té 1114 codificants i 689 no codificants

```
NM_001081765.1
student@ubuntuM0151:~$ cat cab3.txt | awk '{print $2}' | sort | uniq | grep NM |
 head
NM_001081757.1
NM_001081758.1
NM_001081759.1
NM_001081760.1
NM_001081761.1
NM_001081762.1
NM_001081763.1
NM_001081764.1
NM_001081765.1
NM_001081766.1
student@ubuntuM0151:~$ cat cab3.txt | awk '{print $2}' | sort | uniq | grep NM |
 wc -l
1114
student@ubuntuM0151:~$ cat cab3.txt | awk '{print $2}' | sort | uniq | grep NR |
 wc -l
689
student@ubuntuM0151:~$
```

6. **Nombre de trànscrits per cada gen (mitjana):** la variable t fa les vegades de comptador que acumula la suma total. La divisió pel nombre de línies visitades (NR), una vegada finalitzada la lectura del fitxer, genera el valor mitjà.

uniq –c pel nombre de vegades que cada gen apareix en el fitxer.

```
NR_000014.1
student@ubuntuM0151:~$ cat hg38.txt | awk '{print $13}'| sort | uniq -c| awk 'BE
GIN {t=0}{t=t+$1}END{print t/NR}'
4.4814
student@ubuntuM0151:~$ cat hg38.txt | awk '{print $13}'| sort | uniq -c| head
      1 A1BG
      1 A1BG-AS1
     12 A1CF
      5 A2M
      3 A2M-AS1
      9 A2ML1
      1 A2MP1
      1 A3GALT2
```

7. **Nombre d'exons per trànscrit (mitjana):**

```
student@ubuntuM0151:~$ cat hg38.txt | awk '{print $9}'| awk 'BEGIN {t=0}{t=t+$1}
END{print t/NR}'
11.5968
```

8. **Nombre de nucleòtids per trànscrit (mitjana):** columnes 5 i 6

```
student@ubuntuM0151:~$ cat hg38.txt | awk '{print $6-$5}'| awk 'BEGIN {t=0}{t=t+
$1}END{print t/NR}'
76709.7
student@ubuntuM0151:~$
```

Obtenim el resultat:

Genoma	1	2	3	4	5	6	7	8
H. sapiens (hg38)	25	38552	163989	60847	18297	4.48	11.6	76709.7
E. caballus (EcuCab3.0)	33	29517	76581	1114	689	2.6	12.23	66576.4
G. gallus (GRCg6a)	34	23727	62161	6401	776	2.62	12.8	47616.7
S. cerevisiae (sacCer3)	17	6126	6126	5983	123	1	1.06	1466.7

Conclusió: com més complexe és l´organisme, més gens i més trànscrits. L´home té molts més trànscrits, però també té una proporción molt alta de no codificants, com a conseqüència evolutiva. Probablement tenen funcions reguladores.

Exercici 2. Dades d´expressió.

Adjunt amb l'enunciat d'aquesta PAC podreu trobar dos arxius txt del projecte internacional GTEx (https://gtexportal.org/home/). L'objectiu principal d'aquest projecte és construir un repositori d'expressió gènica teixit específic. Mitjançant comandes en bash hauríeu de respondre a les següents preguntes: 1.- Quin és el rang d'edat més freqüent en el que es tenen mostres? 2.- Quin és el tipus de mostra més freqüent per gènere? 3.- Quants tipus de regions del cervell (brain) podem trobar? 4.- Quin és el pacient que més mostres té? 5.- Quines són les 7 mostres que tenen més "Split Reads"? 6.- Mitjançant la comanda "join" respondre a, quantes dones han mort de manera violenta i tenen mostres de sang? i quina és la mitjana de "mapped unique" d'aquesta selecció?

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

El rang d'edat dels pacients correspòn a la tercera columna de Phenotypes.txt. Les contem sense repeticions, ordenades perquè s´obtenen millors resultats. Fem servir -n per considerar les dades com a numèriques, i -r per ordenarles en ordre descendent.

Fem servir GAWK que és la versió GNU de AWK.

```
student@ubuntuM0151:~$  gawk '{print $3}' Phenotypes.txt | sort | uniq -c | sort
 -n -r
    317 60-69
    315 50-59
    153 40-49
     84 20-29
     78 30-39
     33 70-79
      1 AGE
student@ubuntuM0151:~$
```

El rang més freqüent és el de 60-69.

2.- Quin és el tipus de mort més freqüent per gènere?

El gènere correspòn a la columna 2: 1 per homes i 2 per dones. Fem servir el condicional de la columna 2, per obtenir la columna 4, que és la causa de mort.

```
student@ubuntuM0151:~$ gawk '{if ($2 == "1") print $4;}' Phenotypes.txt |sort |u
niq -c |sort -n -r
    317 0
    189 2
     73 4
     37 3
     23 1
     14
student@ubuntuM0151:~$ gawk '{if ($2 == "2") print $4;}' Phenotypes.txt |sort |u
niq -c |sort -n -r
    194 0
     50 2
     46 4
     20 3
     12 1
      5
student@ubuntuM0151:~$
```

La mort més freqüent per a ambdós gèneres és la 0 = Ventilator Case

3.- Quants tipus de regions de cervell (brain) podem trobar?

Hem de fer servir {FS = "\ t"} per indicar el carácter tabulador.

Seleccionem les línies amb la paraula "Brain" de la columna 7.

```
student@ubuntuM0151:~$ gawk 'BEGIN {FS="\t"}{print $7}' Samples.txt | grep "Brai
n*" |sort |uniq | wc -l
13
student@ubuntuM0151:~$ gawk 'BEGIN {FS="\t"}{print $7}' Samples.txt |grep "Brain
*" |sort |uniq
Brain - Amygdala
Brain - Anterior cingulate cortex (BA24)
Brain - Caudate (basal ganglia)
Brain - Cerebellar Hemisphere
Brain - Cerebellum
Brain - Cortex
Brain - Frontal Cortex (BA9)
Brain - Hippocampus
Brain - Hypothalamus
Brain - Nucleus accumbens (basal ganglia)
Brain - Putamen (basal ganglia)
Brain - Spinal cord (cervical c-1)
Brain - Substantia nigra
student@ubuntuM0151:~$
```

El resultat indica 13 tipus de regions de cervell.

4.- Quin és el pacient que més mostres té?

La primera columna de Samples.txt conté l'identificador de mostres, que es separa amb un guió. Substituïm el guió per un caràcter tabulador: gsub (/ - /," \ t ", $ 1), i obtenim els apartats de l'identificador com columnes diferents. Les dues primeres columnes resultants seran els identificadors de pacients.

```
Brain - Substantia nigra
student@ubuntuM0151:~$ gawk 'BEGIN {FS="\t"}{print $1}' Samples.txt | gawk '{gsu
b(/-/, "\t", $1); print}' | gawk '{print $1,$2}' | sort |uniq -c |sort -n -r | h
ead
    217 K 562
     72 GTEX NPJ8
     59 GTEX RU72
     58 GTEX Q2AG
     56 GTEX N7MS
     51 GTEX YEC3
     50 GTEX RNOR
     50 GTEX N7MT
     49 GTEX T5JC
     49 GTEX QDT8
```

Com abans, contem i ordenem numèricament en ordre descendent. El pacient amb més mostres és el K-562 amb 217 mostres.

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

Els "Split Reads" corresponen a la columna 55 de Samples.txt.

Ordenem per ordre descendent la segona columna, utilitzant l'opció -g (g de general) per tractar amb nombres exponencials, no enters.

Podem canviar el format del nombre exponencial a nombre enter: printf "% s % d \n", i ordenem pel nombre enter de major a menor | sort -k2nr | head -7. Obtenim el mateix resultat però mostrant el nombre de "Split reads" com a número enter.

Fem

```
student@ubuntuM0151:~$ gawk 'BEGIN {FS="\t"}{print $1,$55}' Samples.txt | sort -
k2,2gr | head -7
GTEX-14BMU-1526-SM-5TDE6 9.92286e+07
GTEX-14JFF-0526-SM-62LFL 8.17943e+07
GTEX-QMRM-0426-SM-4R1K2 8.06541e+07
GTEX-1JMQI-2026-SM-CMKGP 6.85681e+07
GTEX-1JJE9-0006-SM-CGQEB 6.40318e+07
GTEX-13G51-0011-R8b-SM-5LZZ4 6.07486e+07
GTEX-18D9B-0008-SM-EAZBT 5.7394e+07
student@ubuntuM0151:~$ gawk 'BEGIN {FS="\t"}{printf "%s %d\n", $1, $55}' Samples
.txt | sort -k2nr | head -7
GTEX-14BMU-1526-SM-5TDE6 99228600
GTEX-14JFF-0526-SM-62LFL 81794300
GTEX-QMRM-0426-SM-4R1K2 80654100
GTEX-1JMQI-2026-SM-CMKGP 68568100
GTEX-1JJE9-0006-SM-CGQEB 64031800
GTEX-13G51-0011-R8b-SM-5LZZ4 60748600
GTEX-18D9B-0008-SM-EAZBT 57394000
```

6.- Mitjançant la comanda "join" respondre a, quantes dones han mort de manera violenta i tenen mostres de sang? ¿I quina és la mitjana de "mapped unique" d'aquesta selecció?

Fem servir el condicional $2==”2” de Phenotypes.txt per indicar que han de ser dones, i ens quedem amb les columnes 1 i 4, identificador i tipus de mort respectivament.

Després apliquem el condicional ($2 == "1"), que correspòn a tipus de mort violenta, ara a la columna 2.

Com hem fet abans, substituïm el guió per un caràcter tabulador a la columna1.

El resultat elguardem en un altre arxiu, dones1 .txt .

```
gawk '{if ($2 == "2") print $1, $4;}' Phenotypes.txt |gawk '{if ($2 == "1") print $1;}' | gawk
'{gsub(/-/, "\t", $1);print}'|sort > dones1.txt
```

El tipus de mostra i el nombre de "mapped unique" corresponen a les columnes 6 i 48 de l'arxiu Samples.txt. Seleccionem també la columna 1, per tenir l´identificador comú per fer el join. S´han de seleccionar els registres que contenen la paraula “Blood” i descartar els que contenen “Vessel”. Guardem el resultat en un arxiu: mostra.txt.

```
gawk 'BEGIN {FS="\t"}{print $1,$6,$48}' Samples.txt | grep "Blood" | grep -v "Vessel" | gawk
'{gsub(/-/, "\t", $1);print}'|sort > mostra.txt
```

Fem el join, amb la columna que tenen els dos arxius en comú, l´identificador:

join -1 2 -2 2 dones1.txt mostra.txt | gawk '{ print $1 }' | sort | uniq | wc -l

El resultar és 11 dones.

Per calcular la mitjana dels "mapped unique", a la columna 8, sumem i dividim pel total de linies. Es descarten les linies que no tenen dades.

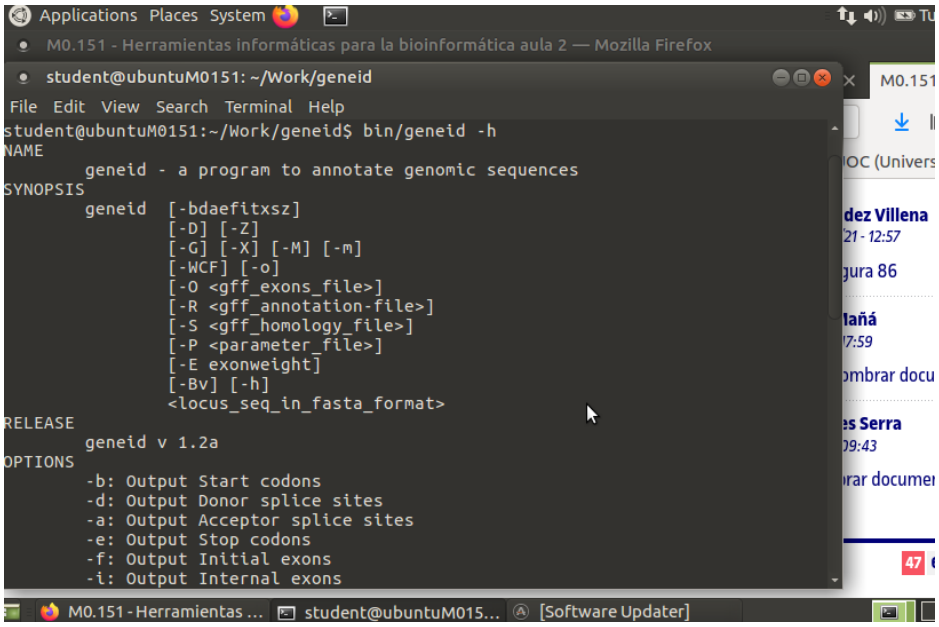
La mitjana és de 9.25807e+07

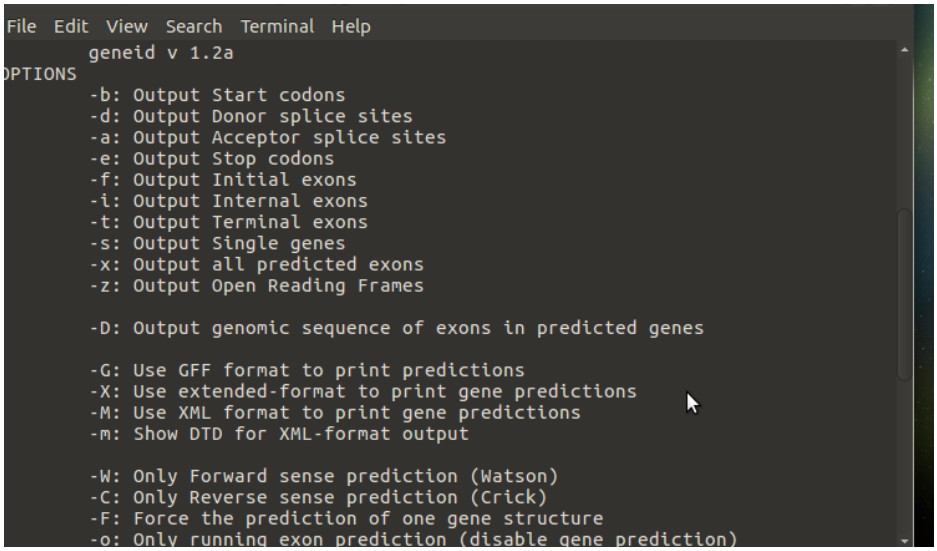
```
student@ubuntuM0151:~$ gawk 'BEGIN {FS="\t"}{print $1,$6,$48}' Samples.txt | gre
p "Blood" | grep -v "Vessel" | gawk '{gsub(/-/, "\t", $1);print}'|sort > mostra.
txt
student@ubuntuM0151:~$ join -1 2 -2 2 dones1.txt mostra.txt | gawk '{ print $1 }
' | sort | uniq | wc -l
11
student@ubuntuM0151:~$ join -1 2 -2 2 dones1.txt mostra.txt | gawk '{if ($8 != "
") print $8;}' |sort | gawk 'BEGIN {t=0}{t=t+$1}END {print t/NR}'
9.25807e+07
student@ubuntuM0151:~$
```

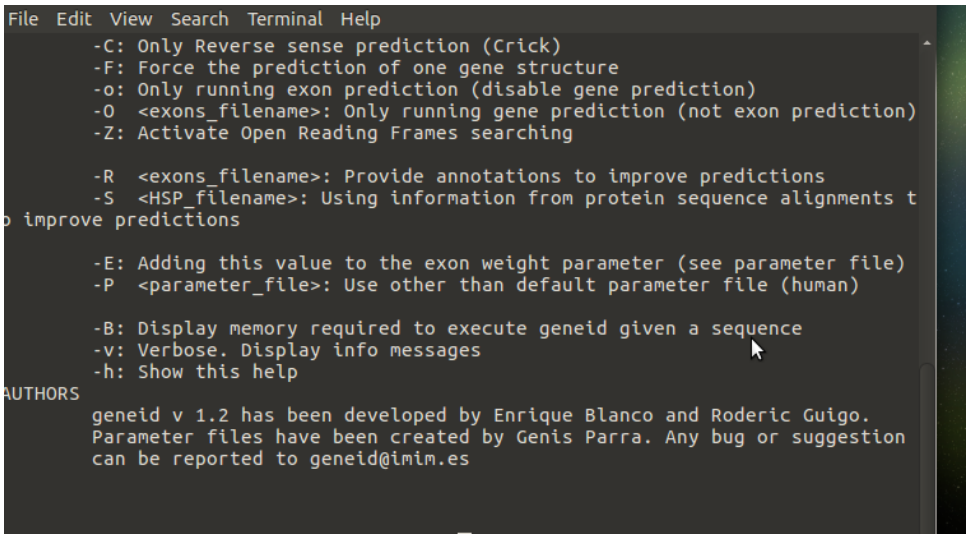
Exercici 3. Predicció computacional de gens.

Un col·laborador ens comenta que la seqüència test.fa (inclosa en aquest enunciat) conté un gen humà. Utilitzarem el programa GENEID en el vostre terminal per identificar una possible estructura exònica i processarem aquesta predicció per visualitzar-la en el navegador genòmic de UCSC.

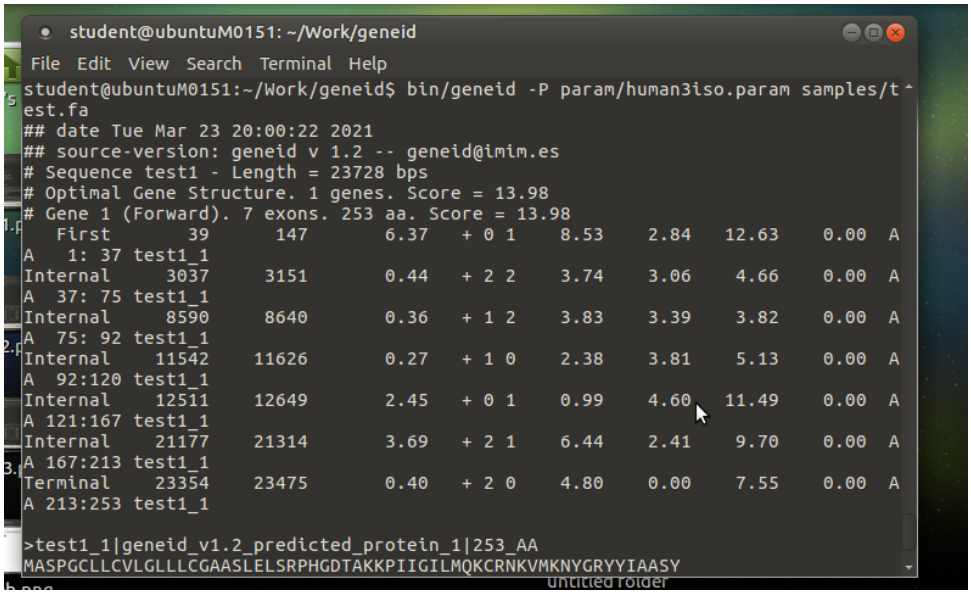
Primer mirem les opcions de geneid amb -h:

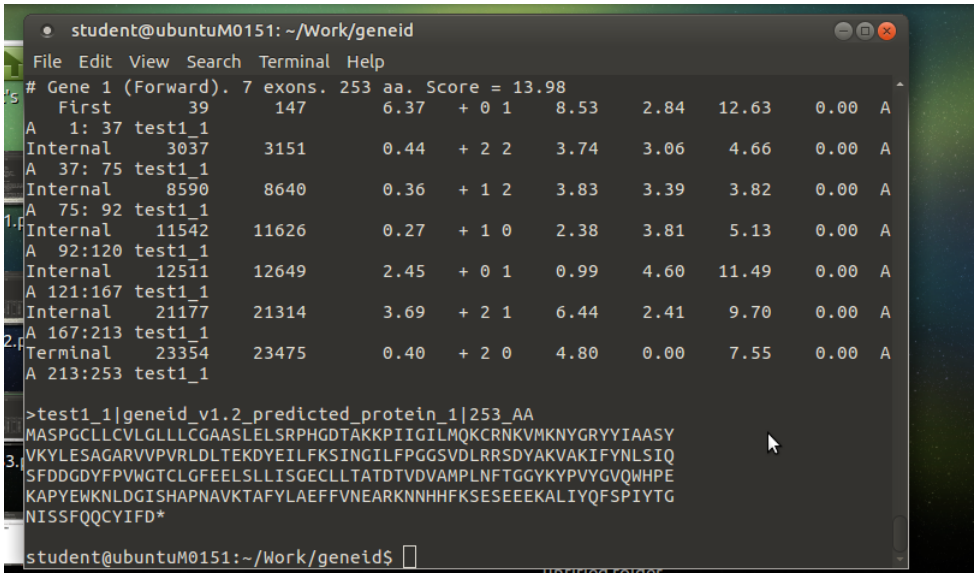

Applications Places System
M0.151 - Herramientas informáticas para la bioinformática aula 2 — Mozilla Firefox
student@ubuntuM0151: ~/Work/geneid
File Edit View Search Terminal Help
student@ubuntuM0151:~/Work/geneid$ bin/geneid -h
NAME
geneid - a program to annotate genomic sequences
SYNOPSIS
geneid [-bdaefitxsz]
[-D] [-Z]
[-G] [-X] [-M] [-m]
[-WCF] [-o]
[-O <gff_exons_file>]
[-R <gff_annotation-file>]
[-S <gff_homology_file>]
[-P <parameter_file>]
[-E exonweight]
[-Bv] [-h]
<locus_seq_in_fasta_format>
RELEASE
geneid v 1.2a
OPTIONS
-b: Output Start codons
-d: Output Donor splice sites
-a: Output Acceptor splice sites
-e: Output Stop codons
-f: Output Initial exons
-i: Output Internal exons
M0.151 - Herramientas ...
student@ubuntuM015...
[Software Updater]


File Edit View Search Terminal Help
geneid v 1.2a
OPTIONS
-b: Output Start codons
-d: Output Donor splice sites
-a: Output Acceptor splice sites
-e: Output Stop codons
-f: Output Initial exons
-i: Output Internal exons
-t: Output Terminal exons
-s: Output Single genes
-x: Output all predicted exons
-z: Output Open Reading Frames
-D: Output genomic sequence of exons in predicted genes
-G: Use GFF format to print predictions
-X: Use extended-format to print gene predictions
-M: Use XML format to print gene predictions
-m: Show DTD for XML-format output
-W: Only Forward sense prediction (Watson)
-C: Only Reverse sense prediction (Crick)
-F: Force the prediction of one gene structure
-o: Only running exon prediction (disable gene prediction)


File Edit View Search Terminal Help
-C: Only Reverse sense prediction (Crick)
-F: Force the prediction of one gene structure
-o: Only running exon prediction (disable gene prediction)
-O <exons_filename>: Only running gene prediction (not exon prediction)
-Z: Activate Open Reading Frames searching
-R <exons_filename>: Provide annotations to improve predictions
-S <HSP_filename>: Using information from protein sequence alignments t
o improve predictions
-E: Adding this value to the exon weight parameter (see parameter file)
-P <parameter_file>: Use other than default parameter file (human)
-B: Display memory required to execute geneid given a sequence
-v: Verbose. Display info messages
-h: Show this help
AUTHORS
geneid v 1.2 has been developed by Enrique Blanco and Roderic Guigo.
Parameter files have been created by Genis Parra. Any bug or suggestion
can be reported to geneid@imim.es

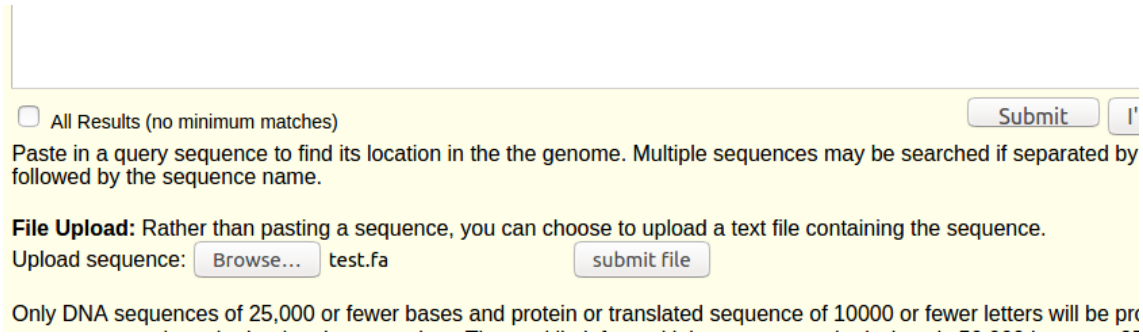
Predicció des del terminal:

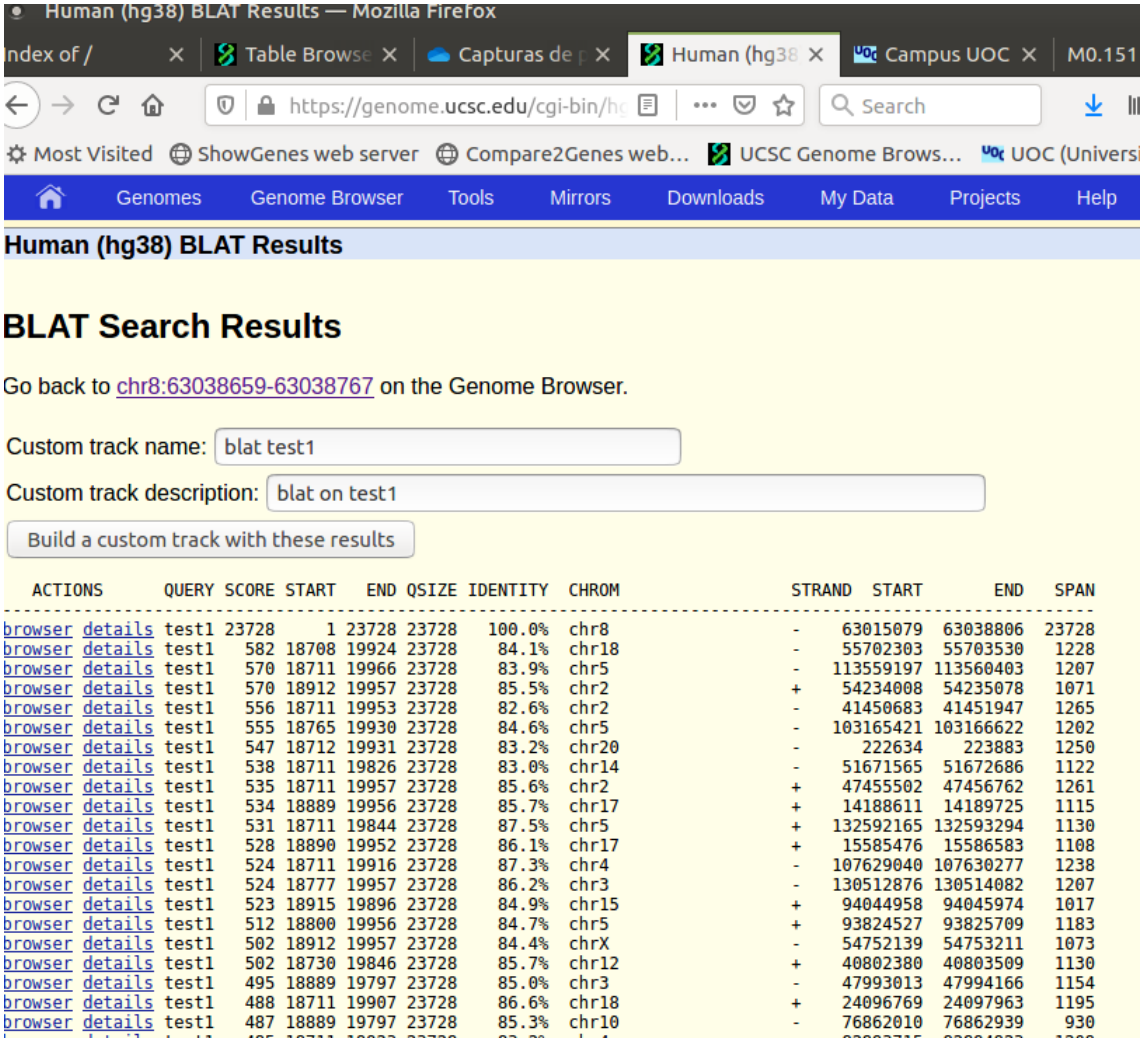




La seqüència conté un únic gen en el bri + (Forward), i aquest gen té un exó inicial (First), 5 exons interns (Internal), i un terminal.

Per visualitzar la predicció a UCSC hem d'esbrinar primer en quina regió del genoma està situada la seqüència completa, i després, adaptar les coordenades de la predicció a aquesta localització. Emprant l'eina BLAT del navegador genòmic UCSC, podem saber la localització en el genoma d'aquesta seqüència (hg38):



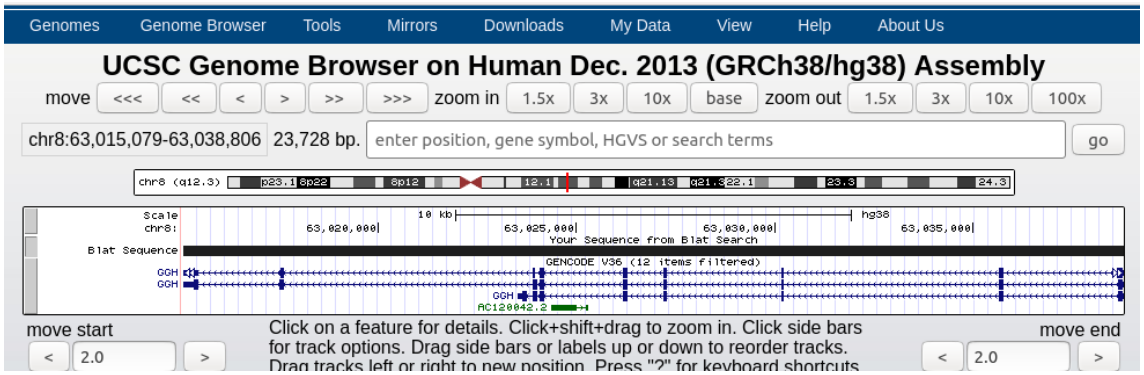


ACTIONS	QUERY	SCORE	START	END	QSIZE	IDENTITY	CHROM	STRAND	START	END	SPAN
browser details	test1	23728	1	23728	23728	100.0%	chr8	-	63015079	63038806	23728
browser details	test1	582	18708	19924	23728	84.1%	chr18	-	55702303	55703530	1228
browser details	test1	570	18711	19966	23728	83.9%	chr5	-	113559197	113560403	1207
browser details	test1	570	18912	19957	23728	85.5%	chr2	+	54234008	54235078	1071
browser details	test1	556	18711	19953	23728	82.6%	chr2	-	41450683	41451947	1265
browser details	test1	555	18765	19930	23728	84.6%	chr5	-	103165421	103166622	1202
browser details	test1	547	18712	19931	23728	83.2%	chr20	-	222634	223883	1250
browser details	test1	538	18711	19826	23728	83.0%	chr14	-	51671565	51672686	1122
browser details	test1	535	18711	19957	23728	85.6%	chr2	+	47455502	47456762	1261
browser details	test1	534	18889	19956	23728	85.7%	chr17	+	14188611	14189725	1115
browser details	test1	531	18711	19844	23728	87.5%	chr5	+	132592165	132593294	1130
browser details	test1	528	18890	19952	23728	86.1%	chr17	+	15585476	15586583	1108
browser details	test1	524	18711	19916	23728	87.3%	chr4	-	107629040	107630277	1238
browser details	test1	524	18777	19957	23728	86.2%	chr3	-	130512876	130514082	1207
browser details	test1	523	18915	19896	23728	84.9%	chr15	+	94044958	94045974	1017
browser details	test1	512	18800	19956	23728	84.7%	chr5	+	93824527	93825709	1183
browser details	test1	502	18912	19957	23728	84.4%	chrX	-	54752139	54753211	1073
browser details	test1	502	18730	19846	23728	85.7%	chr12	+	40802380	40803509	1130
browser details	test1	495	18889	19797	23728	85.0%	chr3	-	47993013	47994166	1154
browser details	test1	488	18711	19907	23728	86.6%	chr18	+	24096769	24097963	1195
browser details	test1	487	18889	19797	23728	85.3%	chr10	-	76862010	76862939	930

El primer hit és el més probable (inclou tota la seqüència, IDENTITY: 100.0%), a més té el SCORE més elevat. Aquest gen es troba situat entre les posicions 63015079 i 63038806 del cromosoma 8, en el bri - , a diferencia de la predicció de geneid.

Si es pitja sobre l'opció **browser** es pot observar que la seqüència problema (BLAT Sequence) coincideix amb el gen GGH.

Fonamentalment, tota la finestra gràfica comprèn la nostra seqüència de treball i la informació es mostra en forma de pistes, on els exons són representats amb caixes. La diferent grossària de les caixes indica si un exó pertany a una regió que codifica una proteïna o no.

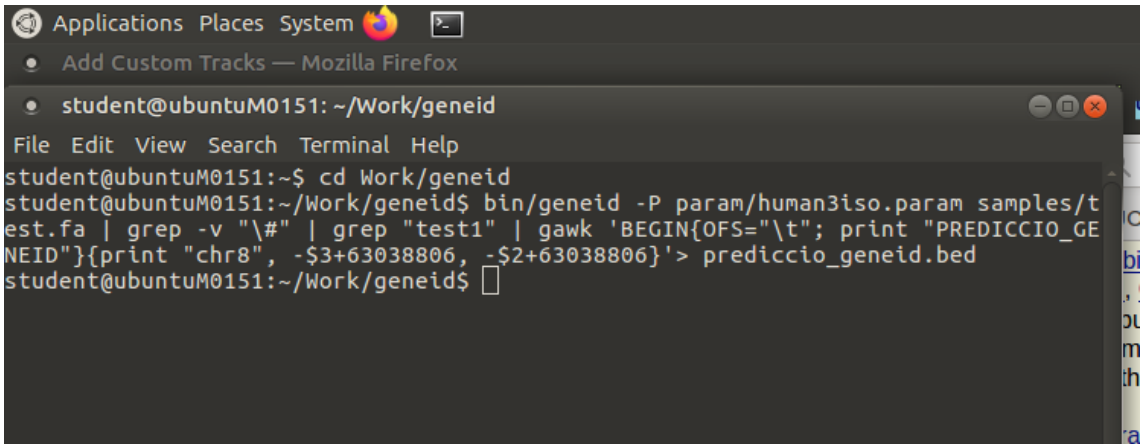


El nostre objectiu serà, per tant, traslladar les coordenades de les prediccions del programa geneid (cada línia és un exó situat de forma relativa dins de la seqüència) cap al sistema de coordenades del navegador (relatiu a la regió del cromosoma que veiem mitjançant BLAT). Heu de modificar el format de sortida dels exons predits per geneid en la seqüència test.fa per incloure'ls com una pista pròpia dins del navegador. En concret, heu de convertir mitjançant comandes GAWK del terminal la sortida de geneid en un nou fitxer de text tabulat, que podreu visualitzar directament en el navegador amb l'opció "Add custom tracks" (situat sota la pantalla gràfica de UCSC). El fitxer de text tabulat que heu d'aconseguir per carregar-ho en el navegador de UCSC ha de respectar el format BED.

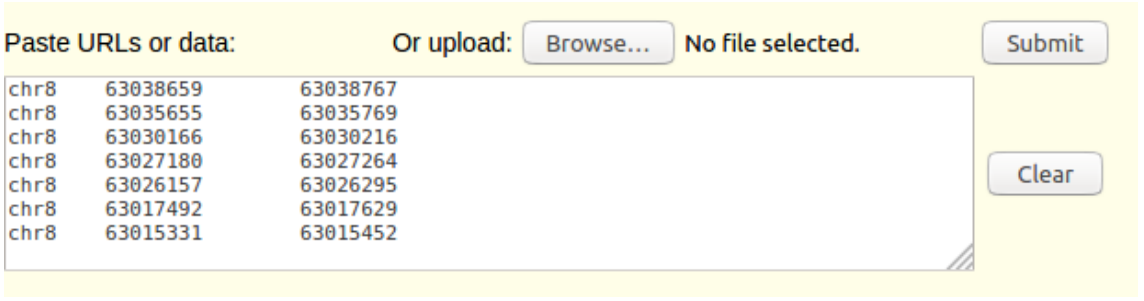
Amb l´opció grep treiem les linies que contenen exons ("test1"), es canvia el signe, i per tant la posición final, 63038806, s´ha de considerar com l´inici:

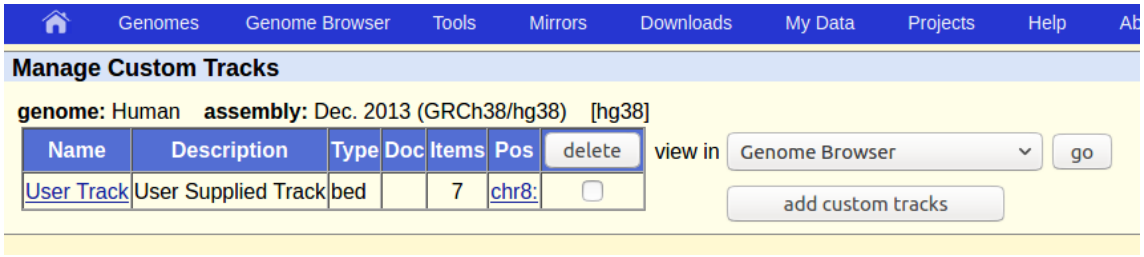
bin/geneid -P param/human3iso.param samples/test.fa | grep -v "\#" | grep "test1" | gawk 'BEGIN{OFS="\t"; print "PREDICCIO_GENEID"}{print "chr8", -$3+63038806, -$2+63038806}'> prediccio_geneid.bed

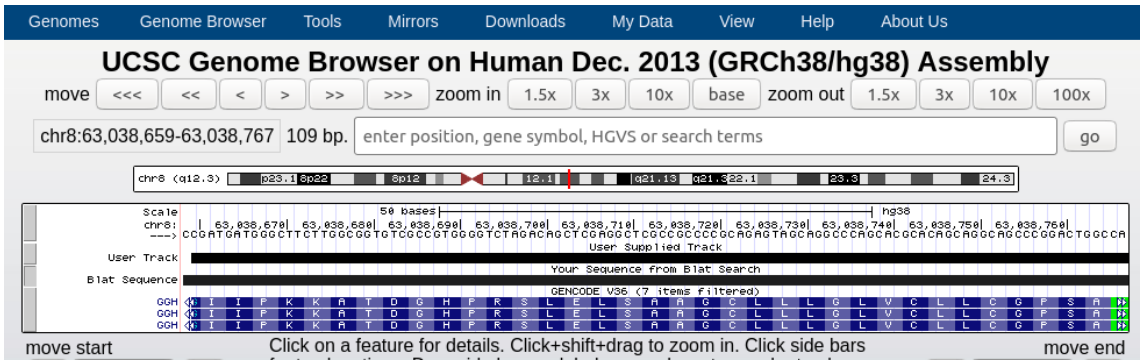
```
student@ubuntuM0151:~/Work/geneid$ bin/geneid -P param/human3iso.param samples/t
est.fa | grep -v "\#" | grep "test1" | gawk 'BEGIN{OFS="\t"; print "PREDICCIO_GE
NEID"}{print "chr8", -$3+63038806, -$2+63038806}'
PREDICCIO_GENEID
chr8    63038659        63038767
chr8    63035655        63035769
chr8    63030166        63030216
chr8    63027180        63027264
chr8    63026157        63026295
chr8    63017492        63017629
chr8    63015331        63015452
chr8    63038806        63038806
student@ubuntuM0151:~/Work/geneid$
```



Els resultats es guarden a **prediccio_geneid.bed**, i els copiem al navegador d´UCSC:







Geneid no encerta el bri, però si 7 dels 9 exons.