

## PAC2. Anàlisi bioinformàtic amb la terminal

Magí Bas

2023-04-23

### Exercici 1 – Descripció dels catàlegs de gens (15%)

Anàlisi comparativa de les col·leccions de gens de diverses espècies. De la mateixa manera que heu vist durant els exercicis pràctics, connecteu-vos al servidor UCSC per accedir als fitxers refSeq.txt de diverses espècies. L'objectiu és que empleneu la següent Taula amb les dades que obtindreu fent servir les comandes apropiades en el vostre terminal sobre els catàlegs de gens per a aquesta versió dels genomes. Us hem anotat dues espècies. **Afegiu una petita interpretació biològica dels resultats anotats (1 punt):**

Genoma	Chr	Gens	Tr	TrC	TrNC	Tr/Gen	Ex/Tr	nuc/Tr
H. sapiens (hg38)								
Zebrafish (danRer11)								
S. cerevisiae (sacCer3)	17	6125	6125	5983	123	1,00	1,058	1467,7
D. melanogaster (dm6)	8	17202	3446	3070	374	2,00	5,4687	9940,3
			3	4	6			

Importem els archius rerSeq des de UCSC:

## Table Browser

Use this tool to retrieve and export data from the Genome Browser annotation track database. You can limit retrieval based on data attributes and intersect or merge with data from another track, or retrieve DNA sequence covered by a track. [More...](#)

### Select dataset

**clade:**

Vertebrate

**genome:**

Zebrafish

**assembly:**

May 2017 (GRCz11/danRer11)

**group:**

Genes and Gene Predictions

**track:**

NCBI RefSeq

**table:**

RefSeq All (ncbiRefSeq)

data format description

### Define region of interest

**region:** ☒ genome ☐ position 

chr6:43,426,669-43,433,274

lookup

define regions

**identifiers (names/accessions):**

paste list

upload list

### Optional: Subset, combine, compare with another track

**filter:**

create

**subtrack merge:**

create

**intersection:**

create

**correlation:**

create

### Retrieve and display data

**output format:**

all fields from selected table

 Send output to ☐ [Galaxy](#) ☐ [GREAT](#)

**output filename:**

danRer\_11.txt

 (add .csv extension if opening in Excel, leave blank to keep output in browser)

**output field separator:** ☒ tsv (tab-separated) ☐ csv (for excel)

**file type returned:** ☒ plain text ☐ gzip compressed

get output

summary/statistics

```
# Donem permís als arxius
$ chmod 777 hg_38.txt
$ chmod 777 danRer_11.txt
```

#bin	name	chrom	strand	txStart	txEnd	cdsStart	cdsEnd	exonCount	exonStart	exonEnd	score	name2	cdsStart	cdsEnd	exonFrame
0	XM_011541469.2	chr1	-	67092164	67109072	67093004	67103382	2	67092164	67095624	67095623	C1orf141	67095624	67109272	-1
0	XM_017001276.2	chr1	-	67092164	67131227	67093004	67127240	9	67092164	67095624	67095623	C1orf141	67095624	67115346	-1
0	XM_011541467.2	chr1	-	67092164	67131227	67093004	67127240	9	67092164	67095624	67095623	C1orf141	67095624	67115346	-1
0	NM_001276352.2	chr1	-	67092164	67134970	67093579	67127240	9	67092164	67095624	67095623	C1orf141	67095624	67125416	-1
0	NM_001276351.2	chr1	-	67092164	67134970	67093004	67127240	8	67092164	67095234	67096251	C1orf141	67115351	671257	-1

```
# Modifiquem la taula perquè ens quedi amb les columnes que ens interessen
$ gawk 'BEGIN {FS="\t"; OFS=","} {print $13 "\t" $1 "\t" $2 "\t" $3 "\t" $4 "\t" $5 "\t" $6"\t"}' hg_38.txt > hg38.txt
```

name2	#bin	name	chrom	strand	txStart	txEnd
C1orf141	0	XM_011541469.2	chr1	-	67092164	67109072
C1orf141	0	XM_017001276.2	chr1	-	67092164	67131227
C1orf141	0	XM_011541467.2	chr1	-	67092164	67131227
C1orf141	0	NM_001276352.2	chr1	-	67092164	67134970

name2	#bin	name	chrom	strand	txStart	txEnd
C1orf141	0	NM_001276351.2	chr1	-	67092164	67134970
C1orf141	0	XM_011541465.3	chr1	-	67092164	67134970
C1orf141	0	XM_011541466.3	chr1	-	67092164	67141646
C1orf141	0	NR_075077.2	chr1	-	67092164	67134970
C1orf141	0	XM_047420474.1	chr1	-	67096250	67131227
name2	#bin	name	chrom	strand	txStart	txEnd

```
# Eliminem les capçaleres
$ sed -i "1d" hg38modif1.txt
```

```
$ head -5 hg38modif1.txt
C1orf141 0 XM_011541469.2 chr1 - 67092164 67109072
C1orf141 0 XM_017001276.2 chr1 - 67092164 67131227
C1orf141 0 XM_011541467.2 chr1 - 67092164 67131227
C1orf141 0 NM_001276352.2 chr1 - 67092164 67134970
C1orf141 0 NM_001276351.2 chr1 - 67092164 67134970
```

```
$ gawk '{print $4}' hg38modif1.txt |sort| uniq | grep -v "_" | wc -l
24
# Porque se tiene en cuenta el comosoma mitocondrial
```

Ens surt un total de 24 parells de cromosomes, però nosaltres sabem que en H. sapiens el número de cromosomes es 22 més el sexual (xy)/(xx). El resultat és de 24 perquè estem comptant també el cromosoma mitocondrial que també està en la base de dades.

```
$ gawk '{print $4}' danRer11_modify.txt |sort| uniq | grep -v "_" | wc -l
26
# Cuenta también el cromosoma Me
```

Aquí passa el mateix. El peix zebra té 24 + 1 parells de cromosomes. I 26 si contem també el mitocondrial.

```
# n° de gens
gawk '{print $1}' hg38modif1.txt |sort| uniq | wc -l
42776

$ gawk '{print $1}' danRer11_modify.txt |sort| uniq | wc -l
30419
```

Veiem com el número de gens es més gran en H. sapiens però sabem que no correlaciona segons la complexitat de l'organisme.

```
# Transcrits
$ gawk '{print $4}' hg38modif1.txt |sort| uniq | wc -l
521
$ gawk '{print $4}' danRer11_modify.txt |sort| uniq | wc -l
1451
```

Com podem veure un número més gran de gens no repercuteix sempre en un número més àmpli de transcrits.

```
# n° transcripts codificants
$ gawk '{print $3}' danRer11_modify.txt |sort| uniq | grep "NM" |wc -l
15393
$ gawk '{print $3}' hg38modif1.txt |sort| uniq | grep "NM" |wc -l
66826
```

No tots els transcrits codifiquen per a una proteïna,

```
# n° transcripts no-codificants
$ gawk '{print $3}' hg38modif1.txt |sort| uniq | grep "NR" |wc -l
20584
$ gawk '{print $3}' danRer11_modify.txt |sort| uniq | grep "NR" |wc -l
476
```

```
# existen campos sin "NR" ni "NM"
$ gawk '{print $3}' hg38modif1.txt |sort| uniq | grep -v "NM" |wc -l
117663
```

```
# Tr/Gen
$ awk '{print $1;}' hg38modif1.txt |sort| uniq -c | gawk 'BEGIN{t=0}{t=t+$1}END{print t/NR}'
```

4.56564

```
$ awk '{print $1;}' danRer11_modify.txt |sort| uniq -c | gawk 'BEGIN{t=0}{t=t+$1}END{print t/NR}'
2.14402
```

# n° exones/transcrito

# n° nucleotidos/transcrito

```
$ gawk 'BEGIN {FS=OFS="\t"; total=0; count=0} {len=$7-$6+1; total+=len; count++} END {print "Número promedio de nu
Número promedio de nucleótidos por transcrito: 76371.7
```

```
$ gawk 'BEGIN {FS=OFS="\t"; total=0; count=0} {len=$7-$6+1; total+=len; count++} END {print "Número promedio de nu
Número promedio de nucleótidos por transcrito: 40588.3
```

Exercici 2 – Extracció de dades del catàleg OMIM (20%)

OMIM (Online Mendelian Inheritance in Man) és un catàleg de gens, trastorns i trets genètics, amb especial atenció en la relació gen-fenotip.

En aquest exercici es subministra tres dels fitxers que conformen aquesta Base de Dades amb els quals podreu contestar les preguntes que es realitzen.

Imagineu que voleu esbrinar:

El nombre de gens associats a la malaltia d’Alzheimer que es troben en els cromosomes sexuals. (5 punts)

```
$ head -10 genemap2.txt
# Chromosome      Genomic Position Start   Genomic Position End     Cyto Location    Computed Cyto Location  MIM Number
chr1      0    27600000    1p36      607413  AD7CNTP Alzheimer disease neuronal thread protein
chr1      0    27600000    1p36      612367  ALPQTL2 Alkaline phosphatase, plasma level of, QTL 2      100196914
chr1      0    123400000   1p        606788  ANON1    Anorexia nervosa, susceptibility to, 1      171514      {Anore
chr1      0    27600000    1p36      605462  BCC1     Basal cell carcinoma, susceptibility to, 1      100307118
chr1      0    27600000    1p36      606928  BMND3    Bone mineral density QTL 3  246259      ?another locus at 3p21
chr1      0    2300000    1p36.33   618815  C1DUPp36.33, DUP1p36.33 Chromosome 1p36.33 duplication syndrome, ATAD3 gene
```

#	Genomic	Genomic		Computed									Mouse
Chro-	Posi-	Posi-	Cyto	Cyto	MIM	Gene	Gene	Approved	Entrez	Ensembl			Gene
mo-	tion	tion	Loca-	Loca-	Num-	Sym-	Name	Sym-	Gene	Gene	Comments	Phenotype	Sym-
some	Start	End	tion	tion	ber	bols		bol	ID	ID			bol/ID
chr1	0	27600000	1p36		612367	ALPQTL2	Alkaline phos- phatase, plasma level of, QTL 2		100196914	linkage with rs1780324	{Alkaline phos- phatase, plasma level of, QTL 2}, 612367 (2)		
chr1	0	123400000	1p		606788	ANON1	Anorexia ner- vosa, suscep- tibility to, 1		171514		{Anorexia ner- vosa, suscep- tibility to, 1}, 606788 (2)		

# Chro- mo- some	Genomic Posi- tion Start	Genomic Posi- tion End	Cyto Loca- tion	Computed Cyto Loca- tion	MIM Num- ber	Gene Sym- bols	Gene Name	Approved Sym- bol	Entrez Gene ID	Ensembl Gene ID	Comments	Phenotype	Mouse Gene Sym- bol/ID
chr1	0	27600000	1p36		605462	BCC1	Basal cell carci- noma, suscep- tibility to, 1		100307118	rs7538876	Associated with cell carcinoma, susceptibility to, 1}, 605462 (2)		
chr1	0	27600000	1p36		606928	BMND3	Bone min- eral den- sity QTL 3	246259				[Bone mineral density QTL 3], 606928 (2)	
chr1	0	2300000	1p36.33		618815	C1DUPp36.33 DUP1p36.33	Chromosome 1p36.33 dupli- cation syn- drome, ATAD3 gene cluster				Chromosome 1p36.33 duplication syndrome, ATAD3 gene cluster, 618815 (4), Auto-somal dominant		
# Chro- mo- some	Genomic Posi- tion Start	Genomic Posi- tion End	Cyto Loca- tion	Computed Cyto Loca- tion	MIM Num- ber	Gene Sym- bols	Gene Name	Approved Sym- bol	Entrez Gene ID	Ensembl Gene ID	Comments	Phenotype	Mouse Gene Sym- bol/ID

```
$ gawk '/Alzheimer/ && ($1 == "chrX" || $1 == "chrY") {print $1}' genemap2.txt | uniq -c
1 chrX
```

El nombre de gens associats a una herència autosòmica dominant i en el fenotip està definida com a síndrome i el gen es troba etiquetat amb el format HGNC. (5 punts)

```
$ gawk '{print $1}' genemap2.txt | cat genemap2.txt | grep "dominant" -A 0 | grep "syndrome" -A 0 | grep -v "^--$"
589
```

Determinar quins són els gens que estan associats a la BRCA2 (5 punts)

```
$ awk -F'\t' '$8 ~ /BRCA2/ || $9 ~ /BRCA2/ {print $7}' genemap2.txt
BCCIP, TOK1
BRCA2, FANCD1, BROVCA2, GLM3, PNCA2
PALB2, FANCN, PNCA3
CNTROB, LIP8
BRCC3, BRCC36
```

Els gens associats amb BRCA2 que apareixen a la columna 7 són:

- BCCIP
- BRCA2

- FANCD1
- BROVCA2
- GLM3
- PNCA2
- PALB2
- FANCN
- PNCA3
- CNTROB
- LIP8
- BRCC3
- BRCC36

Quants gens associats al càncer de pit es troben etiquetats a OMIM (5 punts)

```
$ gawk '{print $6}' genemap2.txt | cat genemap2.txt |  
grep -i "breast" -A 0 | grep -i "cancer" | grep -v "^--$" | wc -l  
36
```

Trobem 36 gens que estan relacionats amb el càncer de pit a OMIM