# PEC3. Análisis bioinformático mediante lenguaje MySQL

bas, magí

2023-05-21

## Ejercicio 1 – Descripción de los catálogos de genes y términos de GO

```
$ wget https://hgdownload.cse.ucsc.edu/goldenPath/hg38/database/ncbiRefSeq.txt
2023/05/18 09:17:32 start download ncbiRefSeq.txt
2023/05/18 09:17:32 total size: 6.939 MB
     7275342/7276366 [================================================ ]       701 kB/s [ FINISHED! ]
2023/05/18 09:17:34 https://hgdownload.cse.ucsc.edu/goldenPath/hg38/database/ncbiRefSeq.txt => ncbiRefSeq.txt
2023/05/18 09:17:34 Time took 3.203725544s

$ wget https://hgdownload.cse.ucsc.edu/goldenPath/hg38/database/ncbiRefSeq.sql
--2023-05-18 09:22:46--  https://hgdownload.cse.ucsc.edu/goldenPath/hg38/database/ncbiRefSeq.sql
Resolving hgdownload.cse.ucsc.edu (hgdownload.cse.ucsc.edu)... 128.114.119.163
Connecting to hgdownload.cse.ucsc.edu (hgdownload.cse.ucsc.edu)|128.114.119.163|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 1994 (1.9K) [application/sql]
Saving to: ‘ncbiRefSeq.sql’

ncbiRefSeq.sql               100%[=============================================>]   1.95K  --.-KB/s    in 0s

2023-05-18 09:22:47 (91.5 MB/s) - ‘ncbiRefSeq.sql’ saved [1994/1994]

$ mv ncbiRefSeq.sql hg38_ncbiRefSeq.sql
```

```
$ wget https://hgdownload.cse.ucsc.edu/goldenPath/galGal6/database/ncbiRefSeq.sql
--2023-05-18 09:24:24--  https://hgdownload.cse.ucsc.edu/goldenPath/galGal6/database/ncbiRefSeq.sql
Resolving hgdownload.cse.ucsc.edu (hgdownload.cse.ucsc.edu)... 128.114.119.163
Connecting to hgdownload.cse.ucsc.edu (hgdownload.cse.ucsc.edu)|128.114.119.163|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 1976 (1.9K) [application/sql]
Saving to: ‘ncbiRefSeq.sql.1’

ncbiRefSeq.sql.1             100%[=============================================>]   1.93K  --.-KB/s    in 0s

2023-05-18 09:24:25 (96.5 MB/s) - ‘ncbiRefSeq.sql.1’ saved [1976/1976]

$ wget https://hgdownload.cse.ucsc.edu/goldenPath/galGal6/database/ncbiRefSeq.txt
--2023-05-18 09:24:31--  https://hgdownload.cse.ucsc.edu/goldenPath/galGal6/database/ncbiRefSeq.txt
Resolving hgdownload.cse.ucsc.edu (hgdownload.cse.ucsc.edu)... 128.114.119.163
Connecting to hgdownload.cse.ucsc.edu (hgdownload.cse.ucsc.edu)|128.114.119.163|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 3315512 (3.2M) [application/x-gzip]
Saving to: ‘ncbiRefSeq.txt.2’

ncbiRefSeq.txt.2             100%[=============================================>]   3.16M  2.23MB/s    in 1.4s

2023-05-18 09:24:33 (2.23 MB/s) - ‘ncbiRefSeq.txt.2’ saved [3315512/3315512]

$ mv ncbiRefSeq.sql galGal6_ncbiRefSeq.sql
$ mv ncbiRefSeq.txt Hg38_ncbiRefSeq.txt
```

| Genoma | Chr | Gens | Tr | TrC | TrNC | Tr/Gen | Ex/Tr | nuc/Tr |
|---|---|---|---|---|---|---|---|---|
| **H. sapiens (hg38)** | 25 | 184489 | 42776 | 15677 | 27099 | 1.06 | 11.71 | 76370.73 |
| **G. gallus (GRCg6a)** | 35 | 62160 | 23726 | 15982 | 7744 | 1.00016 | 12.8121 | 47617.4 |

```
$ systemctl start mysql.service
$ systemctl status mysql.service
 mysql.service - MySQL Community Server
    Loaded: loaded (/lib/systemd/system/mysql.service; disabled; preset: disabled)
    Active: active (running) since Thu 2023-05-18 10:02:49 CEST; 33s ago
      Docs: man:mysqld(8)
            http://dev.mysql.com/doc/refman/en/using-systemd.html
   Process: 13467 ExecStartPre=/usr/share/mysql-8.0/mysql-systemd-start pre (code=exited, status=0/SUCCESS)
  Main PID: 13515 (mysqld)
    Status: "Server is operational"
     Tasks: 38 (limit: 18891)
    Memory: 441.4M
       CPU: 795ms
    CGroup: /system.slice/mysql.service
            13515 /usr/sbin/mysqld
```

```
$ mysql -u root -p
Enter password:
Welcome to the MySQL monitor.  Commands end with ; or \g.
Your MySQL connection id is 8
Server version: 8.0.33 MySQL Community Server - GPL

Copyright (c) 2000, 2023, Oracle and/or its affiliates.

Oracle is a registered trademark of Oracle Corporation and/or its
affiliates. Other names may be trademarks of their respective
owners.

Type 'help;' or '\h' for help. Type '\c' to clear the current input statement.
```

```
mysql> use cataleg;
mysql> source /home/student/hg38_ncbiRefSeq.sql
# igual para el otro archivo .sql
```

```
# Canvi manual en els archius .sql
DROP TABLE IF EXISTS `GalncbiRefSeq`;
/*!40101 SET @saved_cs_client     = @@character_set_client */;
/*!40101 SET character_set_client = utf8 */;
CREATE TABLE `GalncbiRefSeq`

mysql> source /home/nxeet/galGal6_ncbiRefSeq.sql

mysql> show tables;
+-------------------+
| Tables_in_cataleg |
+-------------------+
| GalncbiRefSeq     |
| HgncbiRefSeq      |
+-------------------+
2 rows in set (0.00 sec)
```

**H. sapiens (hg38)**

```
mysql> describe HgncbiRefSeq;
+-------------+-----------------------------------+------+-----+---------+-------+
| Field       | Type                              | Null | Key | Default | Extra |
+-------------+-----------------------------------+------+-----+---------+-------+
| bin         | smallint unsigned                 | NO   |     | NULL    |       |
| name        | varchar(255)                      | NO   | MUL | NULL    |       |
```

```
| chrom        | varchar(255)                        | NO   | MUL | NULL    |       |
| strand       | char(1)                             | NO   |     | NULL    |       |
| txStart      | int unsigned                        | NO   |     | NULL    |       |
| txEnd        | int unsigned                        | NO   |     | NULL    |       |
| cdsStart     | int unsigned                        | NO   |     | NULL    |       |
| cdsEnd       | int unsigned                        | NO   |     | NULL    |       |
| exonCount    | int unsigned                        | NO   |     | NULL    |       |
| exonStarts   | longblob                            | NO   |     | NULL    |       |
| exonEnds     | longblob                            | NO   |     | NULL    |       |
| score        | int                                 | YES  |     | NULL    |       |
| name2        | varchar(255)                        | NO   | MUL | NULL    |       |
| cdsStartStat | enum('none','unk','incmpl','cmpl') | NO   |     | NULL    |       |
| cdsEndStat   | enum('none','unk','incmpl','cmpl') | NO   |     | NULL    |       |
| exonFrames   | longblob                            | NO   |     | NULL    |       |
+--------------+-------------------------------------+------+-----+---------+-------+
16 rows in set (0.00 sec)

mysql> LOAD DATA LOCAL INFILE '/home/student/Hg38_ncbiRefSeq.txt'
    -> into table HgncbiRefSeq
    -> fields terminated by '\t'
    -> lines terminated by '\n';
Query OK, 195300 rows affected (1.68 sec)
Records: 195300  Deleted: 0  Skipped: 0  Warnings: 0

mysql> select distinct substring_index(chrom, '_', 1) as Chrname from HgncbiRefSeq;
+---------+
| Chrname |
+---------+
| chr1    |
| chr10   |
| chr11   |
| chr12   |
| chr13   |
| chr14   |
| chr15   |
| chr16   |
| chr17   |
| chr18   |
| chr19   |
| chr2    |
| chr20   |
| chr21   |
| chr22   |
| chr3    |
| chr4    |
| chr5    |
| chr6    |
| chr7    |
| chr8    |
| chr9    |
| chrUn   |
| chrX    |
| chrY    |
+---------+
25 rows in set (0.05 sec)

# Detecta un cromosoma estraño llamado chrUn y nos falta el chrom 23

mysql> INSERT INTO Hggenoma
    -> SELECT
    ->   COUNT(DISTINCT SUBSTRING_INDEX(chrom, '_', 1)) AS num_cromosomas,
    ->   COUNT(DISTINCT name) AS num_genes,
    ->   COUNT(DISTINCT name2) AS num_transcritos,
```

```
    ->   COUNT(DISTINCT CASE WHEN cdsStartStat = 'cmpl' AND cdsEndStat = 'cmpl' THEN name2 END) AS num_transcritos_
    ->   COUNT(DISTINCT CASE WHEN cdsStartStat != 'cmpl' OR cdsEndStat != 'cmpl' THEN name2 END) AS num_transcritos
    ->   COUNT(name2) / COUNT(DISTINCT name) AS num_transcritos_por_gen,
    ->   AVG(exonCount) AS num_exones_por_transcrito,
    ->   AVG(txEnd - txStart) AS num_nucleotidos_por_transcrito
    -> FROM HgncbiRefSeq;
Query OK, 1 row affected, 3 warnings (0.67 sec)
Records: 1  Duplicates: 0  Warnings: 3

mysql> select * from Hggenoma;
+----------------+-----------+-----------------+-------------------------------+---------------------------------+-
| num_cromosomas | num_genes | num_transcritos | num_transcritos_codificantes | num_transcritos_no_codificantes | n
+----------------+-----------+-----------------+-------------------------------+---------------------------------+-
|             25 |    184489 |           42776 |                         20067 |                           27099 |
+----------------+-----------+-----------------+-------------------------------+---------------------------------+-
1 row in set (0.00 sec)

# Para contar los transcritos tenemos que tener en cuenta que existen 'none' y 'cdsEndStat'
mysql> SELECT COUNT(DISTINCT name2) AS transcritos_solo_codificantes
    -> FROM HgncbiRefSeq
    -> WHERE cdsStartStat <> 'none' AND cdsEndStat <> 'none'
    -> AND name2 NOT IN (
    ->     SELECT name2
    ->     FROM HgncbiRefSeq
    ->     WHERE cdsStartStat = 'none' OR cdsEndStat = 'none'
    -> );

# Problemas con Enable LOAD DATA LOCAL INFILE

ERROR 2068 (HY000): LOAD DATA LOCAL INFILE file request rejected due to restrictions on access.

mysql> SET GLOBAL local_infile=ON;

o

iniciar session con;

$ mysql --local-infile -u usuario -p
--------------------------------------------------------------------------------
# Problemas con el formato del archivo .txt
ERROR 1300 (HY000): Invalid utf8mb4 character string: ''
mv txt.gz
rename again in .txt
```

## Galgal6

```
mysql> describe GalncbiRefSeq;
+--------------+-----------------------------------+------+-----+---------+-------+
| Field        | Type                              | Null | Key | Default | Extra |
+--------------+-----------------------------------+------+-----+---------+-------+
| bin          | smallint unsigned                 | NO   |     | NULL    |       |
| name         | varchar(255)                      | NO   | MUL | NULL    |       |
| chrom        | varchar(255)                      | NO   | MUL | NULL    |       |
| strand       | char(1)                           | NO   |     | NULL    |       |
| txStart      | int unsigned                      | NO   |     | NULL    |       |
| txEnd        | int unsigned                      | NO   |     | NULL    |       |
| cdsStart     | int unsigned                      | NO   |     | NULL    |       |
| cdsEnd       | int unsigned                      | NO   |     | NULL    |       |
| exonCount    | int unsigned                      | NO   |     | NULL    |       |
| exonStarts   | longblob                          | NO   |     | NULL    |       |
| exonEnds     | longblob                          | NO   |     | NULL    |       |
| score        | int                               | YES  |     | NULL    |       |
| name2        | varchar(255)                      | NO   | MUL | NULL    |       |
```

```
| cdsStartStat | enum('none','unk','incmpl','cmpl') | NO   |     | NULL    |       |
| cdsEndStat   | enum('none','unk','incmpl','cmpl') | NO   |     | NULL    |       |
| exonFrames   | longblob                           | NO   |     | NULL    |       |
+--------------+------------------------------------+------+-----+---------+-------+
16 rows in set (0.00 sec)

LOAD DATA LOCAL INFILE '/home/student/galGal6_ncbiRefSeq.txt'
INTO TABLE GalncbiRefSeq
FIELDS TERMINATED BY '\t'
LINES TERMINATED BY '\n';

mysql> select distinct substring_index(chrom, '_', 1) as Chrname from GalncbiRefSeq;
+---------+
| Chrname |
+---------+
| chr1    |
| chr10   |
| chr11   |
| chr12   |
| chr13   |
| chr14   |
| chr15   |
| chr16   |
| chr17   |
| chr18   |
| chr19   |
| chr2    |
| chr20   |
| chr21   |
| chr22   |
| chr23   |
| chr24   |
| chr25   |
| chr26   |
| chr27   |
| chr28   |
| chr3    |
| chr30   |
| chr31   |
| chr32   |
| chr33   |
| chr4    |
| chr5    |
| chr6    |
| chr7    |
| chr8    |
| chr9    |
| chrUn   |
| chrW    |
| chrZ    |
+---------+
35 rows in set (0.02 sec)

mysql> create table genoma (
    -> NumCromosomas INT,
    ->   NumGenes INT,
    ->   NumTranscritos INT,
    ->   NumTranscritosCodificantes INT,
    ->   NumTranscritosNoCodificantes INT,
    ->   PromedioTranscritosPorGen FLOAT,
    ->   PromedioExonesPorTranscrito FLOAT,
    ->   PromedioNucleotidosPorTranscrito FLOAT
    -> );
```

```
INSERT INTO Galgenoma
SELECT
  COUNT(DISTINCT SUBSTRING_INDEX(chrom, '_', 1)) AS NumCromosomas,
  COUNT(DISTINCT name) AS NumGenes,
  COUNT(DISTINCT name2) AS NumTranscritos,
  COUNT(DISTINCT CASE WHEN cdsStartStat = 'cmpl' AND cdsEndStat = 'cmpl' THEN name2 END) AS NumTranscritosCodificar
  COUNT(DISTINCT CASE WHEN cdsStartStat != 'cmpl' OR cdsEndStat != 'cmpl' THEN name2 END) AS NumTranscritosNoCodif
  COUNT(name2) / COUNT(DISTINCT name) AS PromedioTranscritosPorGen,
  AVG(exonCount) AS PromedioExonesPorTranscrito,
  AVG(txEnd - txStart) AS PromedioNucleotidosPorTranscrito
FROM GalncbiRefSeq;

mysql> select * from Galgenoma;
+---------------+----------+----------------+---------------------------+-----------------------------+---------
| NumCromosomas | NumGenes | NumTranscritos | NumTranscritosCodificantes | NumTranscritosNoCodificantes | Promedio
+---------------+----------+----------------+---------------------------+-----------------------------+---------
|            35 |    62160 |          23726 |                     17444 |                        7744 |
+---------------+----------+----------------+---------------------------+-----------------------------+---------
1 row in set (0.00 sec)
```

**El segundo objetivo del ejercicio 1 consiste en añadir una nueva funcionalidad a la DB anterior y que permita interrogar las mismas tablas añadiendo el código Gene Ontology (ie. "GO:0003700") asociado al nombre de cada gen**

**Gene Ontology es un diccionario de conceptos biológicos que son utilizados para definir las funciones de los genes. Analiza las siguientes páginas y describe en pocas líneas en qué consiste.**

**www.geneontology.org**

**http://es.wikipedia.org/wiki/Ontologia_Genica**

**Con este enunciado se suministra el fichero gene_ontology.dat donde se define, para cada código de GO, su traducción en términos biológicos**

**Por otra parte, para cada especie existe un segundo fichero en el que para cada gen de ese organismo se almacenan los correspondientes códigos GO que describen las funciones que desempeña ese gen, según el conocimiento existente en ese momento.**

***[Recordad que para procesar este fichero debemos filtrar los comentarios iniciales y tener en cuenta que, a la hora de separar campos, las columnas están marcadas por el carácter tabulador ("\t")]***

```
https://current.geneontology.org/products/pages/downloads.html

# Descargamos
goa_chiken.gaf.gz
goa_human.gaf.gz

# Descomprimims los archivos
$ gzip -d goa_human.gaf.gz
```

```
# Creamos una tabla para GO
mysql> CREATE TABLE GeneOntology (
    ->   go_code VARCHAR(255),
    ->   description VARCHAR(255)
    -> );
Query OK, 0 rows affected (0.08 sec)

# Importamos los datos del fichero GeneOntology.dat
mysql> load data local infile '/home/student/gene_ontology.dat'
    -> into table GeneOntology
    -> fields terminated by '\t'
    -> lines terminated by '\n';
Query OK, 26083 rows affected (0.57 sec)
Records: 26083  Deleted: 0  Skipped: 0  Warnings: 0
```

Ejemplo formato de los datos en los archivos .gaf:

| Column | Content | Required? | Cardinality | Example |
|---|---|---|---|---|
| 1 | DB | required | 1 | UniProtKB |
| 2 | DB Object ID | required | 1 | P12345 |
| 3 | DB Object Symbol | required | 1 | PHO3 |
| 4 | Qualifier | optional | 0 or greater | NOT |
| 5 | GO ID | required | 1 | GO:0003993 |
| 6 | DB:Reference (\|DB:Reference) | required | 1 or greater | PMID:2676709 |
| 7 | Evidence Code | required | 1 | IMP |
| 8 | With (or) From | optional | 0 or greater | GO:0000346 |
| 9 | Aspect | required | 1 | F |
| 10 | DB Object Name | optional | 0 or 1 | Toll-like receptor 4 |
| 11 | DB Object Synonym (\|Synonym) | optional | 0 or greater | hToll |
| 12 | DB Object Type | required | 1 | protein |
| 13 | Taxon(\|taxon) | required | 1 or 2 | taxon:9606 |
| 14 | Date | required | 1 | 20090118 |
| 15 | Assigned By | required | 1 | SGD |
| 16 | Annotation Extension | optional | 0 or greater | part_of(CL:0000576) |
| 17 | Gene Product Form ID | optional | 0 or 1 | UniProtKB:P12345-2 |

```
# Creamos la tabla relacional
mysql> CREATE TABLE gene_association (
    ->   DB VARCHAR(255) NOT NULL,
    ->   DB_Object_ID VARCHAR(255) NOT NULL,
    ->   DB_Object_Symbol VARCHAR(255) NOT NULL,
    ->   Qualifier VARCHAR(255),
    ->   GO_ID VARCHAR(255) NOT NULL,
    ->   DB_Reference VARCHAR(255) NOT NULL,
    ->   Evidence_Code VARCHAR(255) NOT NULL,
    ->   With_or_From VARCHAR(255),
    ->   Aspect VARCHAR(255) NOT NULL,
```

```
    ->   DB_Object_Name VARCHAR(255),
    ->   DB_Object_Synonym VARCHAR(255),
    ->   DB_Object_Type VARCHAR(255) NOT NULL,
    ->   Taxon VARCHAR(255) NOT NULL,
    ->   Date VARCHAR(255) NOT NULL,
    ->   Assigned_By VARCHAR(255) NOT NULL,
    ->   Annotation_Extension VARCHAR(255),
    ->   Gene_Product_Form_ID VARCHAR(255),
    ->   Species VARCHAR(255) NOT NULL
    -> );
Query OK, 0 rows affected (0.08 sec)

# Cargamos los datos de chicken y human
mysql> load data local infile '/home/student/goa_chicken.gaf'
    -> INTO TABLE gene_association
    -> FIELDS TERMINATED BY '\t'
    -> LINES TERMINATED BY '\n'
    -> (DB, DB_Object_ID, DB_Object_Symbol, Qualifier, GO_ID, DB_Reference, Evidence_Code, With_or_From, Aspect, DB
    -> SET Species = 'Gallus gallus';
Query OK, 126820 rows affected, 65535 warnings (1.75 sec)
Records: 126820  Deleted: 0  Skipped: 0  Warnings: 133908

mysql> load data local infile '/home/student/goa_human.gaf' INTO TABLE gene_association FIELDS TERMINATED BY '\t' L
TERMINATED BY '\n' (DB, DB_Object_ID, DB_Object_Symbol, Qualifier, GO_ID, DB_Reference, Evidence_Code, With_or_Fro
Query OK, 634129 rows affected, 65535 warnings (12.14 sec)
Records: 634129  Deleted: 0  Skipped: 0  Warnings: 642034
```

**Con el fichero gene_ontology.dat (incluido con este enunciado) y los ficheros gene_association para las especies H. sapiens (hg38) y G. gallus (GRCg6a), que tenéis que obtener vosotros mismos de la página web de Gene Ontology, tenéis que contestar a las siguientes preguntas:**

**9. ¿Cuáles son los genes asociados al GO:0003700 en H. sapiens? (1 pto)**

GO:0003700 transcription factor activity

```
# Estos son los genes asociados a GO:0003700
mysql> SELECT DISTINCT DB_Object_Symbol
    -> FROM gene_association
    -> WHERE GO_ID = 'GO:0003700' AND Species = 'H. sapiens';
+------------------+
| DB_Object_Symbol |
+------------------+
| NFILZ            |
| E2F8             |
| NKX2-6           |
| FOXE1            |
| NR5A2            |
| SOX1             |
| E2F3             |
| BHLHE40          |
| PHOX2A           |
| BACH1            |
| IRF6             |
| KLF11            |
| ZNF263           |
| TBXT             |
| TP73             |
| FOXP2            |
| GSC2             |
| CLOCK            |
| NKX2-8           |
| MAFG             |
| CRX              |
| TBX1             |
| KLF4             |
```

| FOXO3 |
| FOXS1 |
| ZBTB14 |
| SOX15 |
| DLX3 |
| ZIC3 |
| CREBL2 |
| FOXD2 |
| LMX1B |
| MAFK |
| MSC |
| ZNF217 |
| FOXH1 |
| NFAT5 |
| ZBTB7A |
| ZIC2 |
| NFATC1 |
| ESRRB |
| FOS |
| ESR1 |
| NR3C1 |
| MYCN |
| TP53 |
| JUN |
| SP1 |
| NR3C2 |
| NFIC |
| POU2F2 |
| HOXB7 |
| GLI2 |
| GLI3 |
| AR |
| RARA |
| NR2F6 |
| NR2F1 |
| THRA |
| THRB |
| EGR2 |
| VDR |
| ESRRA |
| SRF |
| MYCLP1 |
| RARG |
| ZNF35 |
| IRF2 |
| ETS1 |
| ETS2 |
| MYOG |
| ATF2 |
| FOSL1 |
| FOSL2 |
| TCF4 |
| TCF3 |
| GATA1 |
| CREB1 |
| ZNF37A |
| TAL1 |
| ATF7 |
| CEBPB |
| XBP1 |
| SPI1 |
| EGR1 |
| ATF1 |

| ATF3 |
| ATF4 |
| ATF6 |
| ELK1 |
| TFEB |
| TFE3 |
| WT1 |
| RXRA |
| NFKB1 |
| POU3F3 |
| POU3F2 |
| HOXA5 |
| HNF1A |
| MYF6 |
| NFYA |
| PAX7 |
| PAX3 |
| GATA2 |
| GATA3 |
| ZBTB25 |
| NR2F2 |
| NFYB |
| PAX6 |
| ARNT |
| POU1F1 |
| ELK4 |
| TEAD1 |
| MZF1 |
| HOXC13 |
| ELF1 |
| RORA |
| HOXD13 |
| DDIT3 |
| HNF1B |
| SOX5 |
| SOX6 |
| SOX11 |
| ZNF93 |
| AHR |
| SREBF1 |
| PPARG |
| ZEB1 |
| PBX1 |
| STAT3 |
| ETV5 |
| BCL6 |
| ETV6 |
| HNF4A |
| ELK3 |
| STAT1 |
| STAT6 |
| STAT5A |
| GATA4 |
| NKX2-1 |
| RFX2 |
| RFX3 |
| SOX2 |
| SOX9 |
| LHX1 |
| NR2C2 |
| POU3F4 |
| EVX1 |
| CTCF |

| CEBPA |
| MEOX1 |
| MEOX2 |
| ETV1 |
| ASCL1 |
| RORC |
| MECP2 |
| STAT5B |
| AFF3 |
| STAT2 |
| ZNF133 |
| PDX1 |
| GBX2 |
| NKX2-5 |
| FOXA1 |
| FOXA3 |
| PKNOX1 |
| DLX6 |
| SOX10 |
| KLF3 |
| FOXL2 |
| PITX1 |
| POU6F2 |
| ELF3 |
| SIM1 |
| SMAD3 |
| FOXK1 |
| FOXO4 |
| HOXC5 |
| NFKB2 |
| IRF9 |
| E2F1 |
| INSM1 |
| FOXK2 |
| MYT1 |
| FLI1 |
| POU5F1 |
| MEF2A |
| MEF2B |
| CREB5 |
| PAX2 |
| POU3F1 |
| MECOM |
| PPARD |
| ZNF117 |
| ZNF92 |
| ZNF90 |
| RELA |
| REL |
| EGR4 |
| ZNF91 |
| RBPJ |
| MEF2C |
| POU5F1B |
| GABPA |
| PAX8 |
| ZNF33A |
| EGR3 |
| SOX4 |
| DLX2 |
| PPARA |
| FOXO1 |
| TFCP2 |

| NFIA |
| TCF15 |
| FOXF2 |
| FOXC1 |
| FOXD4 |
| FOXI1 |
| FOXL1 |
| PRDM2 |
| ZBTB17 |
| KLF10 |
| REST |
| KLF1 |
| FOXE3 |
| NFATC2 |
| SMAD4 |
| NEUROD1 |
| RUNX3 |
| KLF9 |
| RUNX2 |
| NFYC |
| TFDP1 |
| SIM2 |
| E2F2 |
| NFE2L1 |
| HIC1 |
| HNF4G |
| ZNF268 |
| IRF3 |
| STAT4 |
| MEF2D |
| POU6F1 |
| MTF1 |
| NFIX |
| NR1I3 |
| IRF4 |
| POU4F3 |
| E2F5 |
| SIX1 |
| TEAD4 |
| TEAD2 |
| TGIF1 |
| ZNF174 |
| NEUROD2 |
| SMAD2 |
| SMAD1 |
| USF2 |
| ZFHX3 |
| ZIC1 |
| NFE2L2 |
| E2F4 |
| BATF |
| NFE2 |
| NFIL3 |
| TBR1 |
| HIF1A |
| ZSCAN26 |
| FOXD1 |
| KLF17 |
| ZNF618 |
| AHDC1 |
| PLAG1 |
| GRHL2 |
| ZNF746 |

| ZNF367 |
| ZNF410 |
| GPBP1 |
| ZGPAT |
| ZBTB38 |
| DMBX1 |
| RHOXF1 |
| MAFA |
| BMAL2 |
| CSRNP3 |
| TFAP2B |
| RORB |
| TFAP2C |
| PROX1 |
| GATA6 |
| FOXJ1 |
| DLX4 |
| E2F7 |
| HES6 |
| ZFP42 |
| NR1H4 |
| CSRNP1 |
| ATOH8 |
| ZSCAN10 |
| TCF12 |
| MNT |
| TBX5 |
| TEAD3 |
| PITX2 |
| NPAS1 |
| NPAS2 |
| NKX3-1 |
| PRRX2 |
| ATF6B |
| HINFP |
| MESP1 |
| ZKSCAN3 |
| SOX7 |
| SCRT1 |
| GATA5 |
| SPZ1 |
| FOXP3 |
| NKX6-2 |
| TGIF2 |
| CSRNP2 |
| IKZF4 |
| TP63 |
| ZHX3 |
| NANOG |
| BARX1 |
| ARNT2 |
| TCF7L1 |
| GCM1 |
| MLXIPL |
| SIX2 |
| HEYL |
| TCF7L2 |
| BATF3 |
| SLC2A4RG |
| VSX1 |
| ZNF219 |
| RBPJL |
| HEY2 |

| MLX |
| TRPS1 |
| GTF2IRD1 |
| ZNF639 |
| VAX2 |
| LEF1 |
| FOXD3 |
| ZHX1 |
| TCFL5 |
| ZNF215 |
| SOX13 |
| PLAGL2 |
| LHX6 |
| DMTF1 |
| FOXA2 |
| ATF5 |
| ZKSCAN5 |
| ZNF281 |
| ZNF175 |
| NFE2L3 |
| KLF12 |
| HEY1 |
| KLF2 |
| SOX21 |
| ZFP37 |
| ZNF564 |
| FOXN2 |
| HSFX4 |
| ZNF671 |
| HSFY1 |
| ATOH1 |
| ZNF383 |
| IKZF2 |
| FOXN3 |
| IKZF5 |
| GFI1 |
| BHLHA15 |
| ZNF835 |
| ZNF296 |
| MYRFL |
| FEZF1 |
| BCL11B |
| ZNF395 |
| ZNF649 |
| ZBTB24 |
| ZNF524 |
| ZNF763 |
| ZNF366 |
| ZNF567 |
| HOXD9 |
| BHLHE22 |
| ZNF362 |
| ZNF382 |
| HOXA9 |
| LOC402624 |
| FOXM1 |
| HSFX3 |
| PRDM5 |
| ZNF613 |
| E4F1 |
| HSF2 |
| ZBTB48 |
| HOXC9 |

```
| FOXN4            |
| ZNF821           |
| ZNF691           |
| ATOH7            |
| ZNF667           |
| ZNF707           |
| CDX1             |
| PRDM1            |
| HSF4             |
| ZNF581           |
| ZNF212           |
| ZNF865           |
| ZNF148           |
| ZNF501           |
| THAP1            |
| HOXB9            |
| ZNF704           |
| ZNF358           |
| CDX4             |
| MAX              |
| ZNF350           |
| ZNF660           |
| ZNF710           |
| ZNF683           |
| IKZF1            |
| ZNF662           |
| GTF2I            |
| BHLHE23          |
| BCL11A           |
| HSF5             |
| IKZF3            |
| HSF1             |
| ZNF740           |
| ZNF497           |
| MYRF             |
| THAP10           |
| HSFX1            |
| CDX2             |
| ZNF692           |
+------------------+
440 rows in set (0.41 sec)
```

**10. ¿Qué gen o genes tienen en común las dos especies cuando se analiza el proceso creatine transport? (1 pto)**

GO:0015881 => creatine transport

```
# Buscamos el Go_code
mysql> select go_code from GeneOntology where description like 'creatine transport';
+------------+
| go_code    |
+------------+
| GO:0015881 |
+------------+
1 row in set (0.01 sec)

# Podemos buscar los genes con go_id = GO:0015881
mysql> SELECT DISTINCT gene_association.DB_Object_Symbol
    -> FROM gene_association
    -> WHERE gene_association.GO_ID = 'GO:0015811'
    ->     AND gene_association.Species = 'Gallus gallus'
    ->     AND gene_association.DB_Object_Symbol IN (
    ->         SELECT gene_association.DB_Object_Symbol
    ->         FROM gene_association
    ->         WHERE gene_association.GO_ID = 'GO:0015811'
```

```
    ->             AND gene_association.Species = 'H. sapiens'
    ->     );
+------------------+
| DB_Object_Symbol |
+------------------+
| SLC1A4           |
| SLC3A1           |
| SLC7A11          |
| CTNS             |
| SLC7A9           |
+------------------+
5 rows in set (0.79 sec)
```

**11. Teniendo en cuenta los genes asociados al proceso de tRNA splicing para ambas especies, determina las diferencias y similitudes entre los diferentes cromosomas donde se expresan estos genes. (1 pto)**

```
# Buscams el go_code
mysql> select go_code from GeneOntology where description like 'tRNA splicing';
+------------+
| go_code    |
+------------+
| GO:0006388 |
+------------+
1 row in set (0.01 sec)

# Obtenemos los genes asociados al proceso en H. sapiens
mysql> SELECT DISTINCT gene_association.DB_Object_ID, gene_association.DB_Object_Symbol, gene_association.Taxon, ge
    -> FROM gene_association
    -> JOIN GeneOntology ON gene_association.GO_ID = GeneOntology.go_code
    -> WHERE GeneOntology.description LIKE 'tRNA splicing' AND gene_association.Species = 'H. sapiens';
+--------------+------------------+------------+------------+
| DB_Object_ID | DB_Object_Symbol | Taxon      | GO_ID      |
+--------------+------------------+------------+------------+
| Q52LJ0       | FAM98B           | taxon:9606 | GO:0006388 |
| Q7Z6J9       | TSEN54           | taxon:9606 | GO:0006388 |
| Q8IWT0       | ZBTB8OS          | taxon:9606 | GO:0006388 |
| Q8NCA5       | FAM98A           | taxon:9606 | GO:0006388 |
| Q8NCE0       | TSEN2            | taxon:9606 | GO:0006388 |
| Q8WW01       | TSEN15           | taxon:9606 | GO:0006388 |
| Q92499       | DDX1             | taxon:9606 | GO:0006388 |
| Q92989       | CLP1             | taxon:9606 | GO:0006388 |
| Q9BVC5       | C2orf49          | taxon:9606 | GO:0006388 |
| Q9Y224       | RTRAF            | taxon:9606 | GO:0006388 |
| Q9Y3I0       | RTCB             | taxon:9606 | GO:0006388 |
| Q86TN4       | TRPT1            | taxon:9606 | GO:0006388 |
+--------------+------------------+------------+------------+
12 rows in set (0.43 sec)

# Genes asociados en G. gallus
mysql> SELECT DISTINCT gene_association.DB_Object_ID, gene_association.DB_Object_Symbol, gene_association.Taxon, ge
    -> FROM gene_association
    -> JOIN GeneOntology ON gene_association.GO_ID = GeneOntology.go_code
    -> WHERE GeneOntology.description LIKE 'tRNA splicing' AND gene_association.Species = 'Gallus gallus';
+--------------+------------------+------------+------------+
| DB_Object_ID | DB_Object_Symbol | Taxon      | GO_ID      |
+--------------+------------------+------------+------------+
| A0A8V0YIN1   | TSEN15           | taxon:9031 | GO:0006388 |
| A0A8V1ADV8   | TSEN2            | taxon:9031 | GO:0006388 |
| F1NMM3       | ZBTB8OS          | taxon:9031 | GO:0006388 |
| Q5ZJL4       | CLP1             | taxon:9031 | GO:0006388 |
| Q90WU3       | DDX1             | taxon:9031 | GO:0006388 |
| A0A1D5PCV8   | A0A1D5PCV8       | taxon:9031 | GO:0006388 |
| A0A3Q3A913   | ZBTB8OS          | taxon:9031 | GO:0006388 |
| A0A1D5P7R1   | C5H14orf166      | taxon:9031 | GO:0006388 |
```

```
+--------------+------------------+------------+------------+
8 rows in set (0.37 sec)
```

Les taules no ens proporcionen informació sobre el cromosoma del gen

## Ejercicio 2 – Datos de expresión (10 %)

**Adjunto al enunciado de este ejercicio encontraréis dos archivos texto (files.txt) y dos archivos Excel (files.xlsx) que contienen información generada por el proyecto internacional GTEx. El objetivo principal de este proyecto es construir un repositorio de expresión génica tejido específico: un catálogo con anotaciones funcionales. En la página web del proyecto podéis encontrar información sobre el proyecto (https://gtexportal.org/home/)**

**En este ejercicio las tablas que se os suministran las debéis cargar en la DB MySQL mediante el comando LOAD.**

```
GTEx_Analysis_v8_Annotations_SampleAttributesDD.xlsx
GTEx_Analysis_v8_Annotations_SampleAttributesDS.txt
GTEx_Analysis_v8_Annotations_SubjectPhenotypesDD.xlsx
GTEx_Analysis_v8_Annotations_SubjectPhenotypesDS.txt
```

**El resultado al cargar la tabla se muestra a continuación. Para reconocer si la carga de la tabla ha sido correcta, hay que fijarse/leer la línea que comienza con el keyword Records: *Si al cargar una tabla aparecen números diferentes al cero junto a los keywords deleted or skipped, la carga de datos ha sido ineficiente. Una de las consecuencias de una carga errónea es que las respuestas a las preguntas que se realizan a la DB son incorrectas. De forma que la carga de las tablas han de mostrar valores de deleted and skipped igual a cero para tener la seguridad de que interrogamos de manera correcta, tenerlo en cuenta.***

**1. Enumera por lo menos dos razones por las que los datos de una tabla al cargarse en MySQL no se realiza correctamente (1 pto)**

1. Cuando la tabla de destino no corresponden con las columnas del archivo de origen
2. Si a tabla de destino tiene restricciones de integridad, como claves primarias, foráneas o NOT NULL's

**A partir de los archivos que contienen información generada en el proyecto internacional GTEx y mediante comandos SQL deberéis responder a las siguientes preguntas, aunque antes de resolver las preguntas se tiene que mostrar todos los pasos de transformación de los ficheros, la definición de las variables y la carga de los datos.**

```
## SampleAttributes table
# Creamos la tabla (el archivo .xlsx nos da pistas sobre la cantidad y el tipo de columnas que tiene que contener
mysql> CREATE TABLE sampleattribute (
    ->   SAMPID VARCHAR(255),
    ->   SMATSSCR INT,
    ->   SMNABTCH VARCHAR(255),
    ->   SMNABTCHT VARCHAR(255),
    ->   SMNABTCHD VARCHAR(255),
    ->   SMGEBTCH VARCHAR(255),
    ->   SMGEBTCHD VARCHAR(255),
    ->   SMGEBTCHT VARCHAR(255),
    ->   SMCENTER VARCHAR(255),
    ->   SMPTHNTS VARCHAR(255),
    ->   SMRIN DECIMAL,
    ->   SMTS VARCHAR(255),
    ->   SMTSD VARCHAR(255),
    ->   SMUBRID VARCHAR(255),
    ->   SMTSISCH INT,
    ->   SMTSPAX INT,
    ->   SMAFRZE VARCHAR(255),
    ->   SMGTC VARCHAR(255),
    ->   SME2MPRT DECIMAL,
    ->   SMCHMPRS INT,
    ->   SMNTRART DECIMAL,
    ->   SMNUMGPS INT,
    ->   SMMAPRT DECIMAL,
    ->   SMEXNCRT DECIMAL,
    ->   SM550NRM DECIMAL,
    ->   SMGNSDTC INT,
    ->   SMUNMPRT DECIMAL,
```

```
    ->   SM350NRM DECIMAL,
    ->   SMRDLGTH INT,
    ->   SMMNCPB DECIMAL,
    ->   SME1MMRT DECIMAL,
    ->   SMSFLGTH INT,
    ->   SMESTLBS INT,
    ->   SMMPPD INT,
    ->   SMNTERRT DECIMAL,
    ->   SMRRNANM INT,
    ->   SMRDTTL INT,
    ->   SMVQCFL INT,
    ->   SMMNCV DECIMAL,
    ->   SMTRSCPT INT,
    ->   SMMPPDPR INT,
    ->   SMCGLGTH INT,
    ->   SMGAPPCT DECIMAL,
    ->   SMUNPDRD INT,
    ->   SMNTRNRT DECIMAL,
    ->   SMMPUNRT DECIMAL,
    ->   SMEXPEFF DECIMAL,
    ->   SMMPPDUN INT,
    ->   SME2MMRT DECIMAL,
    ->   SME2ANTI INT,
    ->   SMALTALG INT,
    ->   SME2SNSE INT,
    ->   SMMFLGTH INT,
    ->   SMSPLTRD INT,
    ->   SME1ANTI INT,
    ->   SMBSMMRT DECIMAL,
    ->   SME1SNSE INT,
    ->   SME1PCTS DECIMAL,
    ->   SMRRNART DECIMAL,
    ->   SME1MPRT DECIMAL,
    ->   SMNUM5CD INT,
    ->   SMDPMPRT DECIMAL,
    ->   SME2PCTS DECIMAL
    -> );
Query OK, 0 rows affected (0.07 sec)

# Cargamos el fichero .txt con los datos
mysql> LOAD DATA local  INFILE '/home/student/GTEx_Analysis_v8_Annotations_SampleAttributesDS.txt'
    -> INTO TABLE sampleattributes
    -> FIELDS TERMINATED BY '\t'
    -> LINES TERMINATED BY '\n'
    -> IGNORE 1 LINES;
Query OK, 22951 rows affected, 19946 warnings (0.82 sec)
Records: 22951  Deleted: 0  Skipped: 0  Warnings: 19946

# Parece que los warnings son devidos a truncamientos por la longitud del varcahard en las columnas SMTS y SMNABTC
mysql> ALTER TABLE sampleattributes MODIFY COLUMN SMTS VARCHAR(255);
Query OK, 45902 rows affected (1.34 sec)
Records: 45902  Duplicates: 0  Warnings: 0

mysql> ALTER TABLE sampleattributes MODIFY COLUMN SMNABTCHT VARCHAR(255);
Query OK, 45902 rows affected (1.80 sec)
Records: 45902  Duplicates: 0  Warnings: 0

# Podemos utilizar mysql workbench para visualizar la tabla en una interfaz más gráfica
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content: | Fetch

| # | SAMPID | SMATSSCR | SMNABTCH | SMNABTCHT |
|---|---|---|---|---|
| 1 | GTEX-1117F-0003-SM-58Q7G | | B1 | |
| 2 | GTEX-1117F-0003-SM-5DWSB | | B1 | |
| 3 | GTEX-1117F-0003-SM-6WBT7 | | B1 | |
| 4 | GTEX-1117F-0011-R10a-SM-AHZ7F | | B1, A1 | |
| 5 | GTEX-1117F-0011-R10b-SM-CYK... | | B1, A1 | |
| 6 | GTEX-1117F-0226-SM-5GZZ7 | 0 | B1 | 2 pieces, ~15% vesse |
| 7 | GTEX-1117F-0426-SM-5EGHI | 0 | B1 | 2 pieces, !5% fibrous c |
| 8 | GTEX-1117F-0526-SM-5EGHJ | 0 | B1 | 2 pieces, clean, Moncl |
| 9 | GTEX-1117F-0626-SM-5N9CS | 1 | B1 | 2 pieces, up to 4mm a |
| 10 | GTEX-1117F-0726-SM-5GIEN | 1 | B1 | 2 pieces, no abnormal |
| 11 | GTEX-1117F-1326-SM-5EGHH | 1 | B1 | 2 pieces, diffuse meso |
| 12 | GTEX-1117F-2226-SM-5N9CU | 1 | B1 | 1 piece vascular tissu |

```
## Tabla SubjectPhenotypes
mysql> create table subjectphenotypes (
    -> SUBJID VARCHAR(10) NOT NULL,
    -> SEX INT,
    ->   AGE VARCHAR(10),
    ->   DTHHRDY INT,
    ->   ;
Query OK, 0 rows affected (0.07 sec)

# Insertar datos
mysql> LOAD DATA LOCAL INFILE '/home/student/GTEx_Analysis_v8_Annotations_SubjectPhenotypesDS.txt'
    -> into table subjectphenotypes
    -> FIELDS TERMINATED BY '\t'
    -> LINES TERMINATED BY '\n'
    -> IGNORE 1 LINES;
Query OK, 980 rows affected, 19 warnings (0.05 sec)
Records: 980  Deleted: 0  Skipped: 0  Warnings: 19

# Parece que tenemos problemas con 19 valores en la columna 'DTHHRDY' debido a valores int incorrectos podriamos a
# Parece que hay valores NA en el archivo .txt
```

**2. ¿Cuál es el rango de edad más frecuente en el que se tienen muestras? (1 pto)**

```
# Consultamos en subjectphenotypes
mysql> SELECT AGE, COUNT(*) AS total_muestras
    -> FROM subjectphenotypes
    -> WHERE AGE IS NOT NULL
    -> GROUP BY AGE
    -> ORDER BY total_muestras DESC
    -> LIMIT 1;
+-------+----------------+
| AGE   | total_muestras |
+-------+----------------+
| 60-69 |            317 |
+-------+----------------+
1 row in set (0.00 sec)
```

**3. ¿Cuál es el tipo de muerte más frecuente por género? (1 pto)**

leyenda para 'dthhrdy':

---

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
0) Ventilator Case All cases on a ventilator immediately before death.

---

```
mysql> SELECT
    ->     SEX,
    ->     DTHHRDY,
    ->     COUNT(*) AS frequency
    -> FROM
    ->     subjectphenotypes
    -> WHERE
    ->     DTHHRDY IS NOT NULL
    -> GROUP BY
    ->     SEX, DTHHRDY
    -> ORDER BY
    ->     SEX, frequency DESC;
+------+---------+-----------+
| SEX  | DTHHRDY | frequency |
+------+---------+-----------+
|    1 |       0 |       331 |
|    1 |       2 |       189 |
|    1 |       4 |        73 |
|    1 |       3 |        37 |
|    1 |       1 |        23 |
|    2 |       0 |       199 |
|    2 |       2 |        50 |
|    2 |       4 |        46 |
|    2 |       3 |        20 |
|    2 |       1 |        12 |
+------+---------+-----------+
10 rows in set (0.00 sec)
```

Según los resultados obtenidos, podemos observar lo siguiente:

Para el género masculino y femenino; el tipo de muerte más frecuente es el código 0, que representa los casos en los que el individuo estaba en ventilación mecánica antes de la muerte.

*[Nota: 1=Male / 2=Female]*

**4. ¿Cuántos tipos de regiones del cerebro (brain) se pueden encontrar? (1 pto)**

Para la tabla sampleattributes hay que tener en cuenta que;

SMTS: Tipo de tejido, área de la cual se tomó la muestra de tejido.

SMGEBTCH Genotype or Expression

SMGEBTCHD Date of genotype or expression batch

```
# No sé el motivo pero la columna que contiene los campos del area es 'SMGEBTCHD'
mysql> SELECT COUNT(DISTINCT SMGEBTCHD) AS count_brain_regions
    -> FROM sampleattributes
    -> WHERE SMGEBTCHD LIKE 'Brain%';
+----------------------+
| count_brain_regions |
+----------------------+
|                   13 |
```

```
+---------------------+
1 row in set (0.01 sec)
```

**5. ¿Cuál es el paciente que más muestras tiene? (1 pto)**

```
# Del campo 'SAMPID' filtramos por los caracteres entre los dos primeros guines, ya que son esos los que nos ident
mysql> SELECT SUBSTRING_INDEX(SUBSTRING_INDEX(SAMPID, '-', 2), '-', -1) AS PatientID, COUNT(*) AS TotalRegistros FR
sampleattributes GROUP BY PatientID ORDER BY TotalRegistros DESC LIMIT 2;
+-----------+----------------+
| PatientID | TotalRegistros |
+-----------+----------------+
| 562       |            217 |
| NPJ8      |             72 |
+-----------+----------------+
2 rows in set (0.01 sec)

# Parece que K-562 es un outlier o paceinte no definido asi que vamos a dar como bueno el paciente GTEX-NPJ8
```

**6. ¿Cuáles son las 7 muestras que tienen un mayor valor de "Split Reads"? (1 pto)**

SMSPLTRD Split Reads: The number of reads that span an exon-exon boundary

```
mysql> SELECT SAMPID, SMSPLTRD FROM sampleattributes ORDER BY SMSPLTRD DESC LIMIT 7;
+-------------------------------+-----------+
| SAMPID                        | SMSPLTRD  |
+-------------------------------+-----------+
| GTEX-14BMU-1526-SM-5TDE6      | 145416000 |
| GTEX-14JFF-0526-SM-62LFL      | 103439000 |
| GTEX-1JMQI-2026-SM-CMKGP      |  99701800 |
| GTEX-13G51-0011-R8b-SM-5LZZ4  |  97710000 |
| GTEX-1A32A-3026-SM-72D61      |  94843300 |
| GTEX-QMRM-0426-SM-4R1K2       |  88613400 |
| GTEX-18A7A-2226-SM-7LT8K      |  74265000 |
+-------------------------------+-----------+
7 rows in set (0.01 sec)
```

**7. Mediante el comando "join" responder a, ¿cuántas mujeres han muerto de manera violenta y tienen muestras de sangre? ¿y cuál es la media de "mapped unique" de esta selección? (4 ptos)**

```
# Concatenar las tablas
mysql> describe subjectphenotypes;
+---------+-------------+------+-----+---------+-------+
| Field   | Type        | Null | Key | Default | Extra |
+---------+-------------+------+-----+---------+-------+
| SUBJID  | varchar(10) | NO   | PRI | NULL    |       |
| SEX     | int         | YES  |     | NULL    |       |
| AGE     | varchar(10) | YES  |     | NULL    |       |
| DTHHRDY | int         | YES  |     | NULL    |       |
+---------+-------------+------+-----+---------+-------+
4 rows in set (0.00 sec)

mysql> describe sampleattributes;
+-----------+---------------+------+-----+---------+-------+
| Field     | Type          | Null | Key | Default | Extra |
+-----------+---------------+------+-----+---------+-------+
| SAMPID    | varchar(255)  | YES  |     | NULL    |       |
| SMATSSCR  | int           | YES  |     | NULL    |       |
| SMNABTCH  | varchar(255)  | YES  |     | NULL    |       |
| SMNABTCHT | varchar(255)  | YES  |     | NULL    |       |
| SMNABTCHD | varchar(255)  | YES  |     | NULL    |       |
| SMGEBTCH  | varchar(255)  | YES  |     | NULL    |       |
| SMGEBTCHD | varchar(255)  | YES  |     | NULL    |       |
| SMGEBTCHT | varchar(255)  | YES  |     | NULL    |       |
| SMCENTER  | varchar(255)  | YES  |     | NULL    |       |
| SMPTHNTS  | varchar(255)  | YES  |     | NULL    |       |
| SMRIN     | decimal(10,0) | YES  |     | NULL    |       |
```

```
| SMTS      | varchar(255)  | YES  |     | NULL    |       |
| SMTSD     | varchar(255)  | YES  |     | NULL    |       |
| SMUBRID   | varchar(255)  | YES  |     | NULL    |       |
| SMTSISCH  | int           | YES  |     | NULL    |       |
| SMTSPAX   | int           | YES  |     | NULL    |       |
| SMAFRZE   | varchar(255)  | YES  |     | NULL    |       |
| SMGTC     | varchar(255)  | YES  |     | NULL    |       |
| SME2MPRT  | decimal(10,0) | YES  |     | NULL    |       |
| SMCHMPRS  | int           | YES  |     | NULL    |       |
| SMNTRART  | decimal(10,0) | YES  |     | NULL    |       |
| SMNUMGPS  | int           | YES  |     | NULL    |       |
| SMMAPRT   | decimal(10,0) | YES  |     | NULL    |       |
| SMEXNCRT  | decimal(10,0) | YES  |     | NULL    |       |
| SM550NRM  | decimal(10,0) | YES  |     | NULL    |       |
| SMGNSDTC  | int           | YES  |     | NULL    |       |
| SMUNMPRT  | decimal(10,0) | YES  |     | NULL    |       |
| SM350NRM  | decimal(10,0) | YES  |     | NULL    |       |
| SMRDLGTH  | int           | YES  |     | NULL    |       |
| SMMNCPB   | decimal(10,0) | YES  |     | NULL    |       |
| SME1MMRT  | decimal(10,0) | YES  |     | NULL    |       |
| SMSFLGTH  | int           | YES  |     | NULL    |       |
| SMESTLBS  | int           | YES  |     | NULL    |       |
| SMMPPD    | int           | YES  |     | NULL    |       |
| SMNTERRT  | decimal(10,0) | YES  |     | NULL    |       |
| SMRRNANM  | int           | YES  |     | NULL    |       |
| SMRDTTL   | int           | YES  |     | NULL    |       |
| SMVQCFL   | int           | YES  |     | NULL    |       |
| SMMNCV    | decimal(10,0) | YES  |     | NULL    |       |
| SMTRSCPT  | int           | YES  |     | NULL    |       |
| SMMPPDPR  | int           | YES  |     | NULL    |       |
| SMCGLGTH  | int           | YES  |     | NULL    |       |
| SMGAPPCT  | decimal(10,0) | YES  |     | NULL    |       |
| SMUNPDRD  | int           | YES  |     | NULL    |       |
| SMNTRNRT  | decimal(10,0) | YES  |     | NULL    |       |
| SMMPUNRT  | decimal(10,0) | YES  |     | NULL    |       |
| SMEXPEFF  | decimal(10,0) | YES  |     | NULL    |       |
| SMMPPDUN  | int           | YES  |     | NULL    |       |
| SME2MMRT  | decimal(10,0) | YES  |     | NULL    |       |
| SME2ANTI  | int           | YES  |     | NULL    |       |
| SMALTALG  | int           | YES  |     | NULL    |       |
| SME2SNSE  | int           | YES  |     | NULL    |       |
| SMMFLGTH  | int           | YES  |     | NULL    |       |
| SMSPLTRD  | int           | YES  |     | NULL    |       |
| SME1ANTI  | int           | YES  |     | NULL    |       |
| SMBSMMRT  | decimal(10,0) | YES  |     | NULL    |       |
| SME1SNSE  | int           | YES  |     | NULL    |       |
| SME1PCTS  | decimal(10,0) | YES  |     | NULL    |       |
| SMRRNART  | decimal(10,0) | YES  |     | NULL    |       |
| SME1MPRT  | decimal(10,0) | YES  |     | NULL    |       |
| SMNUM5CD  | int           | YES  |     | NULL    |       |
| SMDPMPRT  | decimal(10,0) | YES  |     | NULL    |       |
| SME2PCTS  | decimal(10,0) | YES  |     | NULL    |       |
+-----------+---------------+------+-----+---------+-------+
63 rows in set (0.01 sec)

# Convertimo el campo 'SAMPID' en la llave primaria de la tabla sampleattributes
mysql> ALTER TABLE sampleattributes
    -> MODIFY COLUMN SAMPID VARCHAR(255) NOT NULL,
    -> ADD PRIMARY KEY (SAMPID);
Query OK, 0 rows affected (4.30 sec)
Records: 0  Duplicates: 0  Warnings: 0

# Creamos un foreing key que referencia a la tabla subjectphenotypes
```

```
mysql> ALTER TABLE sampleattributes
    -> ADD COLUMN subject VARCHAR(10),
    -> ADD FOREIGN KEY (subject) REFERENCES subjectphenotypes(SUBJID);
Query OK, 22951 rows affected (1.26 sec)
Records: 22951  Duplicates: 0  Warnings: 0

# Seleccionamos todos los caracters de SAMPID hasta el segundo guion
mysql> UPDATE sampleattributes
    -> SET subject = SUBSTRING_INDEX(SAMPID, '-', 2);
Query OK, 22951 rows affected (0.78 sec)
Rows matched: 22951  Changed: 22951  Warnings: 0

mysql> SELECT SAMPID, subject
    -> FROM sampleattributes
    -> LIMIT 5;
+--------------------------------+------------+
| SAMPID                         | subject    |
+--------------------------------+------------+
| GTEX-1117F-0003-SM-58Q7G       | GTEX-1117F |
| GTEX-1117F-0003-SM-5DWSB       | GTEX-1117F |
| GTEX-1117F-0003-SM-6WBT7       | GTEX-1117F |
| GTEX-1117F-0011-R10a-SM-AHZ7F | GTEX-1117F |
| GTEX-1117F-0011-R10b-SM-CYKQ8 | GTEX-1117F |
+--------------------------------+------------+
5 rows in set (0.00 sec)
```

DTHHRDY

Hardy Scale

Death Circumstances

integer, encoded value

Death Classification:

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. 0=Ventilator Case 1=Violent and fast death 2=Fast death of natural causes 3=Intermediate death 4=Slow death

```
# Mujeres muertas de manera violenta y con muestras de sangre
# 'Blood' se encuentra en el campo SMGEBTCH
mysql> SELECT COUNT(*) AS count_female_deaths_with_blood_samples
    -> FROM sampleattributes sa
    -> JOIN subjectphenotypes sp ON sa.subject = sp.SUBJID
    -> WHERE sp.SEX = 2
    -> AND sp.DTHHRDY = 1
    -> AND sa.SMGEBTCH LIKE '%Blood%'
    -> ;
+---------------------------------------+
| count_female_deaths_with_blood_samples |
+---------------------------------------+
|                                    66 |
+---------------------------------------+
1 row in set (0.01 sec)
```

SMMPUNRT

Mapped Unique Rate of Total: Ratio of mapping of reads that were aligned and were not duplicates to total reads

```
# Media de "mapped unique" de esta selección
mysql> SELECT AVG(SMMPUNRT) AS average_mapped_unique_rate
    -> FROM sampleattributes sa
    -> JOIN subjectphenotypes sp ON sa.subject = sp.SUBJID
    -> WHERE sp.SEX = 2
    -> AND sp.DTHHRDY = 1
    -> AND sa.SMGEBTCH LIKE '%Blood%'
    -> ;
+----------------------------+
| average_mapped_unique_rate |
+----------------------------+
|                     0.4848 |
+----------------------------+
1 row in set (0.01 sec)
```