

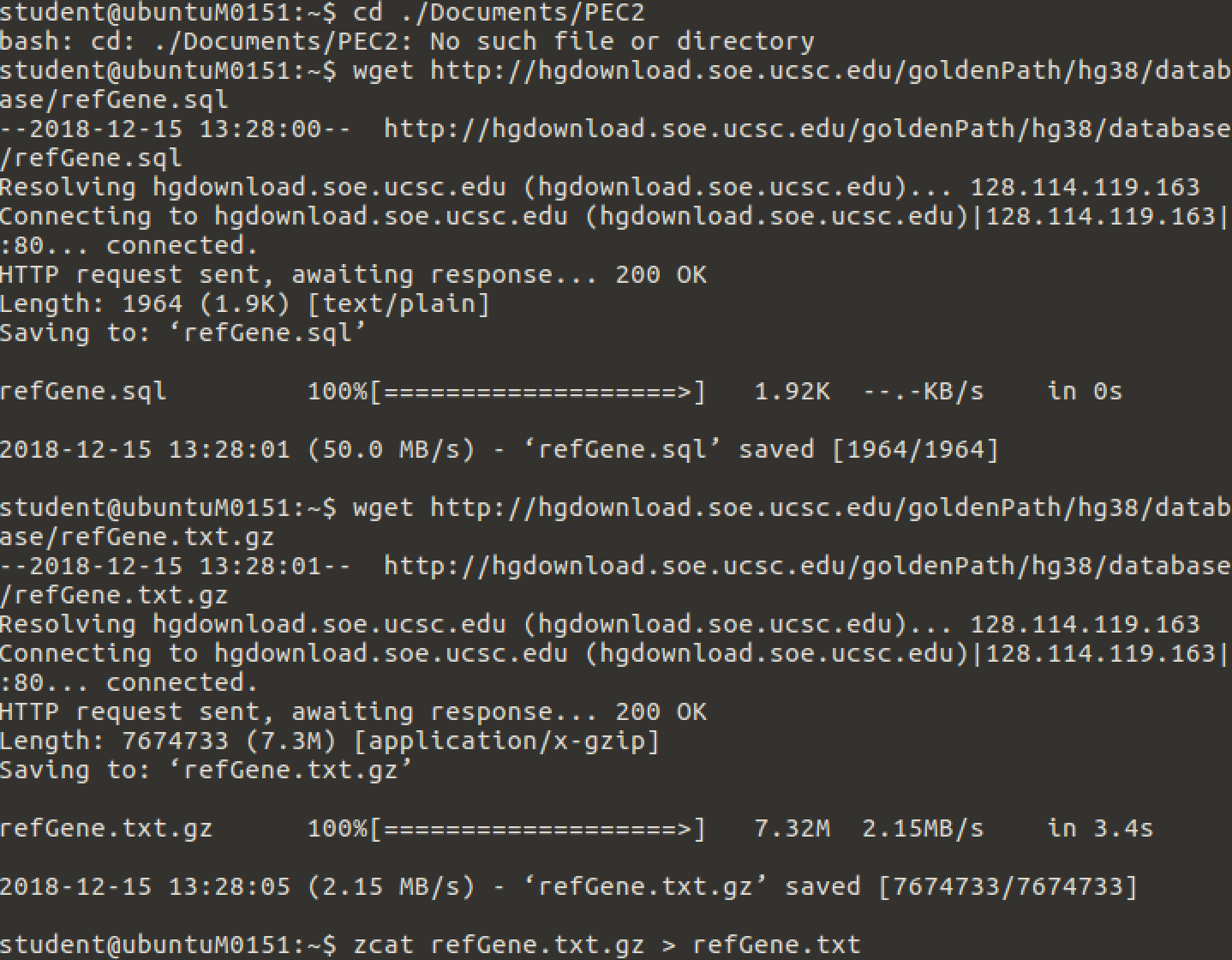
**Herramientas informáticas para la bioinformática**

Víctor Fanjul Hevia noviembre de 2018

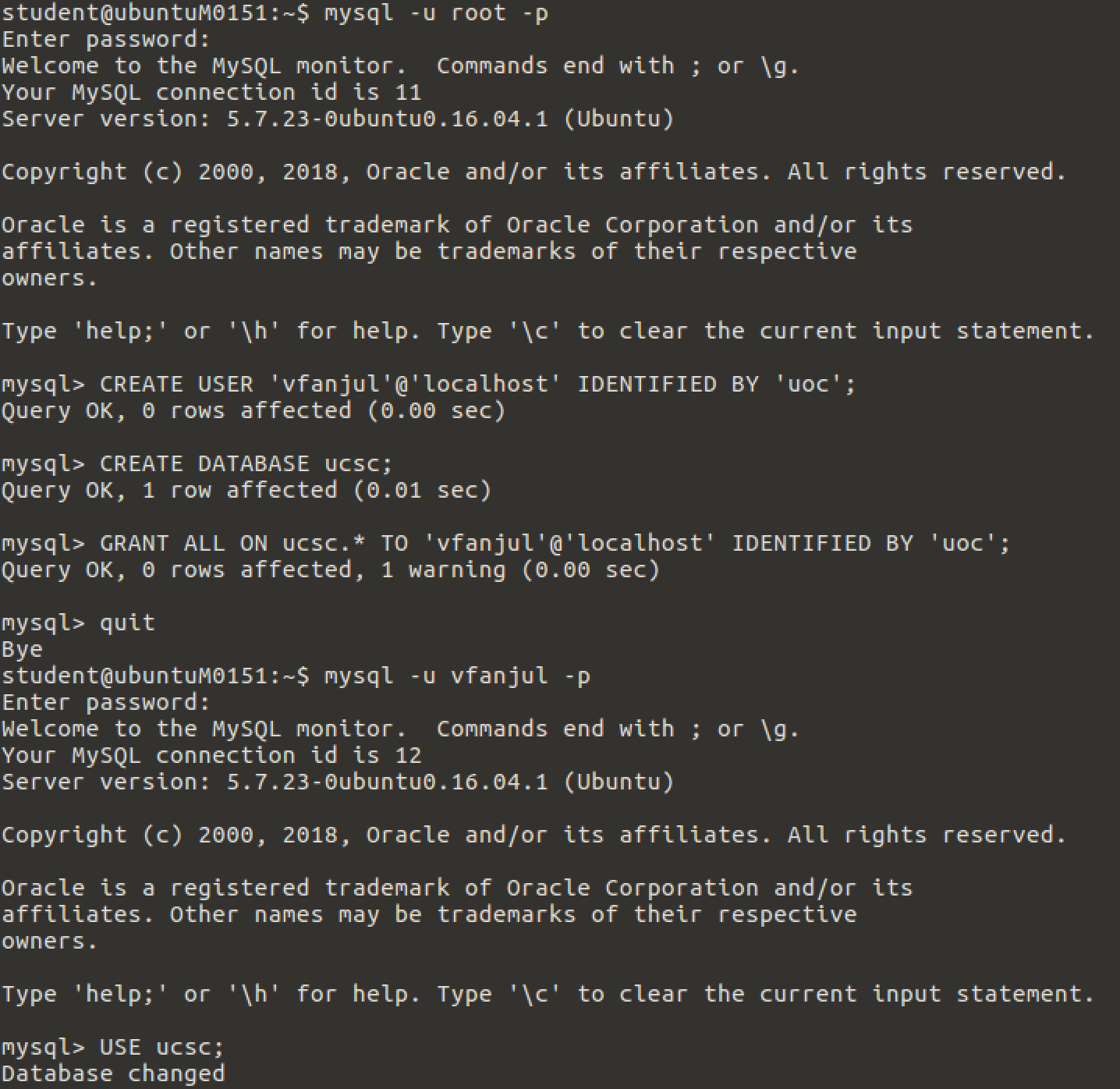
|  |
| --- |
| **PEC2. Bases de datos relacionales y servidores web** |

# Ejercicio 1 – Descripción de los catálogos de genes (20%)

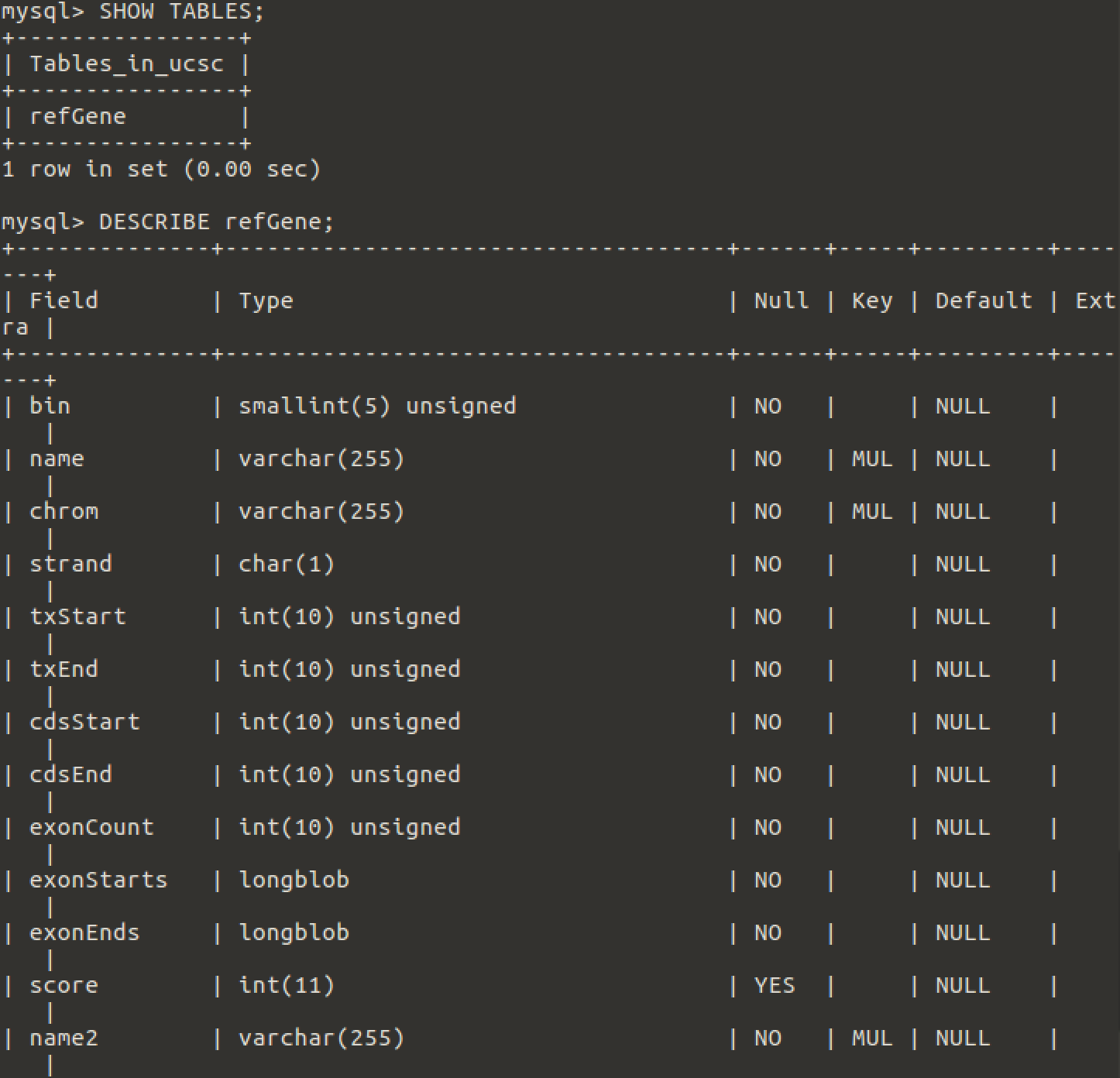
Antes de nada, descargaremos los archivos del servidor y descomprimiremos el fichero “refSeq”. Debido a la limitación de páginas del documento de resultados, mostraremos el procedimiento para el genoma humano.



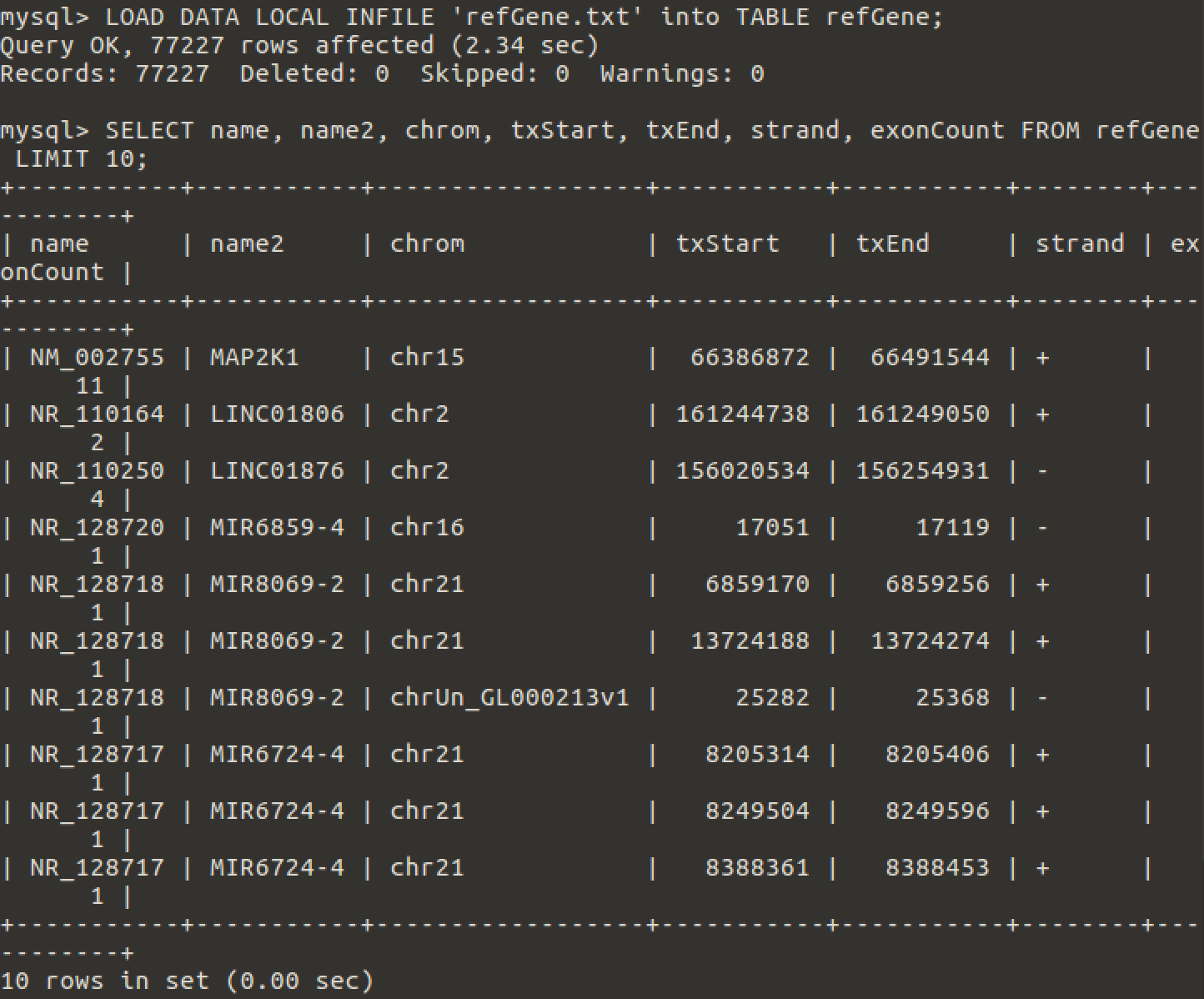
Entramos en “mysql” como administrador, creamos un usuario “vfanjul”, una base de datos “ucsc” y damos permiso al usuario a editar esa base de datos. A continuación, cambiamos de usuario y cargamos la base de datos.



A continuación, importaremos la estructura descargada de UCSC y analizaremos las tablas que contiene (“refGene”) y los atributos de estas.



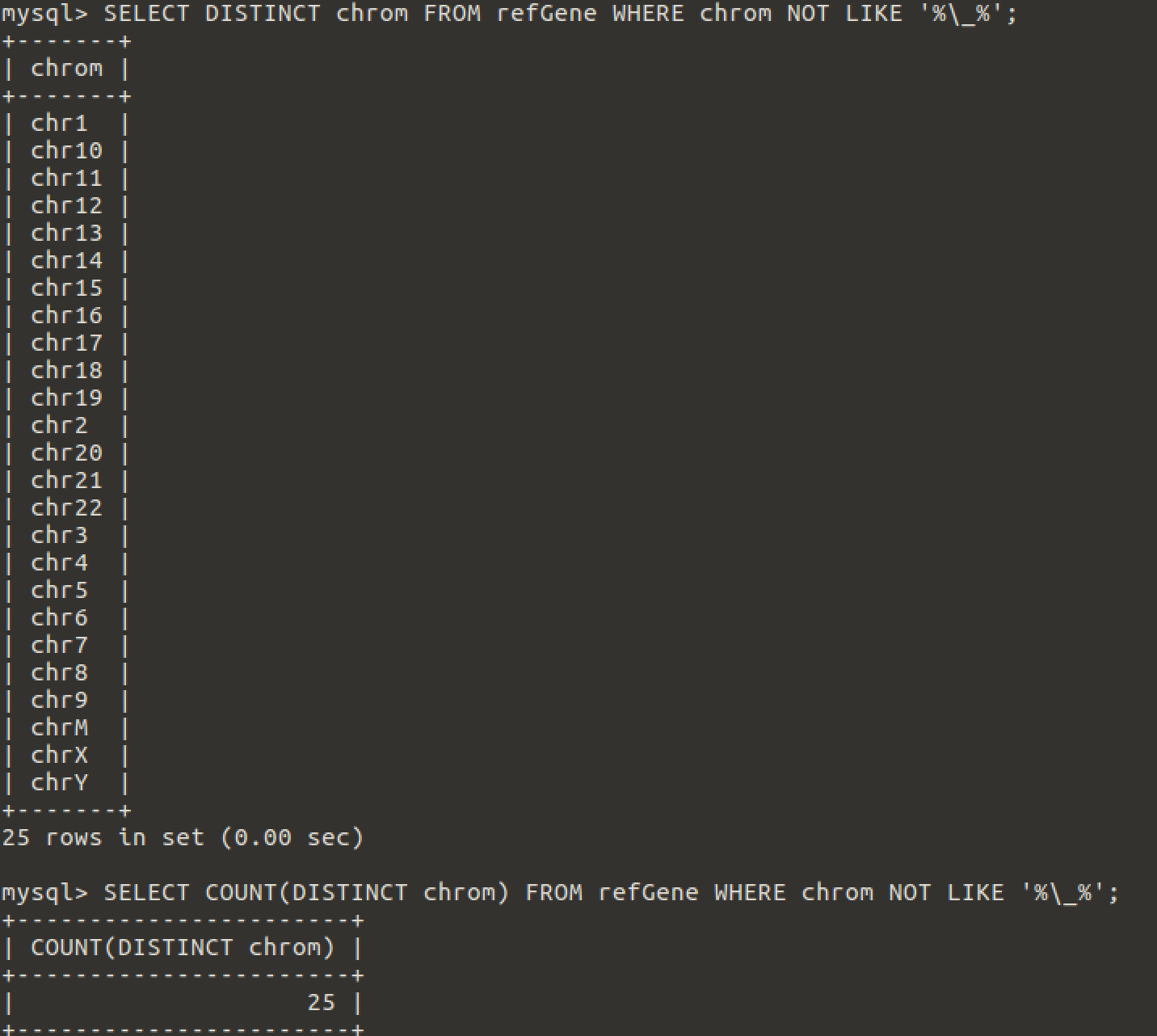
Después poblaremos la tabla “refGene” con los datos de “refSeq.txt” y realizaremos una consulta para comprobar que todo es correcto.



Posteriormente procederemos a realizar los recuentos solicitados:

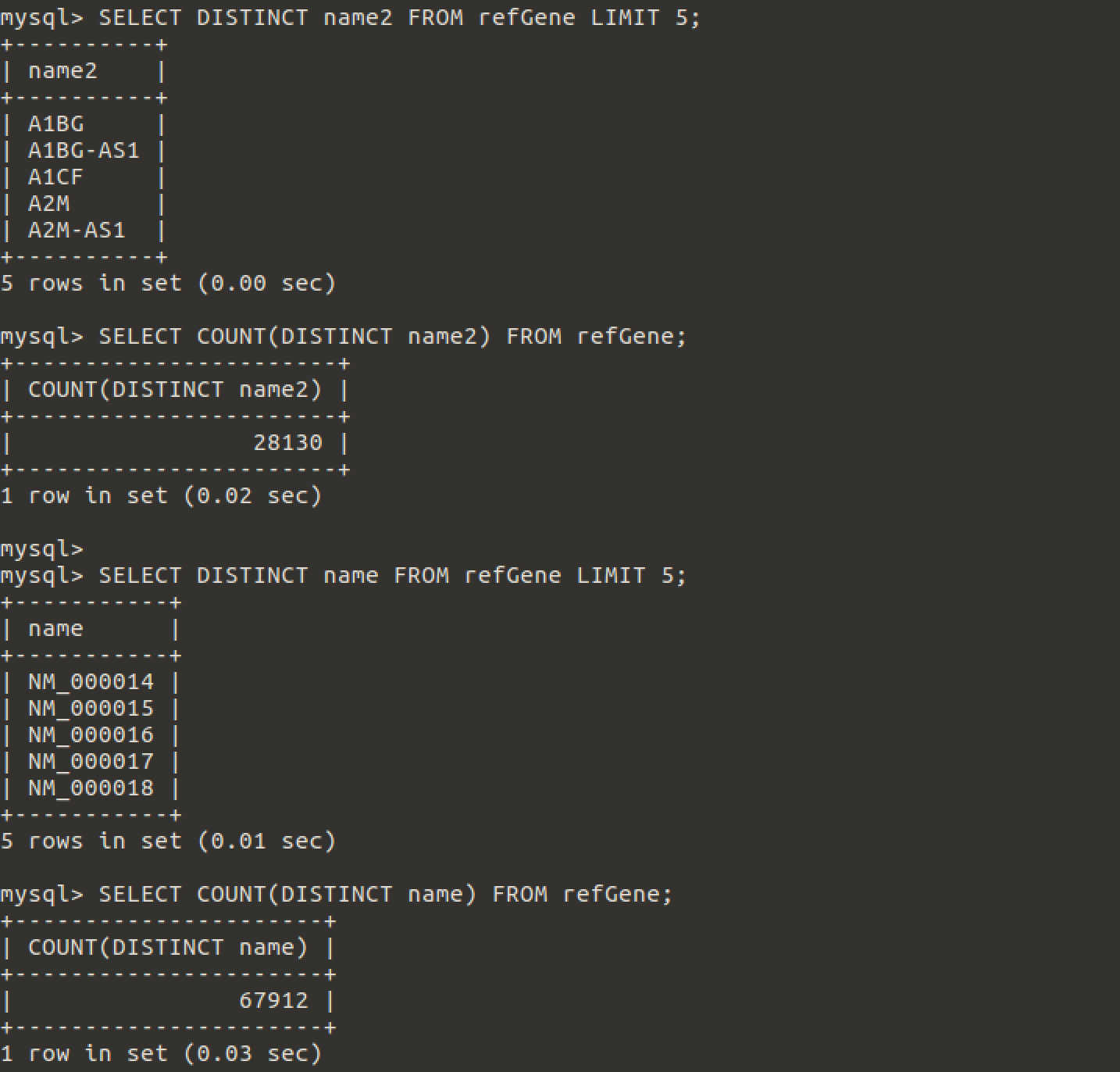
* Número de cromosomas distintos (incluyendo los sexuales y el mitocondrial): 25

Debemos excluir otras construcciones (tienen “\_”)



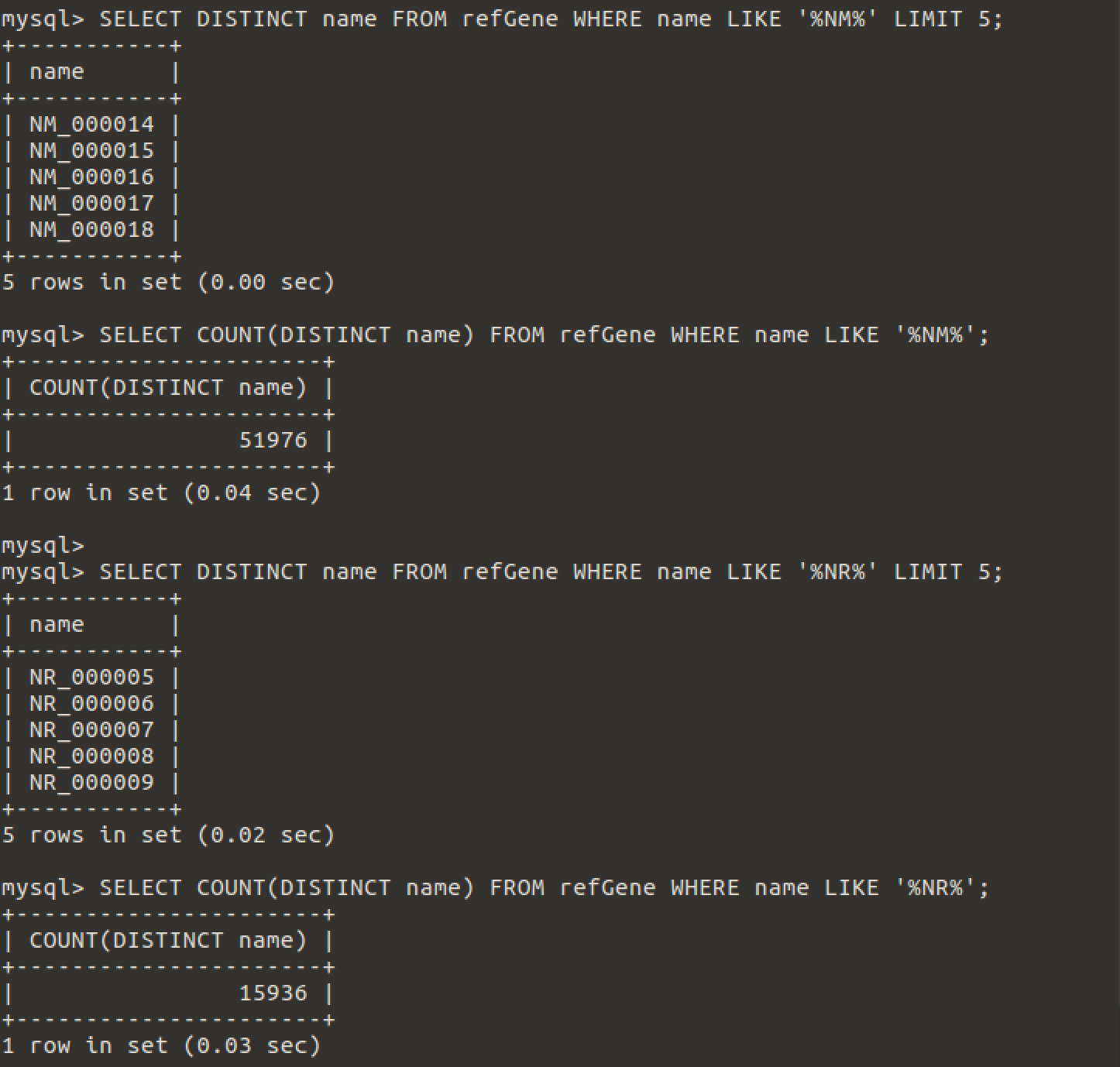
* Número de genes y tránscritos distintos: 28130 y 67912

Debemos quitar los duplicados de las columnas “name2” y “name”



* Número de tránscritos codificantes y no codificantes: 51976 y 15936

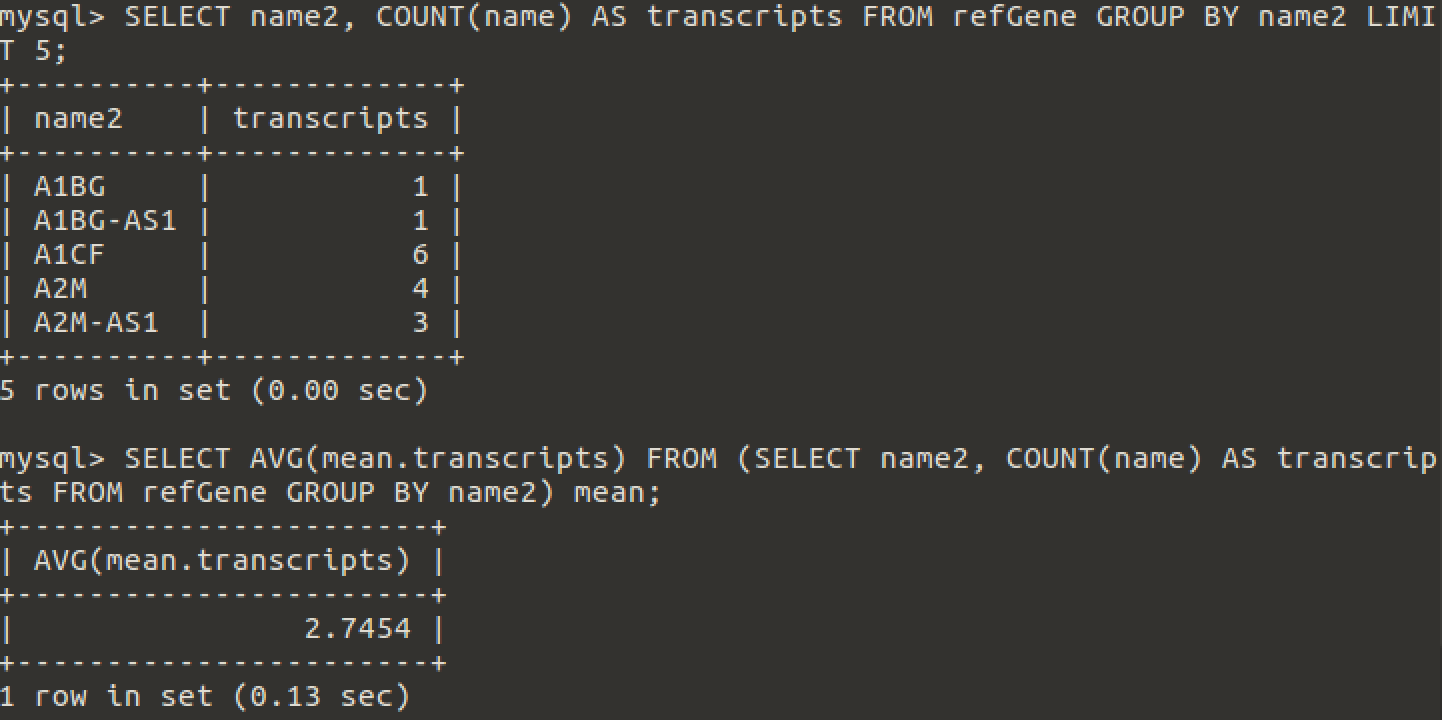
Filtramos los transcritos que contienen “NM” o “NR”



* Número de tránscritos por cada gen (en promedio): 2,7454

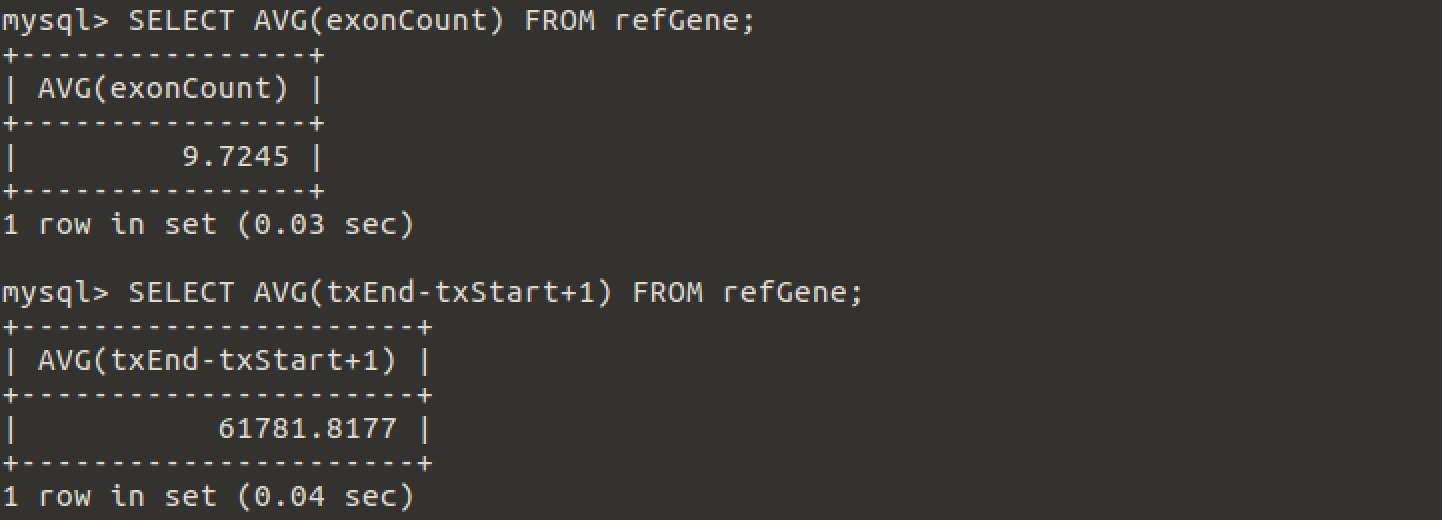
Agrupamos y contamos el número de transcritos por gen

Hacemos una subconsulta calculando la media de los grupos



* Número de exones y de nucleótidos por tránscrito (en promedio): 9,7245 y 61781,8177

Calculamos la media para el campo “exonCount” y para la longitud de los transcritos, que calculamos restando las posiciones de inicio y fin más 1.



Posteriormente, repetimos estas operaciones en el resto de organismos, tras descargar los archivos del servidor y generar diferentes tablas



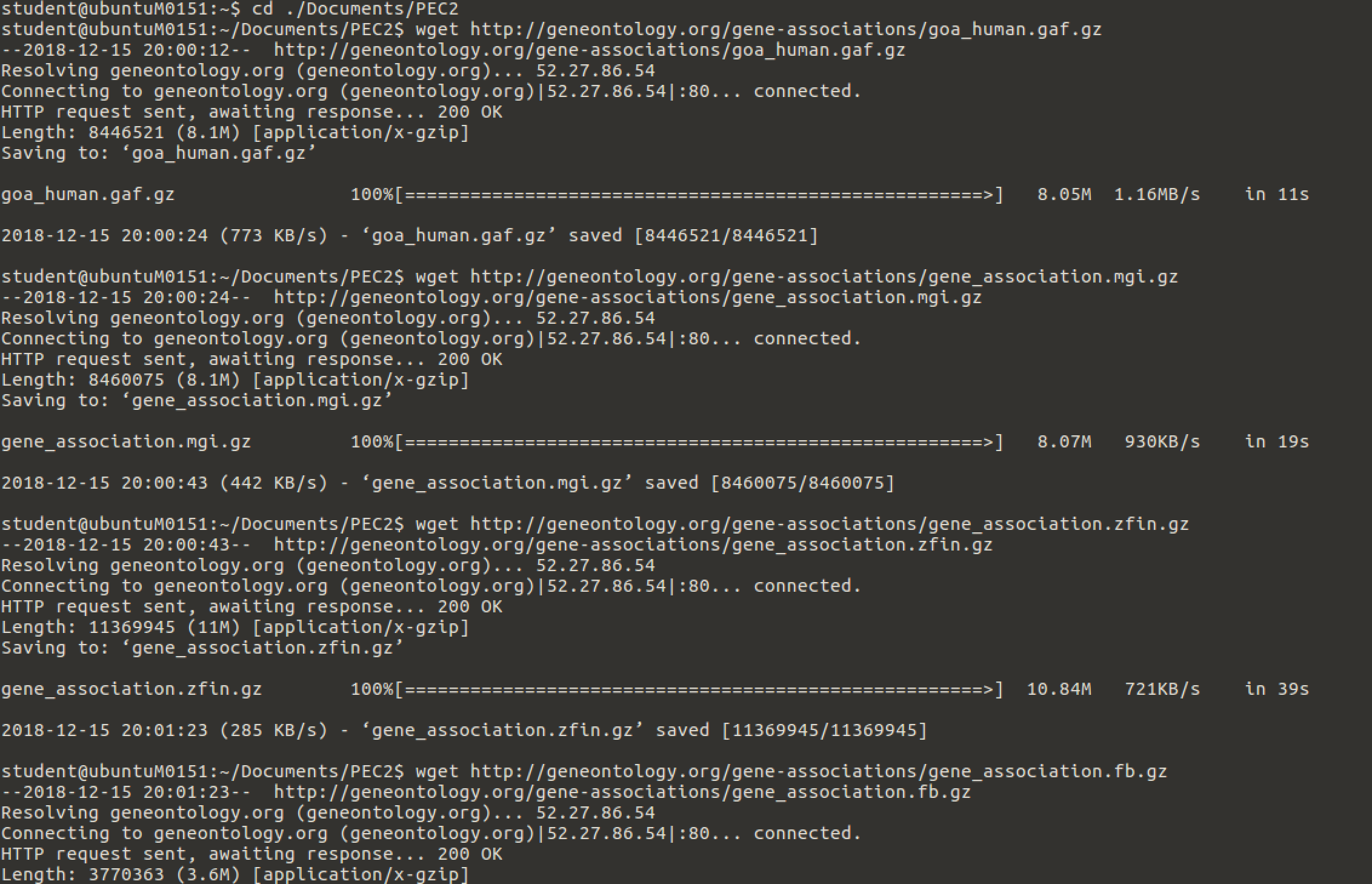
Finalmente mostramos los resultados:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Genoma** | Crom | Gen | Tr | Cod Tr | NC Tr | Tr/Gen | Ex/Tr | Nt/Tr |
| *H. sapiens (hg38)* | 25 | 28130 | 67912 | 51976 | 15936 | 2.74536 | 9.72449 | 61781.8 |
| *M. musculus (mm10)* | 21 | 24984 | 41449 | 35328 | 6121 | 1.70453 | 9.57608 | 50048.9 |
| *D. rerio (danRer10)* | 25 | 15395 | 16268 | 15774 | 494 | 1.09068 | 9.0424 | 27000.7 |
| *D. melanogaster (dm6)* | 7 | 17101 | 34114 | 30480 | 3634 | 2.11496 | 5.22149 | 9336.05 |

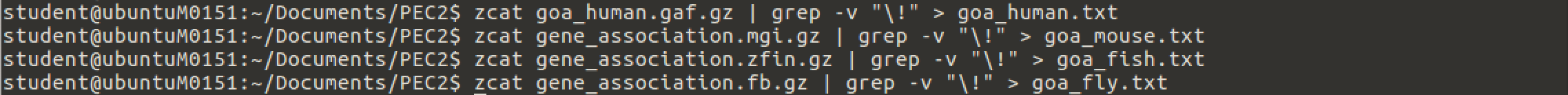
Se observan ligeros cambios en los números respecto a la PEC1 de hace unos meses. Esto se debe a que el proceso de revisión y anotación de genomas es dinámico, está en constante cambio.

# Ejercicio 2 – Anotación funcional de los genes (30%)

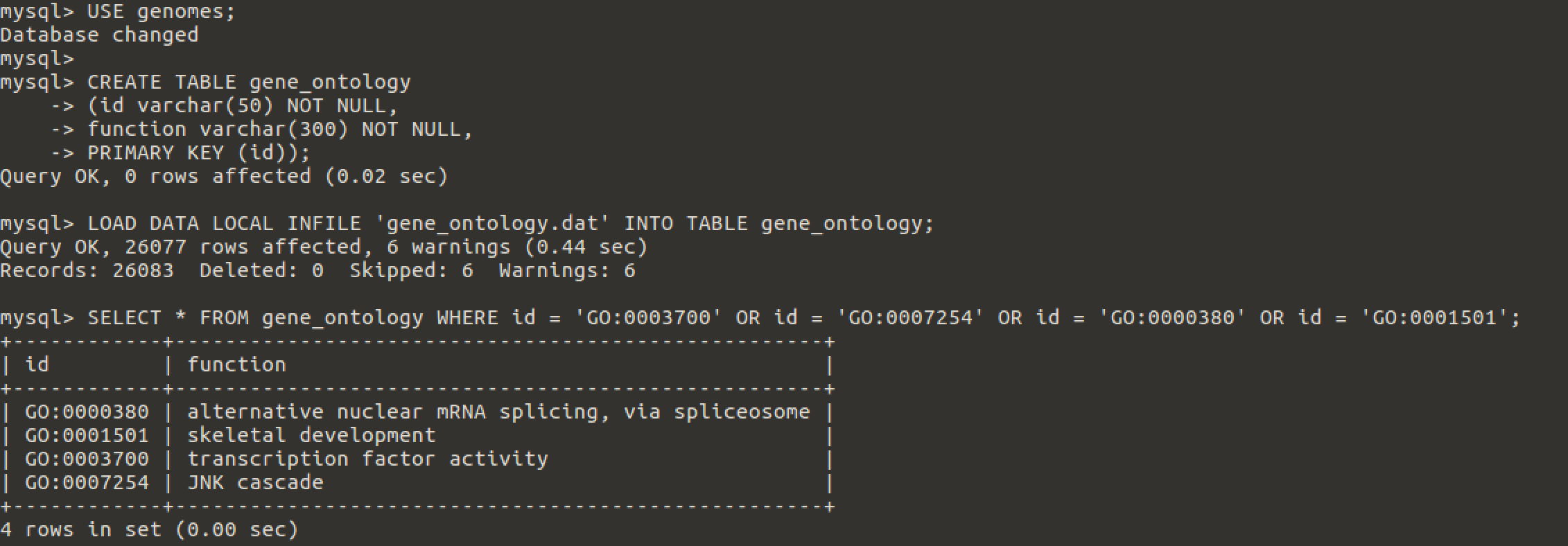
En primer lugar, descargamos los ficheros del servidor (o recuperamos los de la PEC1).



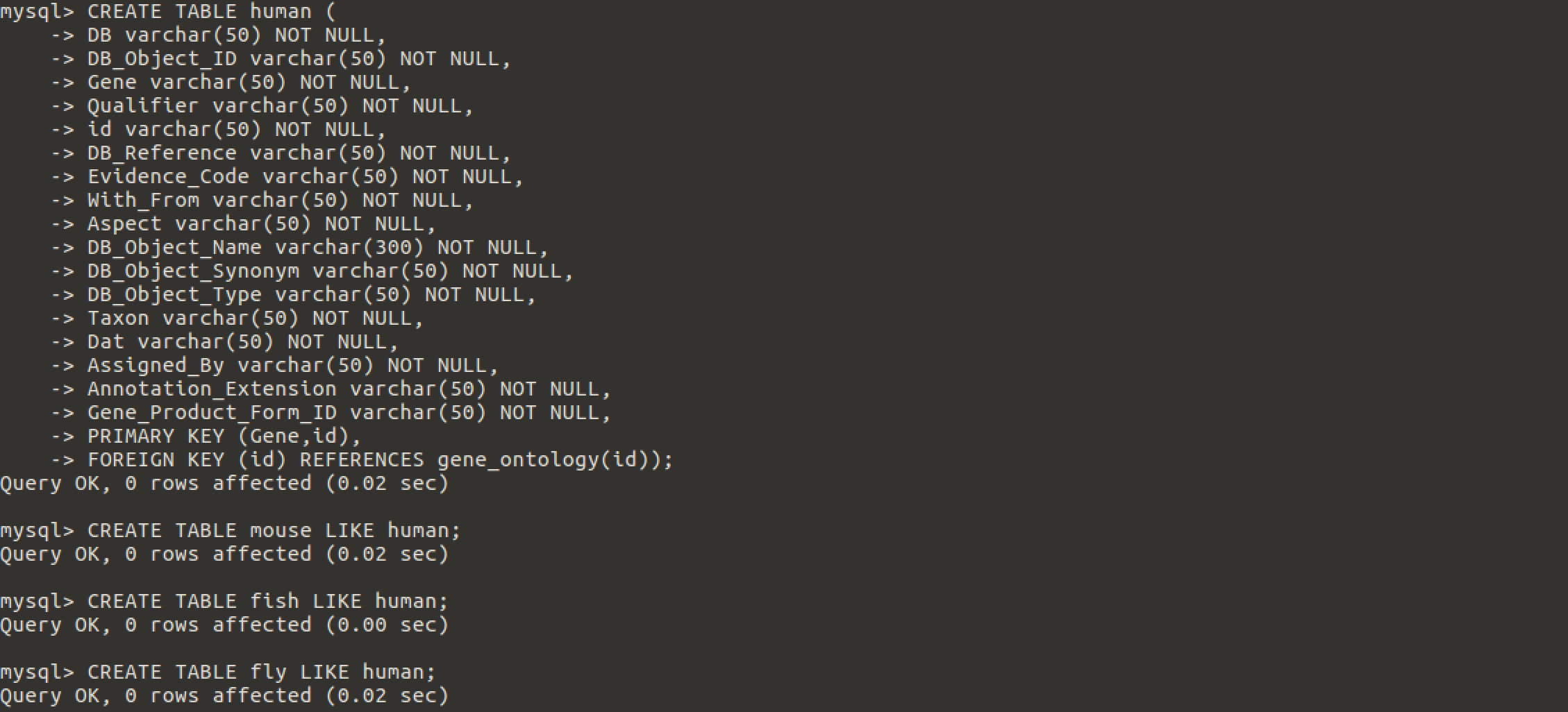
A continuación, debemos descomprimirlos y eliminar las filas que no contienen información sobre genes (comienzan por “!”)



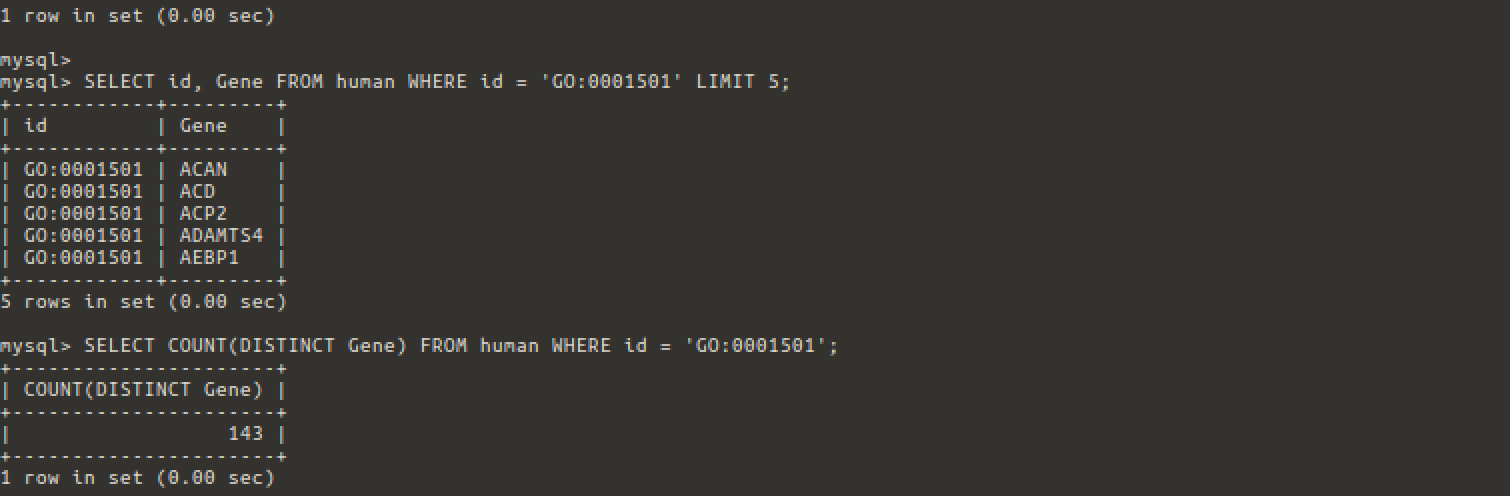
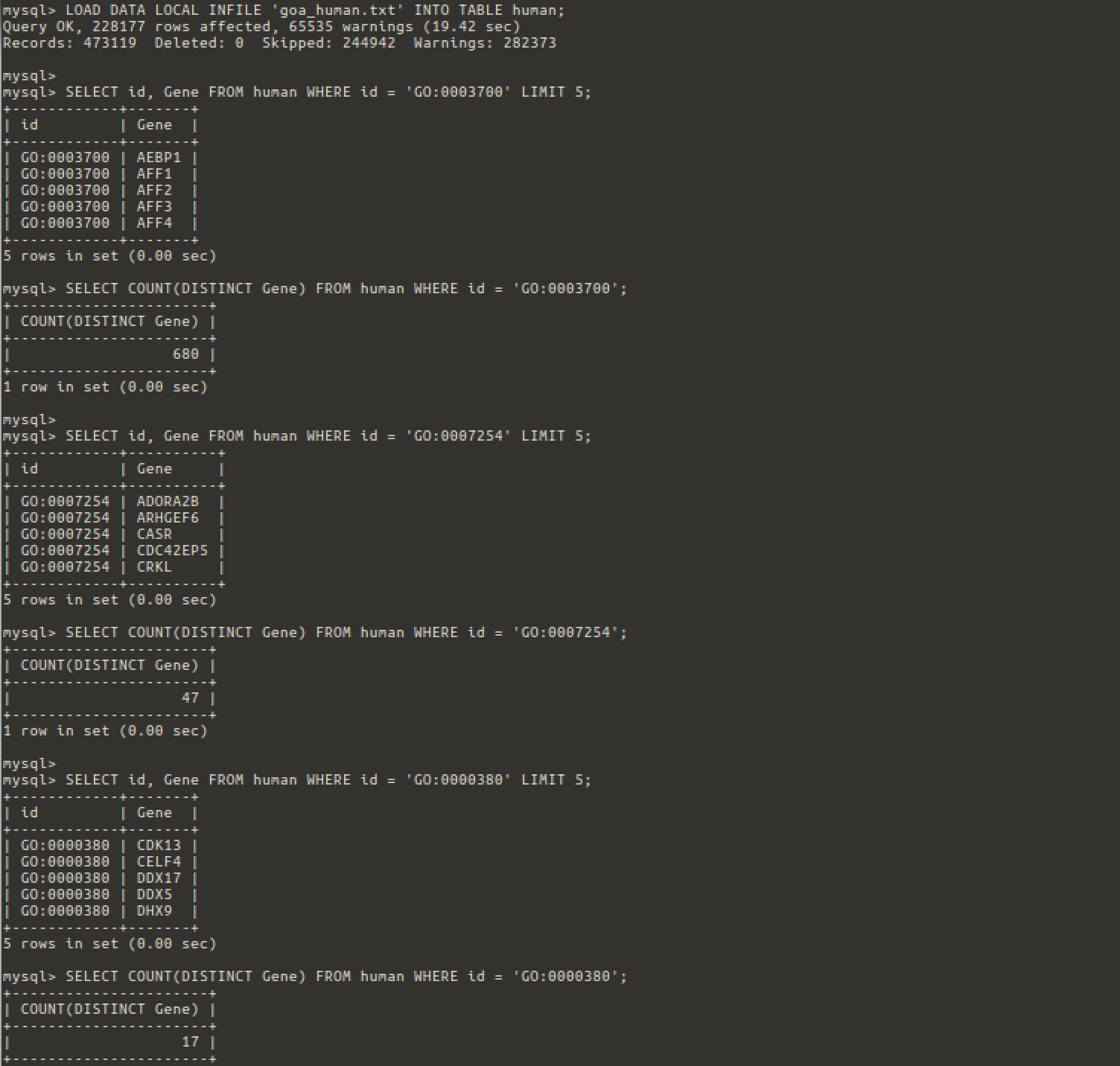
Entramos en “mysql” con nuestro usuario y abrimos la base de datos “genomes” del ejercicio anterior. Creamos una tabla “gene\_ontology” con los atributos “id” (que será la clave) y “function”, y la poblamos con los datos de “gene\_ontology.dat”. Realizamos una consulta buscando los ids de las rutas seleccionadas en la PEC1 para factores de transcripción, JNK, splicing y desarrollo del esqueleto. Comprobamos que obtenemos las descripciones correctas.



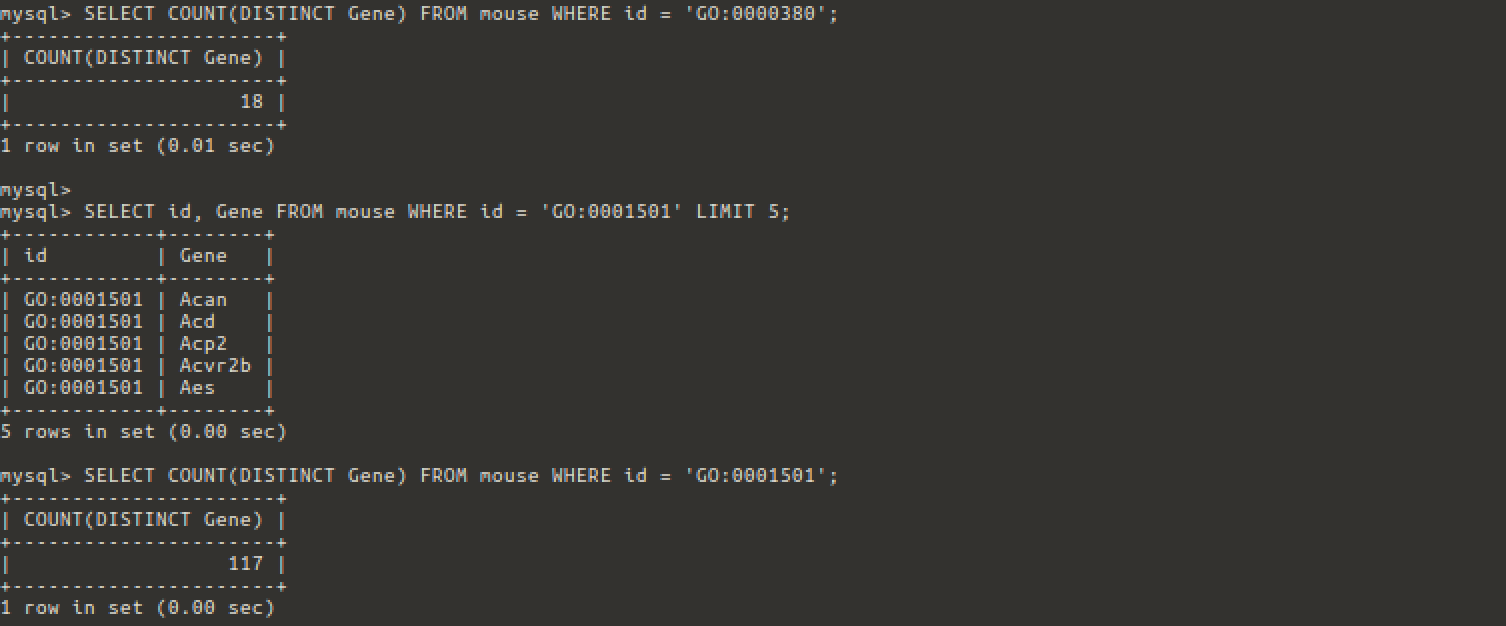
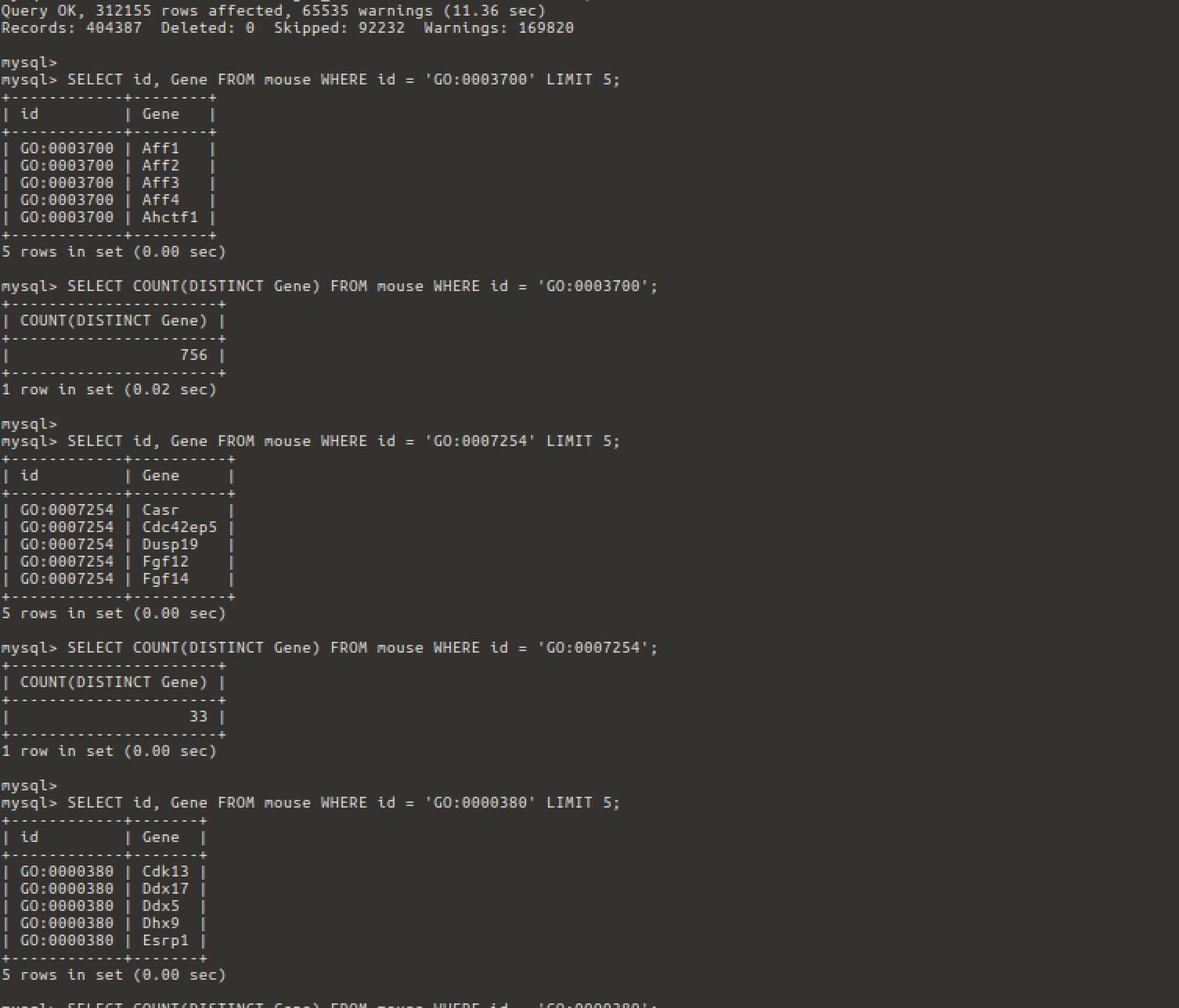
Posteriormente crearemos una tabla “human” con los atributos de los archivos “gaf”. Por comodidad cambiaremos el identificador de GO por “id” y el de los genes por “Gene”. Tengamos en cuenta que la clave en esta tabla debe ser la combinación de “id” y “Gene” para que tratemos con registros únicos. Además, podemos relacionarla con la tabla “gene\_ontology” estableciendo “id” como clave foránea. Después copiaremos la estructura de la tabla para crear las tablas “mouse”, “fish” y “fly” y así estudiar las otras especies.



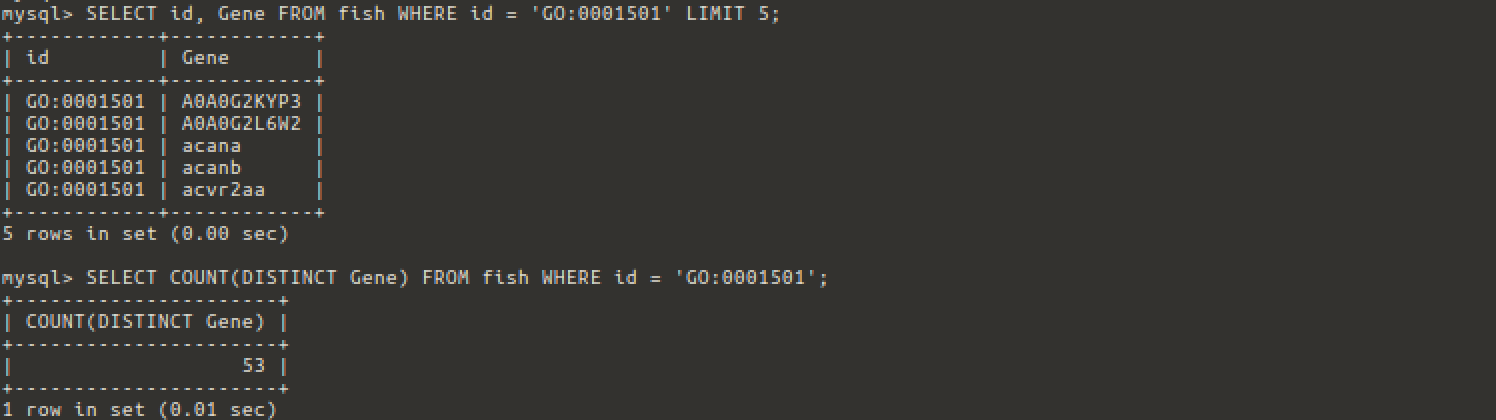
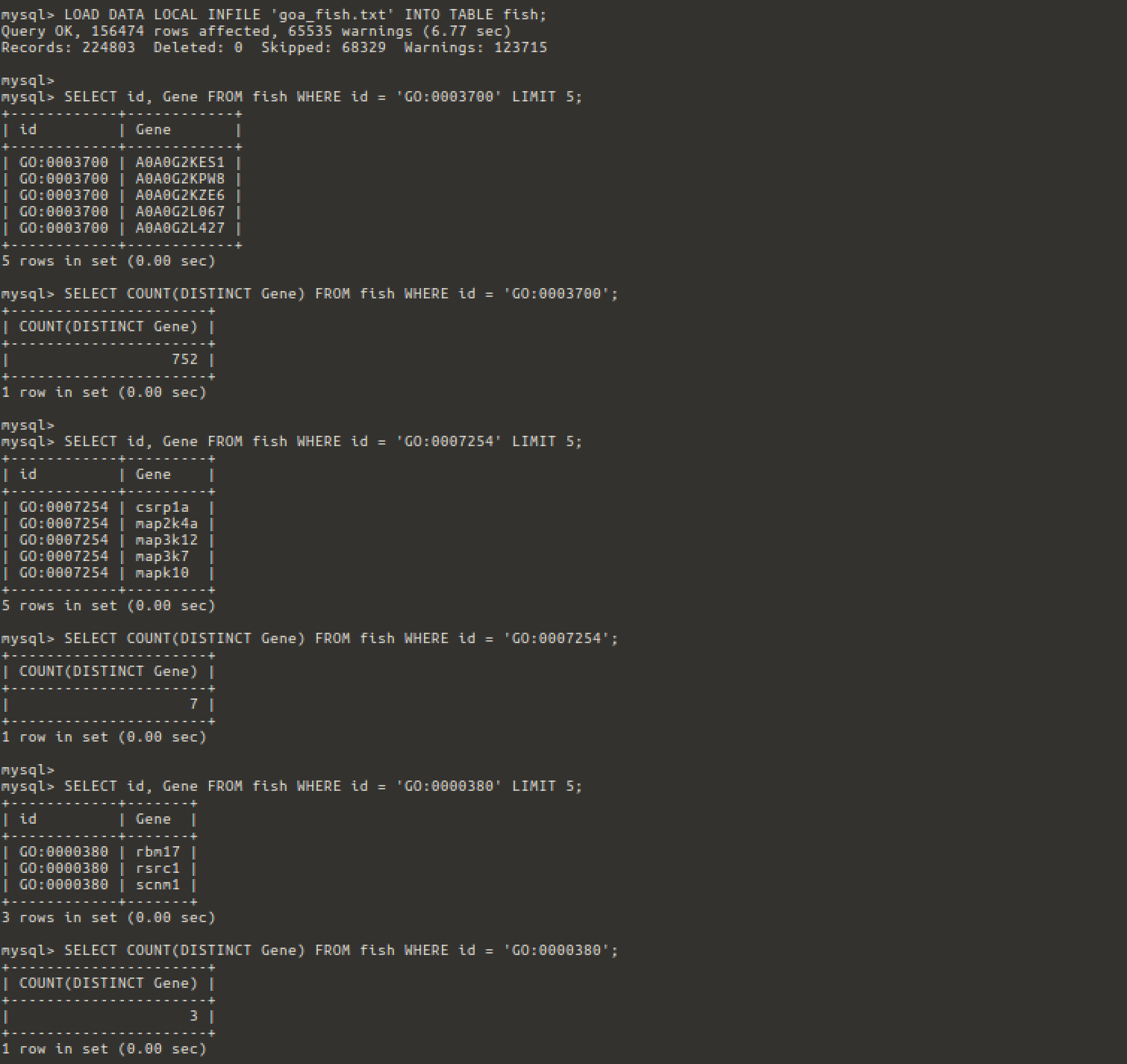
Empezamos el estudio poblando la tabla “human” con los datos del archivo “gaf”. Podríamos relacionarla con la tabla “gene\_ontology” a través de “JOIN” y quedándonos con los valores comunes con “ON human.id = gene\_ontology.id”. Esto nos permitiría buscar con la condición “function” los registros que pertenezcan a las rutas que queremos analizar. No obstante, puesto que conocemos los GO ids, los buscaremos directamente con “WHERE id =”. Para el recuento deberemos eliminar posibles duplicados con “COUNT(DISTINCT Gene)”, aunque no debería haber ninguno. Repetimos el proceso para las 4 rutas.



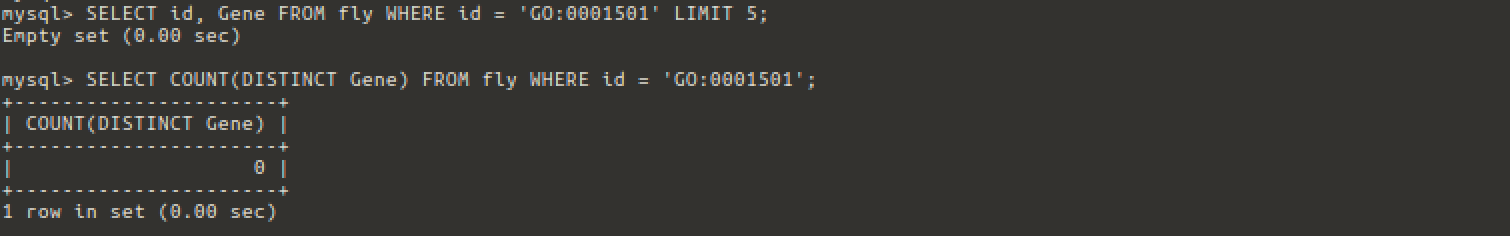
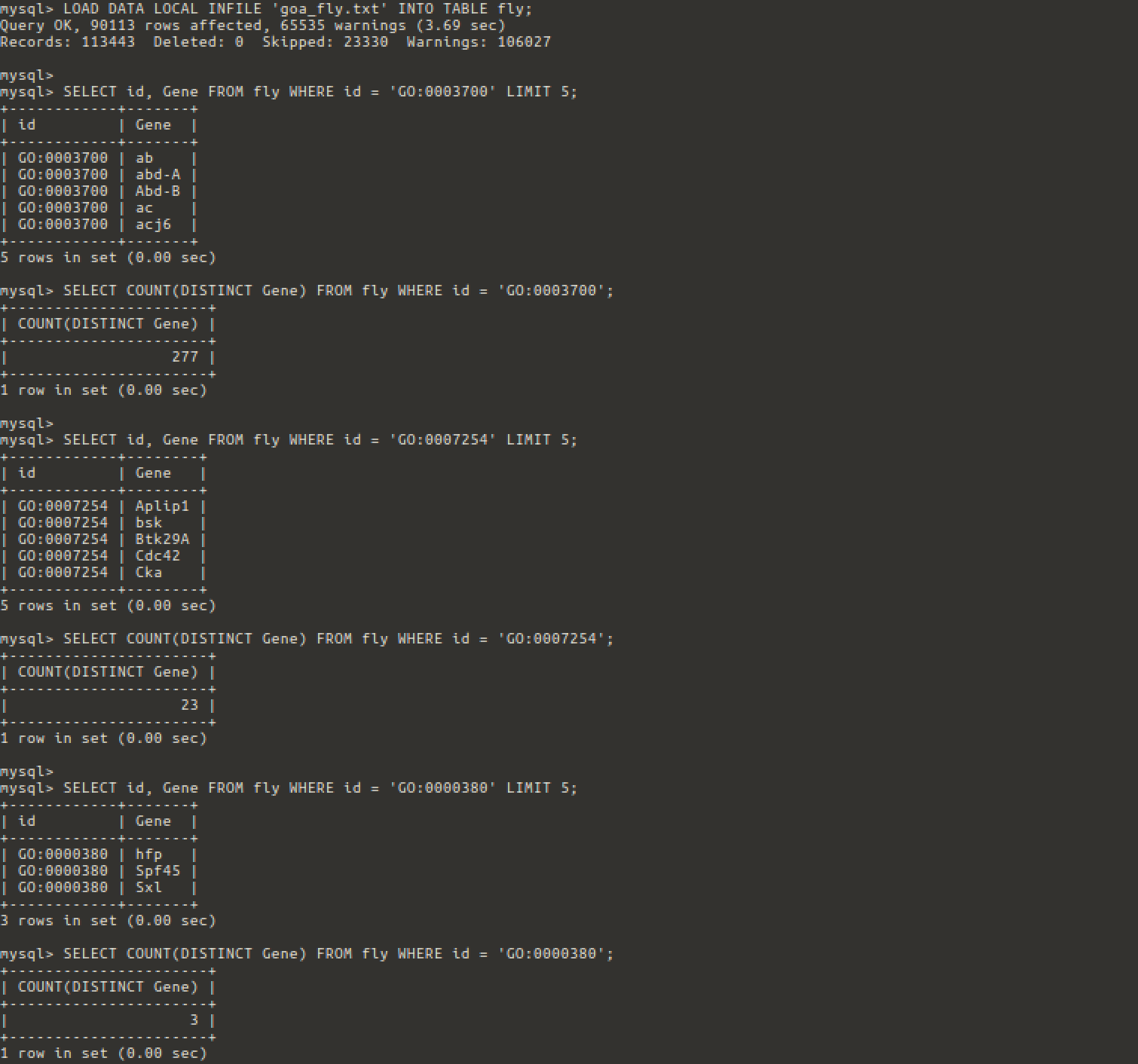
Repetimos el procedimiento con la tabla y archivo de ratón.



A continuación, estudiaremos el pez cebra.



Y por último estudiaremos la mosca del vinagre.

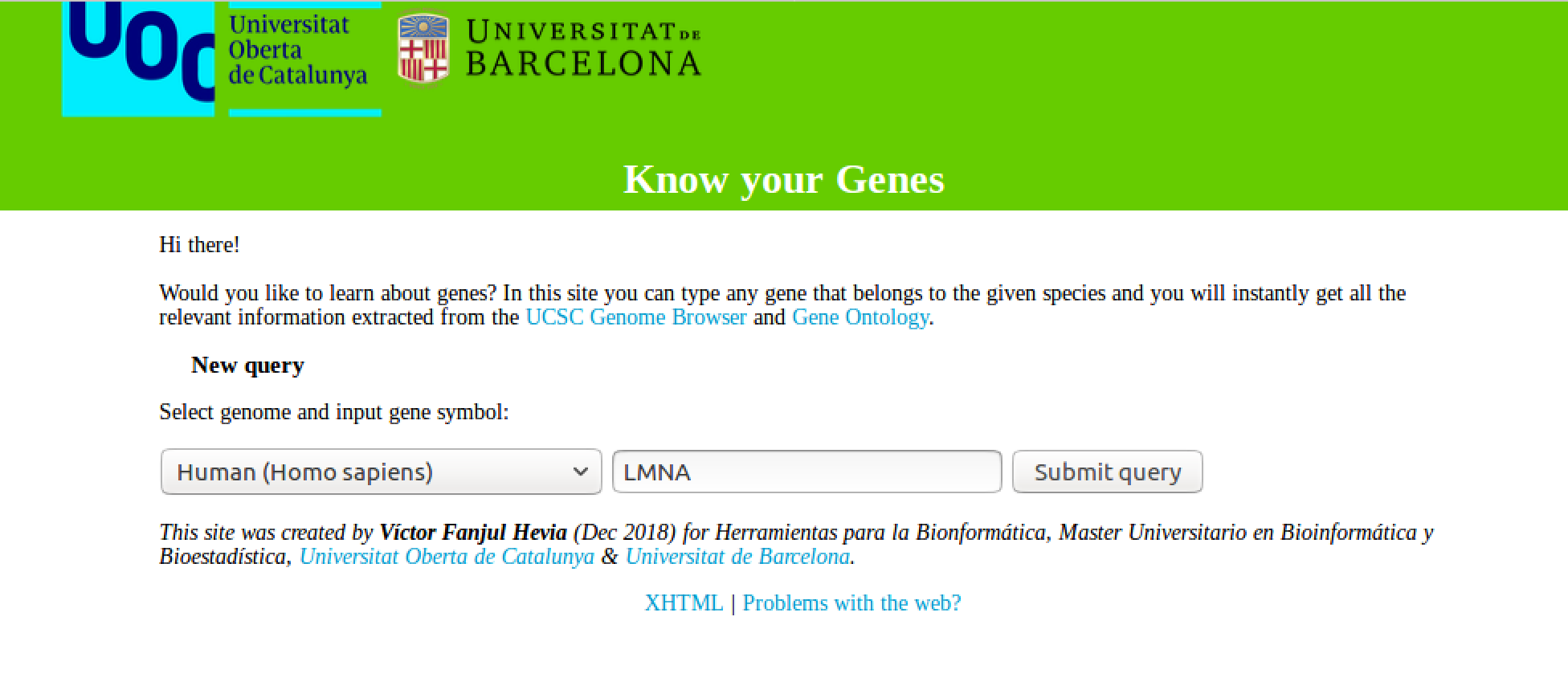


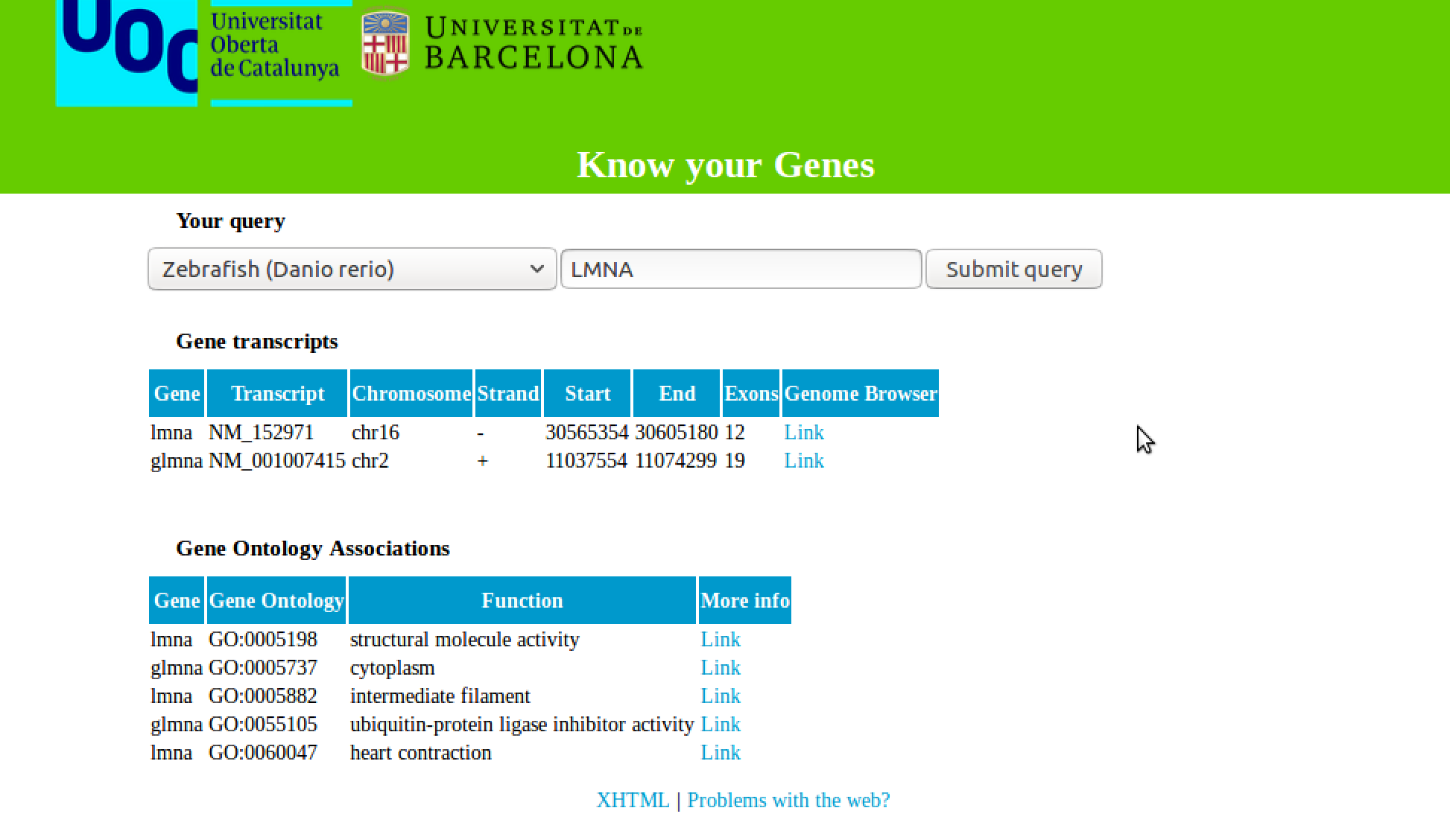
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | TF activity | JNK | Splicing | Skelental dev. |
| *H. sapiens (hg38)* | 680 | 47 | 17 | 143 |
| *M. musculus (mm10)* | 756 | 33 | 18 | 117 |
| *D. rerio (danRer10)* | 752 | 7 | 3 | 53 |
| *D. melanogaster (dm6)* | 277 | 23 | 3 | 0 |

Los resultados son muy similares a los obtenidos en la PEC1. Las variaciones se deben a actualizaciones en los archivos “gaf” de cada especie (se han descargado de nuevo).

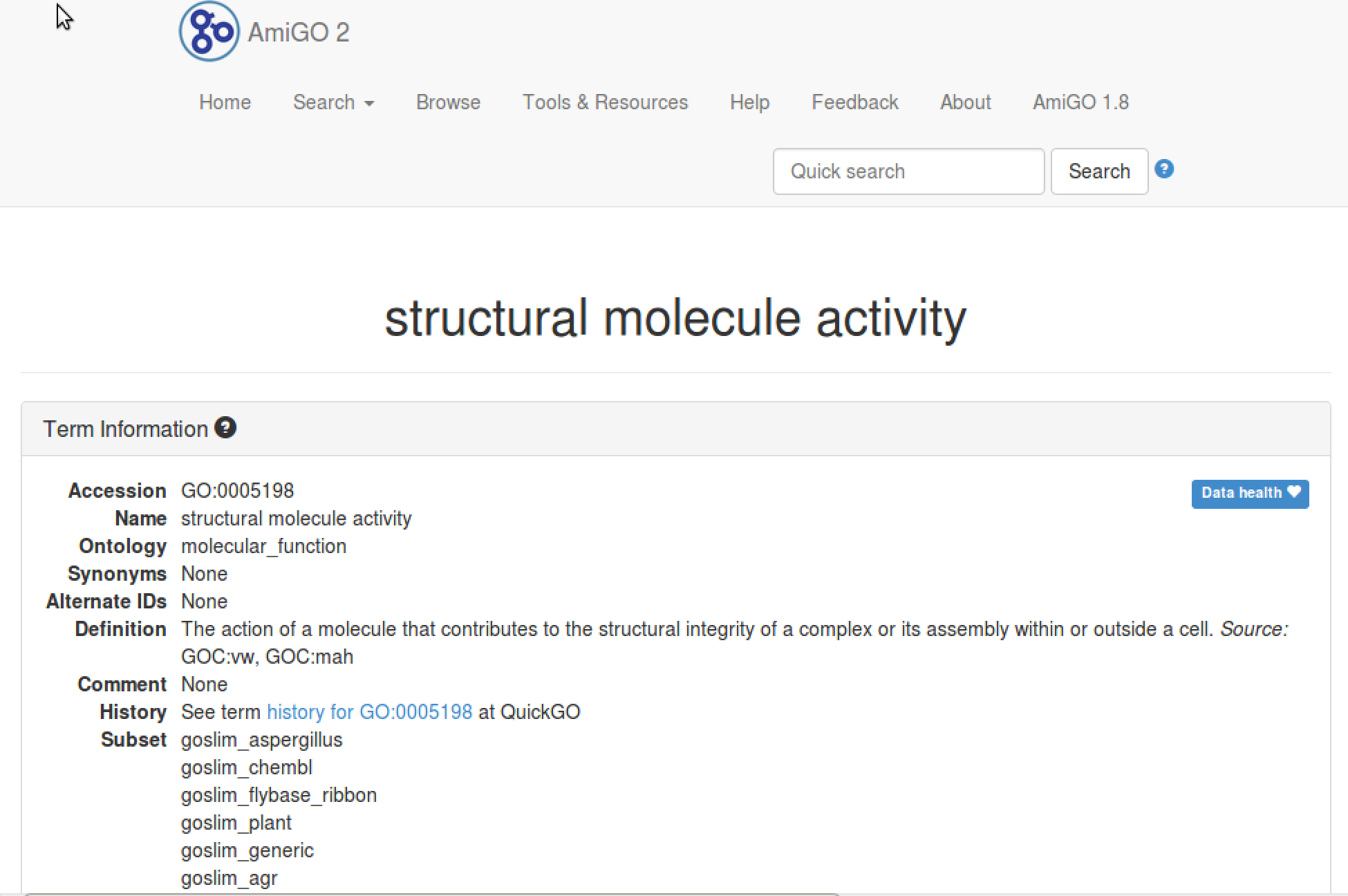
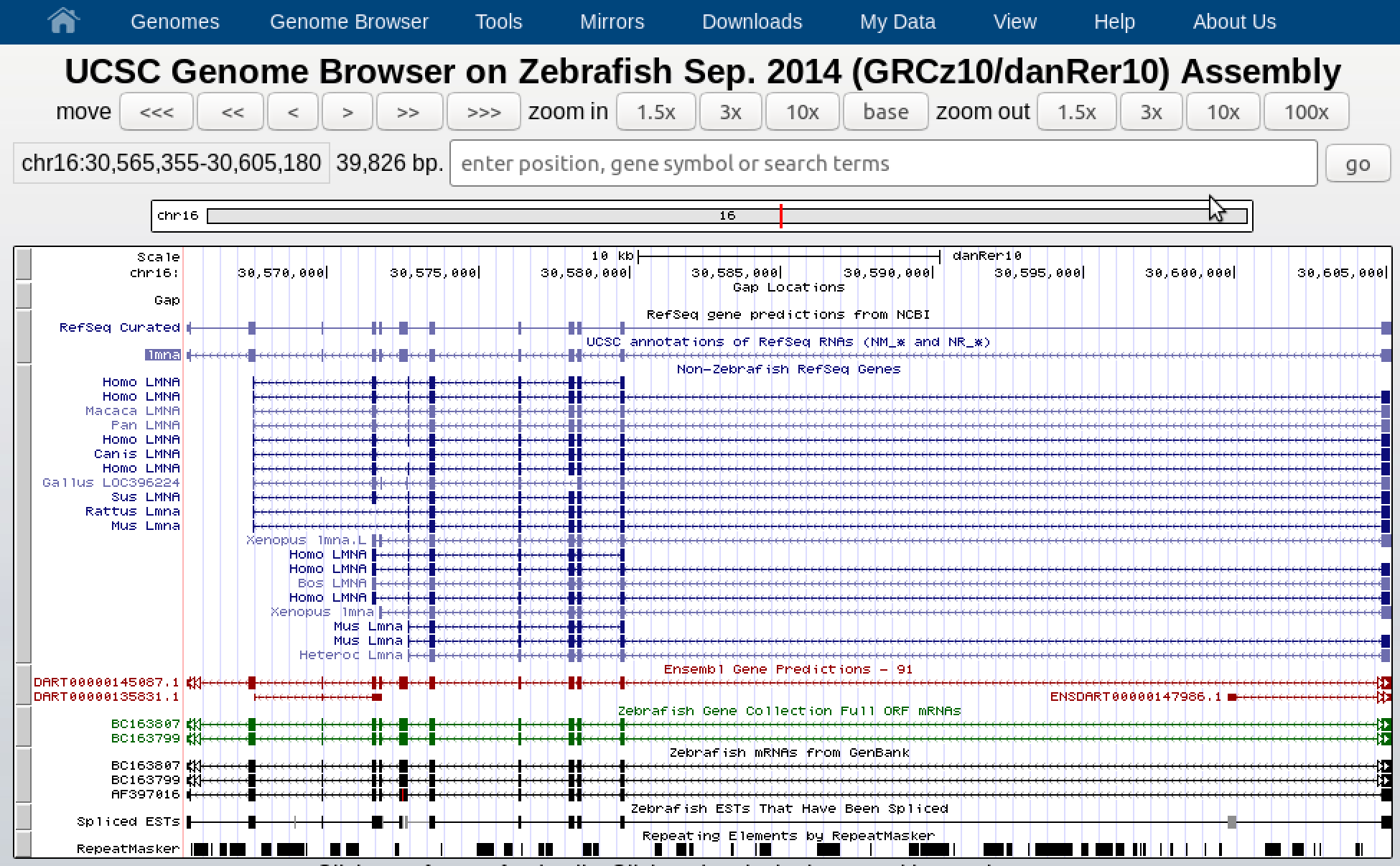
# Ejercicio 3 – Servicio web de consulta de genes (50%)

Utilizando la base de datos “genomes” generada en los ejercicios anteriores se ha creado un servicio web para realizar consultas. El usuario elige un genoma de la lista, teclea un gen y obtiene información relacionada con los transcritos y las anotaciones “Gene Ontology” asociadas al gen.



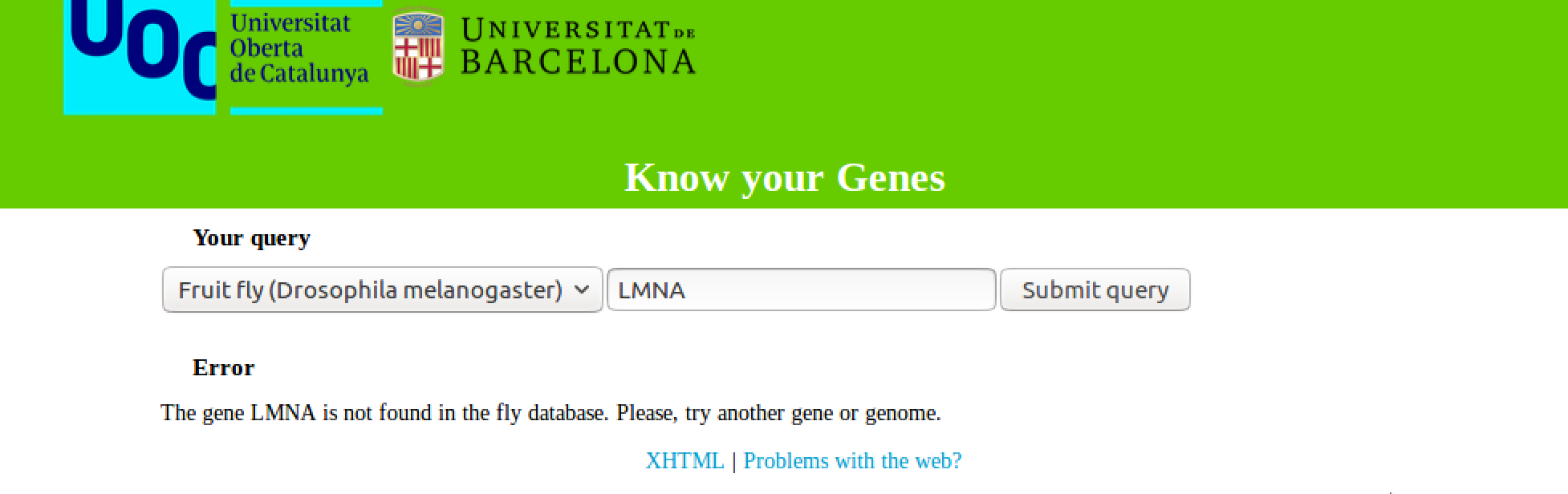


Además, pulsando en los enlaces, puede obtener más información sobre los transcritos y sobre las categorías en las webs oficiales.

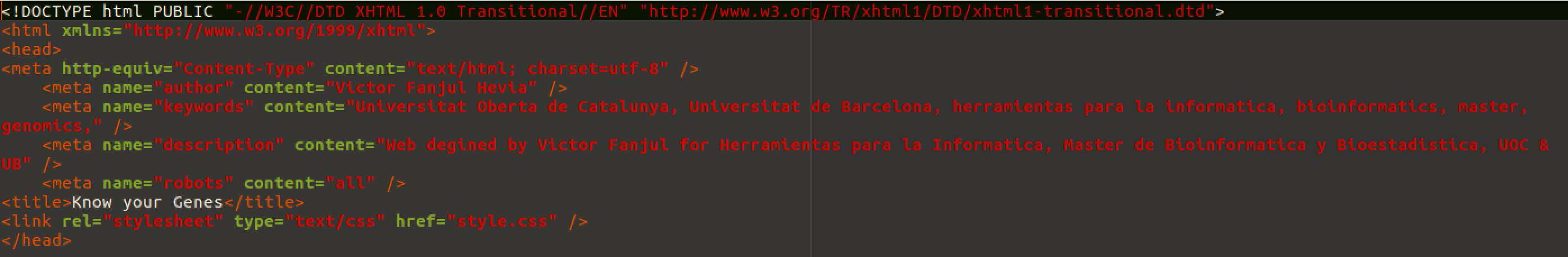


Si el usuario deja el campo gen en blanco o introduce un espacio, el sistema avisará de que hay un error. Lo mismo si se introduce un gen erróneamente o que no está presente en el genoma indicado. Nótese que, tras la consulta, quedan preseleccionados el genoma y el texto utilizados.





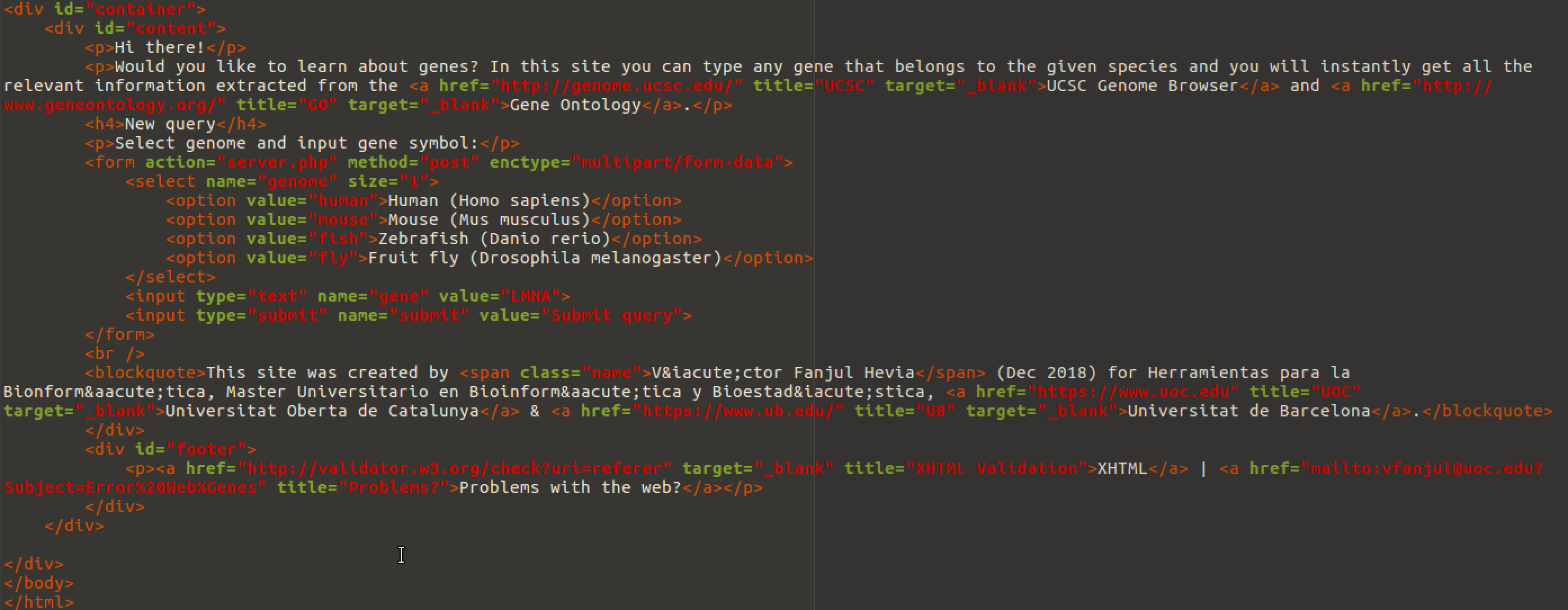
Para llegar a este resultado hemos tenido que generar tres ficheros: “index.html”, “style.css” y “server.php”. El primero contiene la estructura de la web. En la cabecera se introducen diferentes metadatos y el link a la hoja de estilo “style.css”.



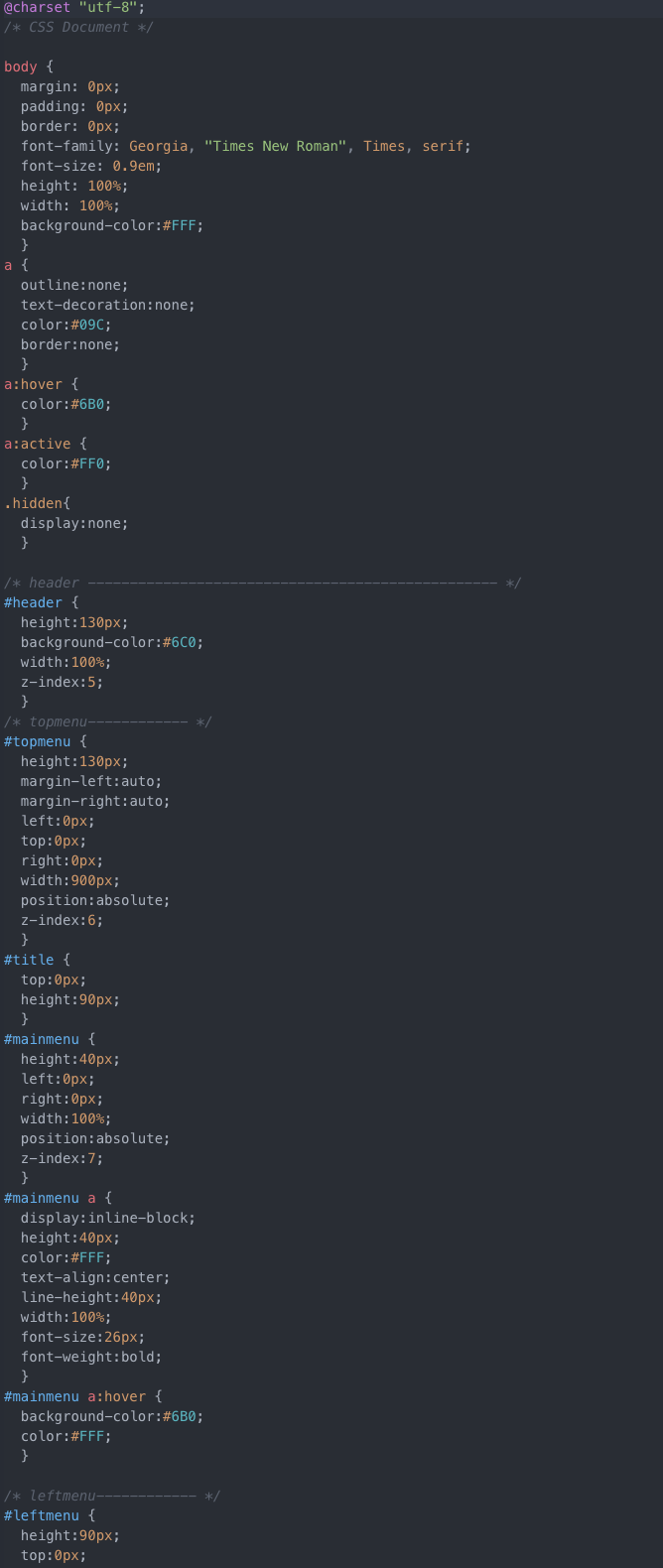
El cuerpo se divide en 2 secciones principales, el encabezado y el contenedor, que engloba al contenido y al pie. El encabezado (fondo verde) contiene links a las instituciones y a la página de inicio.



En el contenedor (fondo blanco) podemos encontrar el cuerpo de la web, donde se explica lo que es, se presenta el formulario y luego se añade información de para qué se ha realizado esta página. Al final, al pie se añaden links para comprobar la calidad del código (no funciona porque la web está *offline*) y para escribir al gestor si hay algún comentario.



La hoja “style.css” personaliza las fuentes, colores, tamaños, alineamientos, animaciones… de las capas, tipos y clases que encontramos en “index.html” y también en la web generada por “server.php” tras la consulta.



Finalmente, el archivo “server.php” se encarga de conectar con la base de datos “genomes” y de generar una página web con los resultados. En esta primera captura se muestra la web que se genera cuando hay un error, que contiene una estructura casi idéntica a la de “intex.html”, con excepción de que la capa de contenido presenta el mensaje de error “$message”.



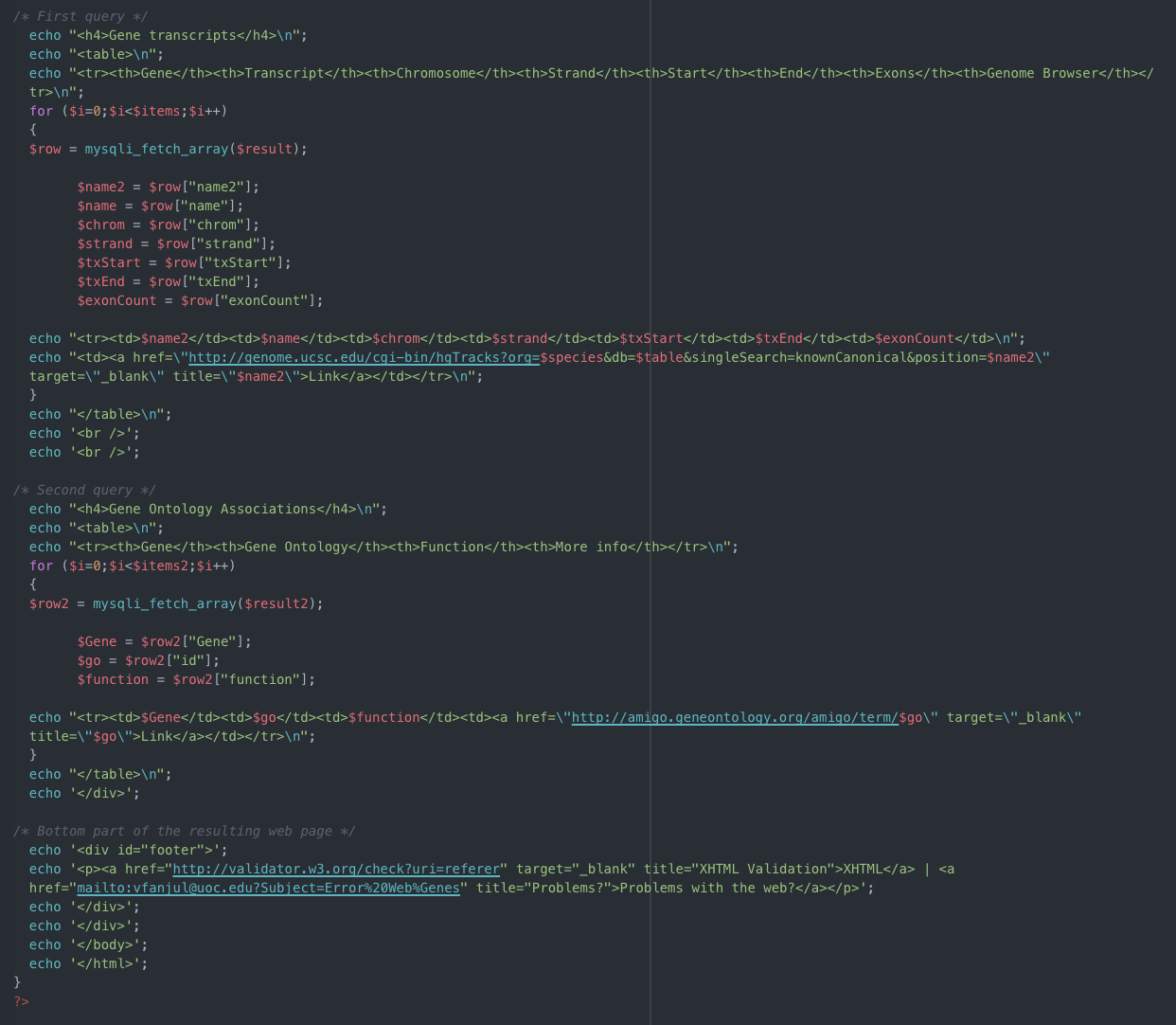
En esta segunda captura se muestra cómo se generan las variables a partir de los datos del formulario (y del momento de la consulta), cómo se crean mensajes de error si la consulta está vacía o es “ “ y cómo se eliminan espacios y caracteres que puedan causar problemas en la consulta. Después, se introducen las variables necesarias para conectar con “MySQL” (usuario, nombre de la base de datos y las tablas en función del genoma elegido). A continuación, se conecta con el servidor y se realizan las consultas oportunas. Si no se obtienen resultados, se genera un mensaje de error.



Si se obtienen resultados en la búsqueda, se genera una nueva página web muy similar a “index.html”. Nótese que en el formulario aparecerán preseleccionados el genoma y el gen que han sido utilizados en la consulta.



La primera consulta genera una tabla en la que encontramos la información relativa a los tránscritos del gen y de genes con el mismo patrón de letras buscadas: Nombre, tránscrito, localización, número de exones y un enlace al explorador de genomas de UCSC, donde lo veremos representado gráficamente. La segunda búsqueda genera otra tabla con las distintas anotaciones “Gene Ontology” descritas para ese gen o genes, con una descripción y un enlace a la web oficial para obtener más información.



Para facilitar la corrección de esta PEC, se adjunta el código de este ejercicio.