Enunciado del trabajo práctico para “Bases de datos SQL” 2022-2023

Introducción

Al final de las clases mostramos brevemente cómo se podía llegar a cargar en una base de datos SQLite, mediante un script de Python, datos de la base de datos ClinVar y CIViC. Los siguientes enlaces contienen una explicación de dónde se obtuvieron los datos, por qué se tomaron varias decisiones y dónde está el script como tal:

- https://bbddmasterisciii.github.io/04-Clinvar/index.html
- https://bbddmasterisciii.github.io/05-Python/index.html
- https://bbddmasterisciii.github.io/files/clinvar_parser.py

El trabajo estará compuesto de una memoria (2 puntos) y dos grandes partes (8 puntos). Tendréis de plazo para entregar el trabajo hasta el 12 de Marzo de 2023 (inclusive). El contenido entregado debe permitir la reproducibilidad de vuestro trabajo, sin incluir ninguna copia de la base de datos. Dentro del contenido entregado deberá ir la memoria, los tres programas en Python a escribir en la primera parte, y un fichero SQL por cada consulta de la segunda parte, todo ello dentro de un archivo .zip , .tar.gz o similar. La memoria debería estar en formato PDF, Word, LibreOffice, o similar. La memoria debe explicar el funcionamiento y parámetros de los programas, así como incluir una explicación de las decisiones tomadas a la hora de realizar los scripts.

Primera parte (4 puntos)

La primera consistirá en crear tres programas, dos de ellos relacionados con `clinvar_parser.py` (un punto cada uno), y un tercero que cargue datos de variantes de CIViC (dos puntos). Por tanto el programa original debería ser tomado como ejemplo e inspiración:

A. Un programa que se encargue de cargar información de ClinVar de los artículos científicos relacionados con las variantes (0.75 puntos).
B. Un programa que se encargue de cargar información de ClinVar de las estadísticas de variantes por gen (0.75 puntos).
C. Un programa que se encargue de cargar información de variantes de CIViC, similar a la información que se carga de ClinVar con clinvar_parser.py (2.5 puntos).

Para este trabajo vamos a trabajar sobre las versiones congeladas de junio de 2022 y diciembre de 2022 de ClinVar, disponibles en https://ftp.ncbi.nlm.nih.gov/pub/clinvar/tab_delimited/archive/ , así como las versiones congeladas de CIViC de junio de 2022 y diciembre de 2022, disponibles en https://civicdb.org/releases/main. Para saber qué ficheros deberéis procesar de ClinVar, y qué

estructura tienen, deberéis leer el fichero https://ftp.ncbi.nlm.nih.gov/pub/clinvar/tab_delimited/README, y en el caso de CIViC, cotejar los nombres de las columnas con la documentatión de campos en https://civic.readthedocs.io/en/latest/model.html. En la memoria deberéis explicar qué ficheros habéis escogido, por qué, y qué programa procesa qué fichero.

Todos los programas deberán crear sus propias tablas sobre la base de datos que se le indique en la entrada, pero deberán ser capaces de trabajar sobre bases de datos ya existentes. Concretamente, sobre las bases de datos generadas al cargar una *release* en concreto de ClinVar usando `clinvar_parser.py`, para poder lanzar consultas combinadas.

El objetivo es lanzar dos tandas de los programas, para generar **dos bases de datos diferentes**, una de datos de junio de 2022 y otra de diciembre de 2022. Esas bases de datos son necesarias para contestar a las preguntas de la segunda parte.

Segunda parte (4 puntos)

La segunda parte consistirá en escribir y realizar las siguientes 10 (0.4 puntos cada una) consultas SQL sobre el conjunto de tablas generado del script `clinvar_parser.py` más los programas que hayáis escrito. Las respuestas deben estar derivadas tanto de los datos de diciembre de 2022 como de junio de 2022, con lo que deberéis trabajar con bases de datos diferentes para esos meses.

1. ¿Cuántas variantes están relacionadas con el gen P53 tomando como referencia el ensamblaje GRCh38 en ClinVar y en CIViC?
2. ¿Qué cambio del tipo “single nucleotide variant” es más frecuente, el de una Guanina por una Adenina, o el una Guanina por una Timina? Usad las anotaciones basadas en el ensamblaje GRCh37 para cuantificar y proporcionar los números totales, tanto para ClinVar como para CIViC.
3. ¿Cuáles son los tres genes de ClinVar con un mayor número de inserciones y deleciones? Usa el ensamblaje GRCh37 para cuantificar y proporcionar los números totales.
4. ¿Cual es la deleción más común en el cáncer hereditario de mama en CIViC? ¿Y en ClinVar? Por favor, incluye en la respuesta además en qué genoma de referencia, el número de veces que ocurre, el alelo de referencia y el observado.
5. Ver el identificador de gen y las coordenadas de las variantes de ClinVar del ensamblaje GRCh38 relacionadas con el fenotipo del *Acute infantile liver failure due to synthesis defect of mtDNA-encoded proteins*.
6. Para aquellas variantes de ClinVar con significancia clínica “Pathogenic” o “Likely pathogenic”, recuperar las coordenadas, el alelo de referencia y el alelo alterado para la hemoglobina (HBB) en el assembly GRCh37.

7. Calcular el número de variantes del ensamblaje GRCh38 que se encuentren en el cromosoma 13, entre las coordenadas 10,000,000 y 20,000,000 , tanto para ClinVar como para CIViC.
8. Calcular el número de variantes de ClinVar para los cuáles se haya provisto entradas de significancia clínica que no sean inciertas (“Uncertain significance”), del ensamblaje GRCh37, en aquellas variantes relacionadas con BRCA2.
9. Obtener el listado de pubmed_ids de ClinVar relacionados con las variantes del ensamblaje GRCh38 relacionadas con el fenotipo del glioblastoma.
10. Obtener el número de variantes del cromosoma 1 y calcular la frecuencia de mutaciones de este cromosoma, tanto para GRCh37 como para GRCh38. ¿Es esta frecuencia mayor que la del cromosoma 22? ¿Y si lo comparamos con el cromosoma X? Tomad para los cálculos los tamaños cromosómicos disponibles tanto en https://www.ncbi.nlm.nih.gov/grc/human/data?asm=GRCh37.p13 como en https://www.ncbi.nlm.nih.gov/grc/human/data?asm=GRCh38.p13 .

Entrega de trabajos

Podéis realizar el trabajo en solitario o por parejas, pero en este último caso deberéis comunicarnos las parejas antes del 31 de Enero de 2023. Por favor, mandadnos los trabajos a eduardo.andres@csic.es y jose.m.fernandez@bsc.es antes del 12 de Marzo de 2023.