**Database documentation**

**Details of each table & What is the database used for.**

This database consists of 8 tables, ranging from gene level to tissue level.

1. **GENE** table:One hundred differentially expressed genes are recorded in this table, each of which has a unique id *(gene\_id)*. Additionally, the names of the gene *(gene\_name)*and ids of the gene in the Ensembl database*(Ensembl\_gene\_id)* are included so that scientists can query more information on the Ensembl database. The *Expr\_quantity* attribute derives from the total median gene expression level across 52 tissues and 2 cell lines obtained from 948 adult post-mortem individuals.
2. **CHROMOSOME** table: *Chrom\_id* represents unique ids for 24 chromosomes, matching the *chromosome* (from chr1 to chrY) attribute. To obtain the relative length of transcripts, lengths of whole chromosomes are recorded *(chrom\_size)*.
3. **GENE\_INTERACTION** table: Some genes interact with other genes to perform specific functions. The disability of some interactions may cause LUAD. Each eligible *gene\_id* corresponds to a set of serial numbers of interactive genes *(inter\_number)* and their names *(inter\_gene\_name)*.
4. **TRANSCRIPT** table: One gene can have several transcripts due to alternative splicing (Ule and Blencowe, 2019). All transcripts derived from 100 genes are logged, each with a unique id *(trans\_id)*. Same as above, the Ensembl transcript ids are provided to acquire more information in the Ensembl database *(Ensembl\_trans\_id).* This table also offers the position *(trans\_start & trans\_end)* of each transcript and whether it is the forward strand or reverse strand *(strand)*. As for the coding information, the position of the initiation codon *(coding\_start)*, termination codon *(coding\_end)*, and the number of exons *(exon\_count)* are provided.
5. **TRANSCRIPT\_CLASS** table: Different types of transcripts have disparate functions like coding protein, catalyzing reactions, silencing expression, and so on. The *trans\_class* attribute notes the 3 main classes of transcripts, while the *trans\_type* attribute documents all subtypes for every main class. Each subtype is allocated a unique id (*trans\_class\_id*).
6. **PROTEIN** table: The protein-coding transcripts can be translated into protein, functioning as mRNA. Each protein has its unique id *(prot\_id)*, short name *(prot\_short)*, and full name *(prot\_full)*. Additionally, protein lengths *(prot\_length)* and protein ids in the Uniprot database *(Uniprot\_id)* are provided for future queries.
7. **PROTEIN\_MUTATION** table: Mutations in genes may lead to protein mutations eventually, causing the disfunction of proteins and thus resulting in cancer. Each type of mutation is allocated a unique id *(mutation\_id)* and the description of the mutation is logged in the *description* attribute. *Prot\_id* attribute is contained in this table to connect mutations to corresponding proteins.
8. **Tissue** table: Some proteins show significant differential expression in specific tissues, therefore, the *description* attribute in the table records detailed information about this to help scientists use specific tissues for gene function studies in the future. As above, the unique id of each description *(tissue\_id)* is created to facilitate queries.

**Normal forms of this database (NF).**

The database is normalized to 1NF, 2NF, and 3NF (Figure 1).

**图片包含 图形用户界面

描述已自动生成**

Figure 1: The schematic diagram of 3NF (created by PowerPoint).

**Entity-Relationship models (ER).**

ER model is a widely used conceptual database modeling method, which can enable the structure of the requirements and provide a way to represent these requirements graphically. The result of ER modeling is shown below (Figure 2).

**图示, 示意图

描述已自动生成**

Figure 2: The ER diagram **(ERD)** of the mini database (created by PowerPoint).

**Five cases to demonstrate examples of usage.**

**Case 1:** In the mini database, some genes do not have interactive genes, which means that more research should be done on these genes to find more functions or pathways of these genes. This query can show the names and Ensembl ids of genes whose interactive genes are empty.

**文本

描述已自动生成图形用户界面, 文本, 应用程序

描述已自动生成**

**Case 2:** Proteins with high expression may have significant functions. Therefore, the top ten genes with high expression are selected in this mini database. Besides the expression quantities, the names of the gene, Ensembl ids, and chromosome positions are provided.

**文本

描述已自动生成**电脑萤幕画面

描述已自动生成

**Case 3:** Open reading frame **(ORF)** can express biologically active proteins that are potential therapeutic targets for cancer. This case shows how to get the ORF length and relative transcript length using gene\_id.

**文本

描述已自动生成电脑萤幕的截图

描述已自动生成**

**Case 4:** Different chromosomes have different numbers of genes. Therefore, finding which chromosomes have relatively more genes is an interesting topic. From the query, I find that chr10 and chr11 have more than 10 genes respectively. The query can also list all genes located on these two chromosomes.

文本

描述已自动生成

**日历

低可信度描述已自动生成**

**Case 5:** Some proteins have tissue specificity, which means they may have particular functions in any tissue. This query list all genes that have lung specificity and the number of their mutation types. Proteins with more mutation types should be paid more consideration because they are inclined to be mutated and become disfunction.

**文本

描述已自动生成电脑屏幕截图

描述已自动生成**