

Contents

Chapter 1 What is Miqualat Database	1
Chapter 2 Setting Up the Working Environment	2
2.1 Install Jupyter Notebook.....	2
2.2 Install MySQL Server.....	2
2.3 Install MySQLclient	3
2.4 Install Required Python Libraries	3
Chapter 3 Folder and Content Description	4
3.1 Description of the FUNCTIONS Folder	4
3.2 Description of the OUTPUT Folder	4
3.3 Description of the DATABASE Folder.....	4
Chapter 4 Miqualat Notebooks Description and Usage	5
Chapter 5 Create Biomart File.....	6
Chapter 6 Create Table File.....	8
Chapter 7 Database Description	9
7.1 Database Schema.....	9
7.2 Database Description.....	10
7.3 Checks Description.....	13

Chapter 1 What is Miqualat Database

With Miqualat Notebooks you have access to an easy-to-use Jupyter Notebook interface.

Miqualat Notebook allows you to save data from publications, genes, variants, technique and other and associate them with descriptions and tags.

A control system has been implemented to maintain flexibility and freedom to import data while maintaining control over them.

You can get this data from different databases as pubmed, ensembl, kegg and make automatically links.

This relational database allows you to link and export information from different tables.

Chapter 2 Setting Up the Working Environment

2.1 Install Jupyter Notebook

TIP: If you use Microsoft Windows you should to use WSL (Windows Subsystem for Linux).

Update and upgrade your system (recommended).

```
$ sudo apt update && sudo apt upgrade
```

Install python3-pip.

```
$ sudo apt install python3-pip
```

Install jupyter notebook.

```
$ sudo apt install jupyter-notebook
```

If you need to configure the configuration file is located in

```
/root/.jupyter/jupyter_notebook_config.py
```

```
$ sudo jupyter-notebook --generate-config
```

To access Jupyter Notebook you may need to set up a password.

```
$ sudo jupyter-notebook password
```

Run jupyter notebook (on port 8888 by default).

```
$ sudo jupyter-notebook
```

2.2 Install MySQL Server

Install apache2 and mysql-server.

```
$ sudo apt install apache2
```

```
$ sudo apt install mysql-server
```

TIP: If you use raspberry or a similar system install mariadb and configure it.

```
$ sudo apt install mariadb-server-10.0
```

Create user and set password for database (this user can be used to access in phpmyadmin, keep your credentials).

```
$ sudo mysql -u root
```

```
$ CREATE USER 'user'@'localhost' IDENTIFIED BY 'password';
```

Grant privileges on all databases.

```
$ GRANT ALL PRIVILEGES ON * . * TO 'user'@'localhost';
```

Grant privileges on one databases (safer).

```
$ GRANT ALL PRIVILEGES ON DATABASE. * TO 'user'@'localhost';  
$ FLUSH PRIVILEGES;
```

!! IMPORTANT (pay attention before continuing with the command below), when you install phpmyadmin (command below) after select apache2 (with TAB) and press SPACE, after leave all empty (press ENTER) !!

```
$ sudo apt install phpmyadmin  
$ sudo systemctl restart apache2
```

Connect phpmyadmin using your browser (on port 80: /phpmyadmin).
[http://192.168.0.\(continue_your_ip\).../phpmyadmin](http://192.168.0.(continue_your_ip).../phpmyadmin)

Show your IP.
\$ hostname -I

<http://localhost/phpmyadmin> (work only on local computer not for remote server)

2.3 Install MySQLclient

Install mysqlclient (connect database remotely using python).

Get more information at: <https://pypi.org/project/mysqlclient/>

```
$ sudo apt-get install python3-dev default-libmysqlclient-dev build-essential  
$ pip3 install mysqlclient
```

Update and upgrade your system (recommended).

```
$ sudo apt update && sudo apt upgrade
```

2.4 Install Required Python Libraries

The following libraries are required for use miqualat notebooks.

```
$ sudo pip3 install biopython  
$ sudo pip3 install ensembl-rest  
$ sudo pip3 install mygene
```

Chapter 3 Folder and Content Description

3.1 Description of the FUNCTIONS Folder

The functions in FUNCTIONS folder, contain python code that is recalled in the notebooks.

1. `ensembl_search.py`
Use the `ensembl-rest` library to get information about genes from the Ensembl database.
2. `ensembl_to_kegg_id.py`
Convert Ensembl gene id to the Kegg id of the gene and related pathways, using `mygene` and `Bio.KEGG.REST` (in `biopython` library).
3. `MIQUALAT_data_import_and_check.py`
Import and check the csv files processed with previous notebooks, using `MySQLdb` (in `mysqlclient` library).
4. `pubmed_search.py`
Carry out searches within the pubmed database using `biopython`.
5. `python_parser_biomart_gene_csv.py`
Convert the biomart file (which contains all biomart genes) into the right format for the GENE table.

3.2 Description of the INPUT Folder

In the INPUT folder are saved all the files processed by the notebook, to be imported into the `miqualat` database.

3.3 Description of the OUTPUT Folder

In the OUTPUT folder are saved all the files exported by the notebook from `miqualat` database.

3.4 Description of the DATABASE_SQL Folder

In the `DATABASE_SQL` folder is present a copy of database in `sql` format.

Chapter 4 Miqualat Notebooks Description and Usage

To use notebooks select the code cell and click on run, below will describe the notebooks and their functions.

Five notebooks have been built, that allow you to search, import and check into the miqualat database.

1. `MIQUALAT_PUBLICATION_data_assess.ipynb`
Carry out searches within the pubmed database, automatize data collection for PUBLICATION table.
Creates PUBLICATION.csv file and move all searches (PUBMED_SEARCH_DATA_Y-m-d_H-M-S.txt format) in the OUTPUT folder.
2. `MIQUALAT_KEGG_data_assess.ipynb`
Download genes and pathways from kegg database (from kegg code), automatize data collection for KEGG table.
Creates ALL_KEGG_GENE.csv (all kegg genes from kegg code) in the OUTPUT folder.
Creates ALL_KEGG_PATH.csv (all kegg pathways from kegg code) in the OUTPUT folder.
3. `MIQUALAT_GENE_data_assess.ipynb`
Allows you to download genes from ensembl database, automatize data collection for GENE table.
Option 1) get information from specific genes (ensembl gene id required).
Option 2) convert all genes downloaded from biomart (GENE.csv required).
In any case creates GENE.csv and move it in the OUTPUT folder.
4. `MIQUALAT_GEN_KEGG_data_assess.ipynb`
Correlates the ensembl gene id to the kegg id (if possible) of the gene and related pathway, automatize data collection for GEN_KEGG table.
Option 1) correlates the ensembl gene id to the kegg id of the gene and related pathways for specific gene (ensembl gene id required), for specific gene.
Option 2) processes the entire GENE.csv file.
Creates ENS_GENE_ID.csv (Ensembl gene id to Kegg gene id) in the OUTPUT folder.
Creates ENS_PATH_ID.csv (Ensembl gene id to Kegg pathways id) in the OUTPUT folder.
5. `MIQUALAT_data_import_and_check.ipynb`
Allows the import of files processed with previous notebooks, once checked.
An error control system prevents the import of incorrect information.

Chapter 5 Create Biomart File

Go to <https://www.ensembl.org/biomart/martview/afe982e758c87b06e672bc93a42a4f30>.

Export all results to File CSV.

Select dataset Ensembl Gene 101.

Select Species (refseq) (ex Cow genes (ARS-UCD1.2)).

Go to Attributes, Gene and create the biomart file following this header (select in the right order).

Gene stable ID, Gene name, Gene description, Chromosome/scaffold name, Gene start (bp), Gene end (bp), Strand

The screenshot shows the Ensembl BioMart interface. At the top, there is a navigation bar with links for BLAST/BLAT, VEP, Tools, and BioMart. Below this, there are tabs for 'New', 'Count', and 'Results'. The 'Dataset' section is active, showing a dropdown menu for 'Ensembl Genes 101' and a text input for 'Cow genes (ARS-UCD1.2)'. The 'Attributes' section is also active, showing a list of attributes: 'Gene stable ID', 'Gene stable ID version', 'Transcript stable ID', and 'Transcript stable ID version'. The 'Dataset' section at the bottom is set to '[None Selected]'.

This screenshot shows the 'Attributes' and 'Structures/Sequences' sections of the Ensembl BioMart interface. The 'Attributes' section on the left lists: 'Gene stable ID', 'Gene name', 'Gene description', 'Chromosome/scaffold name', 'Gene start (bp)', 'Gene end (bp)', and 'Strand'. The 'Dataset' section is set to 'None Selected'. The 'Structures' section is selected, showing a list of attributes with checkboxes: 'Gene stable ID' (checked), 'Gene stable ID version' (unchecked), 'Transcript stable ID' (unchecked), 'Transcript stable ID version' (unchecked), 'Protein stable ID' (unchecked), 'Protein stable ID version' (unchecked), 'Exon stable ID' (unchecked), 'Gene description' (checked), 'Chromosome/scaffold name' (checked), 'Gene start (bp)' (checked), 'Gene end (bp)' (checked), 'Strand' (checked), and 'Karyotype band' (unchecked). The 'Sequences' section is also visible, showing a list of attributes with checkboxes: 'Transcript length (including APPRIS annotation)' (unchecked), 'Gene name' (checked), 'Source of gene name' (unchecked), 'Transcript name' (unchecked), 'Source of transcript name' (unchecked), 'Transcript count' (unchecked), 'Gene % GC content' (unchecked), 'Gene type' (unchecked), 'Transcript type' (unchecked), 'Source (gene)' (unchecked), 'Source (transcript)' (unchecked), and 'Version (gene)' (unchecked).

Results

[★ URL](#)
[XML](#)
[Perl](#)
[Help](#)

CD1.2)

Export all results to

File

CSV

☐ Unique results only

Go

Email notification to

View

10

rows as

HTML

☐ Unique results only

Gene stable ID	Gene name	Gene description	Chromosome/scaffold name	Gene start (bp)	Gene end (bp)	Strand
ENSBTAG00000006648			1	339070	350389	-1
ENSBTAG00000049697	5S_rRNA	5S ribosomal RNA [Source:RFAM;Acc:RF00001]	1	475398	475516	1

fold name

Chapter 6 Create Table File

To pass the check the csv must have the following headers.

!! Create the Table file following this headers (create the header as it is written, do not add quotes or anything else). !!

1. PUB_GEN_TEC_VAR_TAG TABLE

integer_progressive_ID, pubmed_ID, ensembl_gene_ID, variant_name, technique, keyword_tags, relationship_note

2. PUBLICATION TABLE

pubmed_ID, doi, article_title, article_authors, article_journal, publication_year

3. TECHNIQUE TABLE

technique, technique_short_description

4. TAG TABLE

keyword_tags, tags_short_description

5. GENE TABLE

ensembl_gene_ID, gene_name, gene_short_description, refseq, species, chromosome, start_coordinate, end_coordinate, strand

6. KEGG TABLE

kegg_ID, kegg_object_type, kegg_object_name

7. GEN_KEGG TABLE

ensembl_gene_ID, kegg_ID

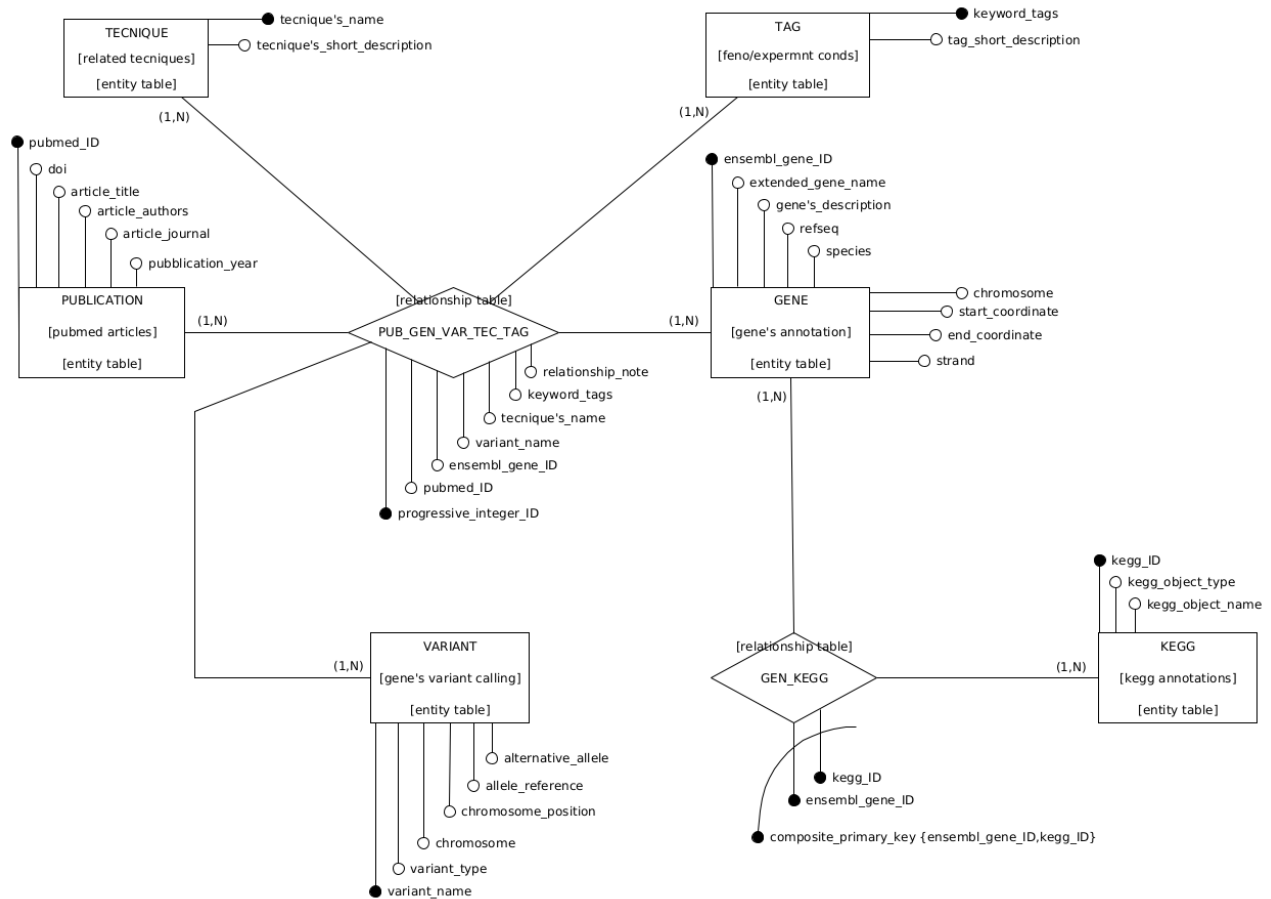
8. VARIANT TABLE

variant_name, variant_type, chromosome, chromosome_position, allele_reference, alternative_allele_reference

Chapter 7 Database Description

7.1 Database Schema

Below the physical Entity-Relationship diagram (ERD) of the Miqualat database.



7.2 Database Description

```
##using the MIQUALAT database and displaying tables##
```

```
mysql> use MIQUALAT;
```

```
Database changed
```

```
mysql> SHOW TABLES;
```

Tables_in_MIQUALAT
GENE
GEN_KEGG
KEGG
PUBLICATION
PUB_GEN_VAR_TEC_TAG
TAG
TECHNIQUE
VARIANT

```
8 rows in set (0.00 sec)
```

```
##description of the structure of the tables and the type of data  
of the attributes of each entity and relation##
```

```
mysql> DESCRIBE GENE;
```

Field	Type	Null	Key	Default	Extra
ensembl_gene_ID	varchar(20)	NO	PRI	NULL	
gene_name	varchar(20)	YES		NULL	
gene_short_description	varchar(300)	YES		NULL	
refseq	varchar(50)	NO		NULL	
species	varchar(50)	NO		NULL	
chromosome	tinyint	NO		NULL	
start_coordinate	int unsigned	NO		NULL	
end_coordinate	int unsigned	NO		NULL	
strand	tinyint	NO		NULL	

```
9 rows in set (0.00 sec)
```

```
mysql> DESCRIBE GEN_KEGG;
```

Field	Type	Null	Key	Default	Extra
ensembl_gene_ID	varchar(20)	NO	PRI	NULL	
kegg_ID	varchar(20)	NO	PRI	NULL	

```
2 rows in set (0.00 sec)
```

```
mysql> DESCRIBE KEGG;
```

Field	Type	Null	Key	Default	Extra
kegg_ID	varchar(20)	NO	PRI	NULL	
kegg_object_type	varchar(30)	NO		NULL	
kegg_object_name	varchar(300)	NO		NULL	

```
3 rows in set (0.00 sec)
```

```
mysql> DESCRIBE PUBLICATION;
```

Field	Type	Null	Key	Default	Extra
pubmed_ID	int unsigned	NO	PRI	NULL	
doi	varchar(100)	YES		NULL	
article_title	varchar(300)	NO		NULL	
article_authors	varchar(300)	NO		NULL	
article_journal	varchar(100)	NO		NULL	
publication_year	year	NO		NULL	

```
6 rows in set (0.00 sec)
```

```
mysql> DESCRIBE PUB_GEN_VAR_TEC_TAG;
```

Field	Type	Null	Key	Default	Extra
integer_progressive_ID	int unsigned	NO	PRI	NULL	auto_increment
pubmed_ID	int unsigned	NO	MUL	NULL	
ensembl_gene_ID	varchar(20)	YES	MUL	NULL	
variant_name	varchar(30)	YES	MUL	NULL	
technique	varchar(50)	YES	MUL	NULL	
keyword_tags	varchar(50)	YES	MUL	NULL	
relationship_note	varchar(200)	YES		NULL	

```
7 rows in set (0.01 sec)
```

```
mysql> DESCRIBE TAG;
```

Field	Type	Null	Key	Default	Extra
keyword_tags	varchar(50)	NO	PRI	NULL	
tags_short_description	varchar(200)	NO		NULL	

```
2 rows in set (0.01 sec)
```

```
mysql> DESCRIBE TECHNIQUE;
```

Field	Type	Null	Key	Default	Extra
technique	varchar(50)	NO	PRI	NULL	
technique_short_description	varchar(300)	NO		NULL	

```
2 rows in set (0.01 sec)
```

```
mysql> DESCRIBE VARIANT;
```

Field	Type	Null	Key	Default	Extra
variant_name	varchar(30)	NO	PRI	NULL	
variant_type	varchar(30)	NO		NULL	
chromosome	tinyint unsigned	NO		NULL	
chromosome_position	int unsigned	NO		NULL	
allele_reference	varchar(50)	NO		NULL	
alternative_allele_reference	varchar(50)	NO		NULL	

```
6 rows in set (0.00 sec)
```

```

##IT IS A RELATIONAL DATABASEIT IS A RELATIONAL DATABASE##
*WITH 6 ENTITY TABLES{
-----» PUBLICATION,
-----» GENE,
-----» VARIANT,
-----» TECHNIQUE,
-----» TAG,
-----» KEGG
}

*AND 2 RELATIONAL TABLES{
-----» PUB_GEN_VAR_TEC_TAG,
-----» GEN_KEGG
}

```

7.3 Checks Description

Below a short resume of database checks description.

- 1) entries fields number
- 2) entry duplicates
- 3) entries already present in the database
- 4) entry fields of a whole numerical nature
- 5) existence of foreign_keys for relational tables
- 6) special check for table PUB_GEN_VAR_TEC_TAG
- 7) special check for table GEN_KEGG

Chapter 8 Database Export Data

File will be exported into OUTPUT folder with name exported_data_Y-m-d_h-d-s.