Vivan Coding Challenge

The following challenge aims to evaluate the candidate's following skills:
· *Knowledge of Python/Mysql*
· *Skills in database management*
· *Using version control software (GIT)*

The first goal of this challenge is to create a python command line tool capable of extracting information from the input files provided and inserting the information into a MySQL database. The main script must have sub-commands capable of performing the following operations:
1) create a MySQL database from scratch
2) read and appropriately extract the information from the input files
3) Enter the information in the MySQL database tables

More Information:
Benchling Electronic Lab Notebook is a cloud-based environment that we use to record all the data about our patients, avatar modeling, and conducted experiments. The following example tables contain information about which genes need to be up/downregulated in a drosophila model to get an avatar for the human patient being examined. The benchling_entries.json file contains the information of the Benchling lab notebook extracted with its API and saved in a JSON object. The candidate must explore this file, identify the entries related to each patient (each patient has an ID such as Pat0XX), and for each patient extract the information present in the "Genes to up-regulate" and "Genes to down-regulate" tables and put them in a table in MySQL. (HS gene refers to the human gene form and Dm genes refer to Drosophila Melanogaster form)

The second file “cnv_procerssed.txt” is a simple TAB file containing the data for the copy number variations of some of the patients seen above. In the column “symbol” it is possible to find the Human Gene symbol.

This file must be processed and inserted into another table of the MySQL database.

As a second milestone, once the database is ready, the candidate must answer the following questions using the appropriate MySQL syntax:
a) Number of patients in Benchling with information for genes to up/down regulate
b) Number of patients with information for copy number variation

c) Identify which patients have both information
d) For each patient found in “c” list the genes present in Benchling and having a copy number variation found with the tool “pipeline_name: sequenza_vivan”

The candidate must at the beginning create a git repository and document with constant commits what she/he is doing. It is also preferable to create a small README.md file that briefly explains how the program works and a little schema/diagram showing the architecture of the database and tables relationship.

Period given to complete: Monday, 9AM UK Time