Initial Plan

Quantifying the different COVID-19 variants present in wastewater within South Wales

Author: Arlyn Miles
Client: Professor Peter Kille
Supervisor: Dr Bailin Deng

Quantifying the different COVID-19 variants present in wastewater within South Wales

Project Description

The aim of this project is to quantify the different variants of COVID-19 present in waste water samples collected within South Wales. I will then determine the proportion of different variants circulating within the population and estimate the number of individuals associated with each variant.

I will be working with Professor Peter Kille (Director of Technology and Bio-Initiatives at the School of Biosciences) as my client, collaborating with the ongoing research into this subject. The data I will be using is RNA data from South Wales wastewater samples that will be sequenced at the School of Bioscience in the coming weeks. I will create an optimised pipeline to process this sequenced data and identify the different COVID-19 variant strains.

New strains of COVID-19 are spreading throughout the population, and as a virus it will continue to mutate. New variants can have significant effects; mutations on the spike protein on variant 501.V2 also known as the “South African variant” may increase its infectivity. It is therefore important to monitor the different COVID-19 variants present in South Wales.

RNA of COVID-19 is present in faecal matter and can therefore be detected within wastewater. It is thus possible to quantify the strains circulating within the wider South Wales population by sequencing viral RNA from the wastewater samples

This data can then be validated and compared against the identified variants present in samples collected from Cardiff University’s screening service. By comparing mutations in the wastewater samples and student samples over a period of time, it may be possible to derive if and how the student population affects the COVID-19 variants present in the wider area.

Project Aims and Objectives

Primary Aims

Minimum required objectives

- Research into existing bioinformatics pipelines and similar research
    - Experiment with bioinformatics libraries and docker images
    - Practice scheduling slurm jobs on the Trinity cluster
    - Research existing methods for identifying COVID-19 variants
    - Update technical objectives and Gantt chart accordingly if any steps present in the initial plan are redundant or if more are needed. I plan to utilise existing packages as much as possible to ensure the project fits within the time scope
- Create initial dataset to be used in first pipeline iteration
    - Create simulated data to use for the first iteration of the pipeline while waiting for the wastewater research data to be sequenced
        - Use simulation tools or online databases such as GISAID to create/modify a simulated data mix which would resemble the sequenced wastewater data
        - Transfer this data to the scratch space on the Trinity cluster
- Develop first iteration of pipeline for processing the sequenced RNA data
    - Setup any necessary tools
        - Create docker images for Charlie Cloud
        - Install packages/modules on the cluster
    - Create a diagram/draft for the planned pipeline
    - Implement the pipeline
        - Able to process the sequenced RNA data by developing a pipeline to identify different COVID-19 variants from the wastewater sample data
- Improve and optimise pipeline
    - Evaluate first iteration pipeline
    - Identify ways to improve and plan second iteration
    - Optimise this pipeline to ensure data can be processed efficiently
    - Review second iteration pipeline with client and adjust if necessary
    - Run pipeline with the real wastewater data

- Quantify the number of infected people from the data
    - From the processed RNA data I have processed and quantitative analysis data from other researchers, I will be able to quantify an estimate of the number of people with each virus variant
    - Create any necessary steps to add to the pipeline for this
    - Write up data findings

Secondary Aims

Additional optional objectives

- Show the diversity in COVID-19 variants across South Wales and how they might differ over time and area
    - Use data science packages to analyse
    - Produce graphs and data models
- Analyse student COVID-19 samples from Cardiff University Screening Service
    - Compare and contrast variants over time with wastewater results
    - Model data in a clear way

Ethics

After discussing with my client and supervisor, we have deemed ethical approval to not be necessary for this project.

There are three possible sources of data I may be using in this project:

- COVID-19 variant RNA samples from online databases such as GISAID and NCBI
- South Wales Wastewater sequenced RNA obtained from researchers in the School of Biosciences
- Sequenced COVID-19 RNA samples obtained from the Student Screening Service

Online databases:
Ethical approval not needed as no personal identifying information is attached to this data, it is only low level RNA information. All data has been published for use within research with credit given.

Wastewater:
As the wastewater is collected from wastewater treatment plants, there is no way to track the COVID-19 samples back to individuals. No ethical approval needed*.

Student Samples:
Any personal identifying information has been stripped. The only data I will have access to is the virus variant sequenced data and the date. No ethical approval needed*.

* Despite lacking any personal identifying information, in theory it could be possible to identify an individual by examining the linearges of that particular strain of COVID-19 and tracking it to a certain location, behaviour, or patient records but doing so would require access to the NHS Track and Trace system data and Cardiff University Screening Service data which I do not and will not have access to. Ethical approval is therefore not necessary.

Work Plan

I will meet with my supervisor, Dr. Bailin Deng, every Friday afternoon, and in addition to this will have two scheduled review meetings which will specifically review the overall progress made towards the project and exist as a milestone to adjust my plan depending on the outcome of these meetings.

I will also be meeting with my client, Professor Peter Kille, when appropriate to coordinate with the ongoing research going into this project such as the sequencing of the wastewater data, and to monitor and give feedback on my progress to ensure my pipeline is suitable.

Please find my modified Gantt chart work plan attached- it is best viewed on the Google Sheets link but screenshots are attached below.

Phases 2,3,4, and 6 will be included in the final report, with phase 5 being an additional aim to include.

https://docs.google.com/spreadsheets/d/19w-k6Jj3wF1DrgnGU-lfkCKch6J0yQOf47LAyABOosM/edit?usp=sharing