Introduction to Focus Areas in Bioinformatics – WS19/20

Prof. Dr. Tim Conrad

4 Project 4

- Deadline: 17.11.2019, 23:59
- All files need to be available through your GIT repository, in the directory “Project 4”.
- You can work in teams up to 3 people. Please state in the report the group member names.
- If your code is in python, I must be able to run your code within a Google Colab notebook. If your code is not in Python or R, you must provide a manual how to compile and run it on a Linux machine.

4.1 Data & Goal

We will use the 1000 Genomes data-set (see https://www.nature.com/collections/dcfqmlgsrw) to get familiar with the BigQuery API.
IMPORTANT: before you execute any SQL query (see below), always check the scan size (estimate_query_size(QUERY) from the helper package).

4.2 Task 1: Introduction

- Go through the tutorial “Advanced guide to analyzing variants using BigQuery” https://cloud.google.com/life-sciences/docs/tutorials/analyze-variants-advanced
- Redo all experiments (SQL queries) using a programming language and the BigQuery API.

4.3 Task 2: Exploring the Data

- Go through the 6 “Data Stories for the 1000 Genomes Project” https://github.com/googlegenomics/bigquery-examples/tree/master/1000genomes/data-stories
- Redo all experiments (SQL queries) using a programming language and the BigQuery API.
- Recreate the figures, if any.

4.4 Deliverables

Your need to upload all source codes and a report to your GIT repository.

- Write a report that lists the SQL queries that you did, what happened “in there” (including scan size) and what was the result of it. (Not the actual result table.)
- For each item in that list, you should describe the “IT” and the “Biological” point of view.
- The report must be delivered in PDF format.
- REMEMBER TO NOT GO OVER THE TIME LIMIT!