Bioinformatics challenge 2022

The objective of the 2022 Amphora Health's Data Challenge is to explore the genomic information that exists in our cells and develop machine learning models that can predict the risk of developing or acquiring a given disease.

The goal of this challenge is to test the candidate's ability to deliver a fully functional computational genomics product, which includes training a machine learning model, evaluating its accuracy, and testing it with new input from patients.

This data challenge is an integral part of Amphora Health's recruitment process and it is designed to identify the top candidates. Thank you for taking the time to participate in this process and for your interest in Amphora Health. We hope that by the end of this challenge you would have learned something novel for you as well.

1. Instructions

Data input. You will be provided with multiple files in two different formats: Variant Call Format (VCF), and raw files in tab-separated values (.tsv). The structure of the files looks as follows:

```
ah-challenge-2022/
 ├── coordination.txt
 ├── vcfs/
     ├── sample1.vcf.gz
     ├── sample2.vcf.gz
       …
     └── sampleN.vcf.gz
 ├
 └── raw-files/
     ├── patient01.tsv
     ├── patient02.tsv
       …
     └── patientM.tsv
```

To learn more about each data format you can read these references:

- VCF:
  - https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3137218/
  - https://samtools.github.io/hts-specs/VCFv4.2.pdf
- Raw Format (TSV).

Genomic Ancestry Challenge

The files provided in the dataset contain genomic information for participants coming from 5 different continental populations: African (AFR), European (EUR), South East Asian (SAS), East Asian (EAS), and American (AMR).

The dataset contains information on more than 1,000 individuals. For 20% of the data, you will know the true continental label, while for the remaining 80% will be unknown.

Task 1. Your first task is to merge all the files into a single table to construct a merged genotype file for several individuals.

Task 2. Create a clustering strategy and visualize them.

Task 3. Provide an evaluation for your clustering assignment.

Deliverable. You are required to generate this challenge in a Github repository for this challenge. This will be your deliverable. In this repository, you should deposit your scripts and code in the programming language of your preference and document it properly. We recommend the use of R or Python, but any other language should be fine. The pipeline should be designed in such a way that we can feed in new data from any other patient and obtain results automatically.

2. Evaluation

We will evaluate this challenge in two ways:

1. **Coding review**. The code must be uploaded to a Github repository and you will provide us with the respective link when you finalize the assignment. Please be cognizant that we will be looking at your code and looking at how clean it is, your variable naming conventions, and documentation. For the entirety of this project please keep your code in English only.
2. **Presentation skills**. The presentation will be held online using the Google Meet platform. We will provide you with a link beforehand so that you can practice. The presentation should be 30 minutes long including intros, presentations, and questions. The Amphora Health Data science team and other external collaborators will be invited. The attendees span a wide range of expertise including Computer Science, Mathematics, Medicine, and Clinical Research.

3. Deadline

You will have a whole week to complete it with the possibility to have an extension if you need it for whatever reason. Once you set up a timeline, we will expect you to deliver it on time.