Optimal region-based reference set design

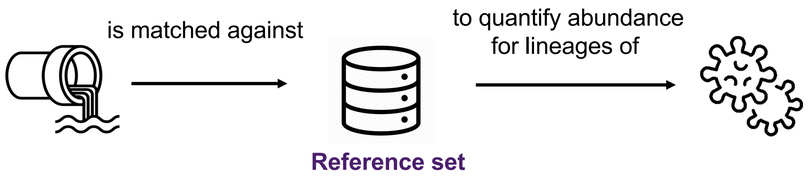
SARS-CoV-2 lineage abundance quantification from wastewater sequencing data

I. Nika (I.Nika@student.tudelft.nl) – Computer Science and Engineering | Supervisor: Jasmijn Baaijens

1. Introduction

- A **lineage** is a collection of virus mutants that share predecessors
- Monitoring existing lineages is crucial for the efforts taken to contain the virus
- SARS-CoV-2 lineage abundance quantification in wastewater helps monitoring existing lineages in cases where clinical sequencing is not feasible

Fig. 1: Abundance qantification pipeline



2. Research question

- How does the reference set design affect prediction accuracy?
- **Focus on:** How should the geographical region where the reference sequences are sourced from be decided?

3. Background

Why does the geographical location matter?

Lineages show different mutations in different geographical regions:

- Random mutations, can be highly represented in a given geographical location.
- Immune responses differ among populations, and change the virus [1] → those differences can be linked to ancestry [2]

4. Hypothesis

Increased performance due to:

- Within lineage variation that becomes common between test set and reference set
- Variation specific to lineages could provide useful variation between different lineages

5. Methods

1. Build reference sets
2. Build test sets containing simulated wastewater sequencing data
3. Evaluate predictions

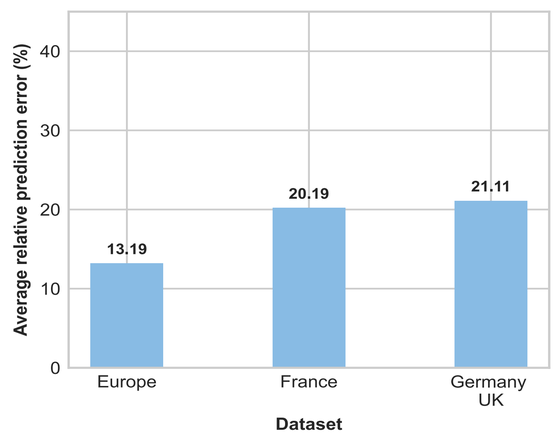
Metric (relative prediction error):

$$\frac{|true\ abundance\ -\ estimated\ abundance|}{true\ abundace} \cdot 100$$

6. Results

Fig. 2: Interactions between populations

Test set: Cyprus, Europe



France : **No tourists reported** [3] **,** Germany, UK : **~50% of tourism** [3]

Fig. 3: Geographical Proximity

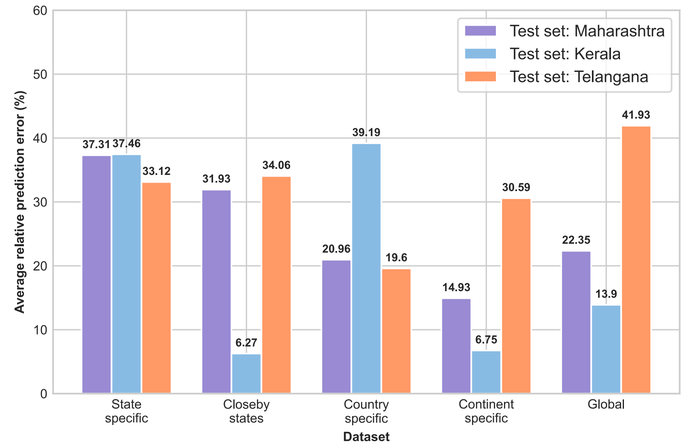
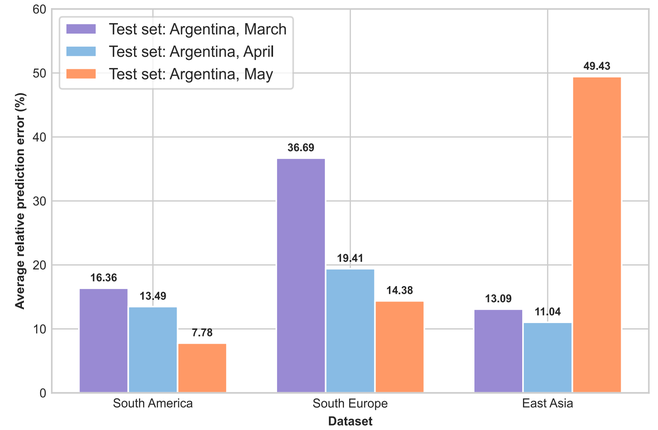


Fig. 4: Ancestry & immune response related mutations

Test set: Argentina, South America



South Europe**: ~67% shared ancestry** [4] , East Asia : **~0.01% shared ancestry** [4]

7. Conclusions

- Continent specific reference sets yield best results
- Overall interactions of a country could be considered
- Ancestry does not influence results

[1] Rui Wang,Yuta Hozumi, Yong-Hui Zheng,Changchuan Yin, and Guo-Wei Wei. Host immune response driving sars-cov-2 evolution.Viruses,12, 10 2020

[2] Yohann Nedlec, Joaquın Sanz, Golshid Baharian, Zachary A. Szpiech, Alain Pacis, Anne Dumaine, Jean-Christophe Grenier, Andrew Freiman, Aaron J. Sams,Steven Hebert, Ariane Pag´e Sabourin, Francesca Luca, Ran Blekhman, Ryan D. Hernandez, Roger Pique-Regi, Jenny Tung, Vania Yotova, and Luis B. Barreiro. Genetic ancestry and natural selection drive population differences in immune responses to pathogens.Cell,167(3):657–669.e21, 2016.

[3] Lucy Panayidou. Press releases, Sep 2020, https://www.pio.gov.cy/en/press-releases-article.html?id=15731, visited on 2022-01-09.

[4] Julian R. Homburger, Andres Moreno-Estrada, Christopher R. Gignoux, Dominic Nelson, Elena Sanchez, Patricia Ortiz-Tello, Bernardo A. Pons-Estel, Eduardo Acevedo-Vasquez, Pedro Miranda, Carl D. Langefeld, Simon Gravel, Marta E. Alarc´on-Riquelme, and Carlos D. Bustamante. Genomic insights into the ancestry and demographic history of south america. PLOS Genetics, 11:e1005602, 2015