Assessing SARS-CoV-2 evolution through the analysis of emerging mutations

Anastasios Mitsigkolas[1], Nikolaos Pechlivanis[2,3], Fotis Psomopoulos[2] & Evert Bosdriesz[4]

[1]Master's student, Bioinformatics and Systems Biology, Faculty of Science, VU Amsterdam, De Boelelaan 1111, Amsterdam 1081 HV, the Netherlands
[2]Institute of Applied Biosciences, Centre of Research and Technology Hellas, Thermi, 57001, Thessaloniki, Greece
[3]Department of Genetics, Development and Molecular Biology, School of Biology, Aristotle University of Thessaloniki, 54124, Thessaloniki, Greece
[4]Bioinformatics, Computer Science, VU Amsterdam, De Boelelaan 1111, Amsterdam 1081 HV, the Netherlands

Contact: ***tasos1109@gmail.com***

1 Highlights

- **A novel method for detecting patterns of SARS-CoV-2 cooccurring mutations.**
- **Detection of evolutionary paths of SARS-CoV-2 virus.**

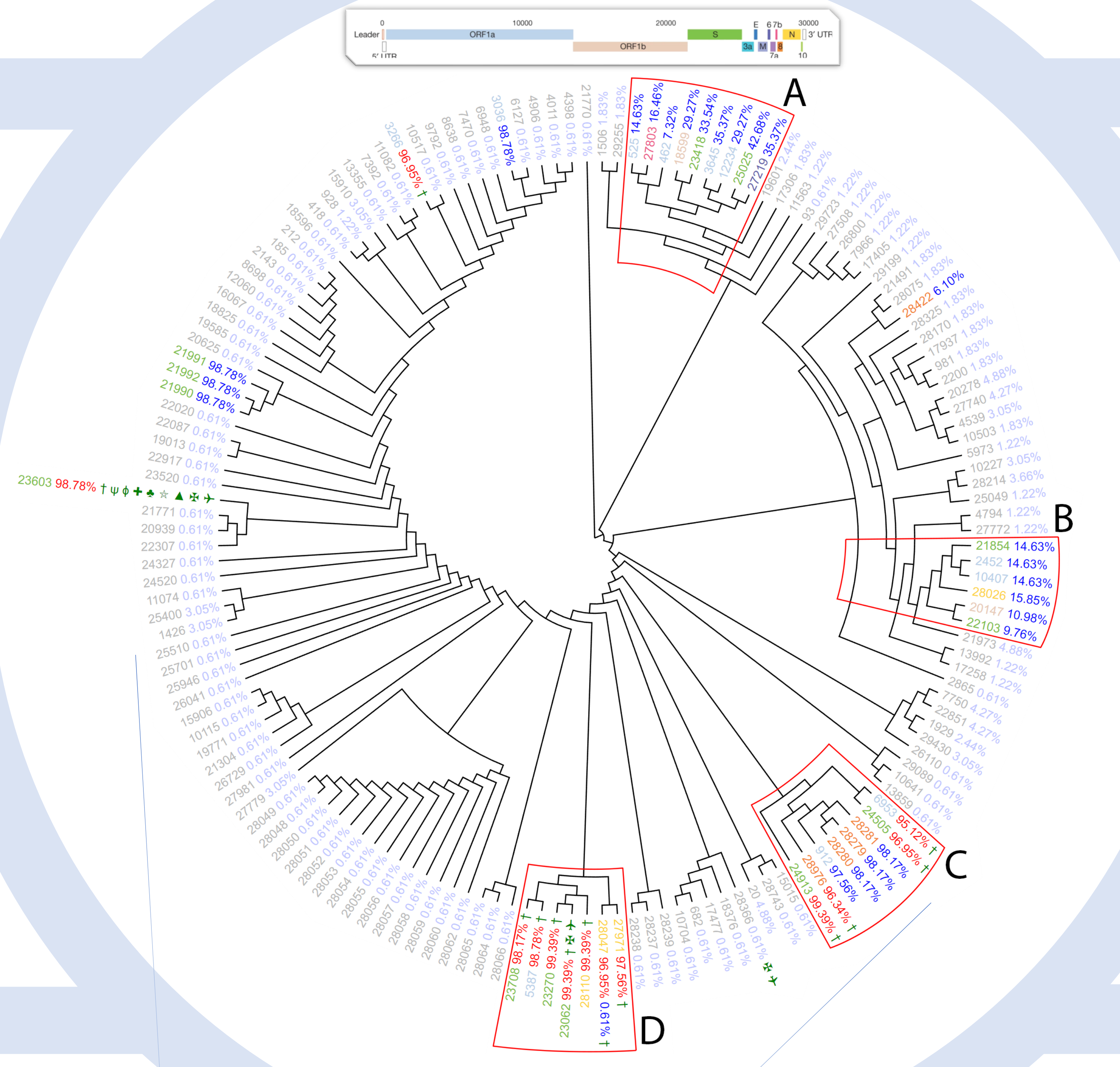
2 Background – Rationale

- Inferring a reliable phylogeny on SARS-CoV-2 is an **inherently complex** task [1].
- Existing classification methods of SARS-CoV-2 populations depend on phylogenetic inference.
- Many novel sub-typing methods fail to determine the phylogenetic relationships among different sub-types [3].

Aim Of The Study 3

- **Can we detect new patterns of co-occurring mutations beyond the strain-specific / strain-defining ones, in SARS-CoV-2 data, through the application of ML methods?**
- **Can we use those patterns in order to groups SARS-CoV-2 populations revealing potentially evolutionary paths?**

*Epidemiologically unrelated individuals could be infected with nearly identical viral genomes. **"Who infected whom"** is tremendously difficult to be determined [2].*

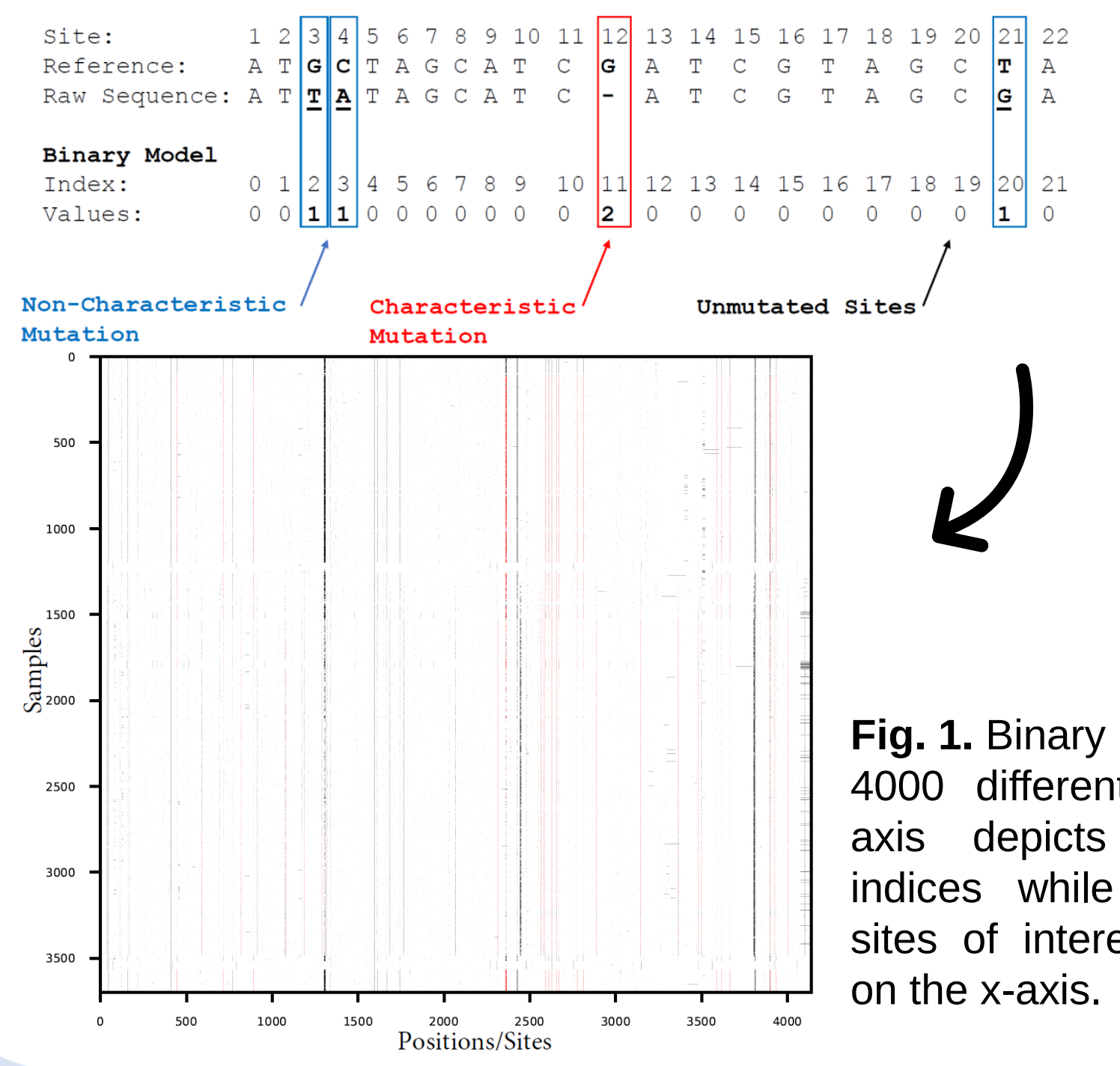


4 Materials

5411 samples and metadata

ENA accession number: PRJEB44141

- Modeling of raw sequences:



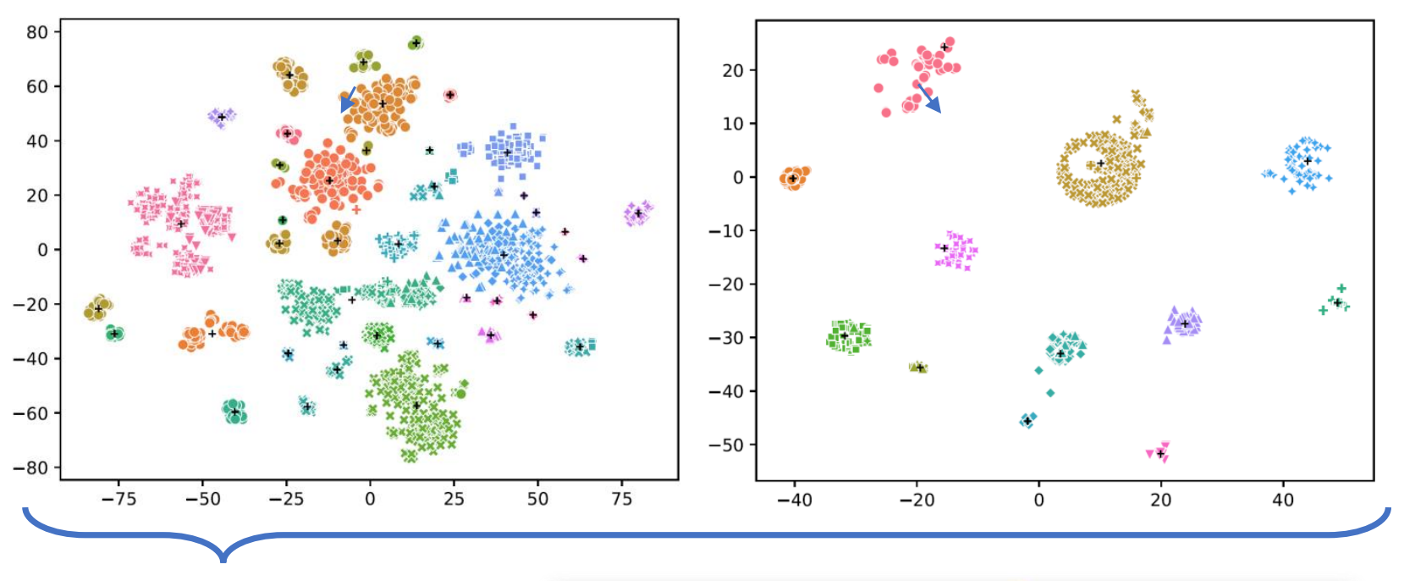
```
Site:          1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22
Reference:     A T G C T A G C A T  C  G  A  T  C  G  T  A  G  C  T  A
Raw Sequence:  A T T A T A G C A T  C  -  A  T  C  G  T  A  G  C  G  A

Binary Model
Index:         0 1 2 3 4 5 6 7 8 9  10 11 12 13 14 15 16 17 18 19 20 21
Values:        0 0 1 1 0 0 0 0 0 0  0  2  0  0  0  0  0  0  0  0  1  0
```

Fig. 1. Binary model Along ~ 4000 different samples. y-axis depicts the sample indices while the mutated sites of interest are shown on the x-axis.

Methods 5

Fig. 2 & 3. Samples clustering on TSNE 2D space based on Non-characteristic & Pango characteristic mutations.



Co-occurring mutations between different clusters of samples of the same lineage were identified.

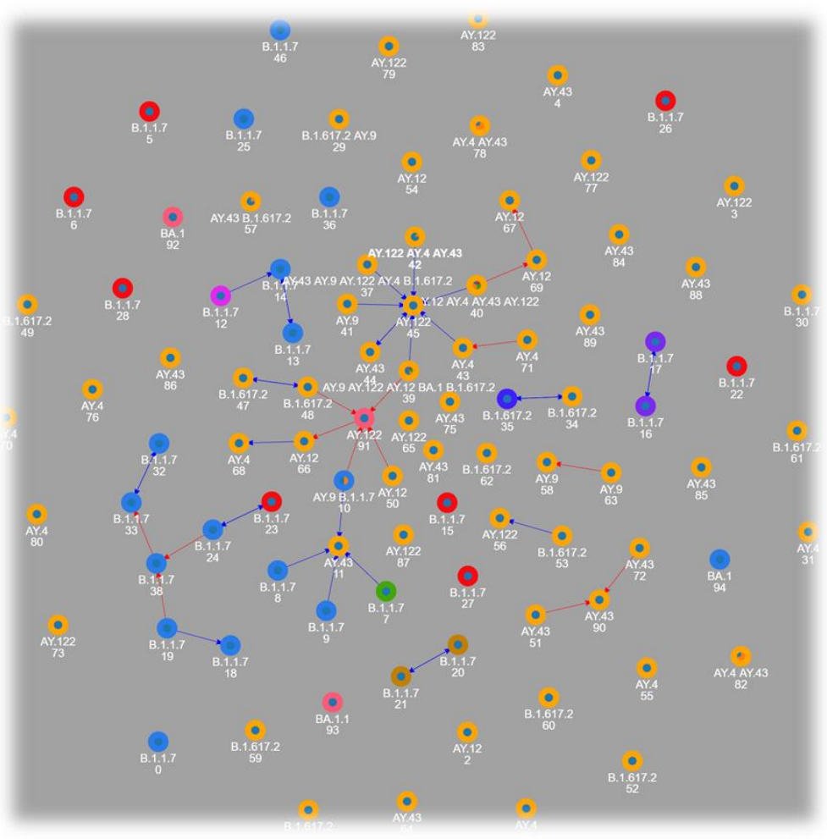


Fig. 4. Directed network. Each node is a cluster, and each arrow implies the existence of at least one co-occurring mutation between two clusters of **Fig. 2.**

6 Results

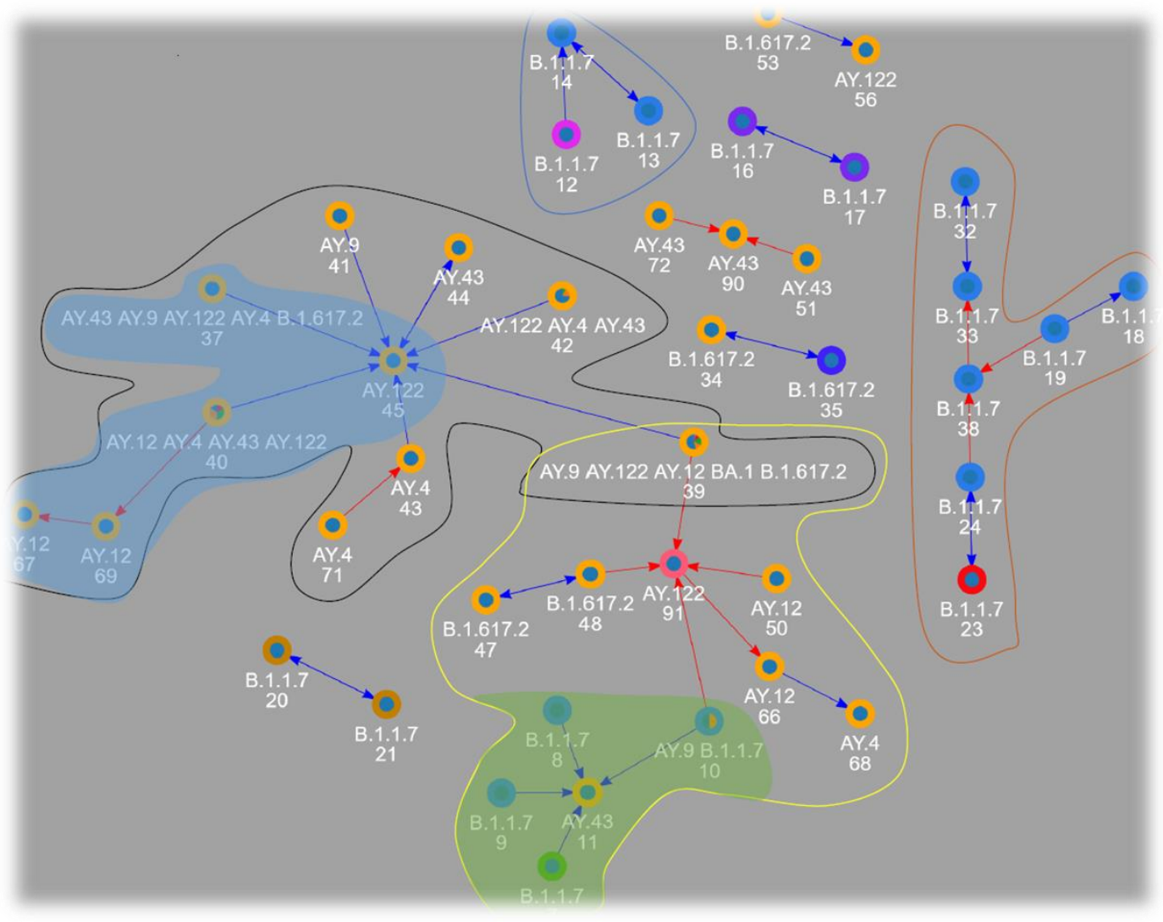
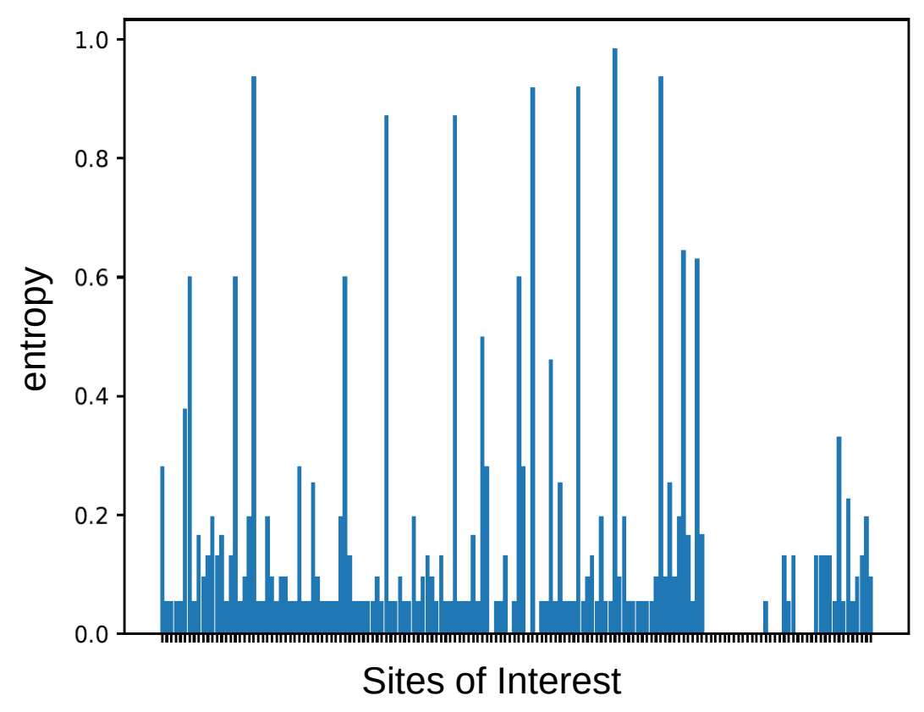
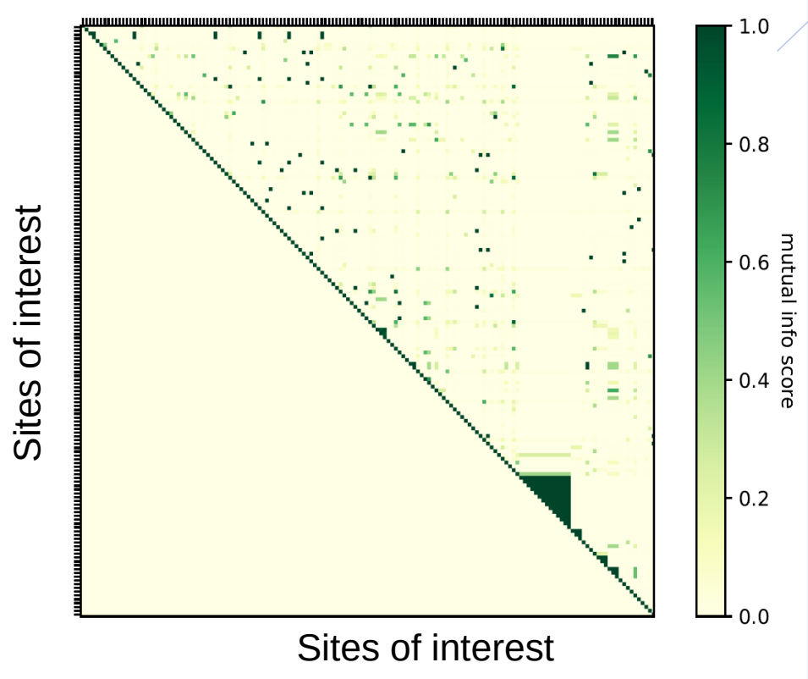


Fig. 5. Paths of interest based on **Fig. 4.**





Nodes : 32 33 38 24 23 19 18

Validation:

For each path:

- MSA of the reported samples was obtained.
- Mutual info between all pairs of sites was calculated.
- Hierarchical clustering.

According to the central dendrogram:

- Group, close to the root of the tree with medium to high rates of non-characteristic mutated sites.
- Sites of non-characteristic mutations appear to be mutated at lower percentages than those in A.
- High prevalence of non-characteristic and characteristic mutations that belong to B.1.1.7 lineage, at very high rates.
- Presence of the B.1.1.7 - characteristic mutation (23062) at very high rates that belongs to the BA.1 and BA.1.1 and it is strongly correlated to other B.1.1.7-characteristic mutations.

Conclusions 7

- We present **a computational method for detecting patterns of co-occurring mutations potentially revealing the evolution of SARS-CoV-2**.
- Evolutionary pressure could lead to new B.1.1.7 sub-lineages, forcing those mutations to prevail.
- **Circulation of non-characteristic mutations closely related to characteristic mutations** could potentially reveal useful patterns.
- Could help us identify potentially important mutations in future lineages.
- Correlation is not causation though, and thus further research is needed to be done on drivers of evolution and the emergence of new mutations.

References 8

1. B. Morel, P. Barbera, L. Czech, B. Bettisworth, L. Hübner, S. Lutteropp, D. Serdari, E.-G. Kostaki, I. Mamais, A. M. Kozlov, P. Pavlidis, D. Paraskevis, and A. Stamatakis. Phylogenetic Analysis of SARS-CoV-2 Data Is Difficult. Molecular Biology and Evolution, 38(5):1777–1791, May 2021.
2. A. L. Valesano, K. E. Rumfelt, D. E. Dimcheff, C. N. Blair, W. J. Fitzsimmons, J. G. Petrie, E. T. Martin, and A. S. Lauring. Temporal dynamics of SARS-CoV-2 mutation accumulation within and across infected hosts. PLOS Pathogens, 17(4):e1009499, 2021. Publisher: Public Library of Science.
3. Z. Zhao, B. A. Sokhansanj, C. Malhotra, K. Zheng, and G. L. Rosen. Genetic grouping of SARS-CoV-2 coronavirus sequences using informative subtype markers for pandemic spread visualization. PLoS computational biology, 16(9):e1008269, Sept. 2020.

GenOptics

*This research has been co-financed by the European Union and Greek national funds through the Operational Program Competitiveness, Entrepreneurship and Innovation, under the call RESEARCH – CREATE – INNOVATE (**GenOptics**, project code: T1EDK-00407).*

Source code

VU UNIVERSITY AMSTERDAM

INAB INSTITUTE OF APPLIED BIOSCIENCES

CERTH CENTRE FOR RESEARCH & TECHNOLOGY HELLAS

ECCB2022