A SARS-CoV-2 Phylogenetic tree based on the Normalized Compression Distance

Presented by: Aloysio Galvão Lopes (aloysio.galvao-lopes@ip-paris.fr)

Professor: Jean-Louis Dessalles

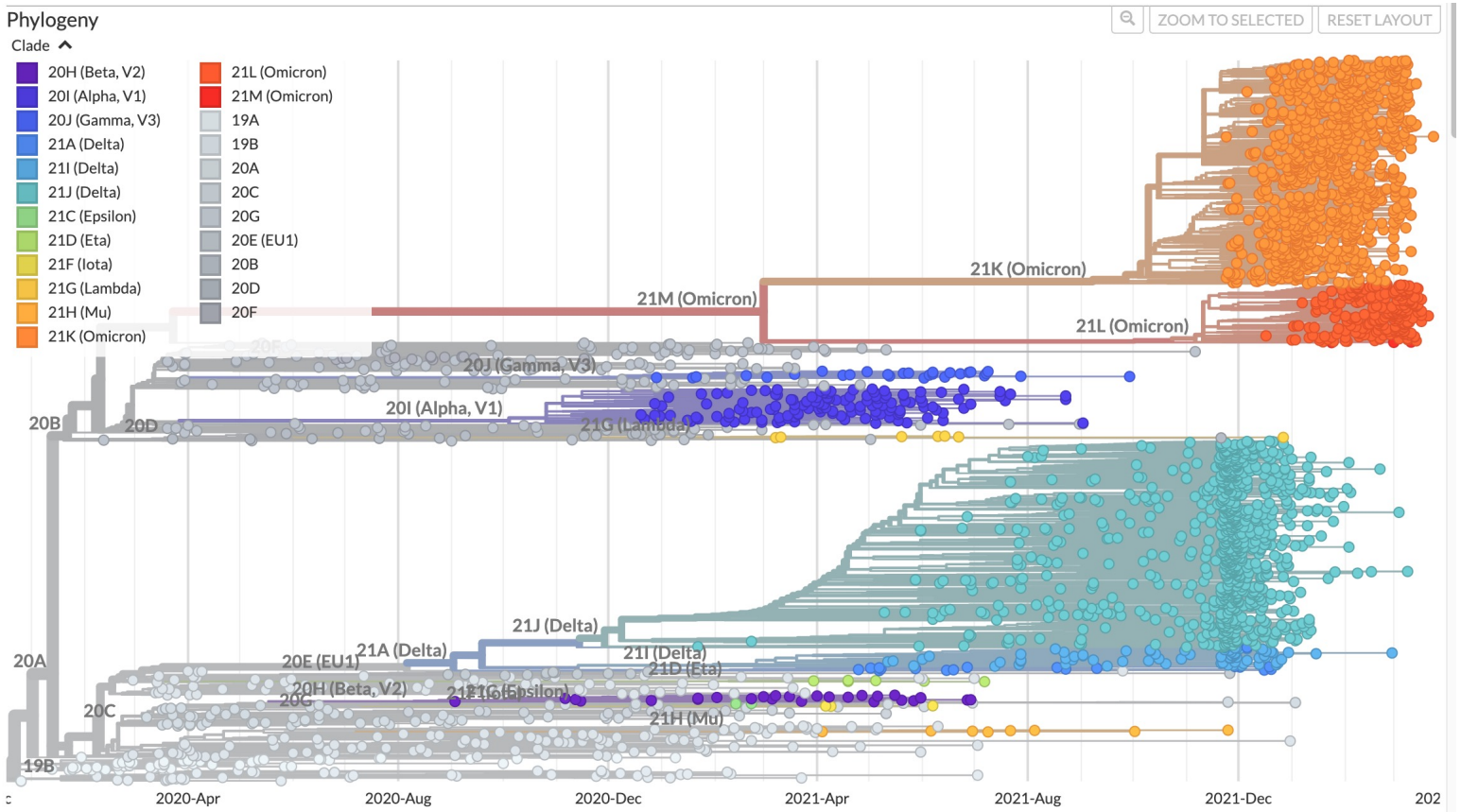
Key topics: Bioinformatics, Normalized Compression Distance, Algorithmic Information, Pangolin Classification, COVID-19

Summary

- Introduction to SARS-CoV-2 Pango naming and lineage classification
- Approach
- Results and comparison with SOTA
- Conclusions

Introduction

- How are SARS-CoV-2 lineages named?



Source: https://nextstrain.org/ncov/gisaid/global

Introduction

- The Pango nomenclature system

  A, A.1, B.1, B.1.1

Introduction

- The Pango nomenclature system

  A, A.1, B.1, B.1.1, B.1.529 (WHO name: Omicron)

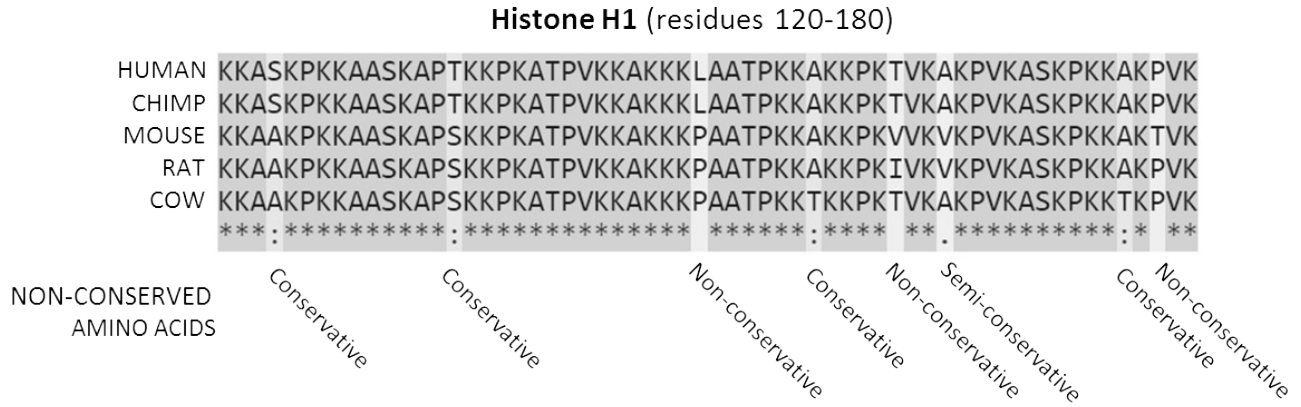
More info: https://www.pango.network/

Introduction

- The Pango nomenclature system

  A, A.1, B.1, B.1.1, B.1.529 (WHO name: Omicron) More info: https://www.pango.network/

- Pangolin uses a ML model to assign lineages, this process needs sequence alignment, they're aligned based on A



Souces: https://cov-lineages.org/ and https://en.wikipedia.org/wiki/Sequence_alignment

Approach

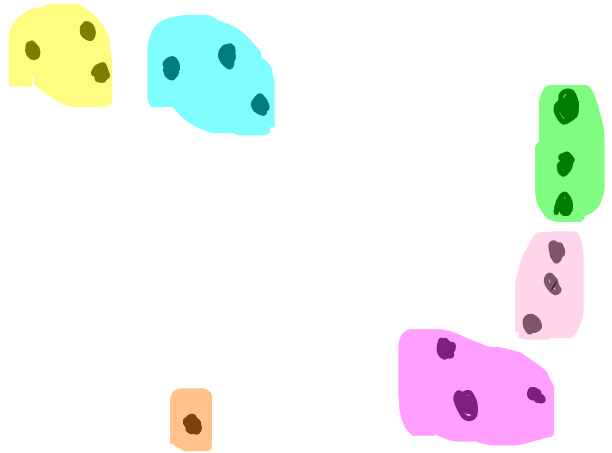
- All sequences taken from NCBI, a little over 52 GB of genome data. Total 883,020 genomes
- Sampled max 2 genomes per day totaling 1247 genomes
- Distance matrix computed in parallel using Python's zlib

Approach

- The distance used is a slightly changed version of the Normalized Compression distance (NCD)

$$NCD(x,y) = \frac{\frac{Z(xy)+Z(yx)}{2} - min(Z(x), Z(y))}{max(Z(x), Z(y))}$$

- Samples are clustered using hierarchical clustering with complete linkage to make for vizualisation and computation purposes



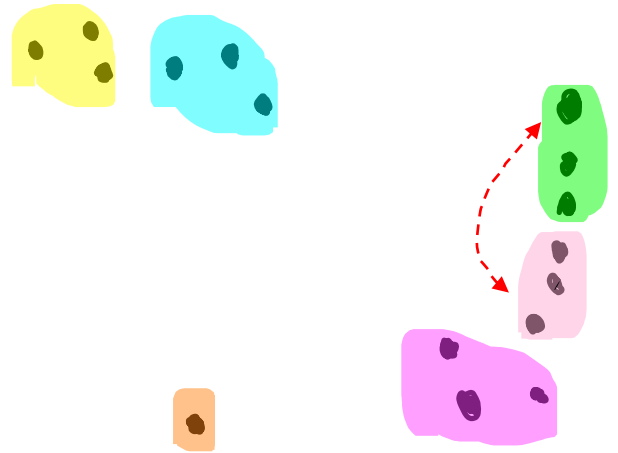
Approach

- The distance used is a slightly changed version of the Normalized Compression distance (NCD)

$$NCD(x,y) = \frac{\frac{Z(xy)+Z(yx)}{2} - min(Z(x), Z(y))}{max(Z(x), Z(y))}$$

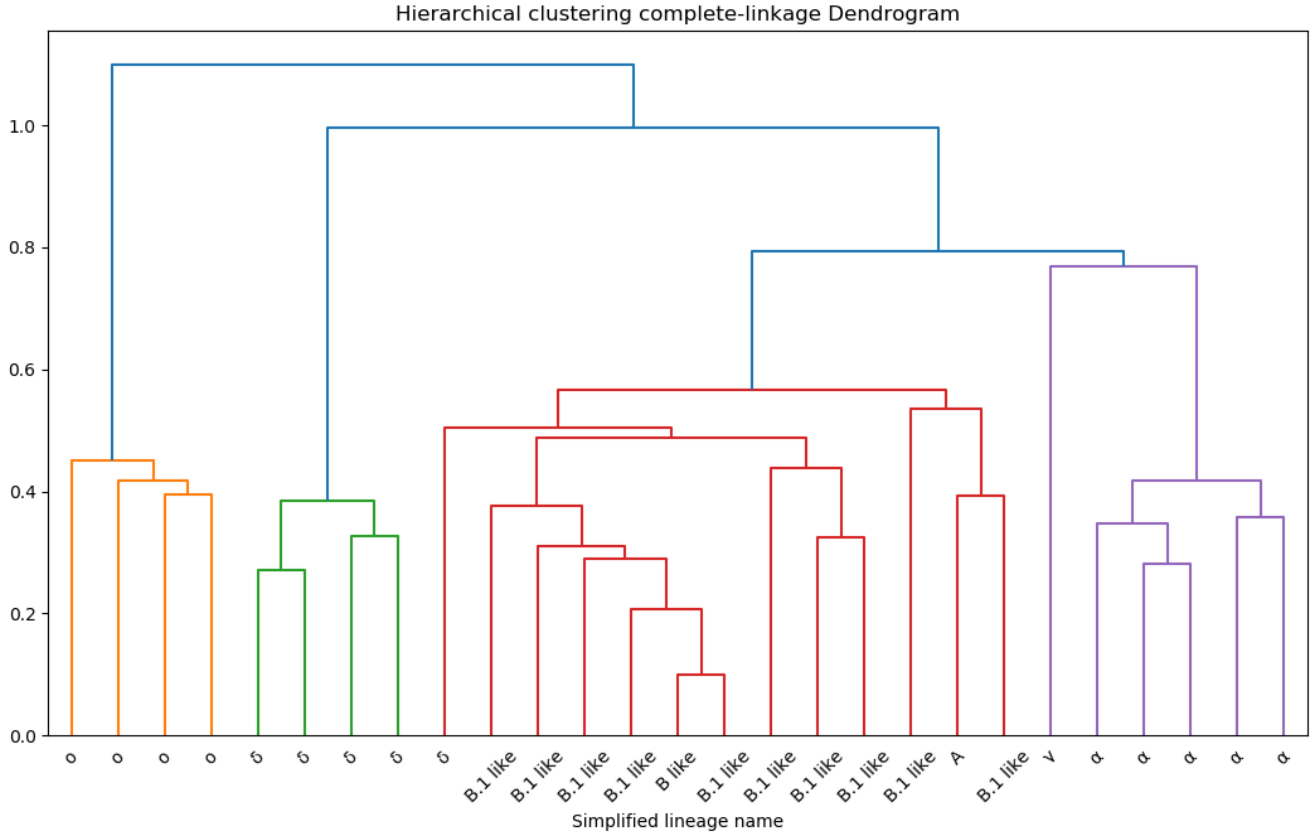
- Samples are clustered using hierarchical clustering with complete linkage to make for vizualisation and computation purposes

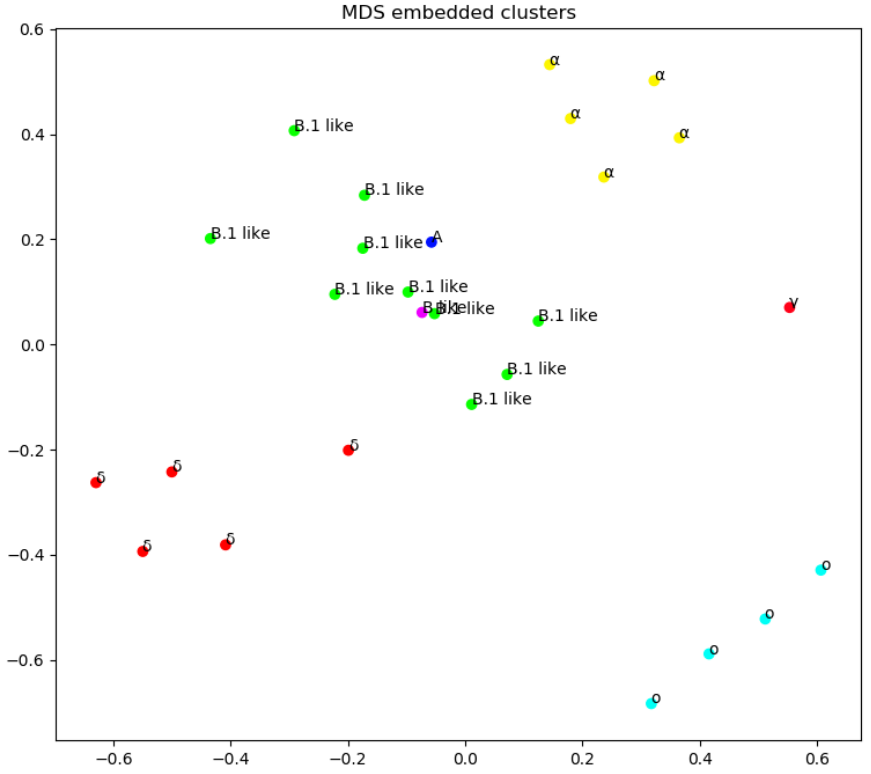
- 27 cluters were generated and used for the phylogenetic tree



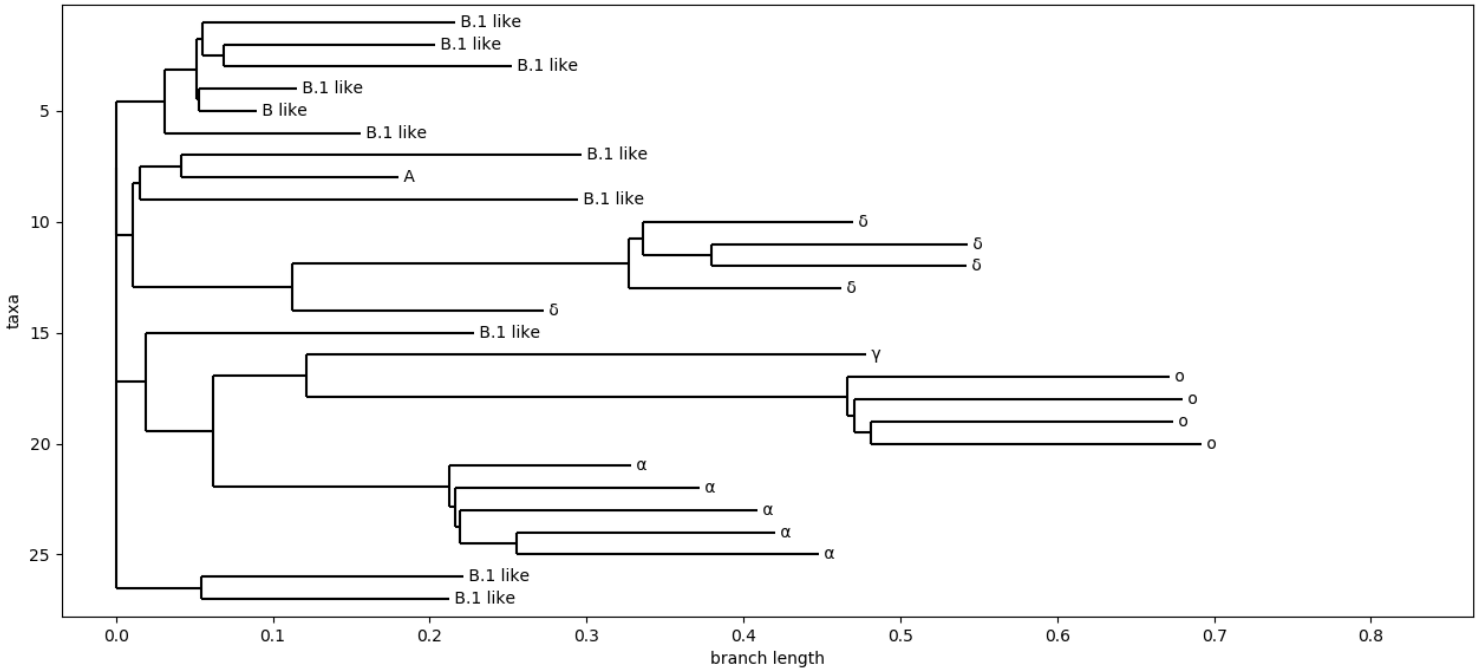
Results

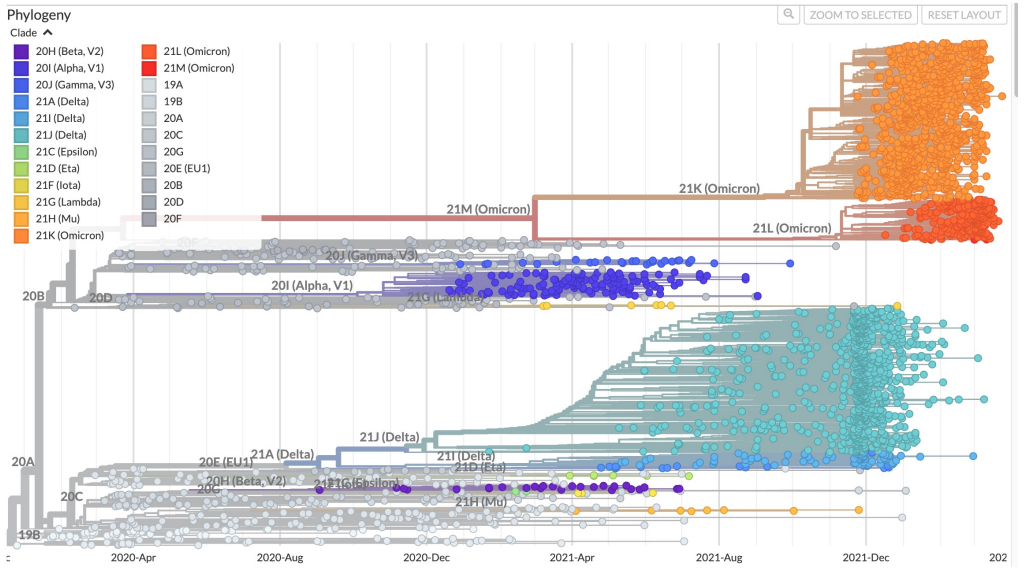
Variant	Start date	Approx onset date	End date
δ	2021-07-14	2021-05-XX	2022-03-04
α	2021-01-05	2020-11-XX	2021-11-24
o	2021-12-31	2021-11-XX	2022-03-21
γ	2021-05-05	2021-02-XX	2022-03-18

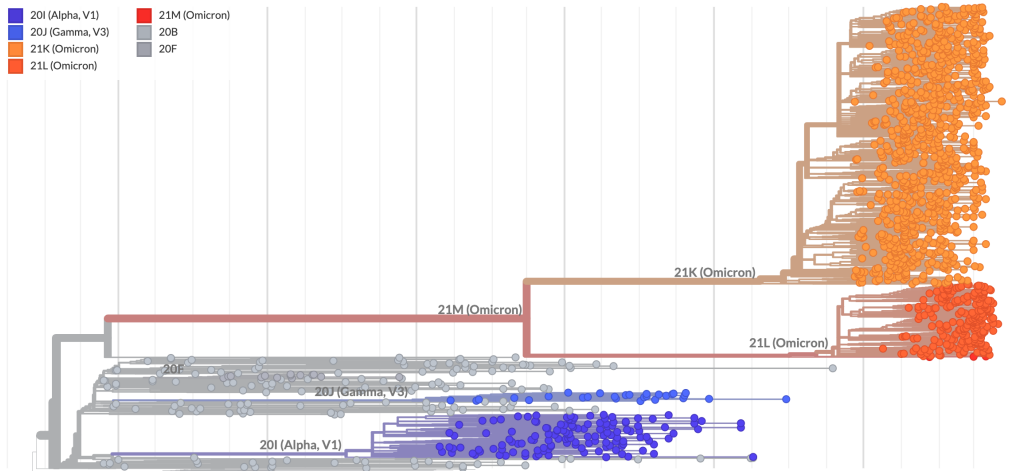




Results


B.1 like
B.1 like
B.1 like
B.1 like
B like
B.1 like
B.1 like
A
B.1 like
δ
δ
δ
δ
B.1 like
γ
ο
ο
ο
ο
α
α
α
α
α
B.1 like
B.1 like
taxa
5
10
15
20
25
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
branch length


Phylogeny
Clade
ZOOM TO SELECTED
RESET LAYOUT
20H (Beta, V2)
20I (Alpha, V1)
20J (Gamma, V3)
21A (Delta)
21I (Delta)
21J (Delta)
21C (Epsilon)
21D (Eta)
21F (Iota)
21G (Lambda)
21H (Mu)
21K (Omicron)
21L (Omicron)
21M (Omicron)
19A
19B
20A
20C
20G
20E (EU1)
20B
20D
20F
21K (Omicron)
21M (Omicron)
21L (Omicron)
20J (Gamma, V3)
20I (Alpha, V1)
20B
20D
21G (Lambda)
20A
21J (Delta)
21A (Delta)
20E (EU1)
21I (Delta)
21D (Eta)
20H (Beta, V2)
21C (Epsilon)
20C
21H (Mu)
19B
2020-Apr
2020-Aug
2020-Dec
2021-Apr
2021-Aug
2021-Dec


20I (Alpha, V1)
20J (Gamma, V3)
21K (Omicron)
21L (Omicron)
21M (Omicron)
20B
20F
21K (Omicron)
21M (Omicron)
21L (Omicron)
20F
20J (Gamma, V3)
20I (Alpha, V1)

Conclusions

Summary

- Using the normalized compression distance (NCD), I've compted distances between a subsample of all covid cases
- I've created aglomerative clusters with complete linkage
- I've used those clusters to build the final tree

Takeaways

- The general genealogy of the virus has been successfully found
- AIT methods don't require sequence alignment
- The NCD can be difficult to intepret, as well as we loose some information when working only with distances