OAK RIDGE National Laboratory

Data-driven whole-genome clustering to detect geospatial, temporal, and functional trends in SARS-CoV-2 evolution

Jean Merlet[1], John Lagergren[2], Verónica G. Melesse Vergara[2], Mikaela Cashman[3], Christopher Bradburne[4], Raina Plowright[5], Emily Gurley[4], Wayne Joubert[2], Daniel Jacobson[2]

[1]University of Tennessee, [2]Oak Ridge National Laboratory, [3]Berkeley National Laboratory, [4]Johns Hopkins University, [5]Cornell University

ORNL is managed by UT-Battelle LLC for the US Department of Energy

U.S. DEPARTMENT OF ENERGY

OAK RIDGE National Laboratory

Covid-19 pandemic

- More than 750 million cases
- Over 200,000 new cases / week

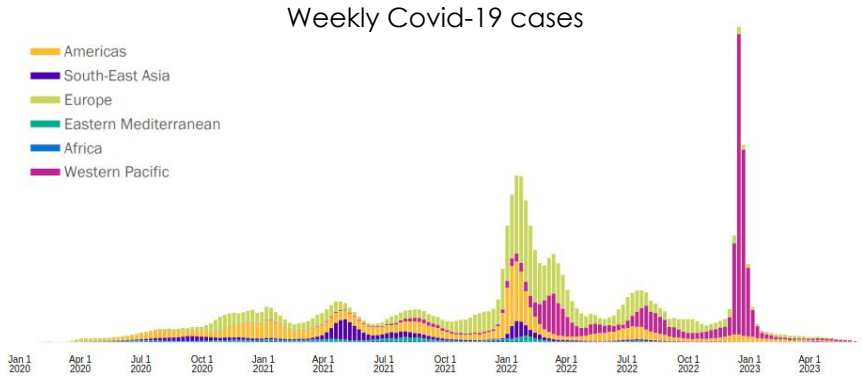
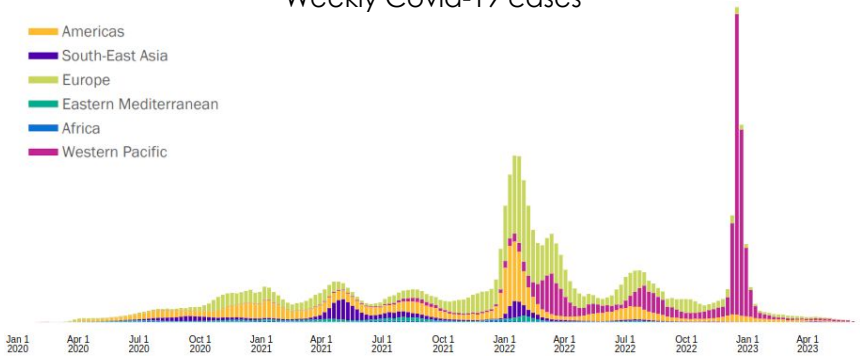


Image: WHO

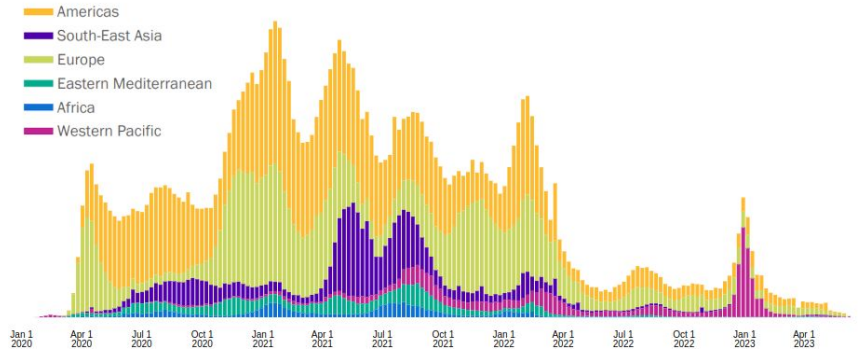
Covid-19 pandemic

- More than 750 million cases
- Over 200,000 new cases / week
- More than 7 million deaths
- Over 1,300 deaths / week

Weekly Covid-19 cases



Weekly Covid-19 deaths

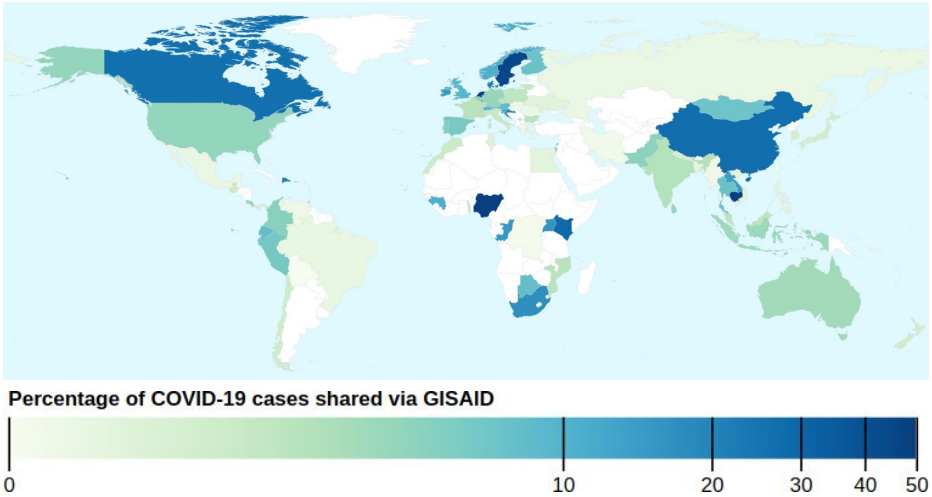


Images: WHO

Covid-19 pandemic

- More than 750 million cases
- Over 200,000 new cases / week
- More than 7 million deaths
- Over 1,300 deaths / week
- Nearly 16 million SARS-CoV-2 sequences on GISAID
- Over 17,000 uploads / month

Last 3 months of sequence uploads



Images: gisaid.org

SARS-CoV-2 genome sequences

- ~30,000 nucleotides
  - A, T, C, G, N, D

Image: astrochem.org

SARS-CoV-2 genome sequences

- ~30,000 nucleotides
  - A, T, C, G, N, D
- Known sequencing biases
  - Geospatial
  - Lag time
  - Selective sampling

Last 3 months of sequence uploads

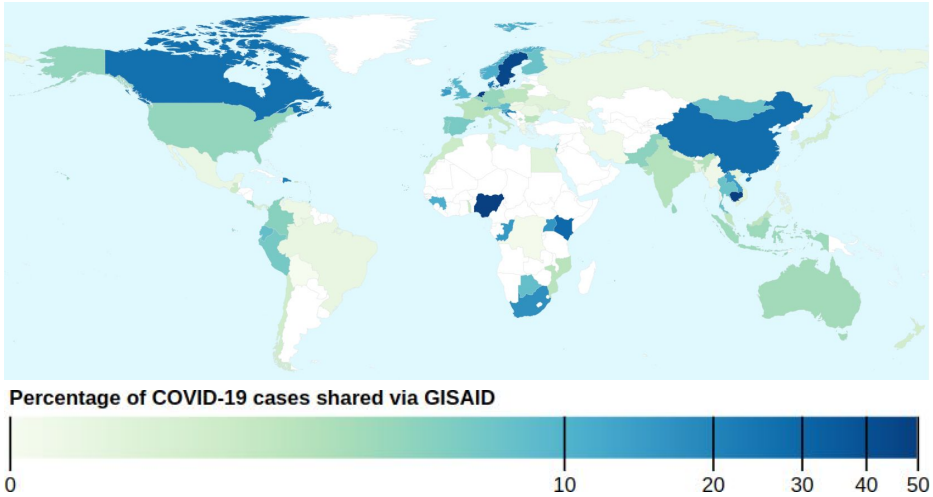


Image: gisaid.org

SARS-CoV-2 genome sequences

- ~30,000 nucleotides
  - A, T, C, G, N, D
- Known sequencing biases
  - Geospatial
  - Lag time
  - Selective sampling
- ~24,000 nucleotide positions have shown mutations

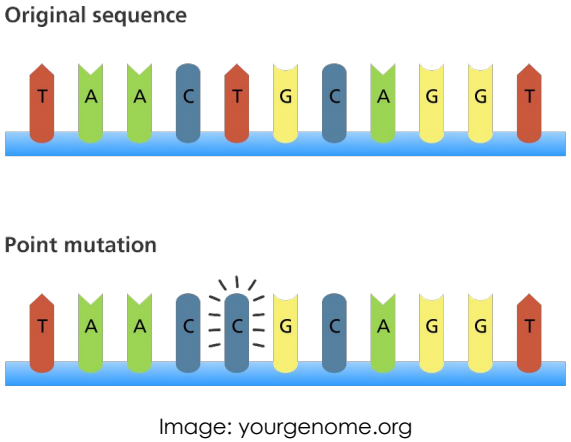


Image: yourgenome.org

Lineage classification

- **P**hylogenetic **A**ssignment of **N**amed **G**lobal **O**utbreak

Lineage classification

- **P**hylogenetic **A**ssignment of **N**amed **G**lobal **O**utbreak
- 3,053 lineages

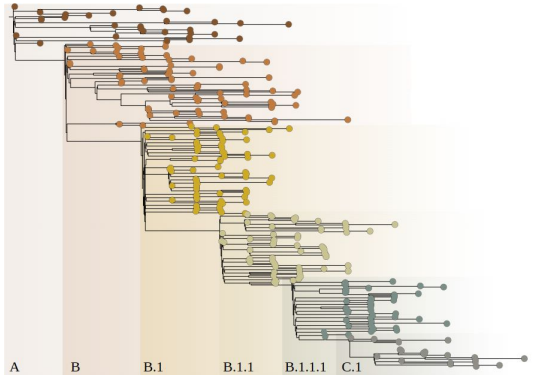


Image: pango.network

Lineage classification

- **P**hylogenetic **A**ssignment of **N**amed **G**lobal **O**utbreak
- 3,053 lineages
- Automated lineage assignment through pangolin

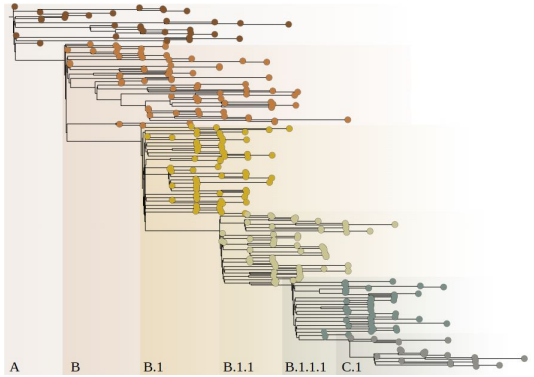


Image: pango.network

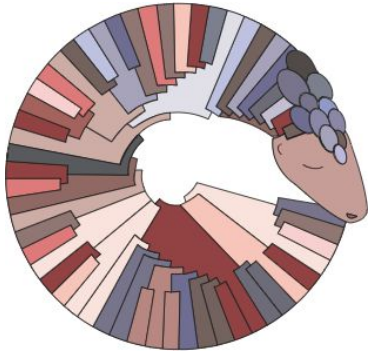


Image: cov-lineages.org

Lineage classification

- **P**hylogenetic **A**ssignment of **N**amed **G**lobal **O**utbreak
- 3,053 lineages
- Automated lineage assignment through pangolin
- Lineages are based on small, hand-curated sets of mutations

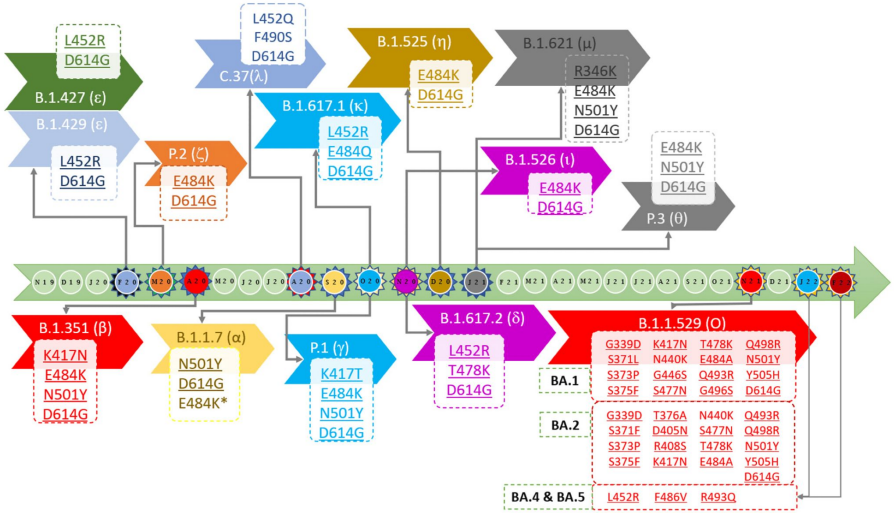


Image adapted from: Bhadane, R. and Salo-Ahen, O.M.H. 2022. High-Throughput Molecular Dynamics-Based Alchemical Free Energy Calculations for Predicting the Binding Free Energy Change Associated with the Selected Omicron Mutations in the Spike Receptor-Binding Domain of SARS-CoV-2. Biomedicines. MDPI AG.

Lineage classification

- **P**hylogenetic **A**ssignment of **N**amed **G**lobal **O**utbreak
- 3,053 lineages
- Automated lineage assignment through pangolin
- Lineages are based on small, hand-curated sets of mutations
- Need a data-driven, whole-genome model

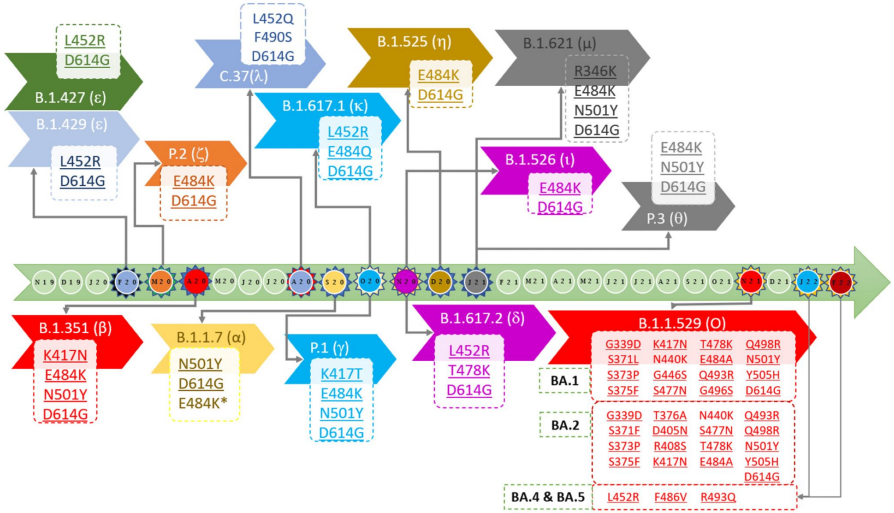


Image adapted from: Bhadane, R. and Salo-Ahen, O.M.H. 2022. High-Throughput Molecular Dynamics-Based Alchemical Free Energy Calculations for Predicting the Binding Free Energy Change Associated with the Selected Omicron Mutations in the Spike Receptor-Binding Domain of SARS-CoV-2. Biomedicines. MDPI AG.

Overview: data-driven strains

Millions of genomes

```
seq. 1: ATTAAAG …
seq. 2: ATTAAAG …
seq. 3: ATTAAAG …
        …
```

Combinatorial **Met**rics (CoMet)

```
seq. 1: 0000101 …
seq. 2: 0000101 …
seq. 3: 0000101 …
        …
```

High **P**erformance **C**omputing (HPC)

Genome-genome similarity network

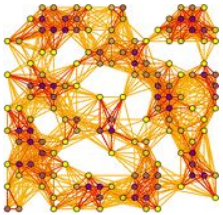


Image: micans.org/mcl

High **P**erf. **M**arkov **Cl**ustering (HipMCL)

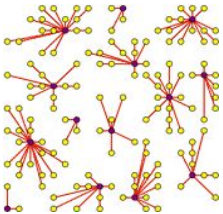
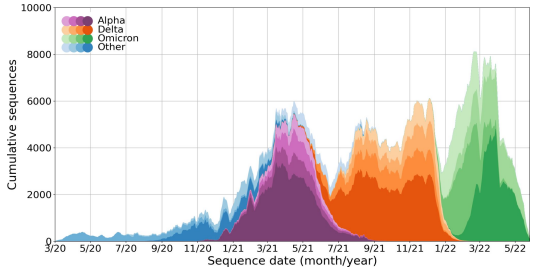
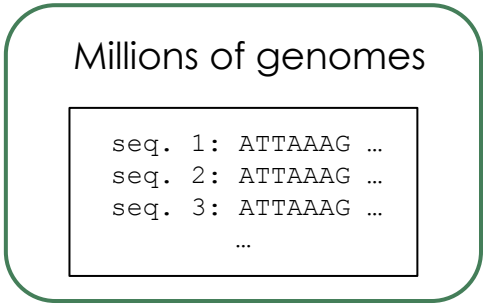


Image: micans.org/mcl

Spatiotemporal trajectories for data-driven lineages



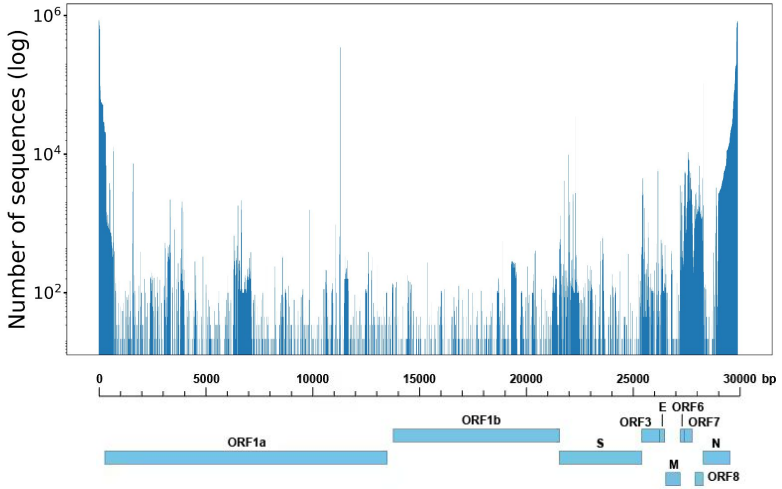
Overview: data-driven strains



Data pre-processing

- Challenges
  - Low-quality sequences

Distribution of deletions (Ds)



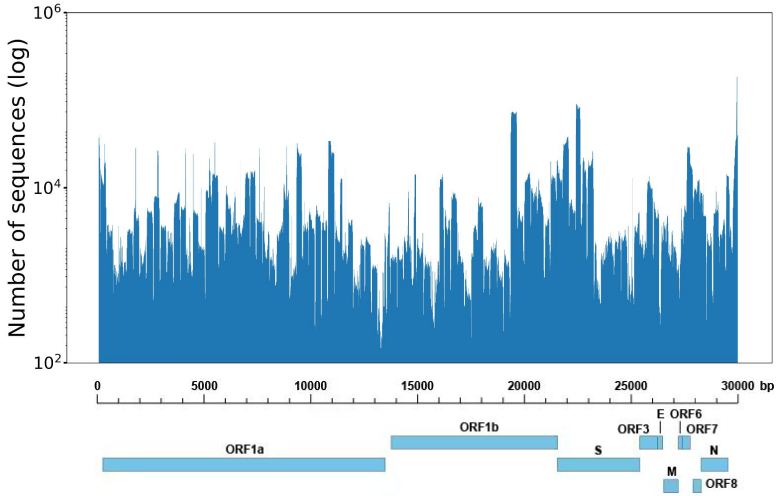
x-axis: SARS-CoV-2 genome

Genome adapted from novusbio.com

Data pre-processing

- Challenges
  - Low-quality sequences
  - Sequencing error

Distribution of Ns



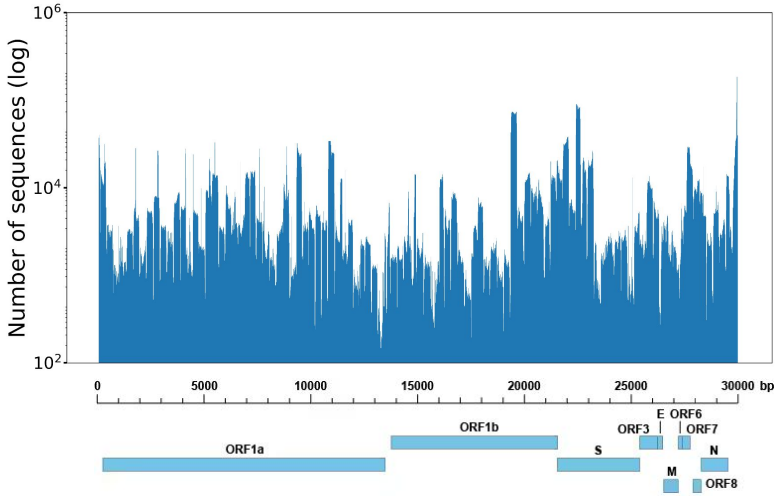
x-axis: SARS-CoV-2 genome

Genome adapted from novusbio.com

Data pre-processing

- Challenges
  - Low-quality sequences
  - Sequencing error
  - Incomplete metadata

Distribution of Ns



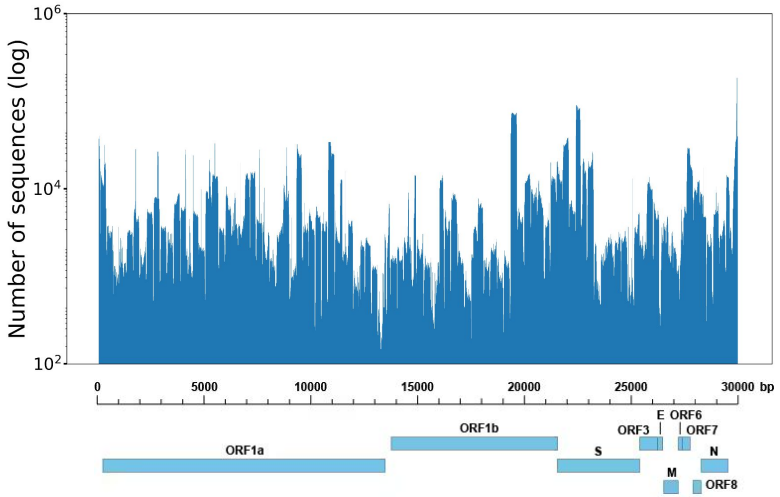
x-axis: SARS-CoV-2 genome

Genome adapted from novusbio.com

Data pre-processing

- Challenges
  - Low-quality sequences
  - Sequencing error
  - Incomplete metadata
- 11.0 -> 7.7 million sequences

Distribution of Ns



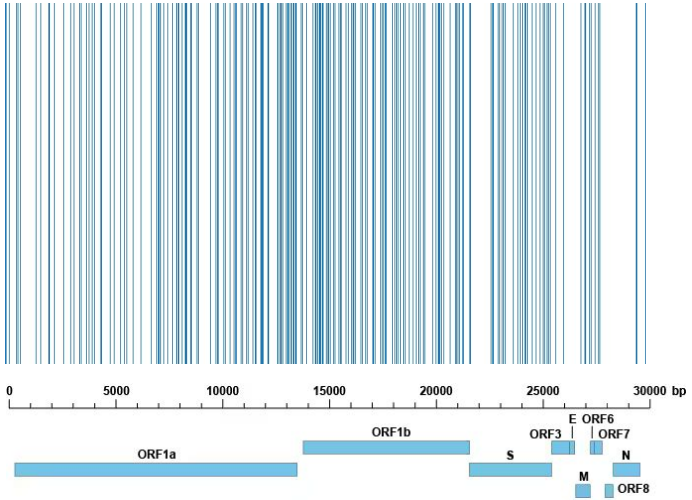
x-axis: SARS-CoV-2 genome

Genome adapted from novusbio.com

Data pre-processing

- Challenges
  - Low-quality sequences
  - Sequencing error
  - Incomplete metadata
- 11.0 -> 7.7 million sequences
- Detect significant mutation positions

Distribution of Ns



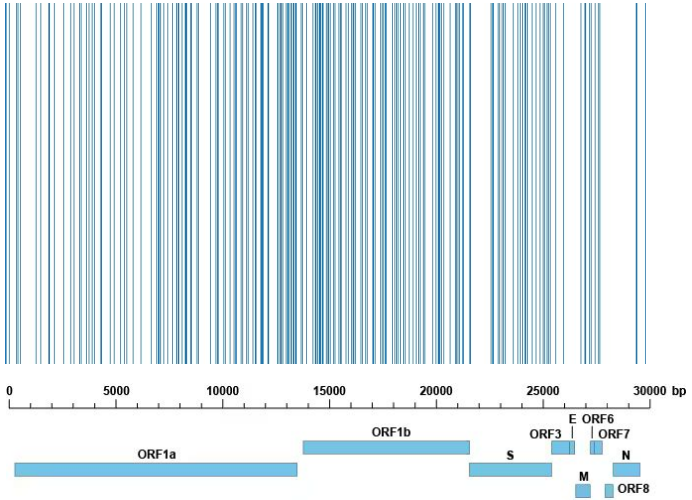
x-axis: SARS-CoV-2 genome

Genome adapted from novusbio.com

Data pre-processing

- Challenges
  - Low-quality sequences
  - Sequencing error
  - Incomplete metadata
- 11.0 -> 7.7 million sequences
- Detect significant mutation positions
- 29.9 -> 21.4 thousand nucleotides

Distribution of Ns



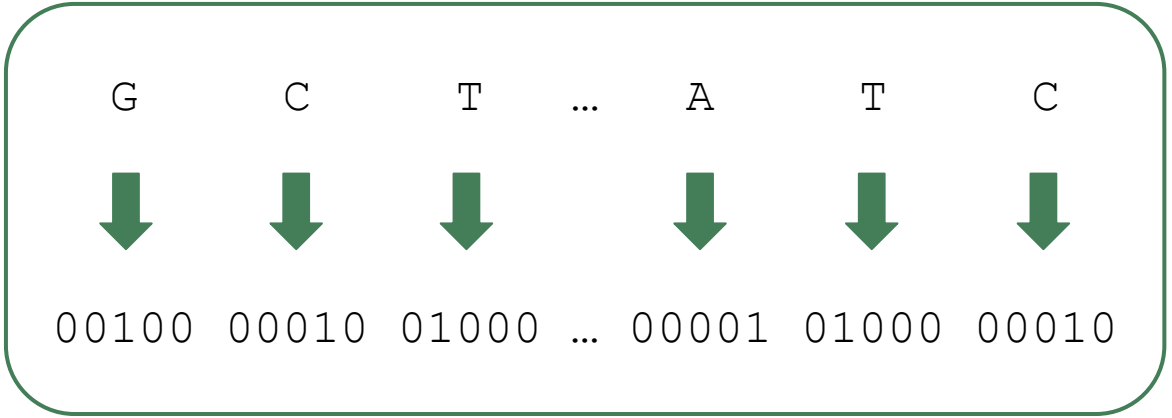
x-axis: SARS-CoV-2 genome

Genome adapted from novusbio.com

SARS-CoV-2 genome vector binarization

Binarization scheme

A: 00001
C: 00010
G: 00100
T: 01000
D: 10000
N: 00000



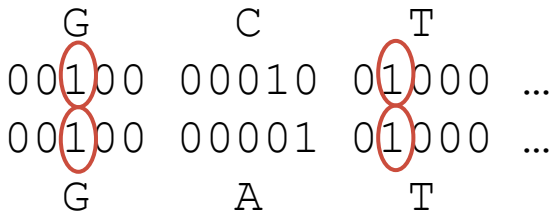
21,400 **nucleotides**

↓

107,000 **bits**

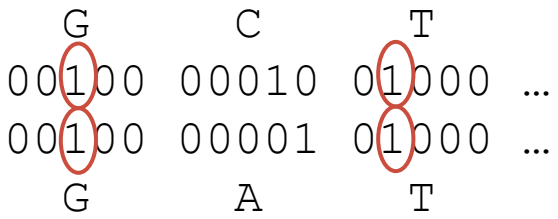
Duo similarity metric

- Compares binary vectors



Duo similarity metric

- Compares binary vectors
- 2-way
  - i and j binary vectors
  - N total vectors
  - $q = \frac{2}{3}$ scaling constant
  - D_{ij} comparison frequency
  - f_i vector i binary frequency

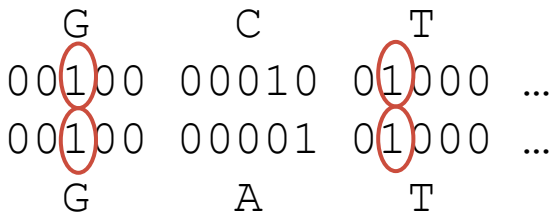


$$\mathrm{Duo}_{ij} = 5D_{ij}\left(1 - \frac{f_i}{q}\right)\left(1 - \frac{f_j}{q}\right)$$

$$D_{ij} = \frac{1}{N}\sum_{n=1}^{N} \mathbb{1}\{i_n = 1, j_n = 1\}$$

Duo similarity metric

- Compares binary vectors
- 2-way
  - i and j binary vectors
  - N total vectors
  - $q = ⅔$ scaling constant
  - D_{ij} comparison frequency
  - f_i vector i binary frequency
- Choice of $q = 0$ converts to Sørensen-Dice index

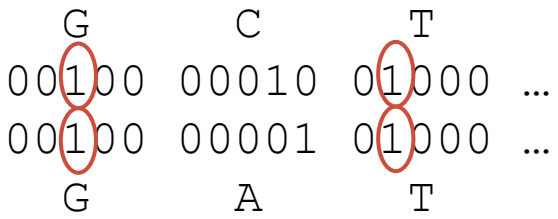


$$\mathrm{SD}_{ij} = 5D_{ij}$$

$$D_{ij} = \frac{1}{N}\sum_{n=1}^{N} \mathbb{1}\{i_n = 1, j_n = 1\}$$

Duo similarity metric

- Compares binary vectors
- 2-way
  - i and j binary vectors
  - N total vectors
  - $q = ⅔$ scaling constant
  - D_{ij} comparison frequency
  - f_i vector i binary frequency
- Choice of $q = 0$ converts to Sørensen-Dice index
- Also have 3-way capability



$$\mathrm{SD}_{ij} = 5D_{ij}$$

$$D_{ij} = \frac{1}{N}\sum_{n=1}^{N} \mathbb{1}\{i_n = 1, j_n = 1\}$$

Overview: data-driven strains

Millions of genomes

```
seq. 1: ATTAAAG …
seq. 2: ATTAAAG …
seq. 3: ATTAAAG …
        …
```

Combinatorial **Met**rics (CoMet)

```
seq. 1: 0000101 …
seq. 2: 0000101 …
seq. 3: 0000101 …
        …
```

High **P**erformance **C**omputing (HPC)

Genome-genome similarity network

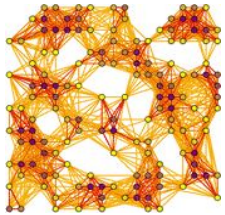
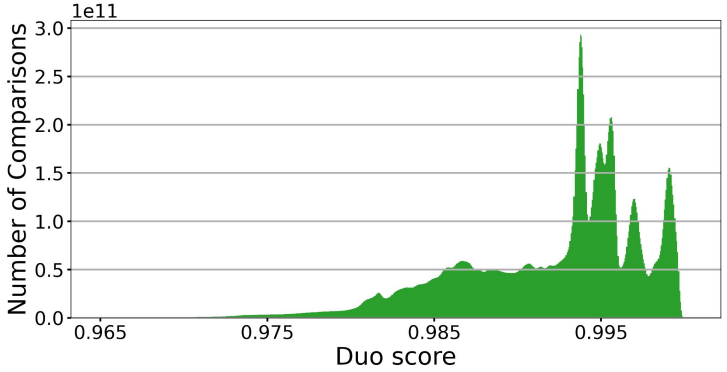


Image: micans.org/mcl

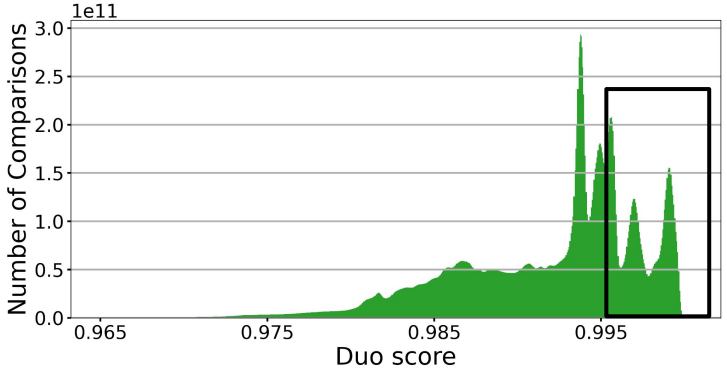
CoMet: Network generation

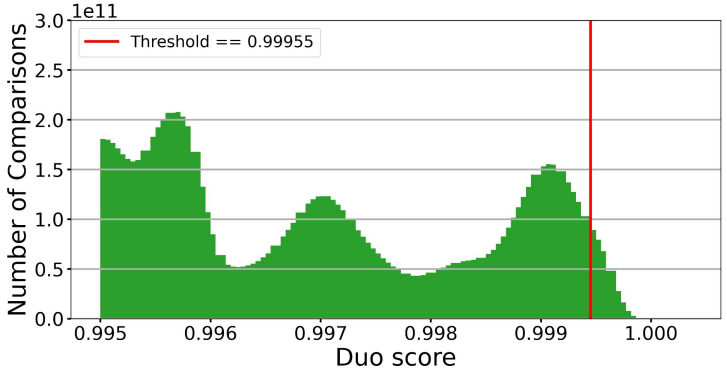
- 7.7 million binarized vectors
- Vectors each 107,000 bits
- 2.98 x 10^{13} vector comparisons



CoMet: Network generation

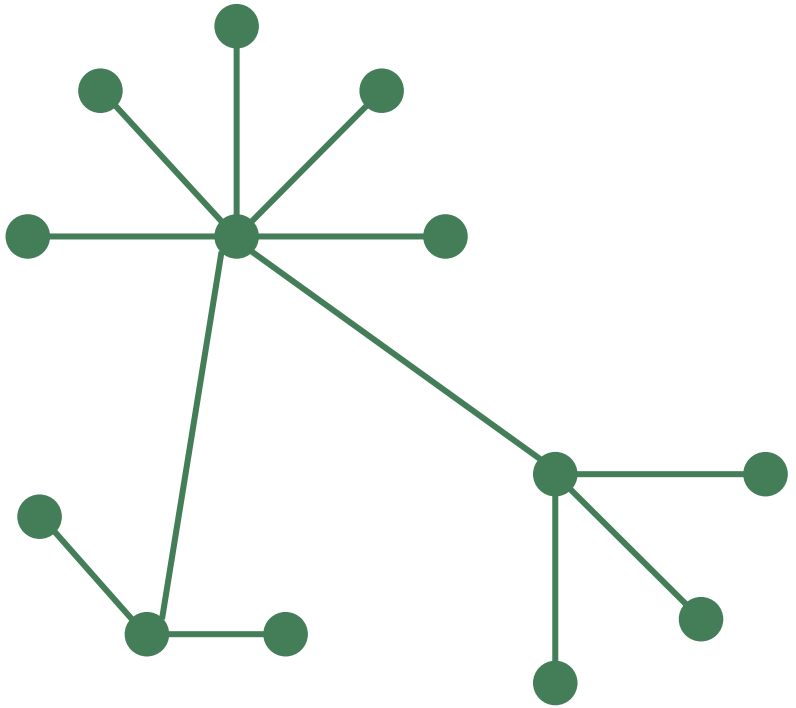
- 7.7 million binarized vectors
- Vectors each 107,000 bits
- 2.98 x 10^{13} vector comparisons
- Top 0.1% most similar vectors
  - 5.3 million nodes
  - 778 trillion edges
  - 147 thousand edges / node





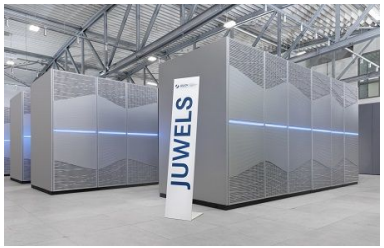
CoMet: Network generation

- 7.7 million binarized vectors
- Vectors each 107,000 bits
- 2.98 x 10^{13} vector comparisons
- Top 0.1% most similar vectors
  - 5.3 million nodes
  - 778 trillion edges
  - 147 thousand edges / node
- Genome-to-genome network



CoMet

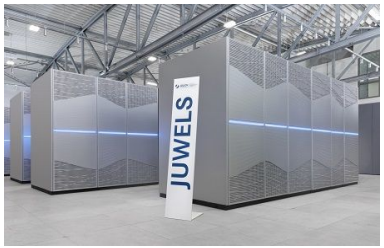
- Deployed on various HPC systems
  - **Summit** (#5 TOP500 list)
  - **Frontier** (#1 TOP500 list)
  - JUWELS Booster
  - Perlmutter
- Awarded Gordon Bell in 2018



Images: olcf.ornl.gov, fz-juelich.de, nersc.gov

CoMet

- Deployed on various HPC systems
  - **Summit** (#5 TOP500 list)
  - **Frontier** (#1 TOP500 list)
  - JUWELS Booster
  - Perlmutter
- Awarded Gordon Bell in 2018
- Application agnostic



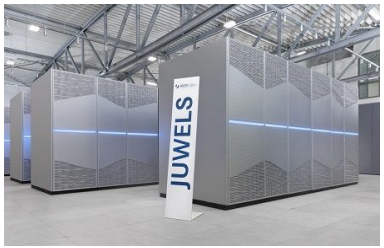
Images: olcf.ornl.gov, fz-juelich.de, nersc.gov

CoMet

- Deployed on various HPC systems
  - **Summit** (#5 TOP500 list)
  - **Frontier** (#1 TOP500 list)
  - JUWELS Booster
  - Perlmutter
- Awarded Gordon Bell in 2018
- Application agnostic
- Record-setting computations

2.34 ExaFLOPs

6.6 ExaFLOPs



9.37 ExaFLOPs

TBD ExaFLOPs

Images: olcf.ornl.gov, fz-juelich.de, nersc.gov

Overview: data-driven strains

Millions of genomes

```
seq. 1: ATTAAAG …
seq. 2: ATTAAAG …
seq. 3: ATTAAAG …
        …
```

Combinatorial **Met**rics (CoMet)

```
seq. 1: 0000101 …
seq. 2: 0000101 …
seq. 3: 0000101 …
        …
```

High **P**erformance **C**omputing (HPC)

Genome-genome similarity network

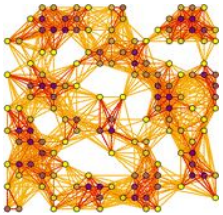


Image: micans.org/mcl

High **P**erf. **M**arkov **Cl**ustering (HipMCL)

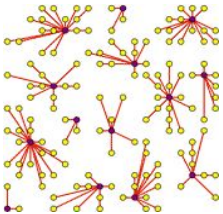


Image: micans.org/mcl

Markov clustering with HipMCL*

- Unsupervised algorithm

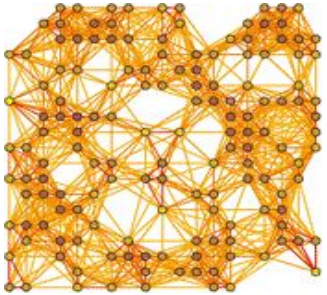
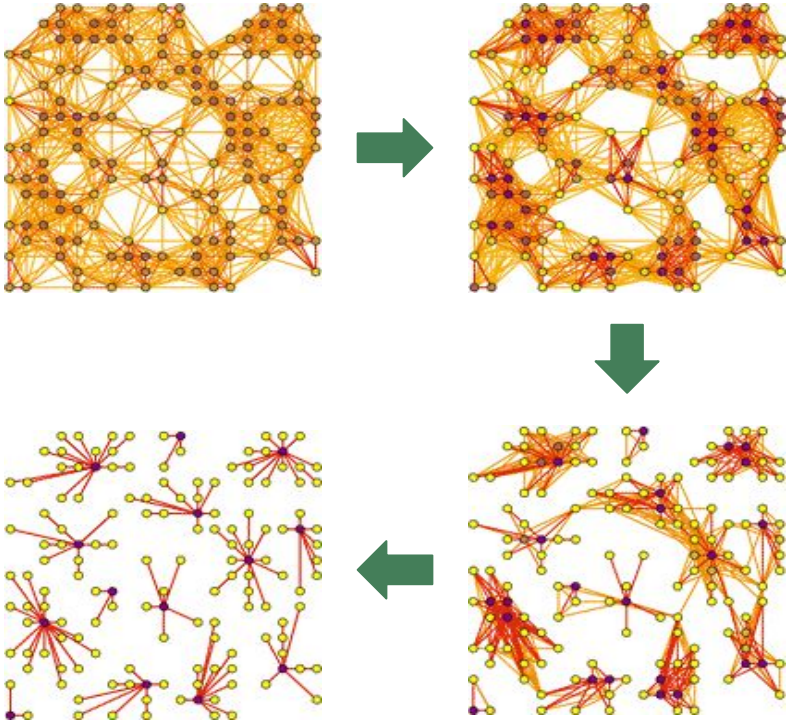


Image: micans.org/mcl

* Azad, A. et al. 2018. HipMCL: a high-performance parallel implementation of the Markov clustering algorithm for large-scale networks. Nucleic Acids Research. Oxford University Press (OUP).

Markov clustering with HipMCL

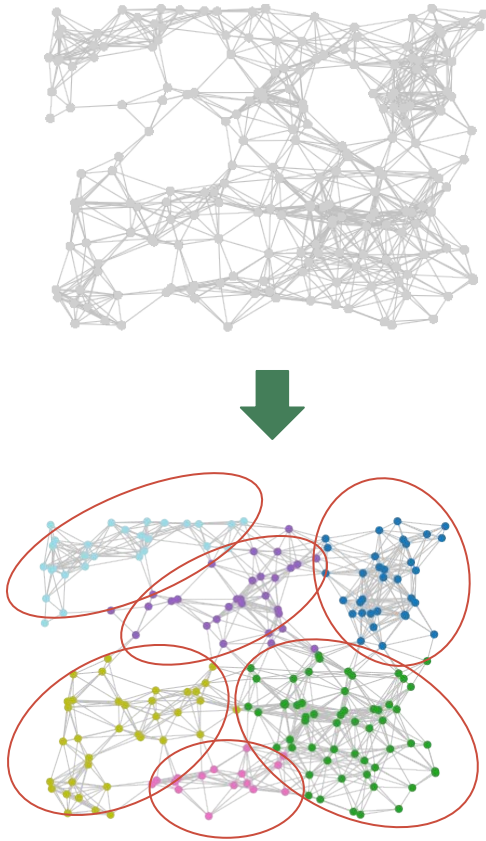
- Unsupervised algorithm
- Random walk
  - expansion (flow spreads)
  - inflation (flow recedes)



Images: micans.org/mcl

Markov clustering with HipMCL

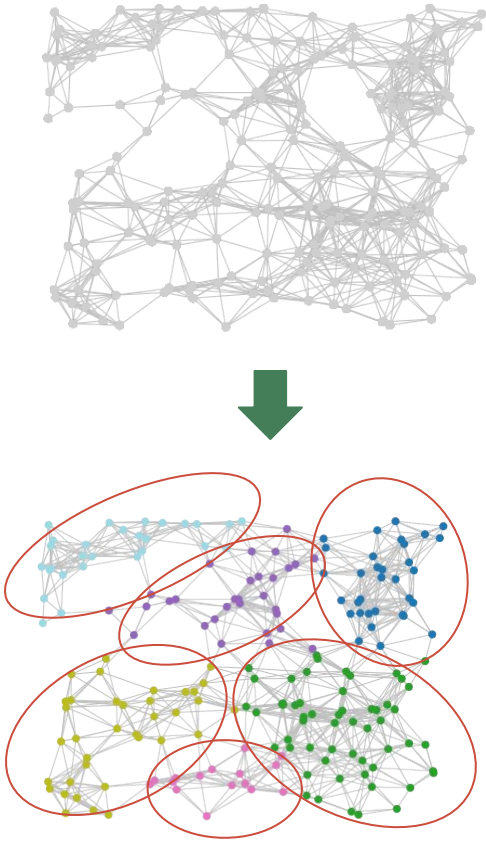
- Unsupervised algorithm
- Random walk
  - expansion (flow spreads)
  - inflation (flow recedes)
- Network -> clusters



Images: github.com/GuyAllard/markov_clustering

Markov clustering with HipMCL

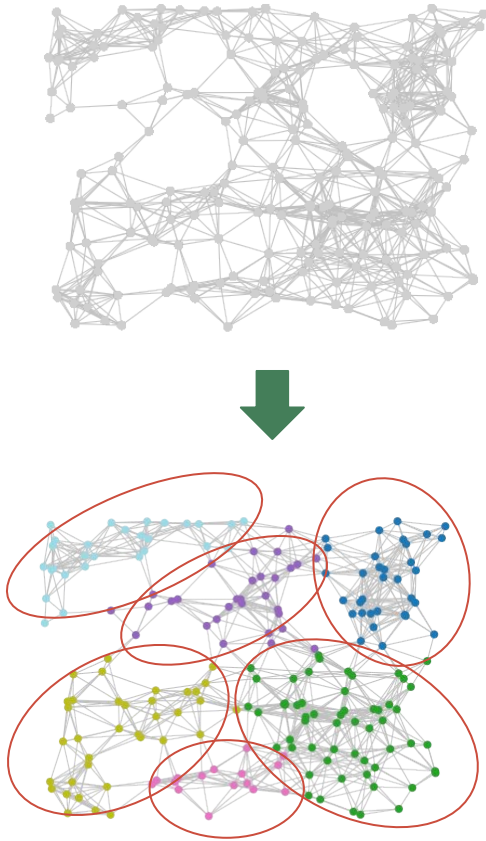
- Unsupervised algorithm
- Random walk
  - expansion (flow spreads)
  - inflation (flow recedes)
- Network -> clusters
- 22,000 SARS-CoV-2 strains



Images: github.com/GuyAllard/markov_clustering

Markov clustering with HipMCL

- Unsupervised algorithm
- Random walk
  - expansion (flow spreads)
  - inflation (flow recedes)
- Network -> clusters
- 22,000 SARS-CoV-2 strains
- Cluster number driven by graph topology



Images: github.com/GuyAllard/markov_clustering

Overview: data-driven strains

Millions of genomes

```
seq. 1: ATTAAAG …
seq. 2: ATTAAAG …
seq. 3: ATTAAAG …
        …
```

Combinatorial **Met**rics (CoMet)

```
seq. 1: 0000101 …
seq. 2: 0000101 …
seq. 3: 0000101 …
        …
```

High **P**erformance **C**omputing (HPC)

Genome-genome similarity network

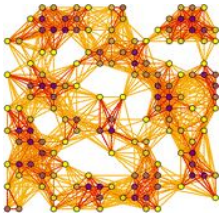


Image: micans.org/mcl

High **P**erf. **M**arkov **Cl**ustering (HipMCL)

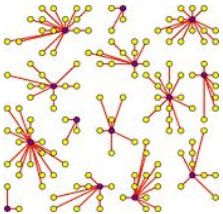
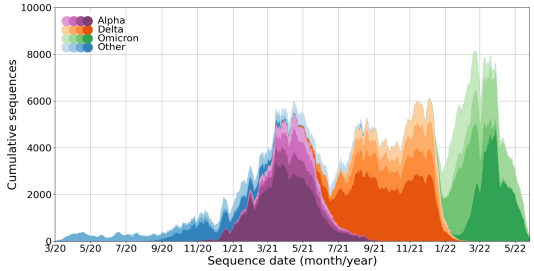


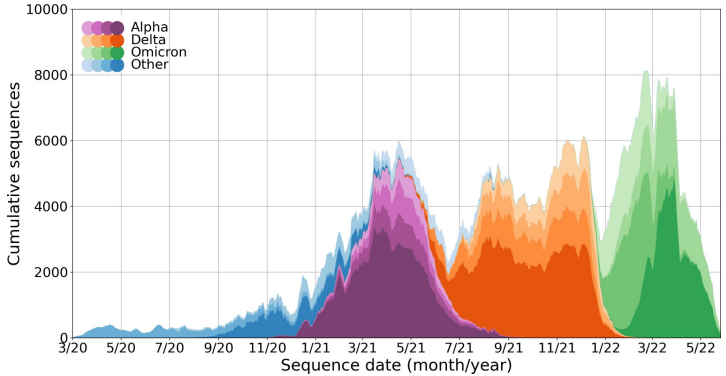
Image: micans.org/mcl

Spatiotemporal trajectories for data-driven lineages



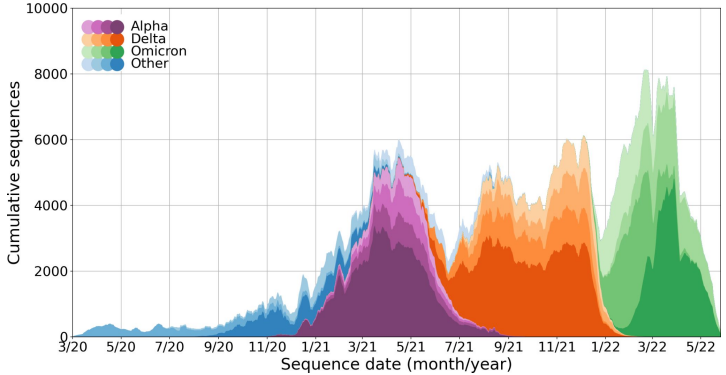
Our SARS-CoV-2 strains and WHO variants

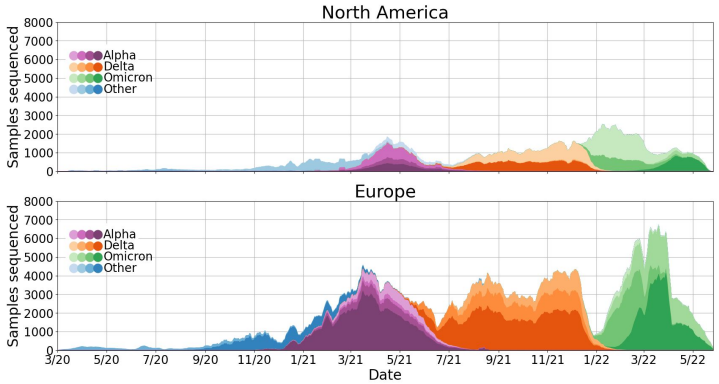
- Mapped clusters to WHO variant
- Four largest clusters per WHO variant
- Longitudinal view



Our SARS-CoV-2 strains and WHO variants

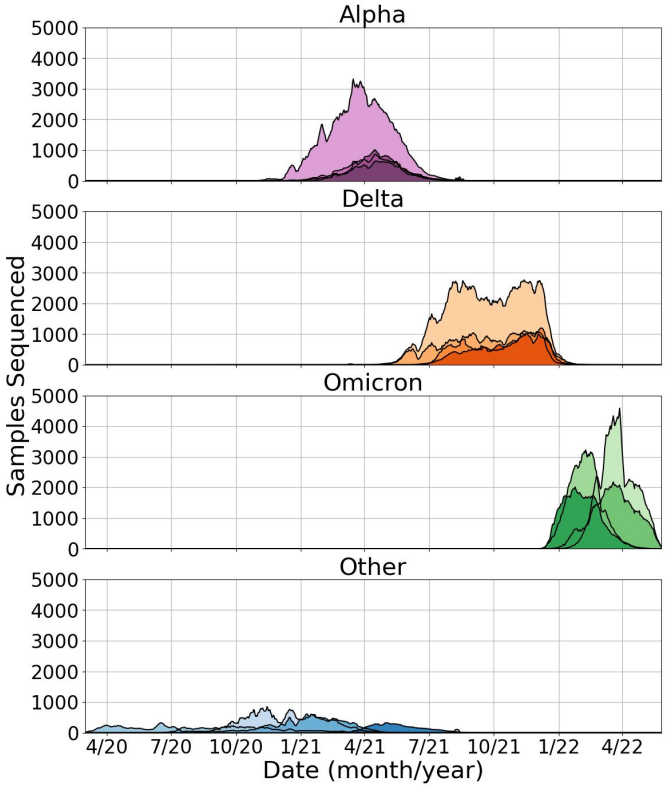
- Mapped clusters to WHO variant
- Four largest clusters per WHO variant
- Longitudinal view
  - By continent





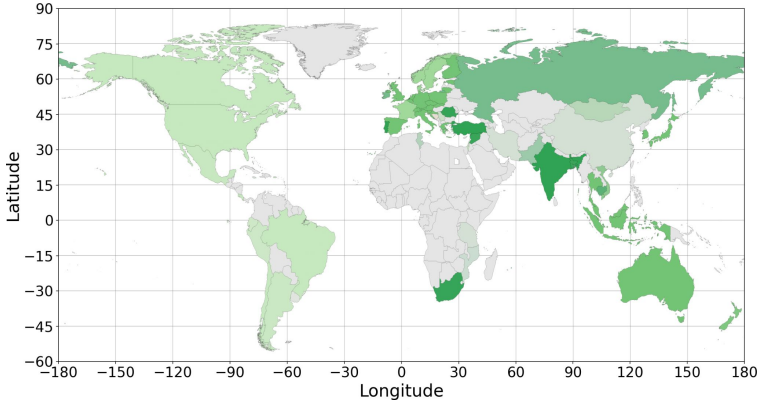
Our SARS-CoV-2 strains and WHO variants

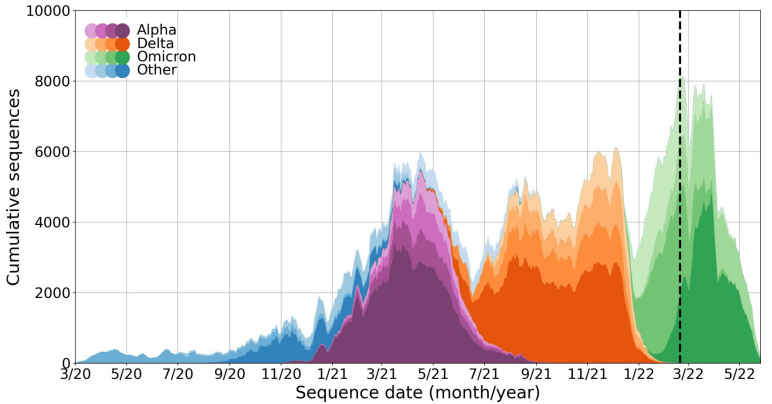
- Mapped clusters to WHO variant
- Four largest clusters per WHO variant
- Longitudinal view
  - By continent
  - By WHO variant



Our SARS-CoV-2 strains and WHO variants

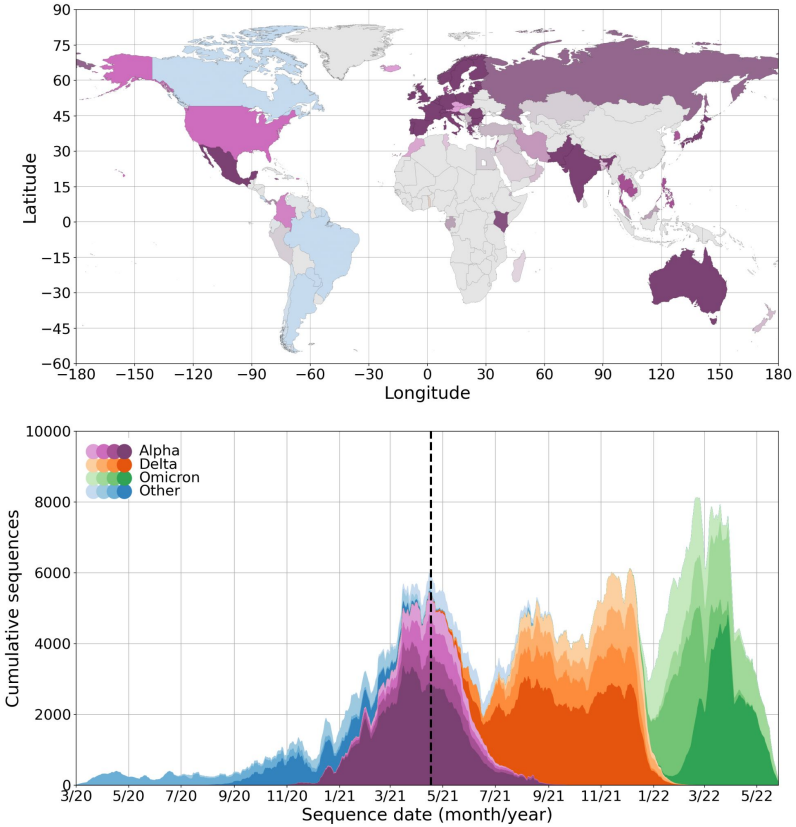
- Mapped clusters to WHO variant
- Four largest clusters per WHO variant
- Longitudinal view
  - By continent
  - By WHO variant
- Geospatial view





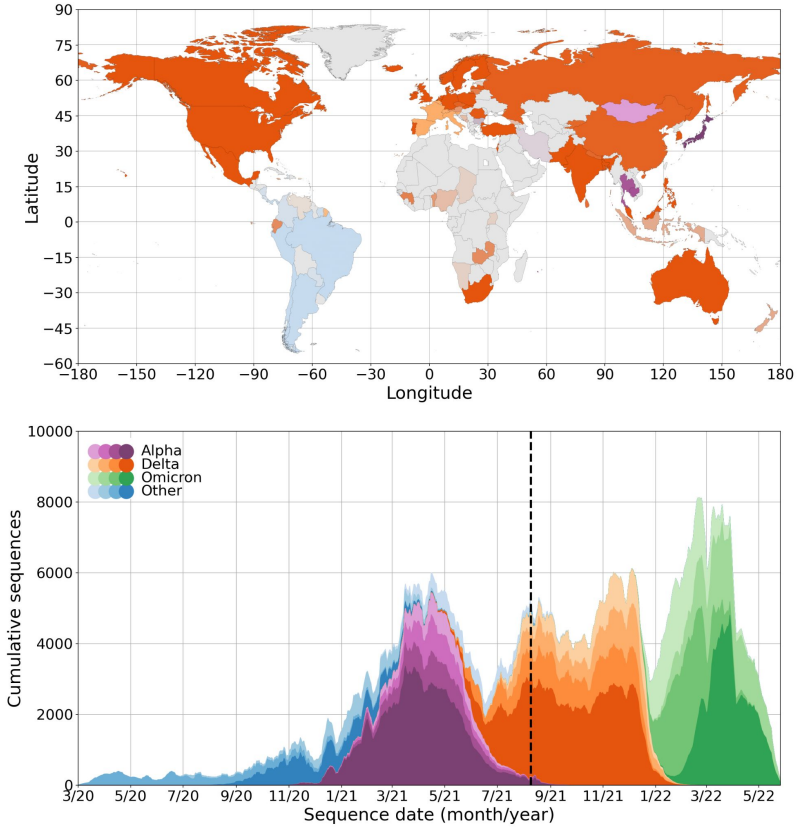
Geospatial asynchronicity - Alpha

- Geospatial - longitudinal view
- 10 largest clusters (dominant strains with most samples)
- Comprised x % of all sequences
- Shared country borders, different strains



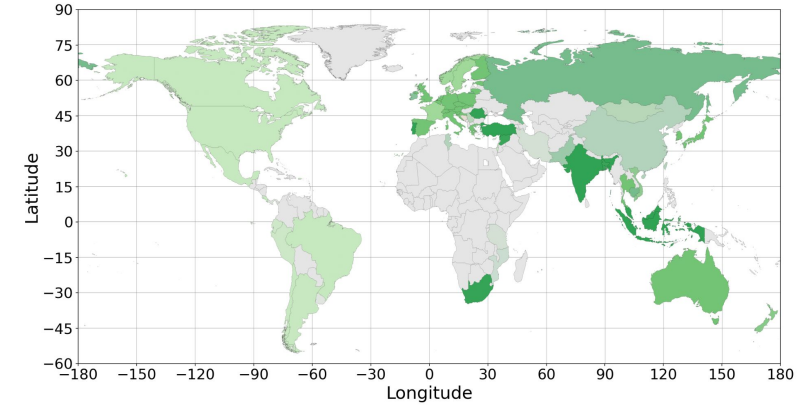
Geospatial asynchronicity - Delta

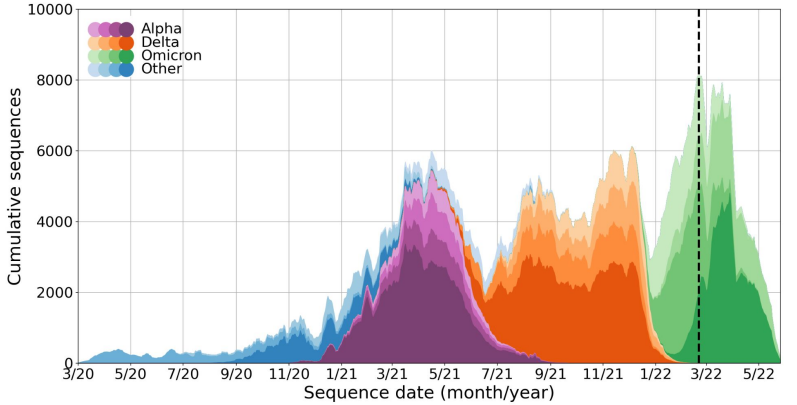
- Geospatial - longitudinal view
- 10 largest clusters (dominant strains with most samples)
- Comprised x % of all sequences
- Shared country borders, different strains



Geospatial asynchronicity - Omicron

- Geospatial - longitudinal view
- 10 largest clusters (dominant strains with most samples)
- Comprised x % of all sequences
- Shared country borders, different strains



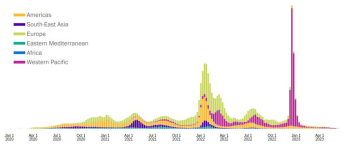


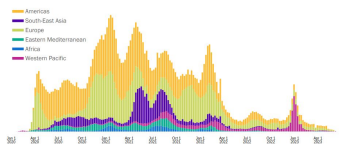
Geospatial asynchronicity - early pandemic

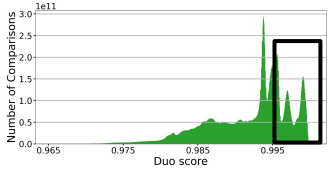
- Early pandemic view on Frontier
- 10 largest clusters (dominant strains with most samples)
- Shared country borders, different strains

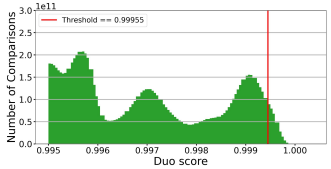
Summary

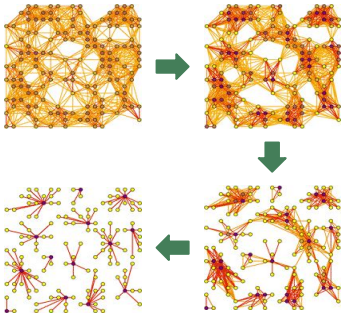
- All-against-all, whole-genome vector comparison
- > 22,000 SARS-CoV-2 strains
- Geospatial and temporal trajectories for each strain
- Spatially asynchronicity of dominant strains
- Possibly different pattern of epidemiological risk

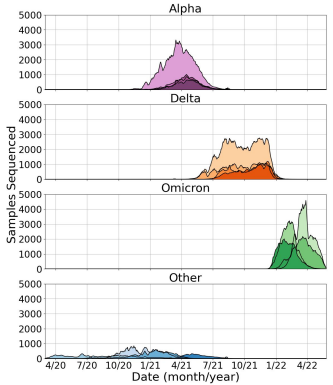






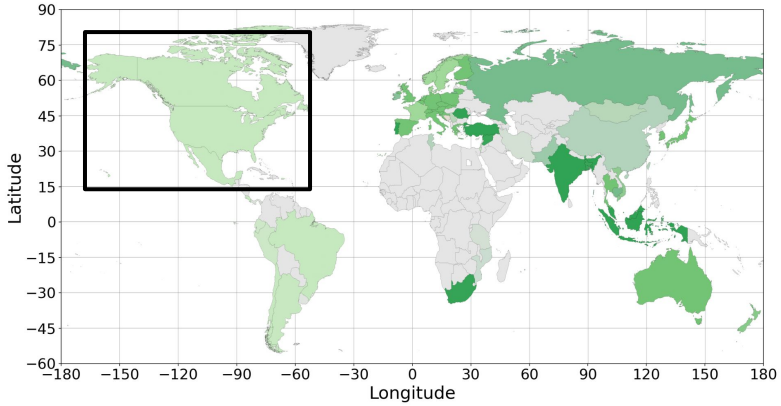


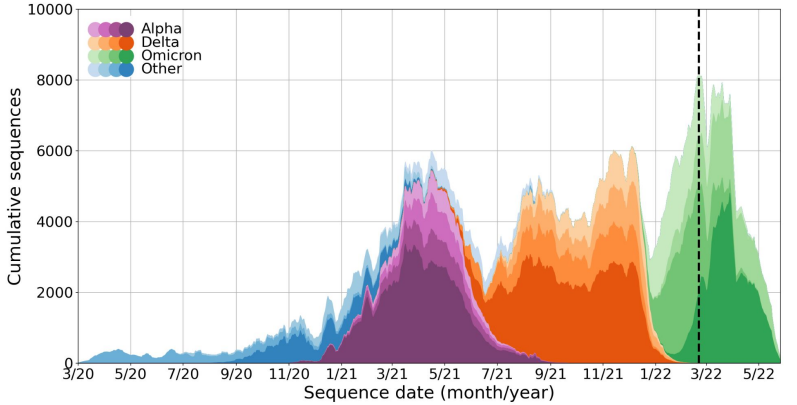




Future work

- County or city geo-resolution
- Triplets of vectors (3-way)
- Assimilation of new samples
- Recombination analyses
  - Phylogenetic trees exclude recombination
- Relationship to other variables
  - Environmental
  - Demographic
  - Mortality rate





Acknowledgements

Oak Ridge National Lab

Dan Jacobson

John Lagergren

Verónica G. Melesse Vergara

Wayne Joubert

Johns Hopkins University

Christopher Bradburne

Emily Gurley

Berkeley National Lab

Mikaela Cashman

Cornell University

Raina Plowright

OAK RIDGE National Laboratory | LEADERSHIP COMPUTING FACILITY

National Science Foundation

U.S. DEPARTMENT OF ENERGY

Funding

DE-AC02-05CH11231
DE-AC02-06CH11357
DE-AC05–00OR22725
DE-AC02-98CH10886
EF-2133763
DOE INCITE

Questions?

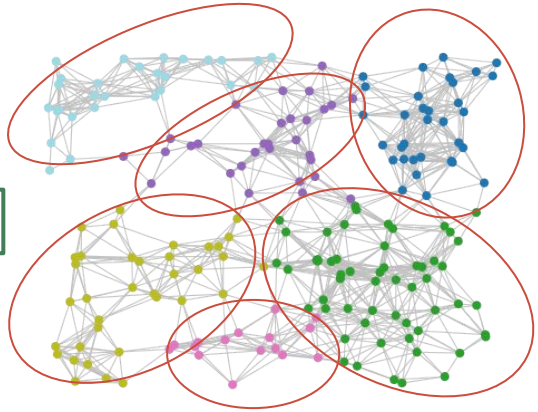
merletjj@ornl.gov

jacobsonda@ornl.gov

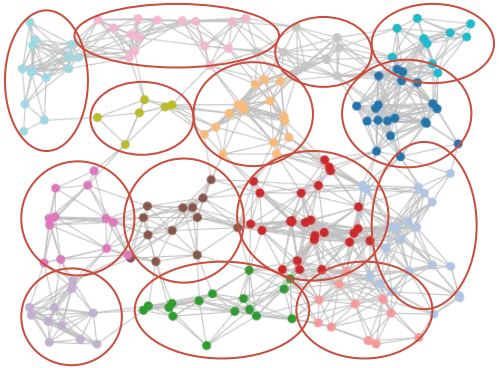
Markov clustering with HipMCL

- Unsupervised algorithm
- Random walk
  - expansion (flow spreads)
  - inflation (flow recedes)
- Network -> clusters
- 22,000 SARS-CoV-2 strains
- Cluster number affected by
  - graph topology
  - inflation value

Low inflation



High inflation



Images: github.com/GuyAllard/markov_clustering