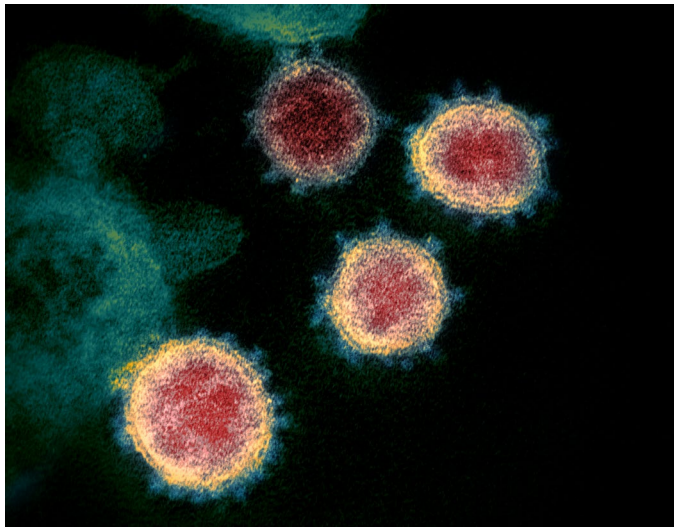
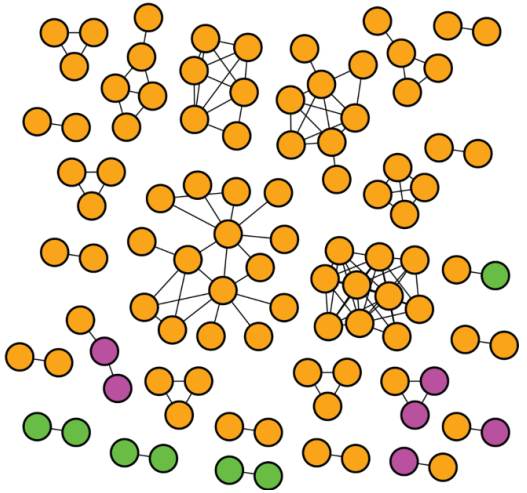
Whole Genome Sequence Analysis of SARS-CoV-2

- Insights and Challenges

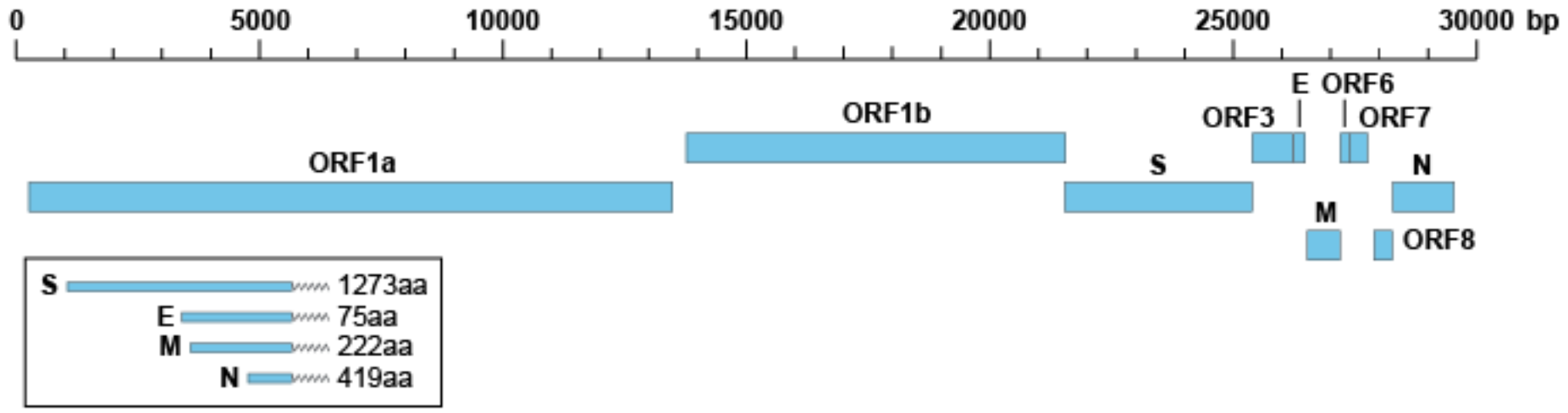




Dr. Benjamin Sobkowiak
Department of Mathematics, SFU

The SARS-CoV-2 genome

- Positive single-stranded RNA genome
- Genome shares ~82% sequence identity with SARS-CoV and MERS-CoV
- ~30,000bp sequence – Wuhan-Hu-1 reference strain 29,903bp



https://www.novusbio.com/support/sars-cov-research-resources

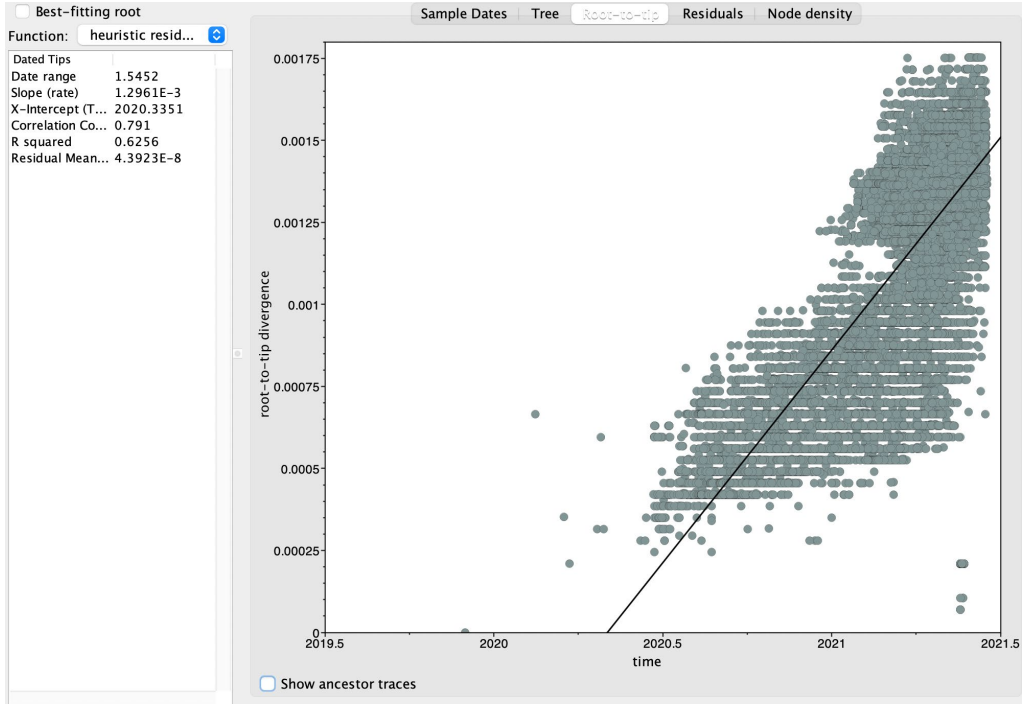
SARS-CoV-2 evolution

- Fast mutation rate (~$8x10^{-4}$ substitutions/site/year Boni *et. al.* 2020) – though relatively slow for an RNA virus
  - HIV (~$5x10^{-3}$ substitutions/site/year)
  - Influenza (~$4x10^{-3}$ substitutions/site/year)

- No evidence for mutation rate differences between lineages

- Large number of isolates and fast mutation rate has led to emergence of a number of lineages

- Multiple convergent mutations including many in the spike protein

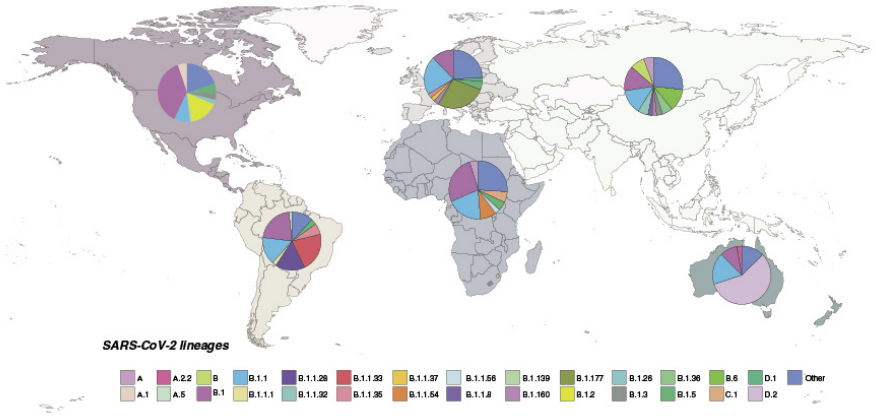


What can we learn from genomic analysis of SARS-CoV-2?

- First whole genome sequence - Wuhan-Hu-1 - published early in the pandemic (3rd February 2020)

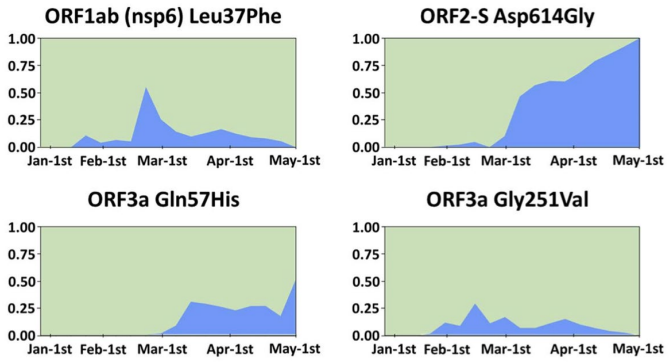
- GISAID initiative holds 2,232,941 SARS-CoV-2 whole genome sequences (as of 5th July 2021)

Lineage assignment



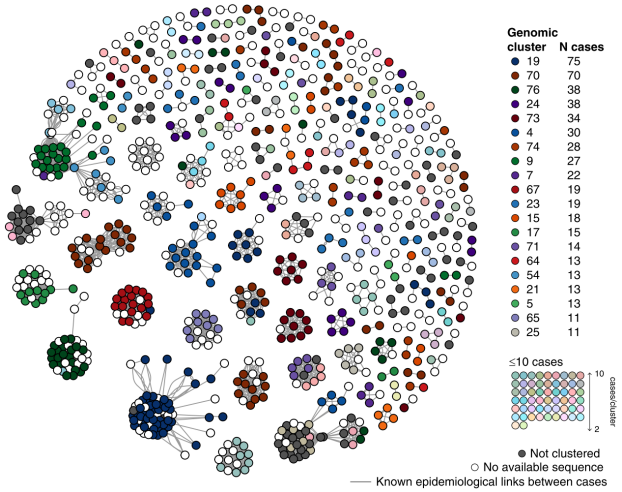
Cella *et. al.* 2021

Mutation detection



Cortey *et. al.* 2020

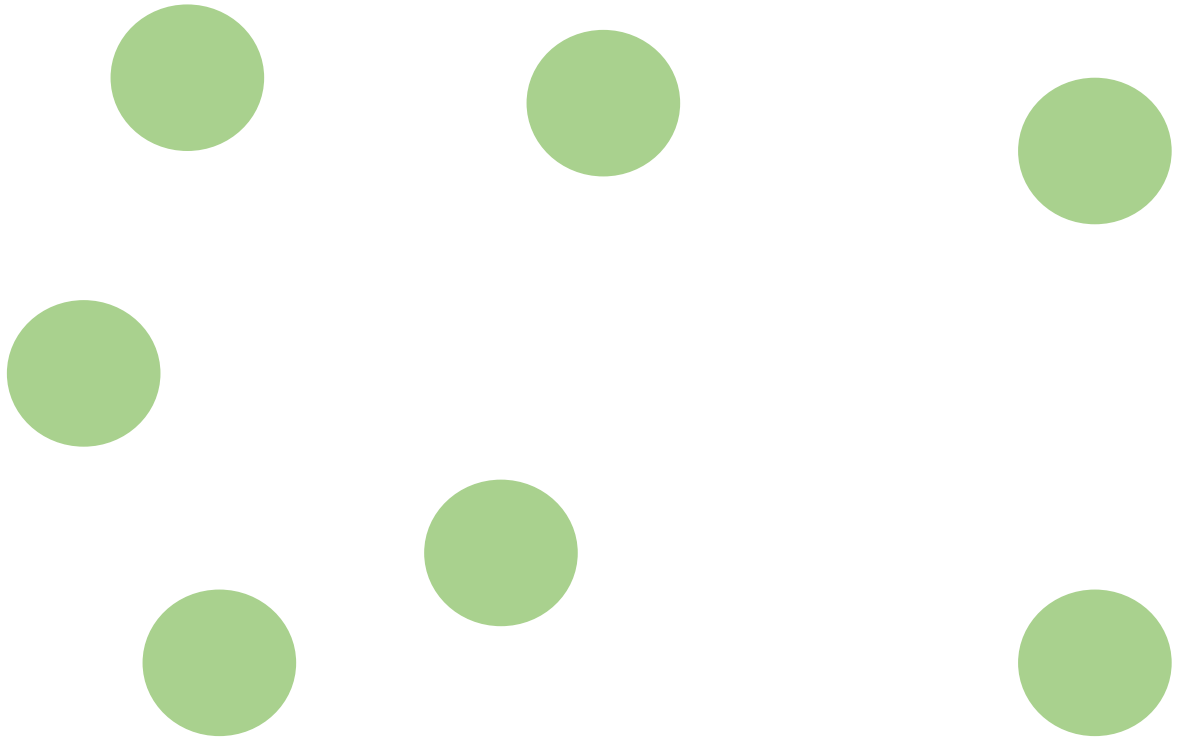
Clustering



Seemann *et. al.* 2020

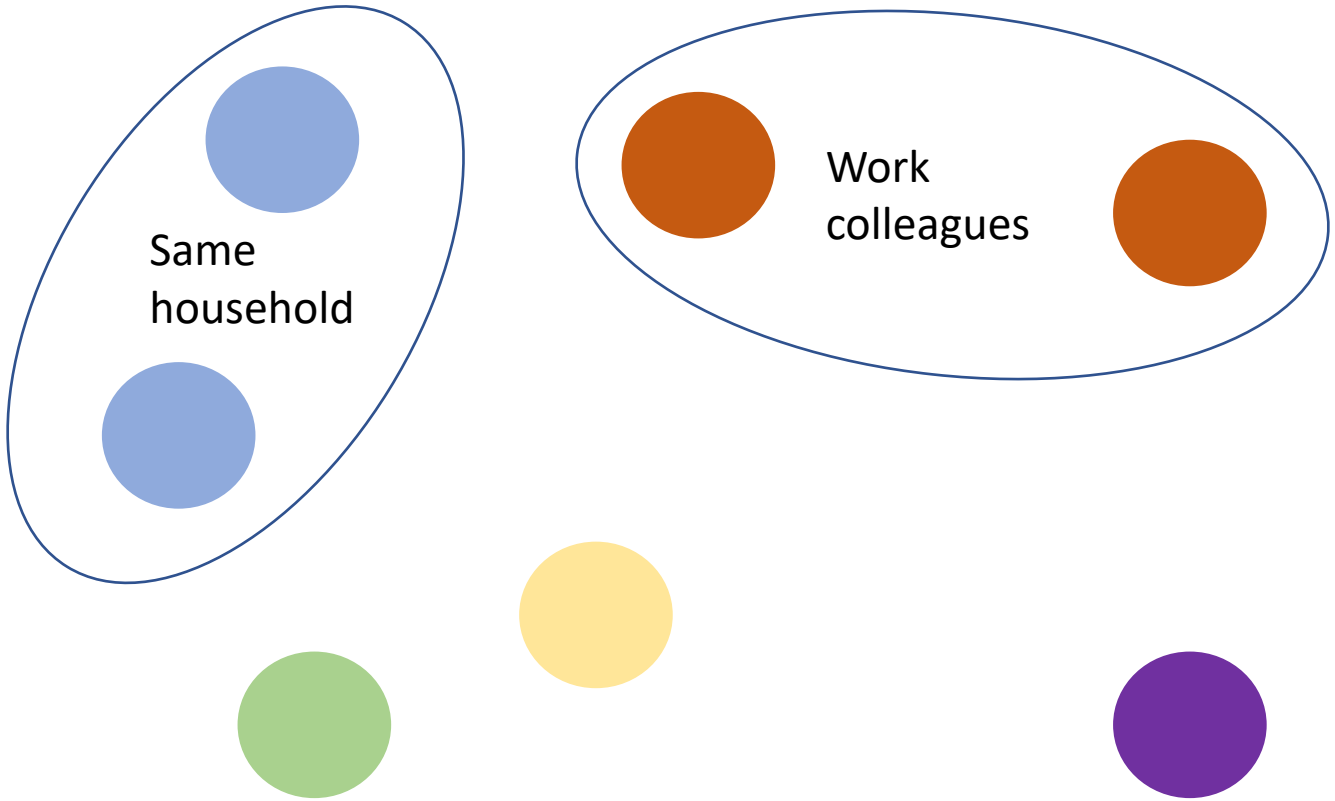
Clustering of SARS-CoV-2 cases

- Traditional epidemiological approach – contact tracing



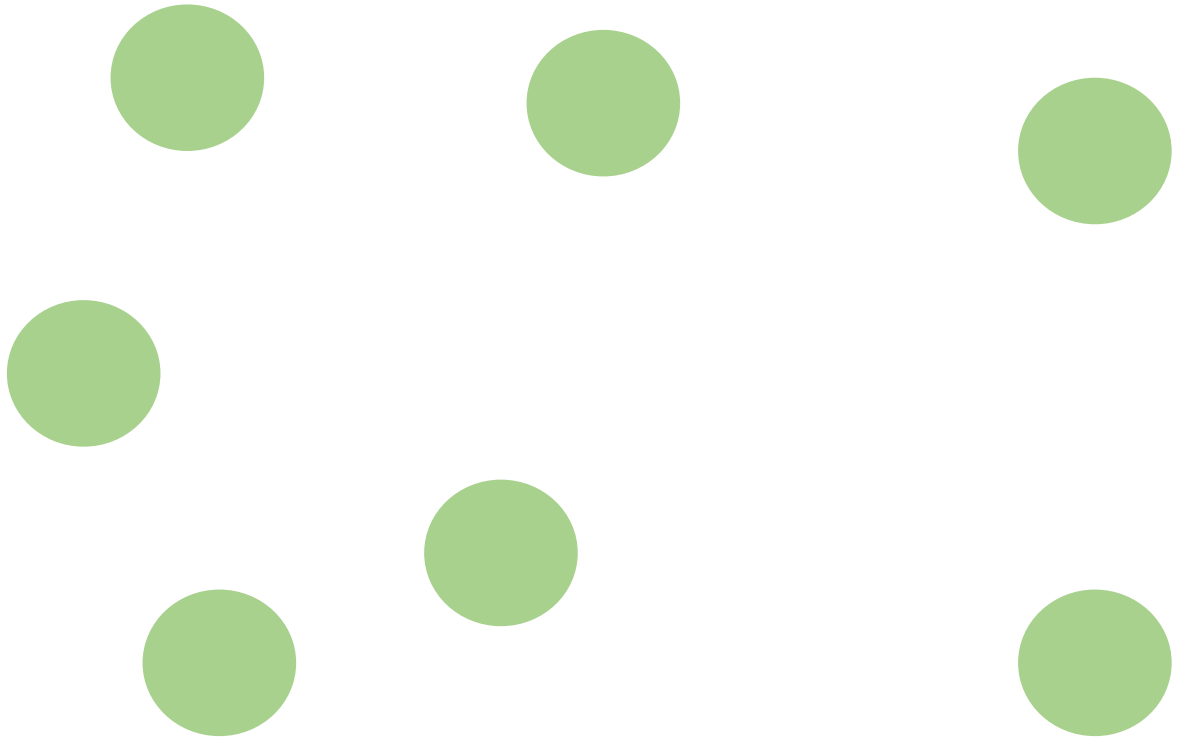
Clustering of SARS-CoV-2 cases

- Conduct questionnaire – 48-hour contacts, household members etc.



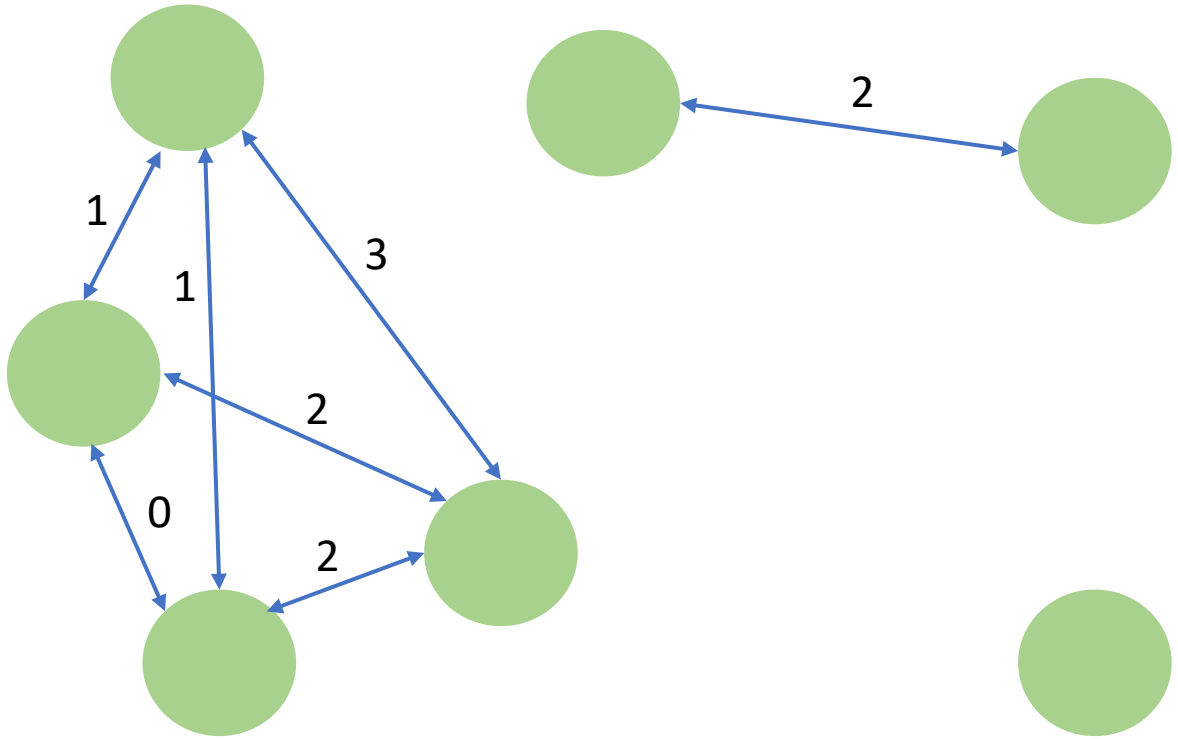
Clustering of SARS-CoV-2 cases

- Simple genomic approach – SNP threshold



Clustering of SARS-CoV-2 cases

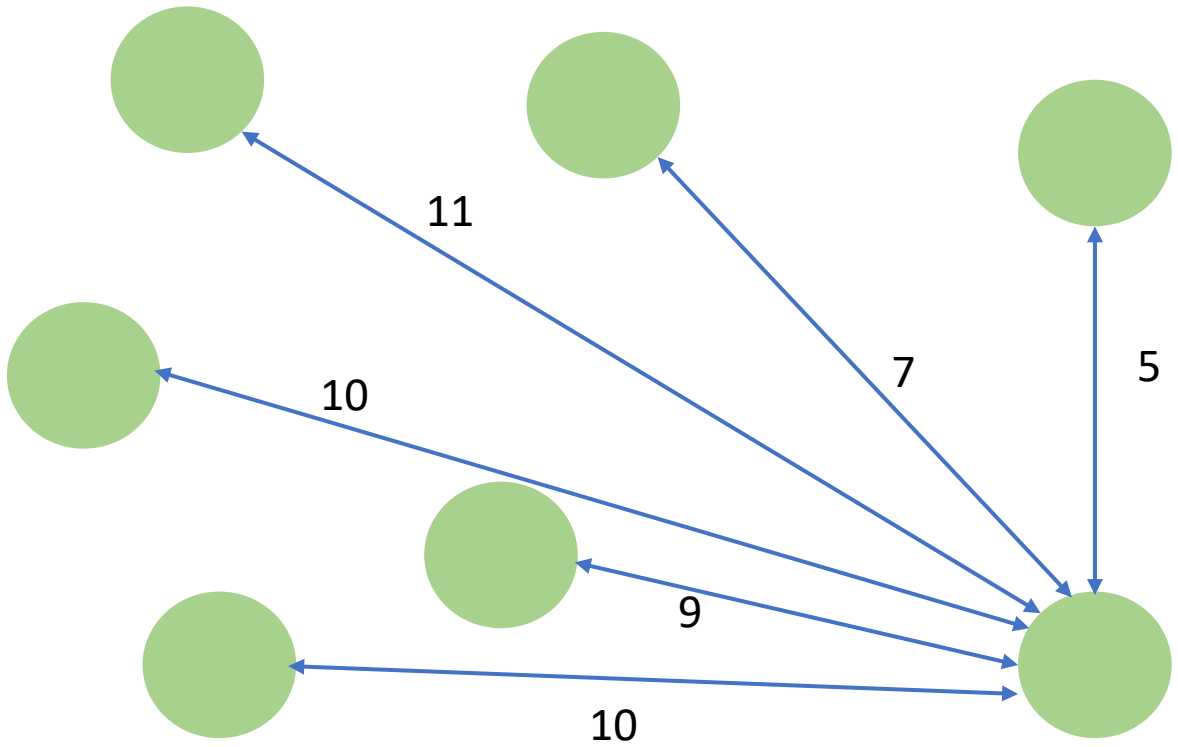
- Simple genomic approach – SNP threshold



Threshold
2 SNPs

Clustering of SARS-CoV-2 cases

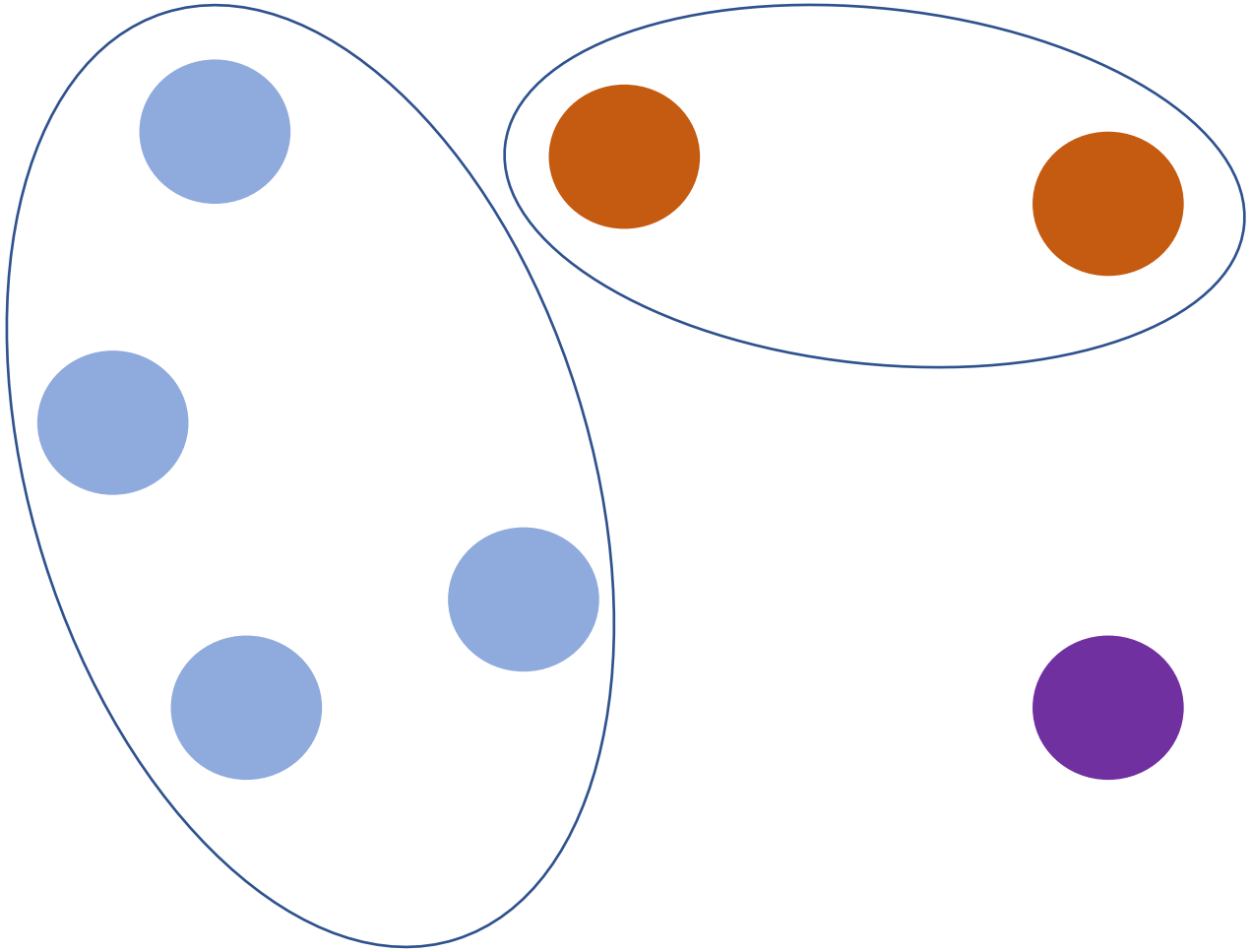
- Simple genomic approach – SNP threshold



Threshold
2 SNPs

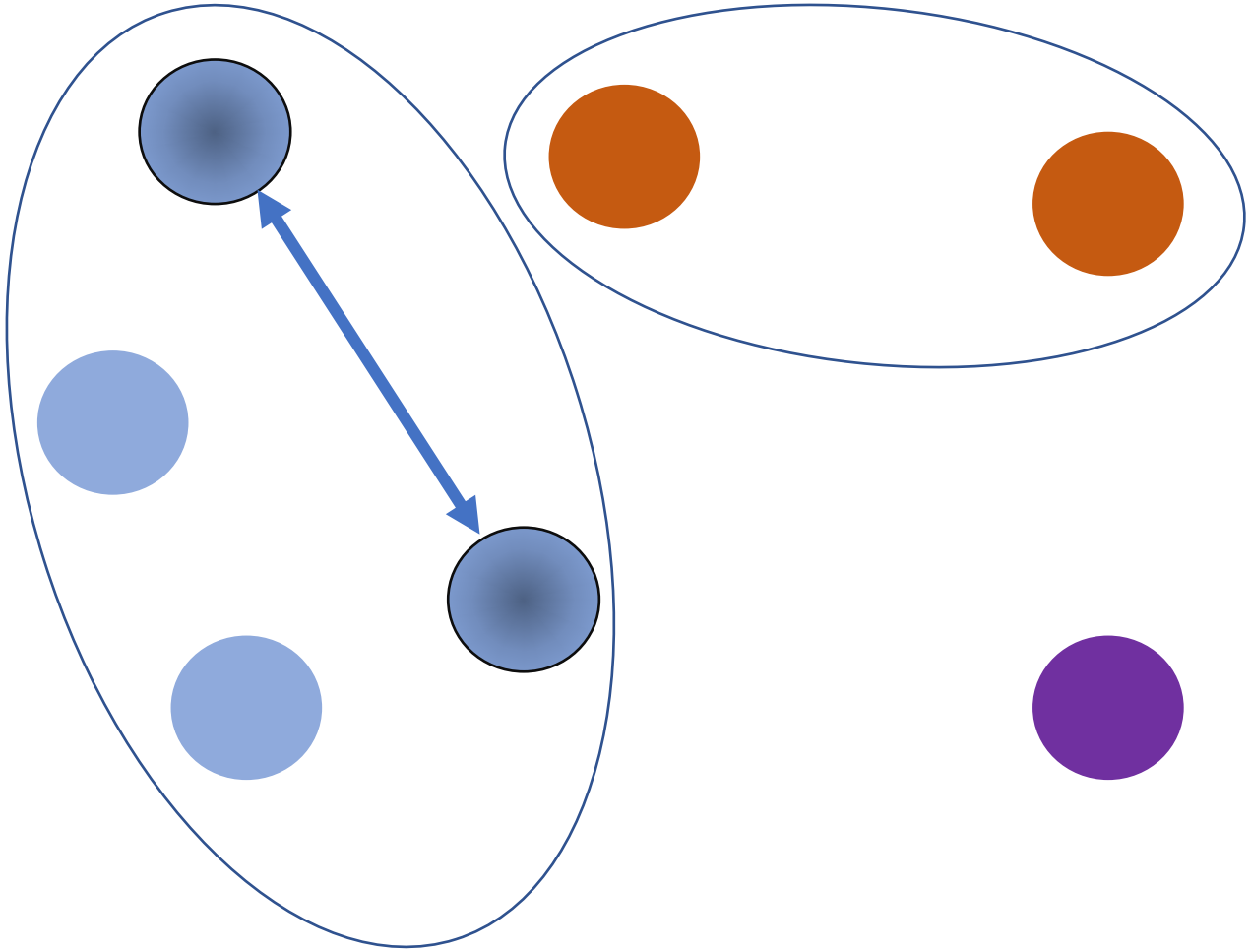
Clustering of SARS-CoV-2 cases

- Link cases under SNP threshold



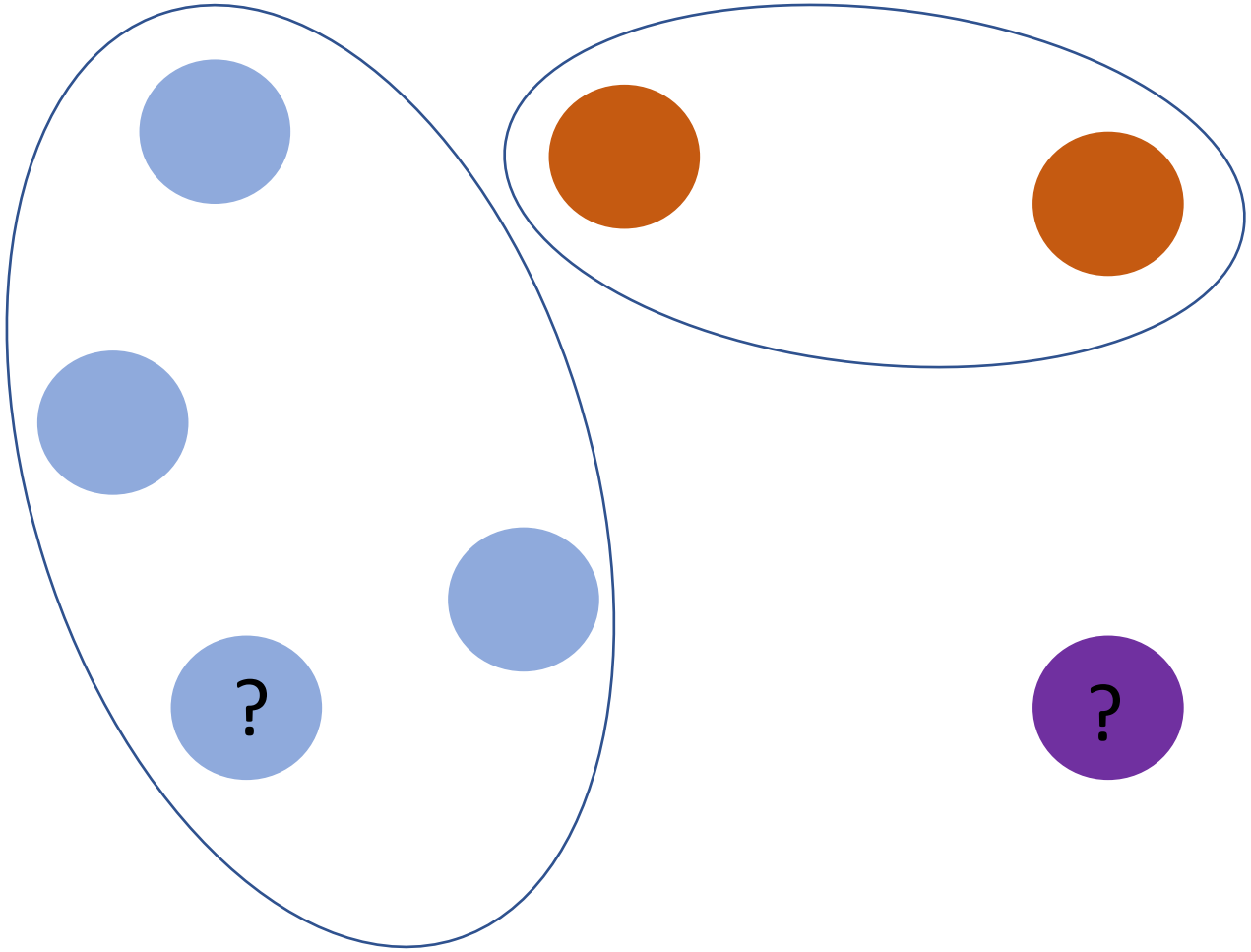
Clustering of SARS-CoV-2 cases

- May find links between cases – same restaurant etc.



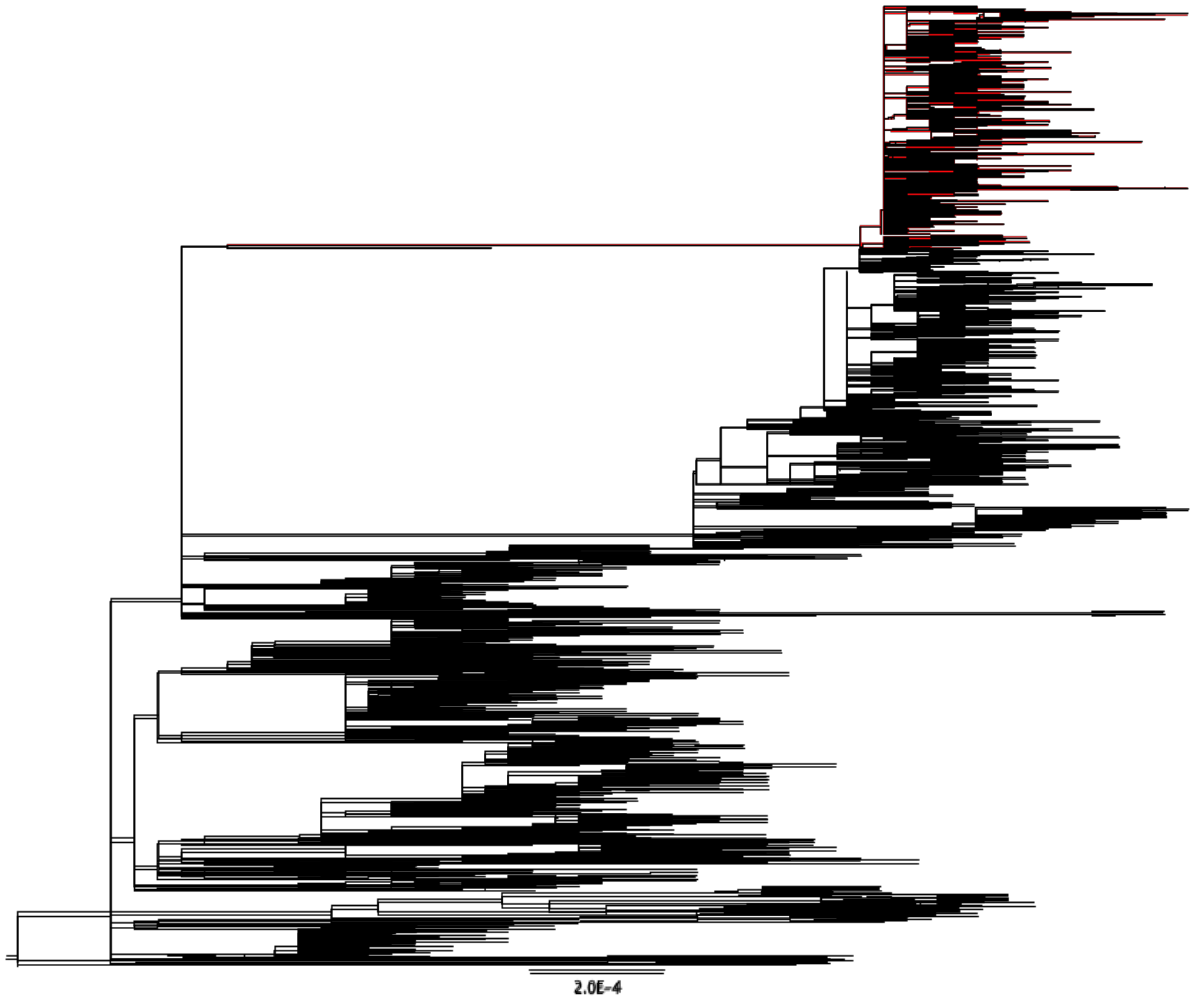
Clustering of SARS-CoV-2 cases

- Still may require some epi investigation but can reduce or direct efforts



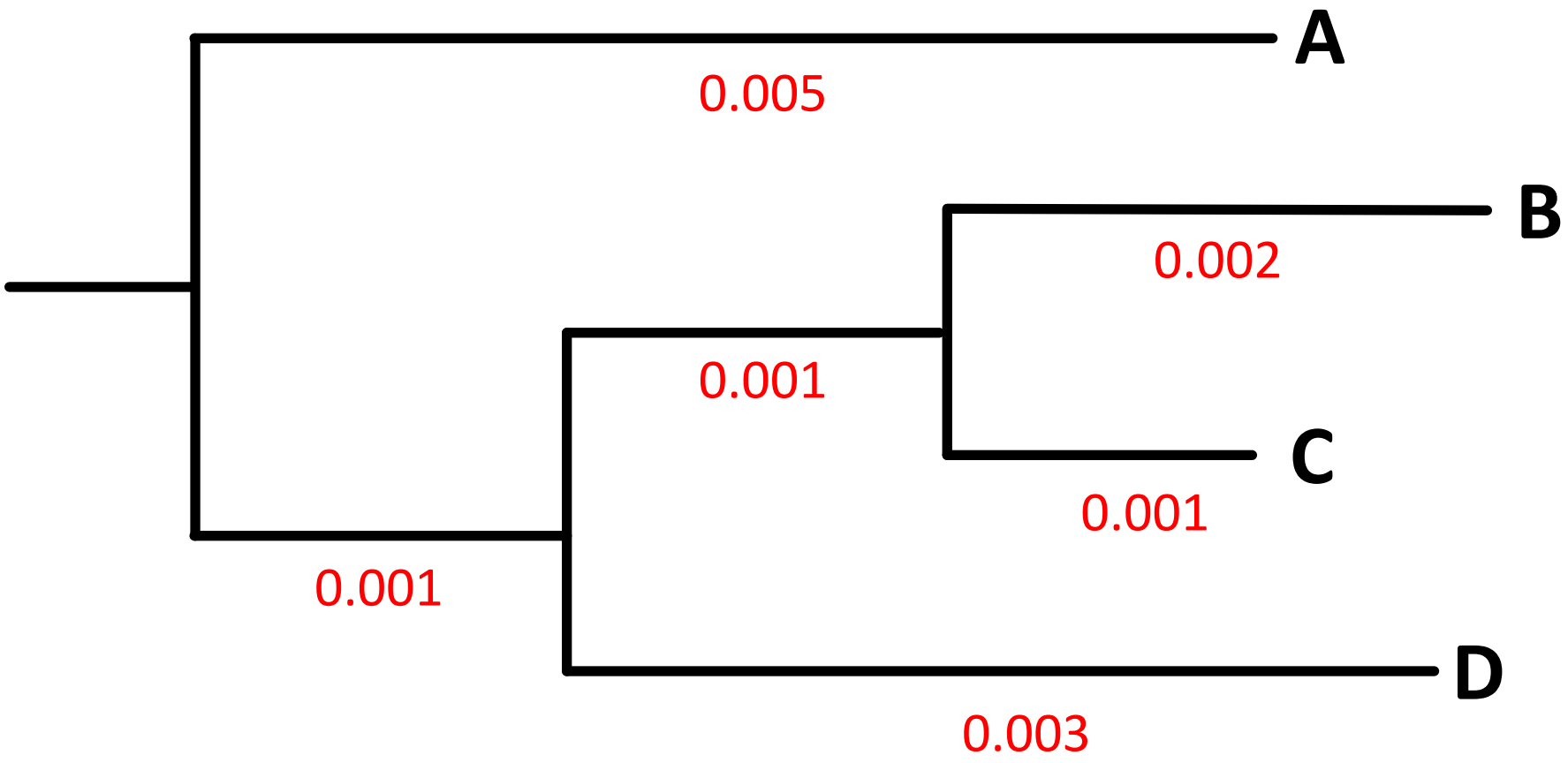
Clustering of SARS-CoV-2 cases using genomics (phylogenetics)

- Maximum likelihood (ML) tree
- Branch length represents genetic diversity
- Uses models of nucleotide substitution to infer evolutionary relationships - not all mutations are the same (transitions vs transversions etc.)



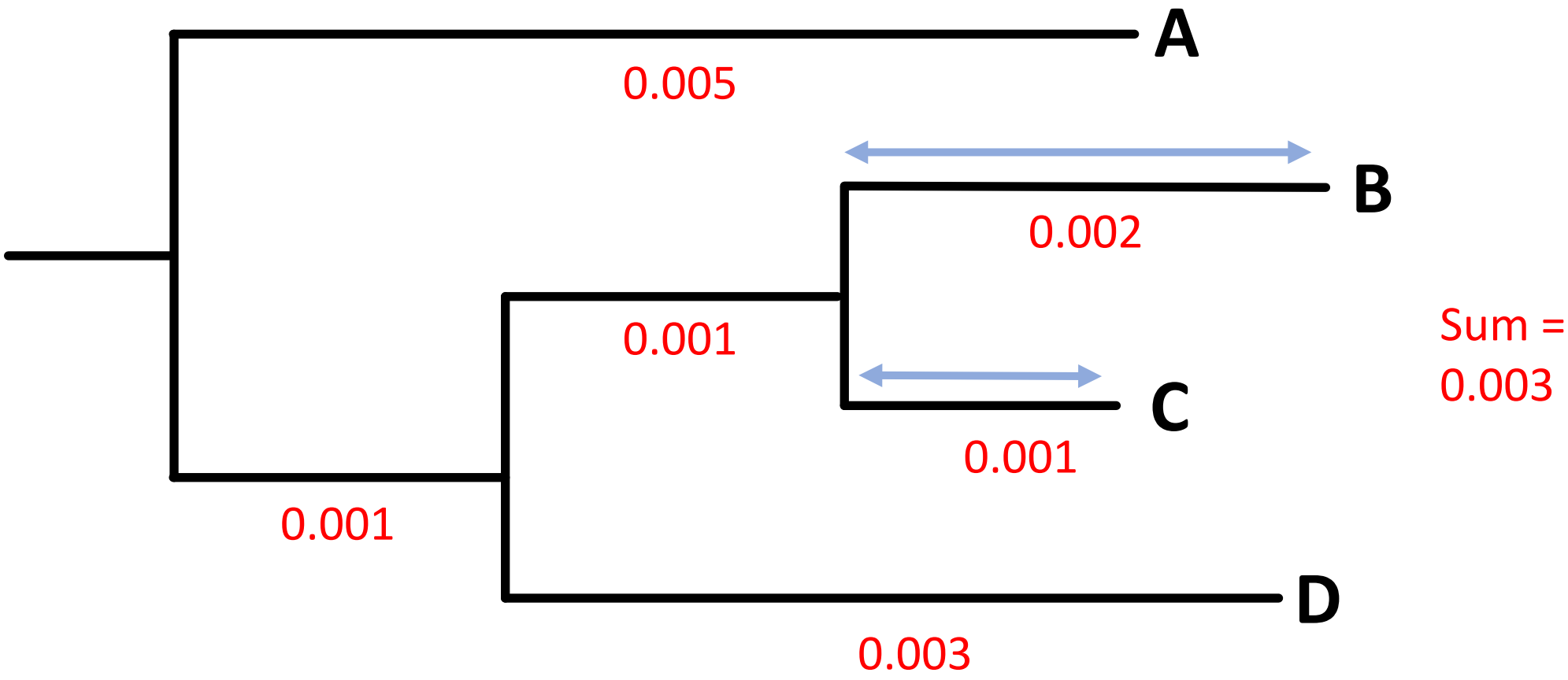
Phylogenetic clustering

- TreeCluster (Balaban et. al. 2020)



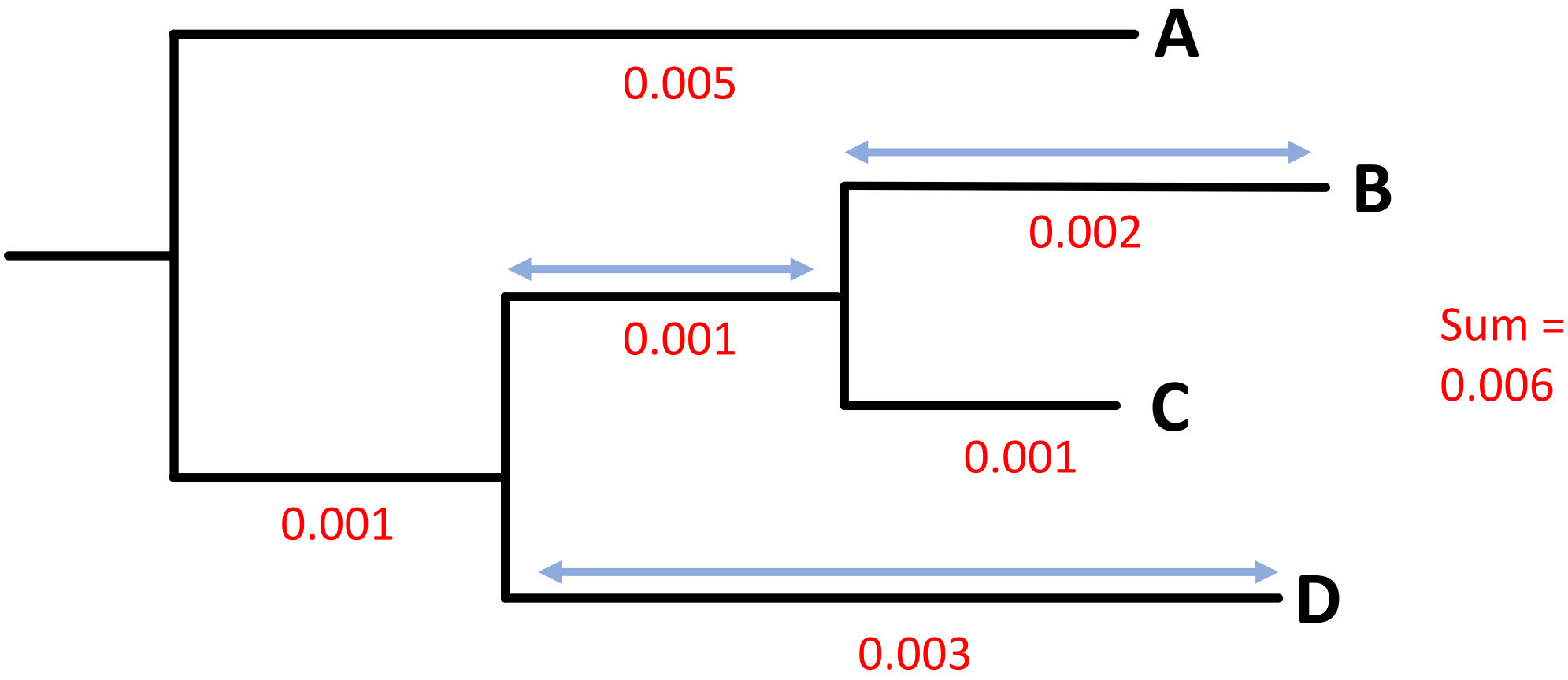
Phylogenetic clustering

- TreeCluster (Balaban et. al. 2020) – patristic distance (sum of branch lengths)



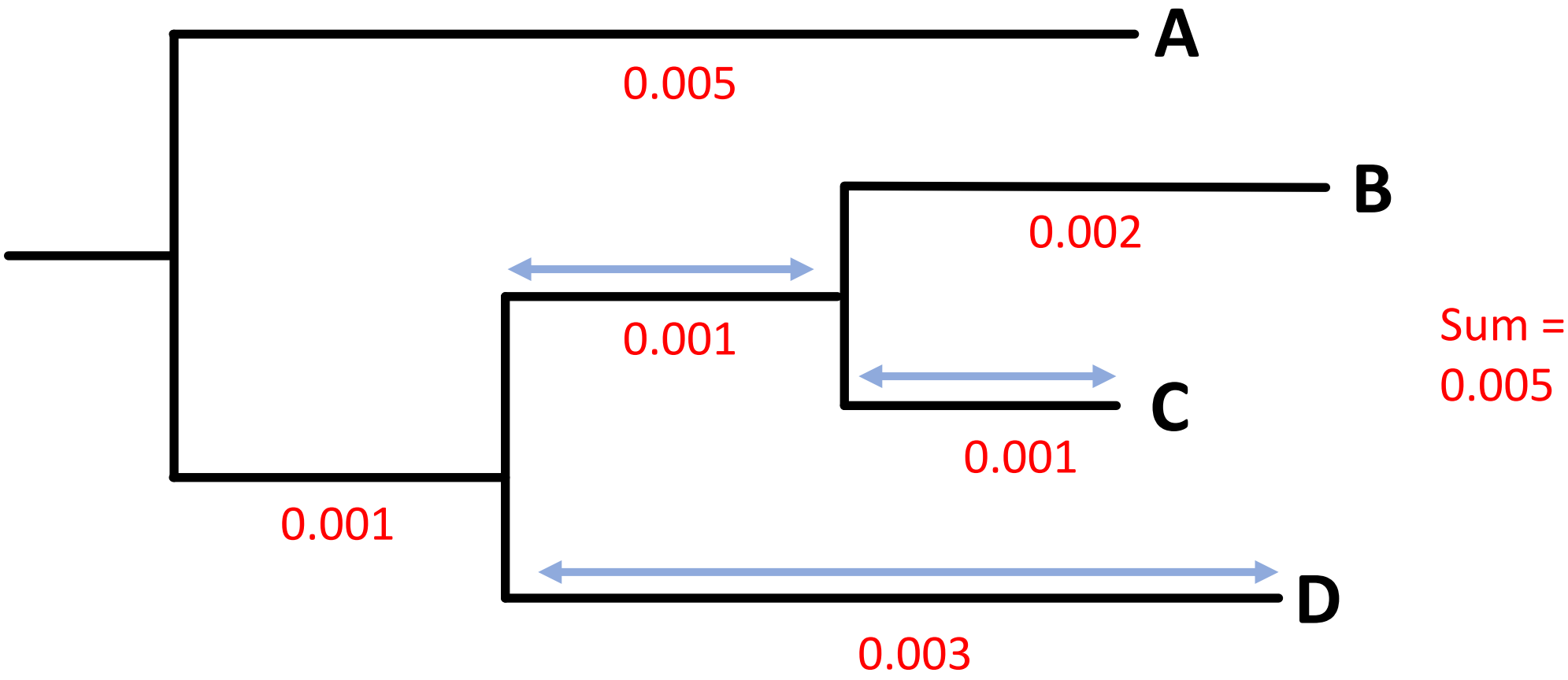
Phylogenetic clustering

- TreeCluster (Balaban et. al. 2020)



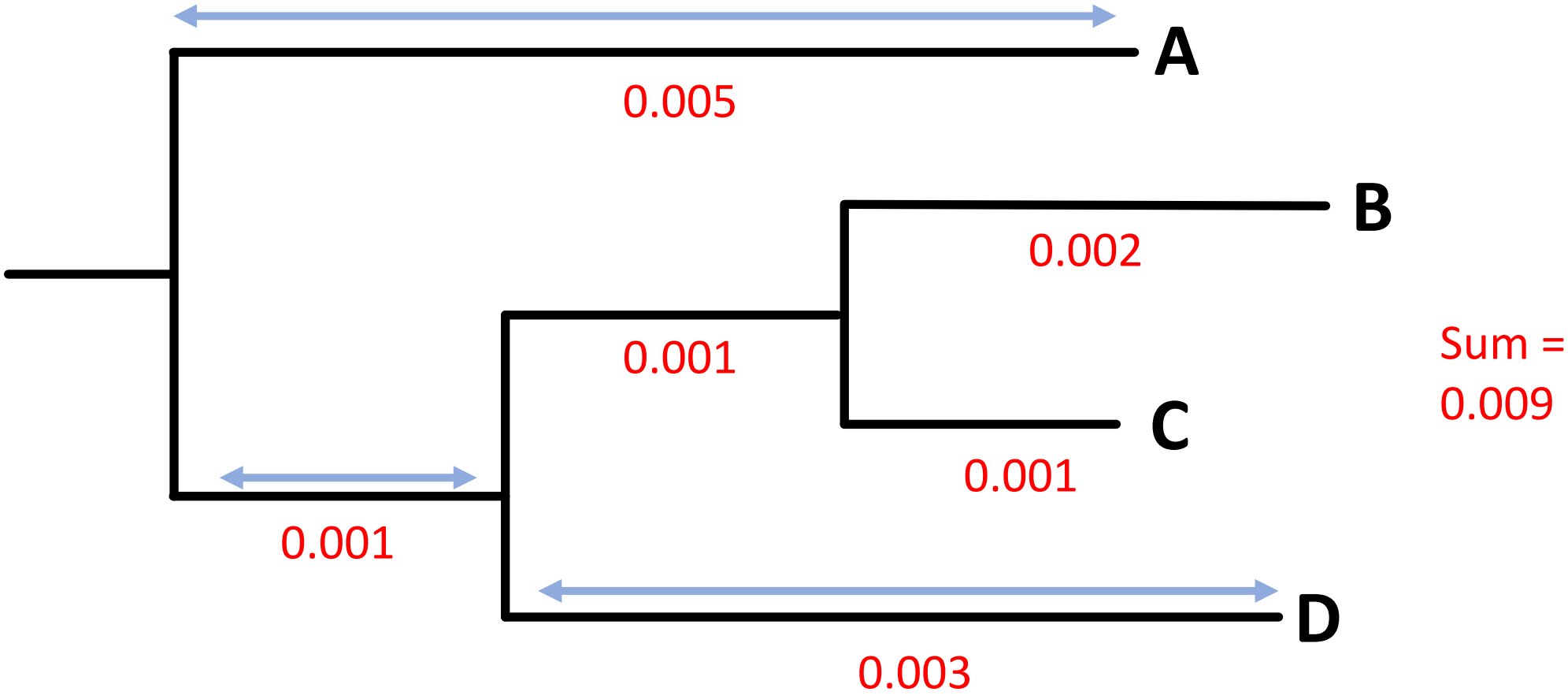
Phylogenetic clustering

- TreeCluster (Balaban et. al. 2020)



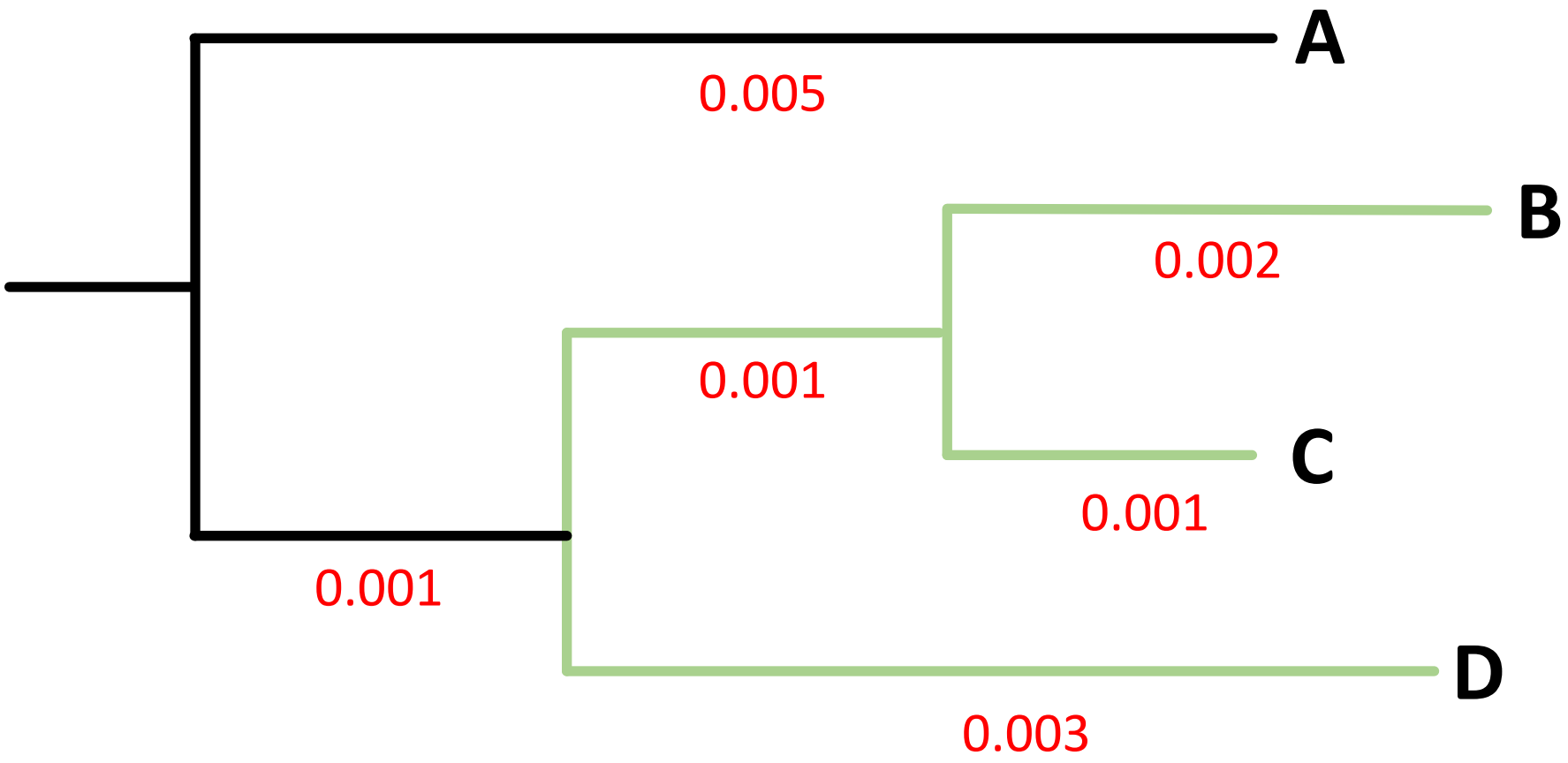
Phylogenetic clustering

- TreeCluster (Balaban et. al. 2020)



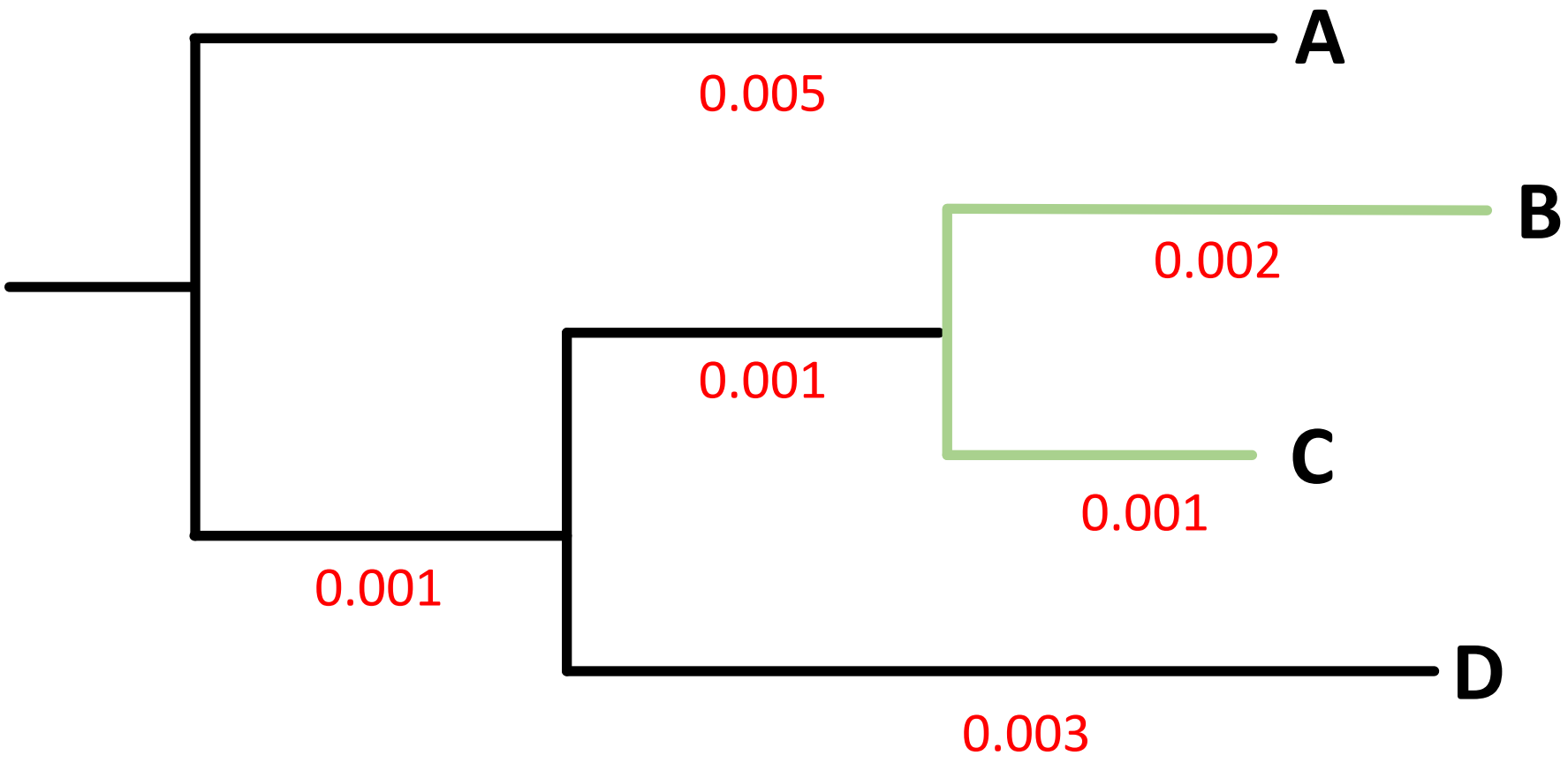
Phylogenetic clustering

- TreeCluster (Balaban et. al. 2020) – pairwise threshold = 0.005



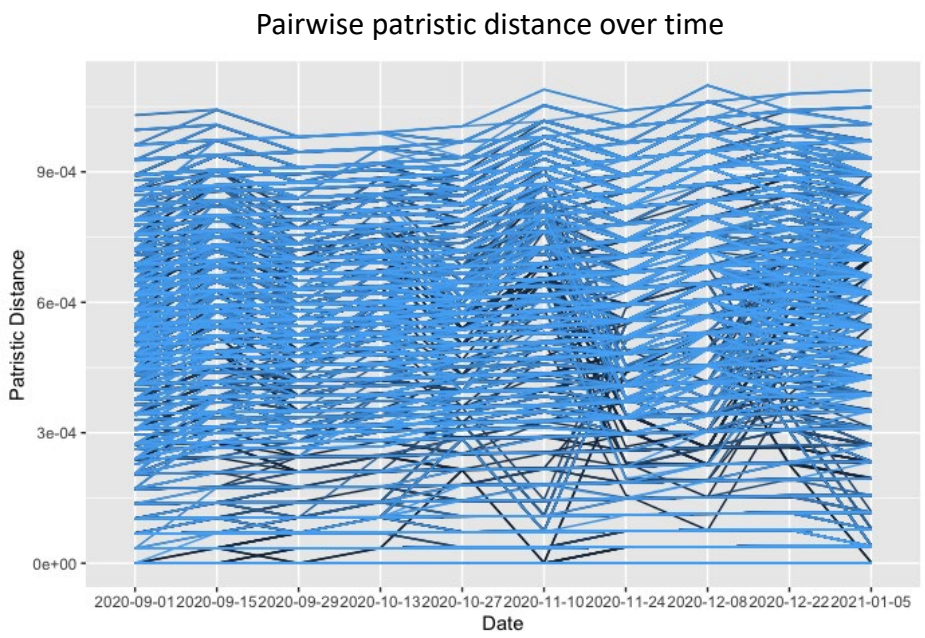
Phylogenetic clustering

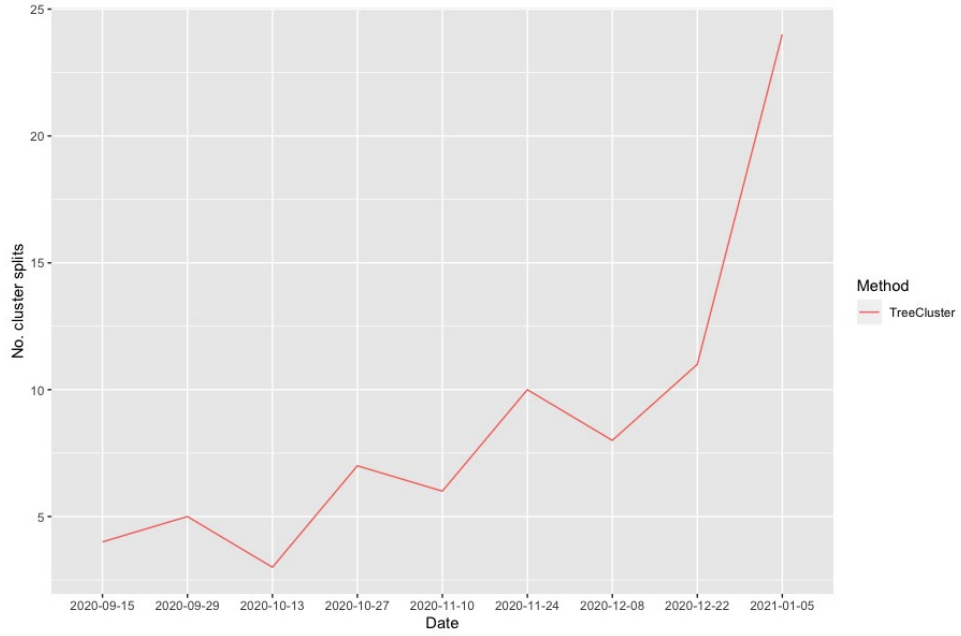
- TreeCluster (Balaban et. al. 2020) – max clade threshold = 0.005



Issues with phylogenetic clustering

- Developed for HIV – higher diversity through time – more stable phylogeny
- SARS-CoV-2 often low diversity in tree – very large clusters
- Instability of clusters with max clade threshold





Genomic clustering with logit function

- Uses pairwise patristic distance from phylogenetic tree, coupled with other variables (e.g. dates) to calculate probability of cases being linked - P ∈ (0,1)

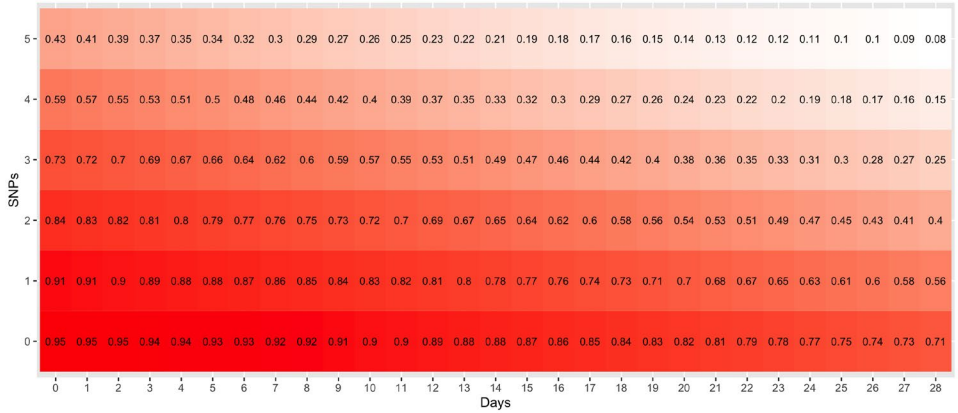
$$\ell = \log \frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_n x_n \quad (1)$$

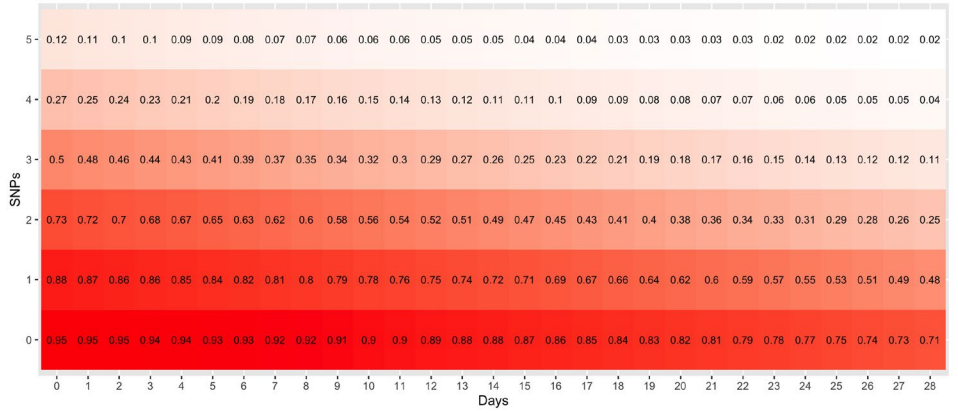
ℓ = log-odds
p = probability of Y = 1 (sequences are linked)
β_i = beta coefficient parameters
x_1 = distance predictor variable (pairwise SNP or genetic distance)
x_2 = time predictor variable (days between collection dates)

- Set max pairwise threshold to link cases and form clusters

Genomic clustering with logit function

- Can set betas accordingly and adjust for different lineages (e.g. P1 very low diversity, stricter beta on genomic diversity)



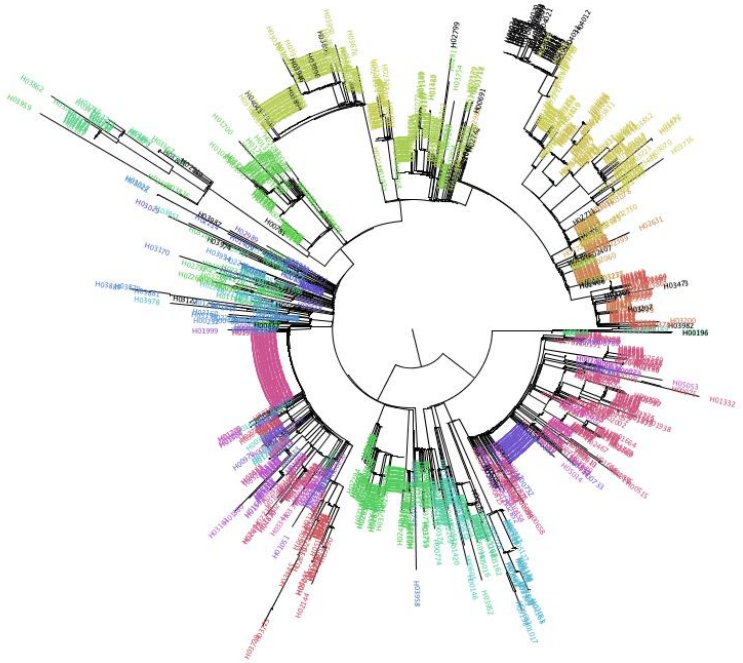


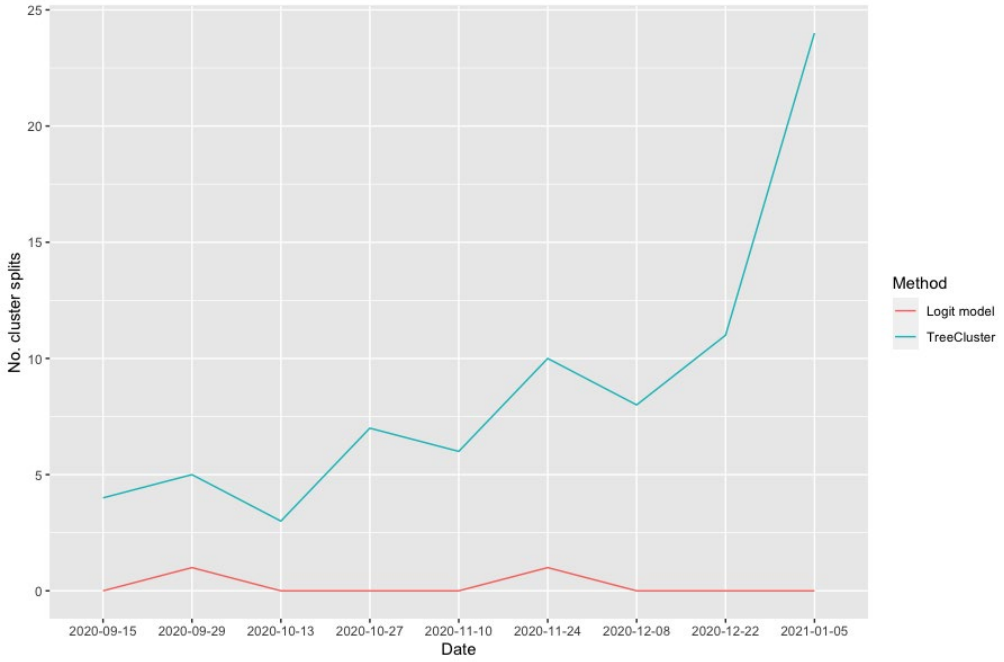
- Add any important explanatory variables that may help define clusters

Results – logit model

- Stable SARS-CoV-2 clusters that can inform epidemiological investigation

- Easy to implement and interpret in public health settings





Limitations and improvements – logit model

- Long chain clustering still causes large clusters if only considering patristic distance - still need additional data to refine clusters

- Slower to run than TreeCluster - around 5 minutes for 20,000 sequences compared to a few seconds

- Requires good quality data – genomic and epidemiological

Acknowledgements

SFU
Caroline Colijn

BCCDC
Natalie Prystajecky
Linda Hoang
Kimia Kamelian
John Tyson

University of Melbourne
Anders Gonçalves da Silva
Courtenay Lane
Anne Watt

BC Centre for Disease Control