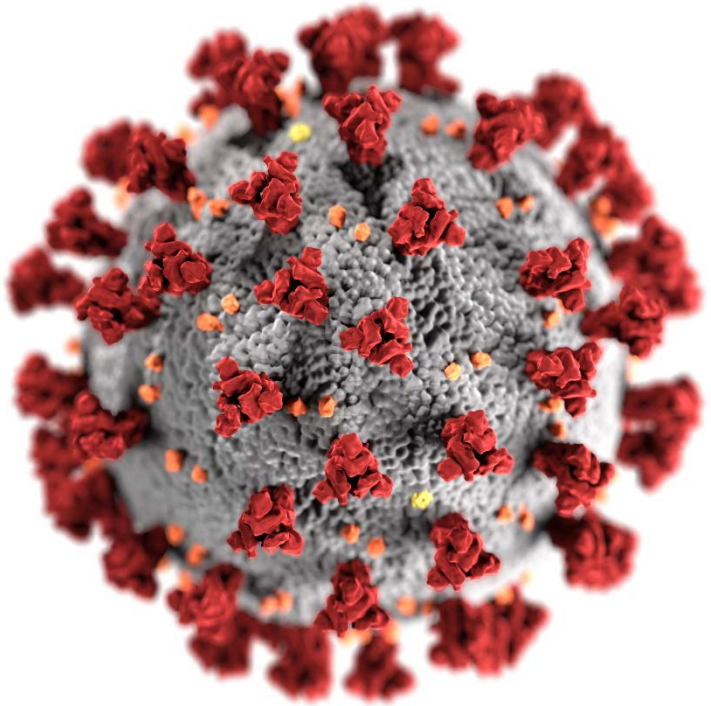
How to read a phylogenetic tree

COVID-19 Genomic Epidemiology Toolkit: Module 1.3



Michael Weigand, PhD
Bioinformatician
Centers for Disease Control and Prevention

cdc.gov/coronavirus

Toolkit map

Part 1: Introduction

1.1 What is genomic epidemiology?

1.2 The SARS-CoV-2 genome

1.3 How to read phylogenetic trees

Part 2: Case Studies

2.1 SARS-CoV-2 sequencing in Arizona

2.2 Healthcare cluster transmission

2.3 Community Transmission

Part 3: Implementation

3.1 Getting started with Nextstrain

3.2 Getting started with MicrobeTrace

3.3 Linking epidemiologic data

Sampling transmission networks for sequencing

From Module 1.1: *What is genomic epidemiology?:* Only some individuals (blue) from the transmission network are selected for sequencing.

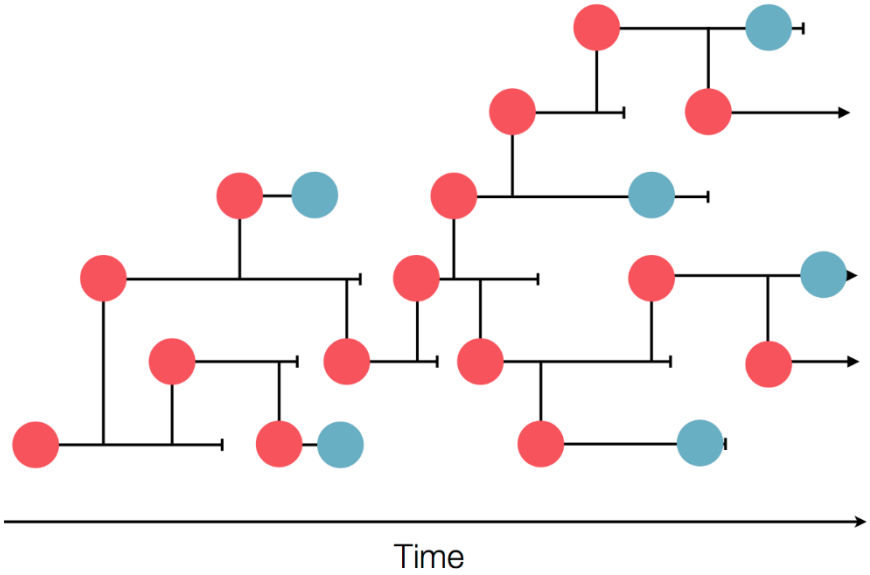
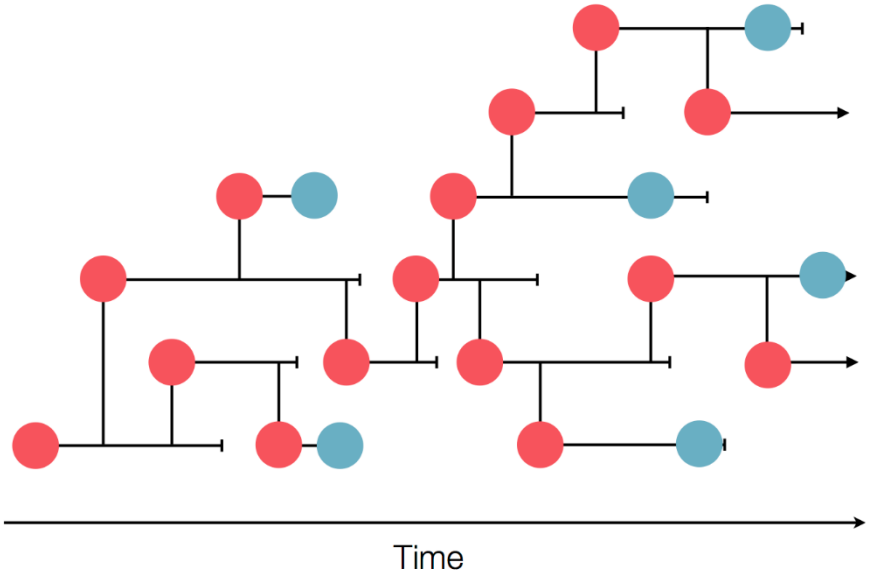
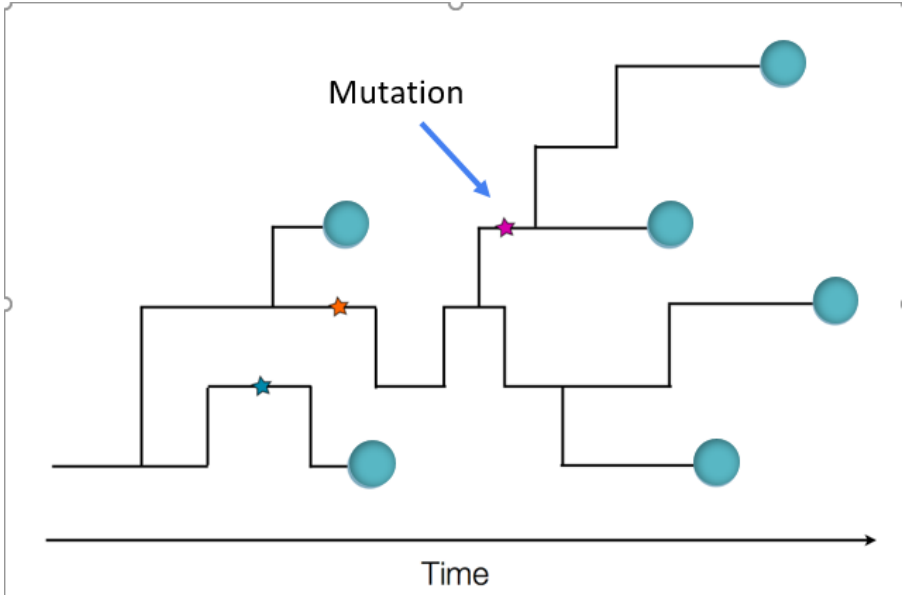


Image from Trevor Bedford Group: https://docs.nextstrain.org

Genetic fingerprinting

Viruses mutate as they spread, providing a “fingerprint” that can be used to infer ancestral relationships among sampled individuals.

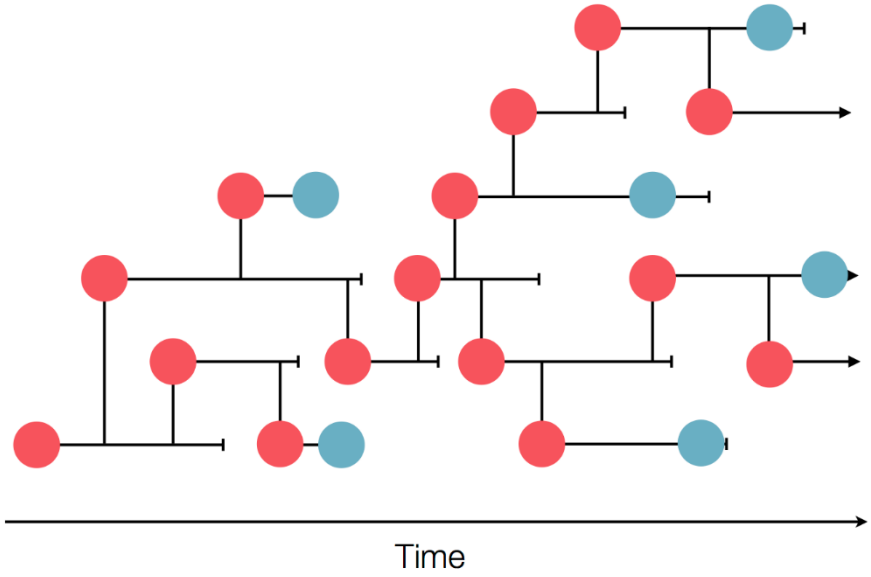


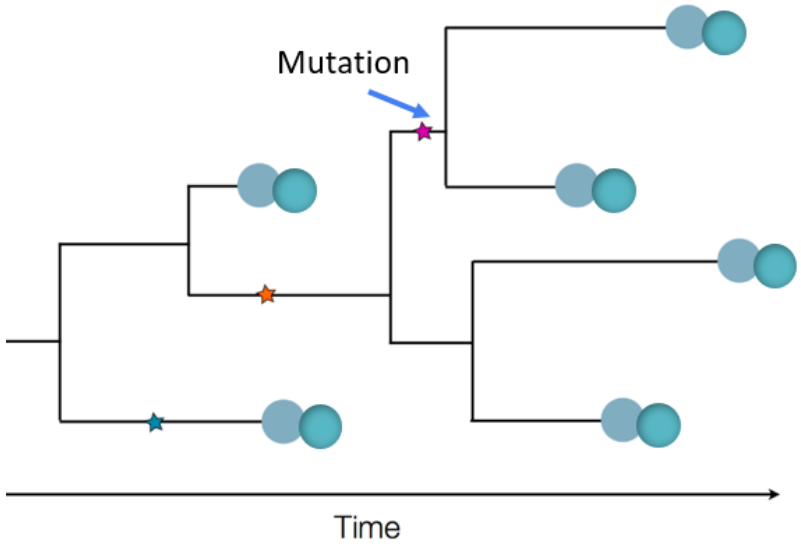


Images from Trevor Bedford Group: https://docs.nextstrain.org

Planting trees

Using phylogenetics, those relationships can be visualized as a “tree” that is always an approximation of the true network.

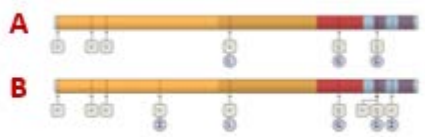




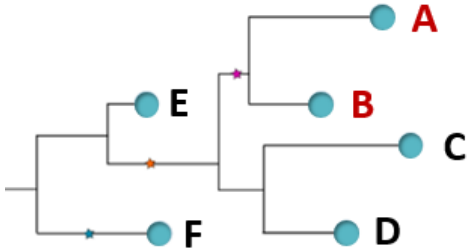
Images from Trevor Bedford Group: https://docs.nextstrain.org

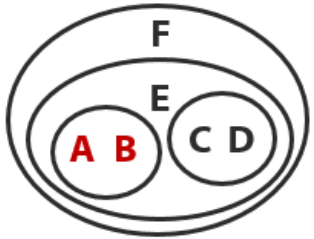
“Phylogeny approximates epidemiology” – Lee Katz

Strains that are phylogenetically closer are more likely to share an epidemiological association.



- Building trees from genetic fingerprints
- Parts of a phylogenetic tree
- Tree interpretation
- Limitations



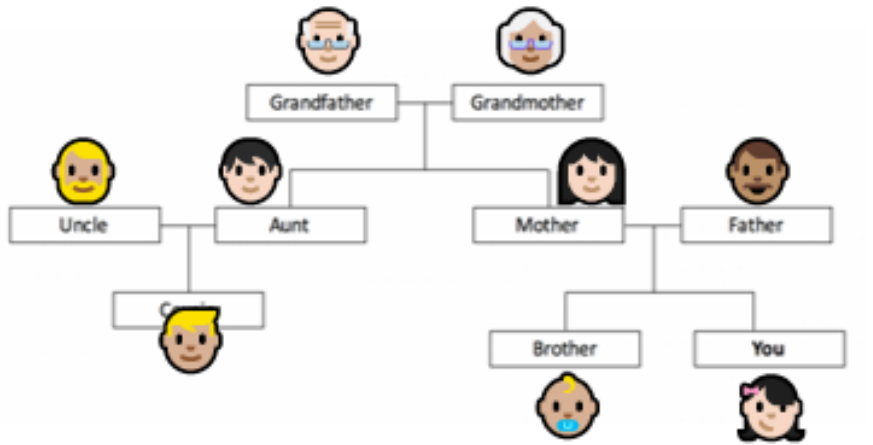


Genome image from *The New York Times* www.nytimes.com/interactive/2020/04/30/science/coronavirus-mutations.html
Phylogenic tree from Trevor Bedford Group: https://docs.nextstrain.org

What is phylogenetics?

- The study of **evolutionary relations** among biological entities (populations, organisms, genes)

- Such relationships are almost always inferred from **molecular sequence data**



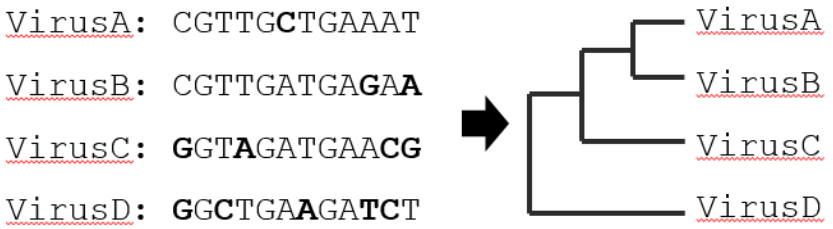


Image adapted from Nathan Grubaugh

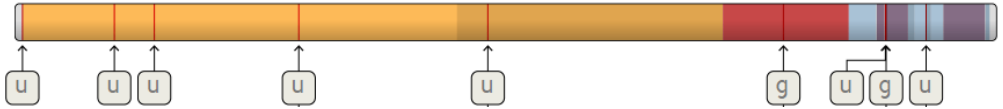
Basic unit of difference: Single nucleotide polymorphisms

- SNP = Single Nucleotide Polymorphism
  - ATGTT**C**CTC sequence
  - ATGTT**G**CTC reference
- SNPs occur across the full genome, with varied frequency:

Genome image adapted from *The New York Times* www.nytimes.com/interactive/2020/04/30/science/coronavirus-mutations.html

Multiple sequence alignment

- SNP profiles are genetic fingerprints



- Combine SNP profiles into a **multiple sequence alignment (MSA)** of multiple genomes
- MSAs are used to:
  - Measure relatedness
  - Build phylogenetic trees

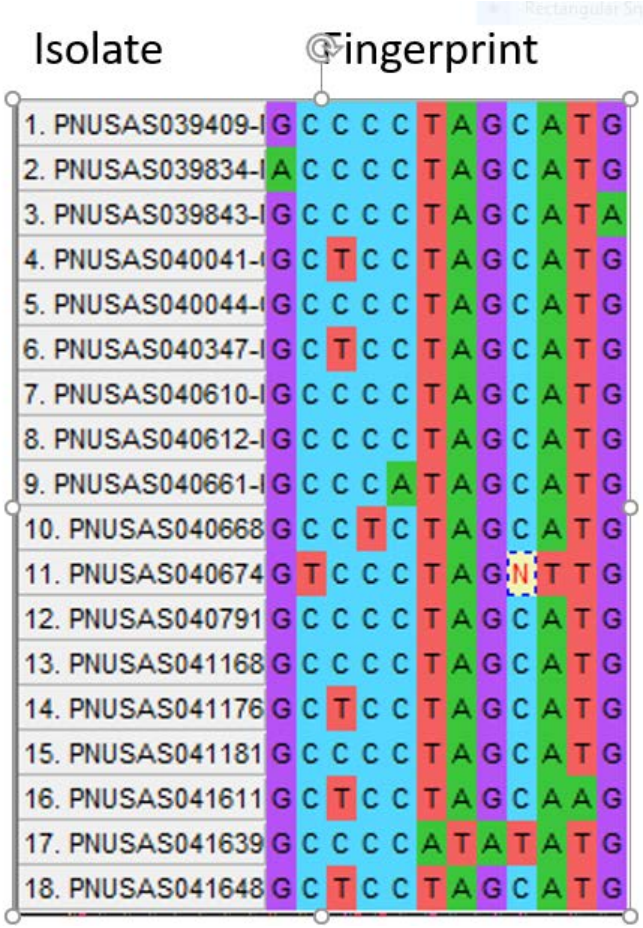


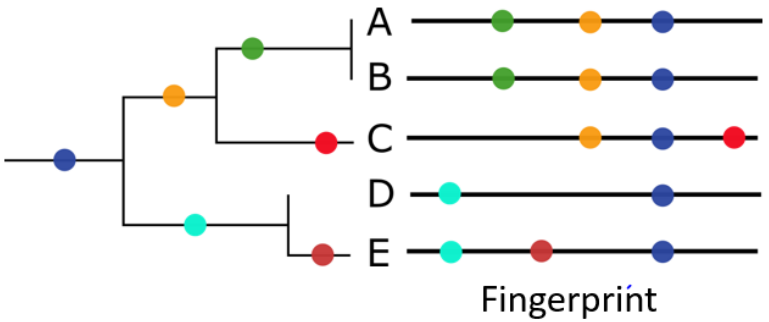
Image adapted from *The New York Times* www.nytimes.com/interactive/2020/04/30/science/coronavirus-mutations.html

Multiple sequence alignment

Genetic relatedness (SNP differences)

	2016K-0438	PNUSAS010651	PNUSAS011968	PNUSAS013510	PNUSAS013903	PNUSAS013904	PNUSAS013905
2016K-0438	0	35	37	36	25	28	26
PNUSAS010651	35	0	2	41	31	35	34
PNUSAS011968	37	2	0	41	30	34	33
PNUSAS013510	36	41	41	0	28	33	30
PNUSAS013903	25	31	30	28	0	1	1
PNUSAS013904	28	35	34	33	1	0	2
PNUSAS013905	26	34	33	30	1	2	0

Phylogenetic trees



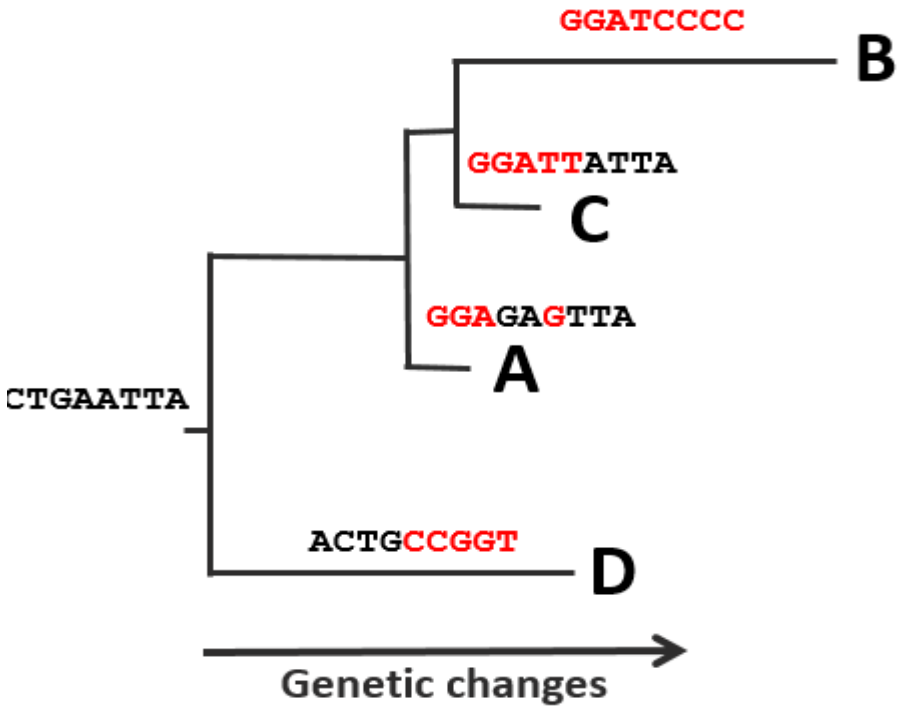
Tree image from Trevor Bedford Group: https://docs.nextstrain.org

Growing trees from MSA

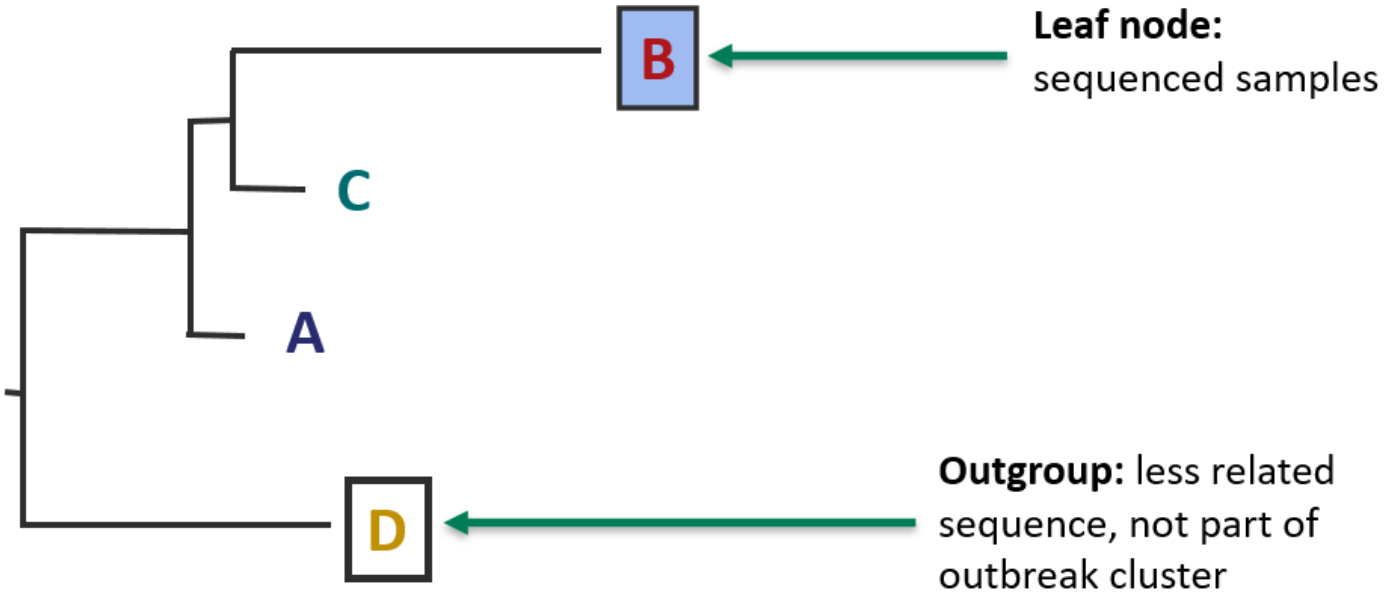
Isolate	Fingerprint
Ancestor	**ACTGAATTA**
A	**GGAGAGTTA**
B	**GGATCCCCC**
C	**GGATTATTA**
D	**ACTGCCGGT**

Growing trees from MSA

Isolate	Fingerprint
Ancestor	**ACTGAATTA**
A	**GGAGAGTTA**
B	**GGATCCCCC**
C	**GGATTATTA**
D	**ACTGCCGGT**



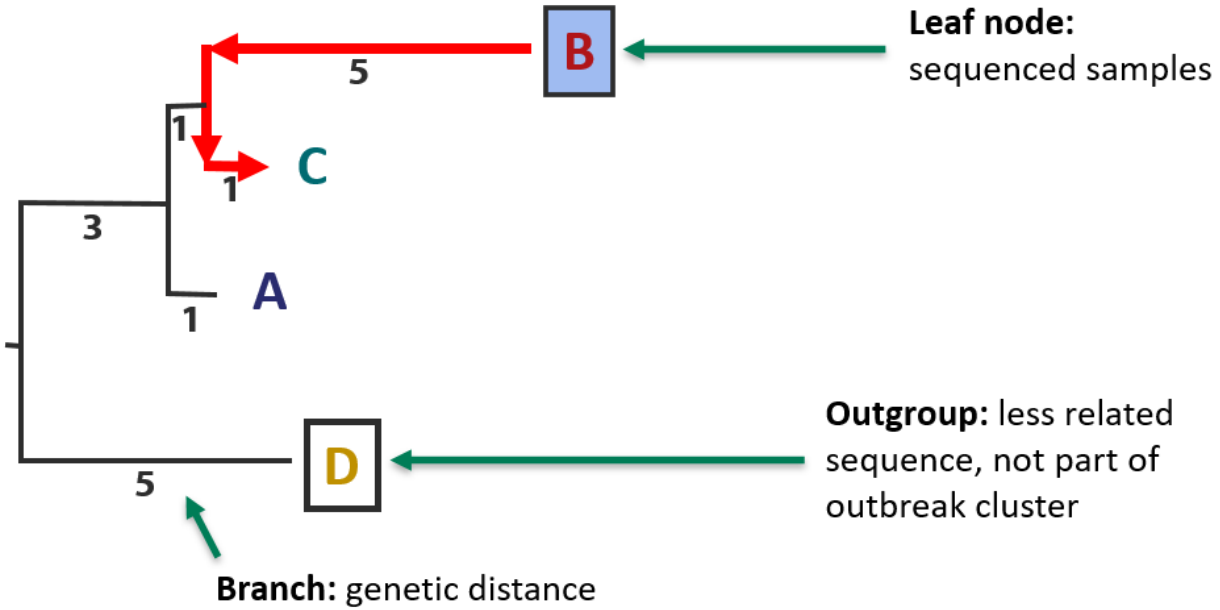
Anatomy of a phylogenetic tree



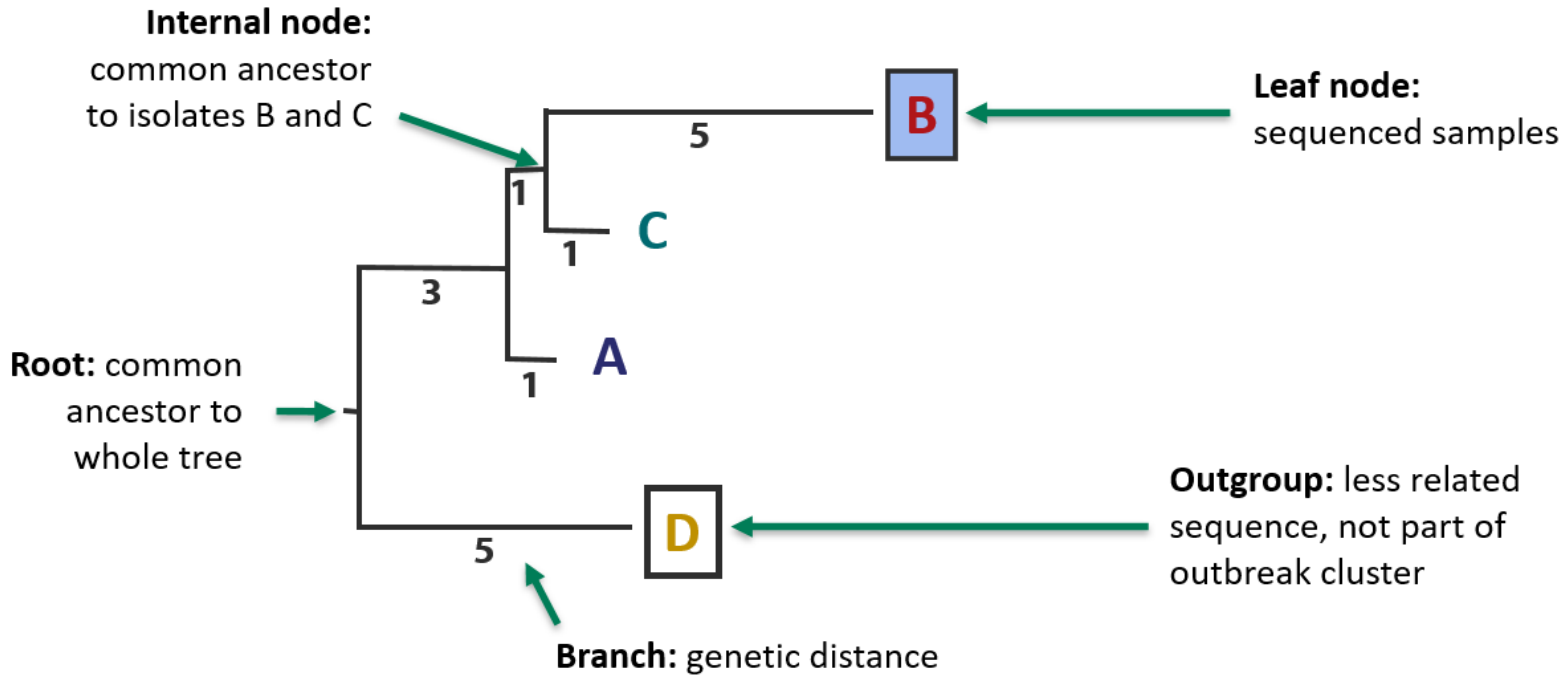
Anatomy of a phylogenetic tree

Genetic distance

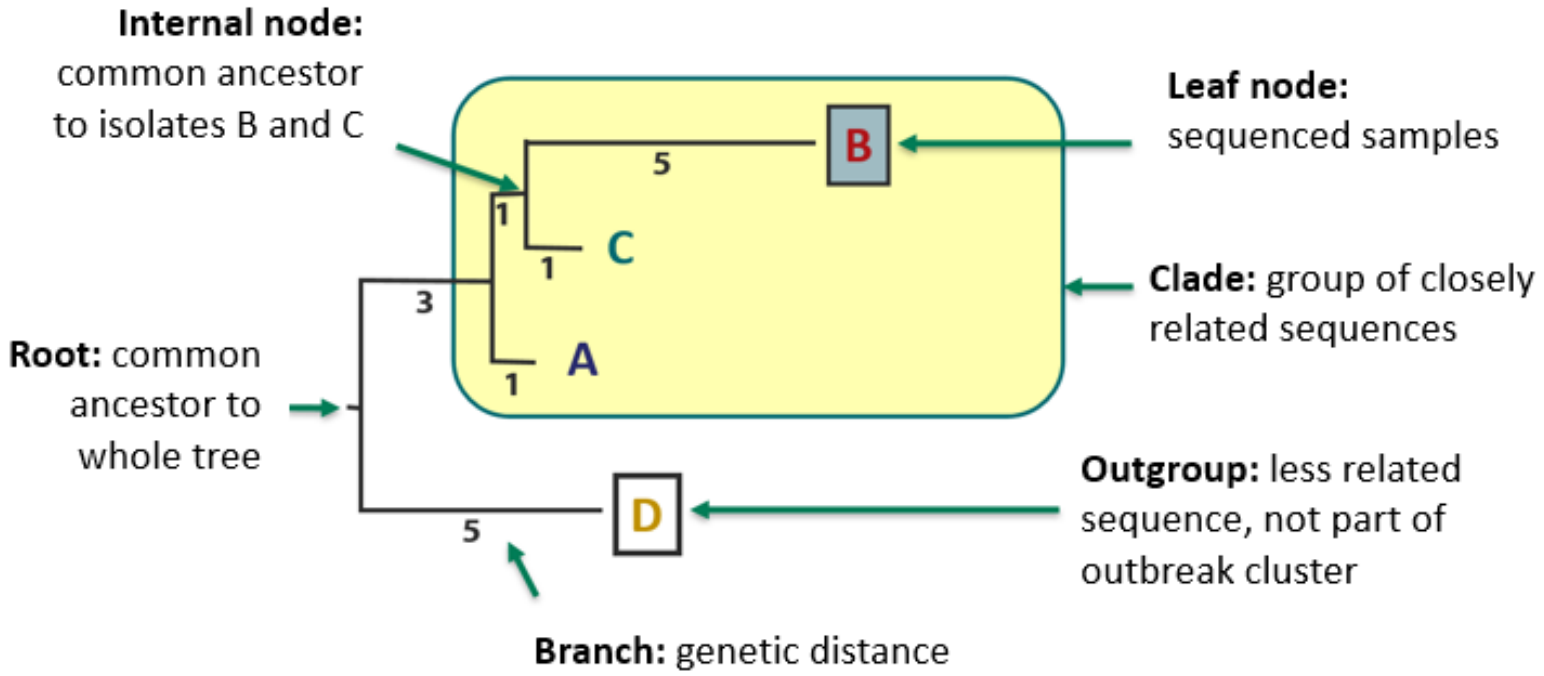
	A	B	C	D
A		7	3	9
B	7		6	14
C	3	6		10
D	9	14	10	



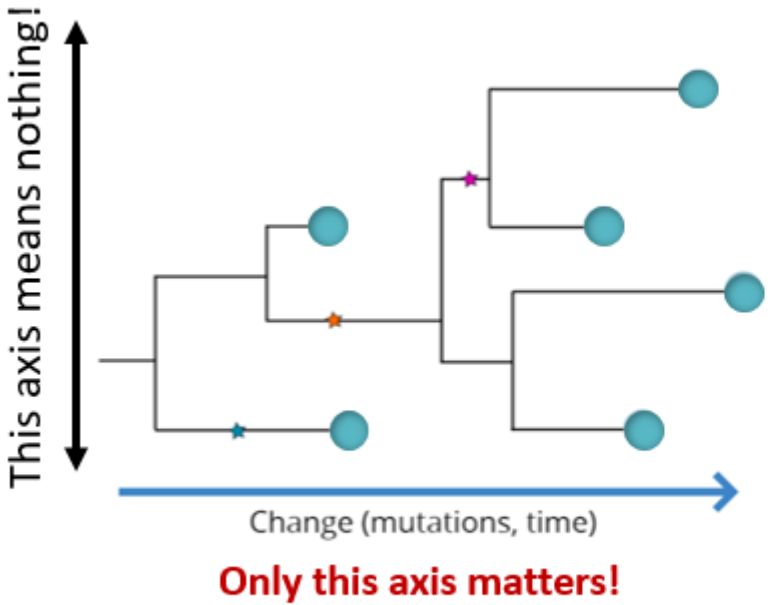
Anatomy of a phylogenetic tree

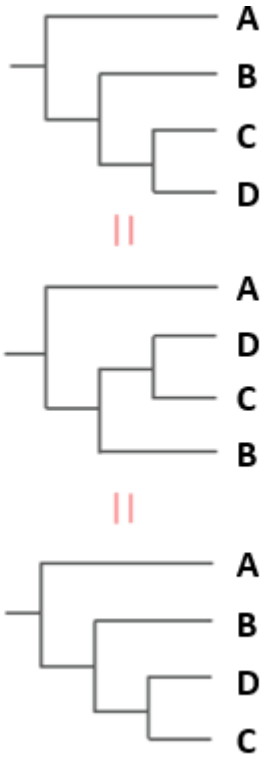


Anatomy of a phylogenetic tree



Branch rotations don’t change the tree

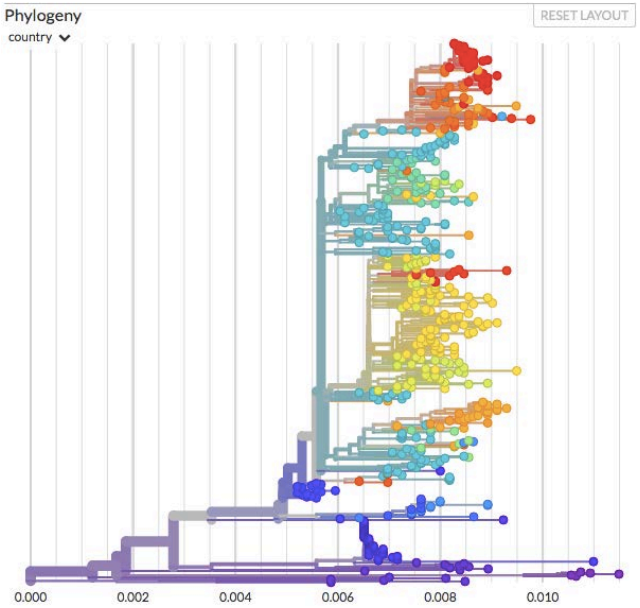




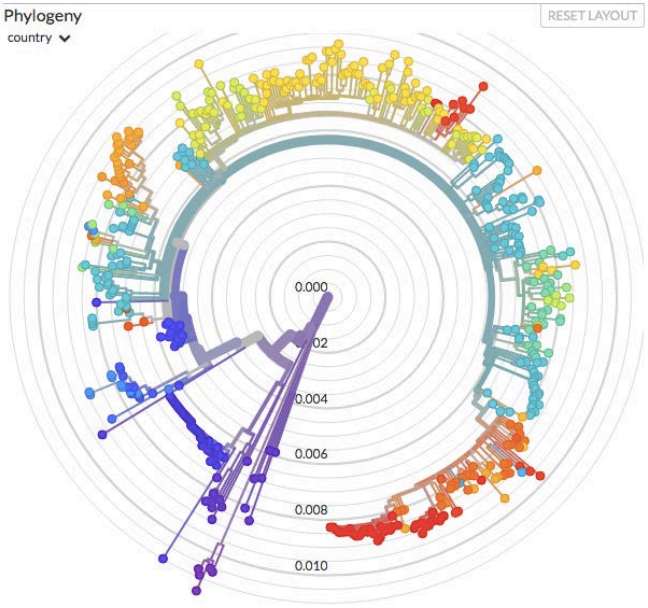
All three trees are the same

Adapted from Robert Bear

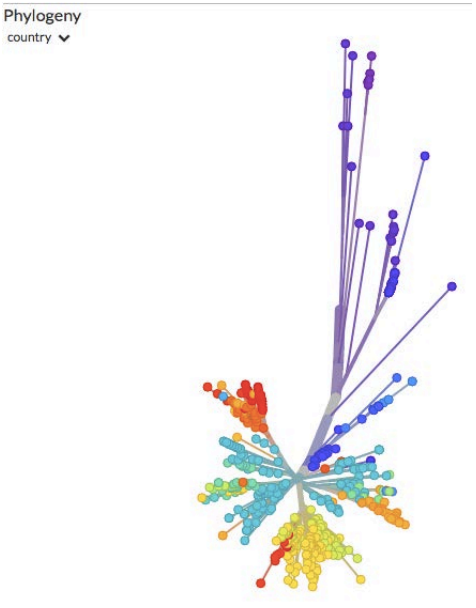
Same tree, different representations



Rectangular Rooted trees
(when outgroup is known)



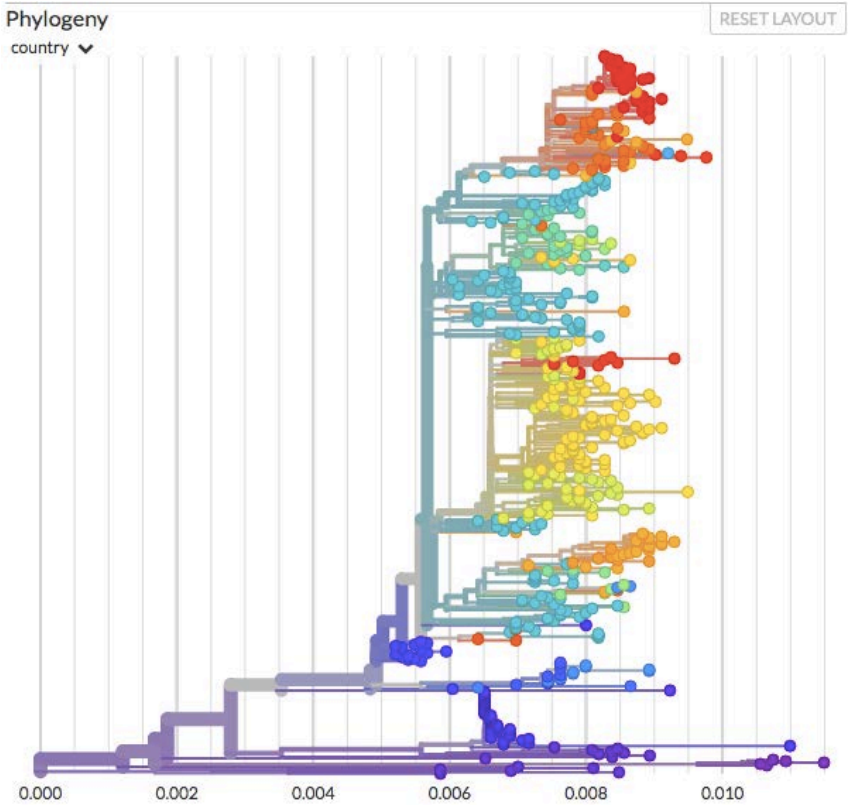
Radial Rooted trees
(when outgroup is known)



Unrooted tree
(direction of evolution unknown)

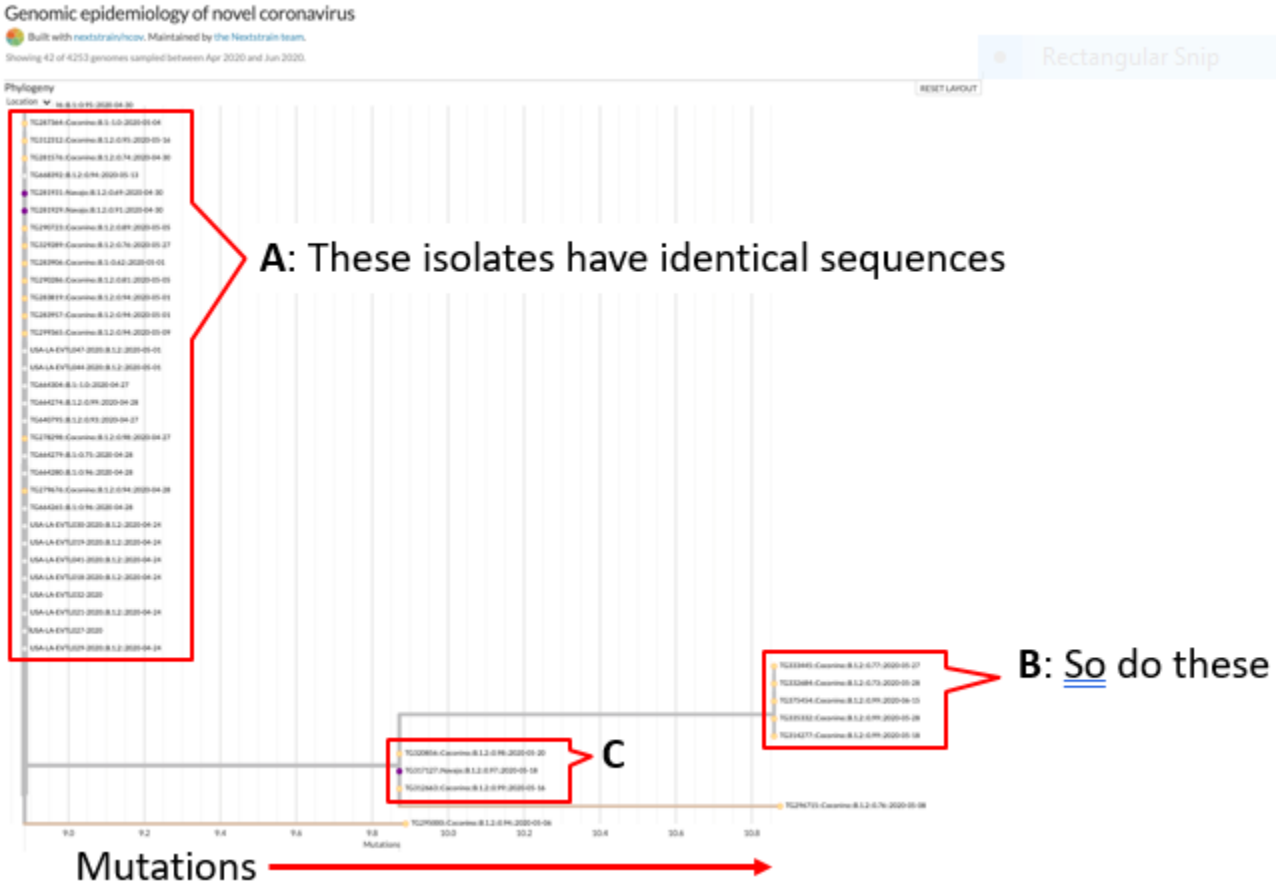
Adapted from Nathan Grubaugh Source: nextstrain.org

Visualizing trees: Nextstrain.org



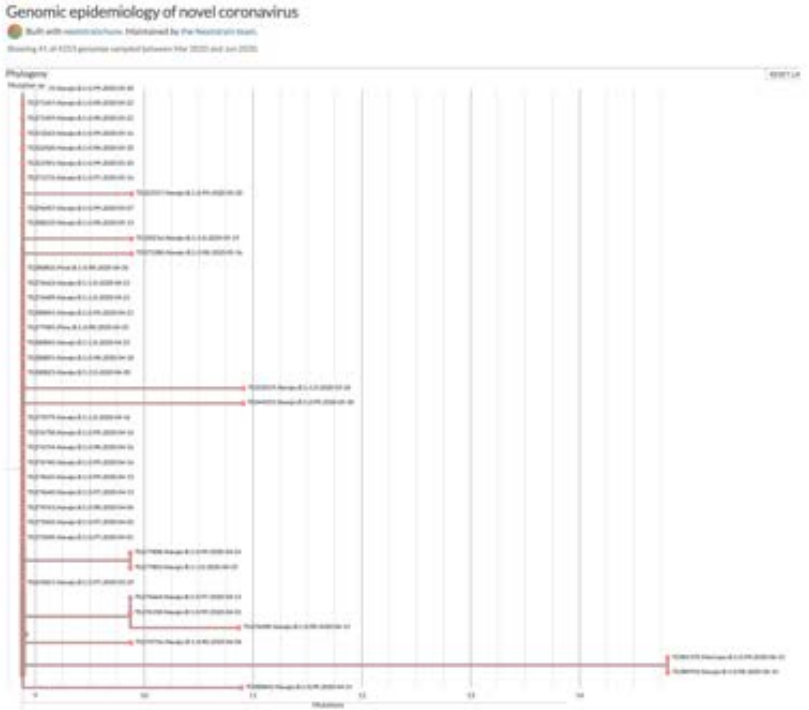
- Powerful and popular web app for visualizing phylogenetic trees
- Easily color leaf nodes with case metadata (e.g., location)
- Designed to aid epidemiological understanding
- Widely used for SARS-CoV-2
  - Case studies in this toolkit
  - Learn more in Module 3.1

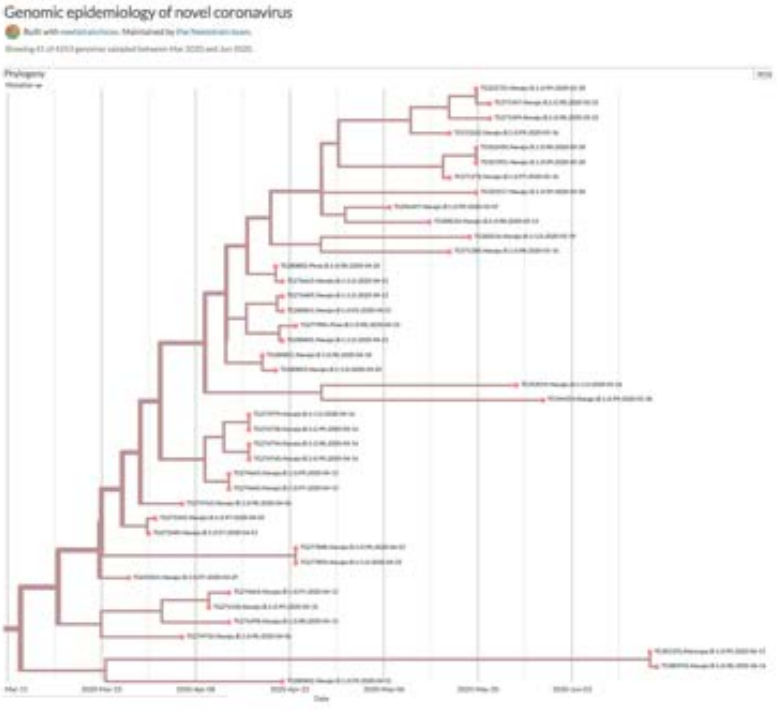
Images from Trevor Bedford Group: https://docs.nextstrain.org



Tree from Hayley Yaglom

Mutations vs Collection date





Trees from Hayley Yaglom

Visualizing trees: other tools

- FigTree (download, free): http://tree.bio.ed.ac.uk/software/figtree/
- Geneious (download, $$): https://www.geneious.com/
- UGENE (download, free): http://ugene.net/
- TreeView (download, free): http://jtreeview.sourceforge.net/
- iTOL (online, free or $$): https://itol.embl.de/
- ETE Toolkit (online, free): http://etetoolkit.org/treeview/
- MicroReact (online, free): http://microreact.org/

Listed for identification only and does not imply endorsement by the Centers for Disease Control and Prevention or the US Department of Health and Human Services.

Adapted from Nathan Grubaugh

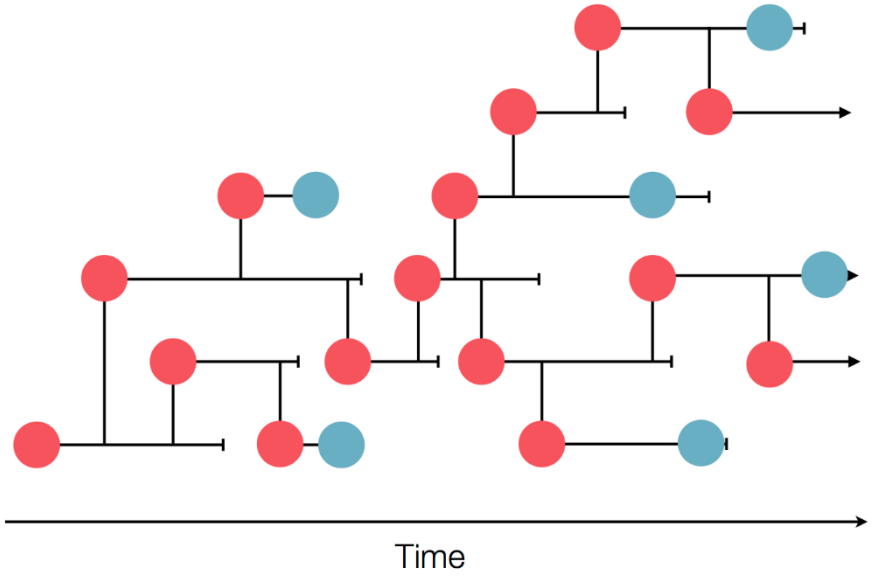
Limitations of core assumptions: implications

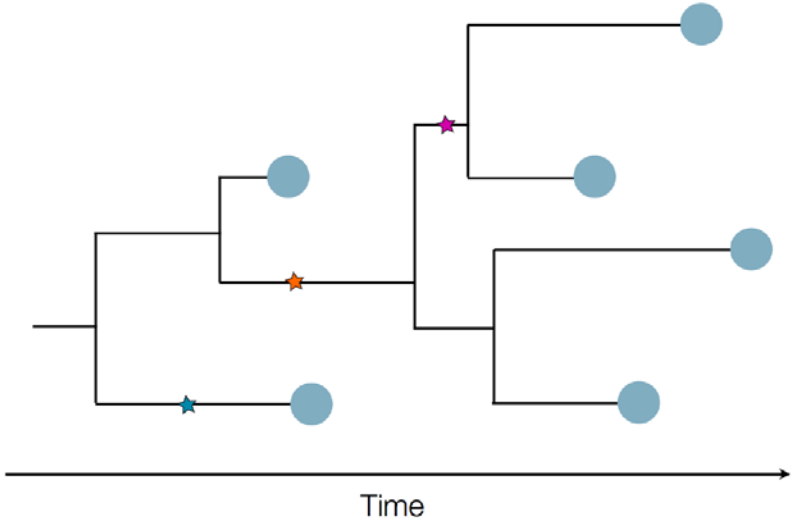
Strains that are phylogenetically closer are more likely to share an epidemiological association. BUT…

- Transmission pathways (and the direction of transmission) cannot be assumed to mirror phylogeny (without other data)
- Causal links (e.g., between cases and exposures) cannot be assumed from sequence data alone
- Trees are only an approximation of the true story!

Limitation: Phylogeny =/= Transmission

Strains that are phylogenetically closer are more likely to share an epidemiological association. BUT…





Images from Trevor Bedford Group: https://docs.nextstrain.org

Limitation: Phylogeny =/= Transmission

Interpret with caution because topology depends on sampling:

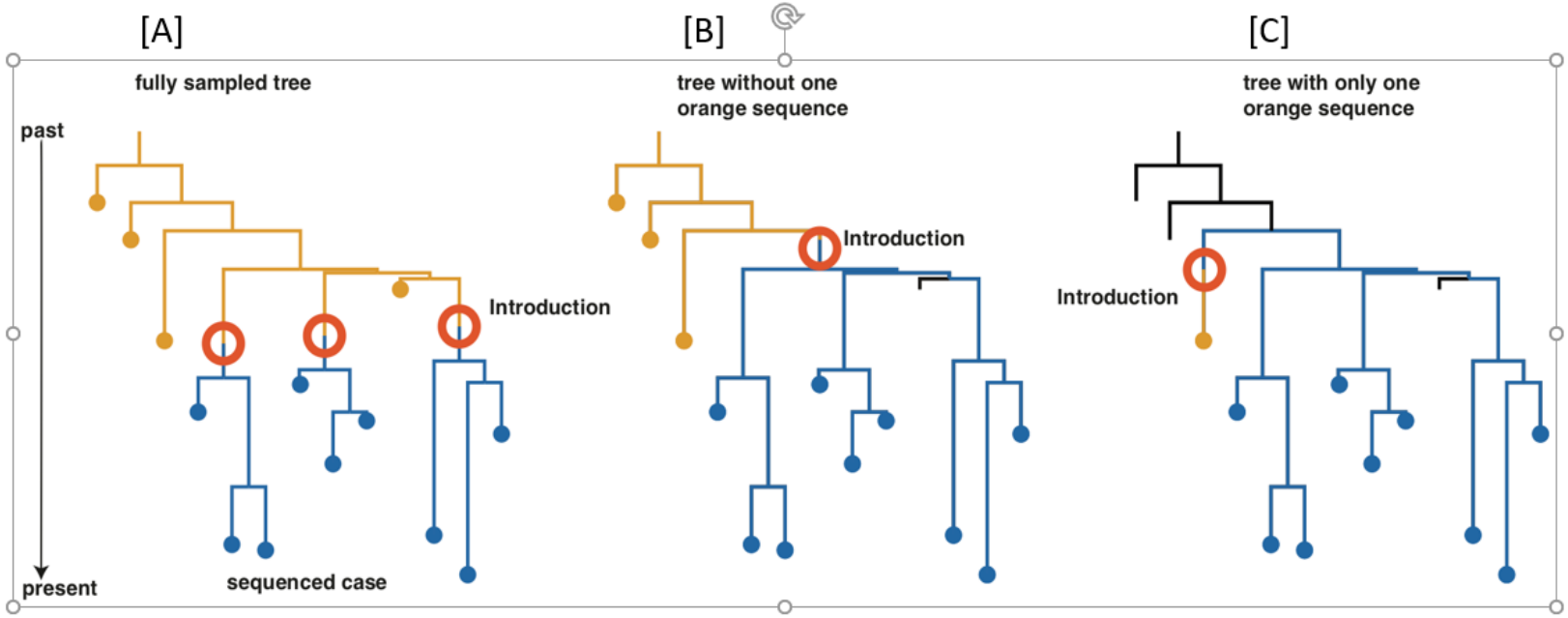


Figure from Nextstrain.org

Summary

- Viruses mutate as they spread, producing a genetic fingerprint (SNPs)
- Fingerprints from many sequenced viral isolates can be combined into a multiple sequence alignment for comparison
- The ancestral relationships among sequences can be represented in phylogenetic trees
- Strains that are phylogenetically closer are *more likely* to share an epidemiological association
- Interpret with caution, all trees are an approximation of the truth!

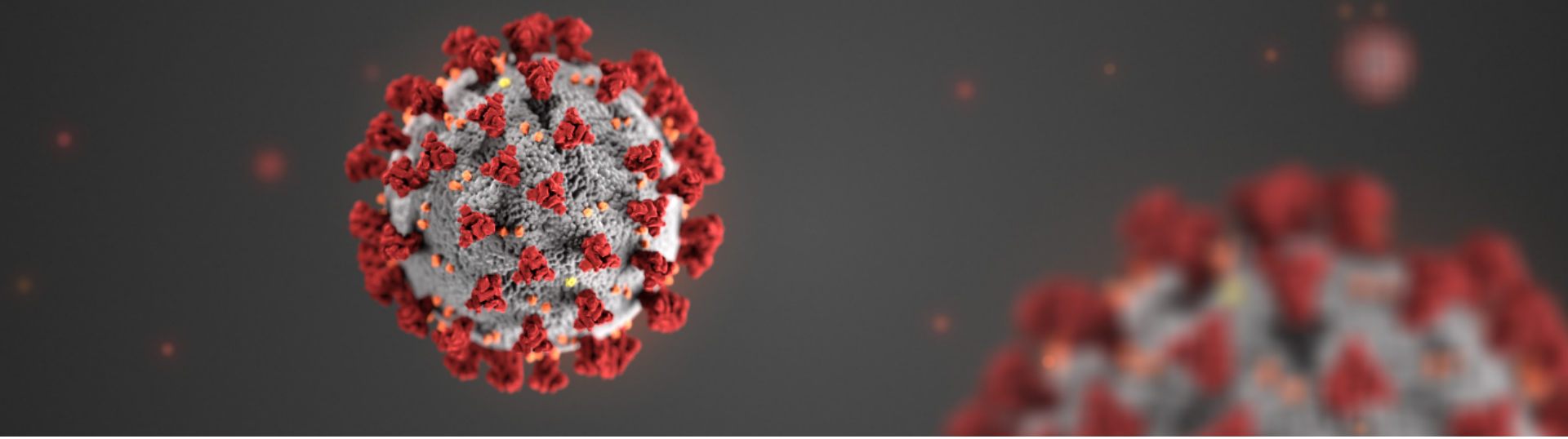
"Phylogenetic trees can be beautifully dangerous in their interpretation."

- Emma Hodcroft

Learn more

- Other introduction modules
  - What is genomic epidemiology? – Module 1.1
  - The SARS-CoV-2 genome – Module 1.2

- COVID-19 Genomic Epidemiology Toolkit
  - Find further reading
  - Subscribe to receive updates on new modules as they are released
    go.usa.gov/xAbMw



For more information, contact CDC
1-800-CDC-INFO (232-4636)
TTY: 1-888-232-6348 www.cdc.gov

The findings and conclusions in this report are those of the authors and do not necessarily represent the official position of the Centers for Disease Control and Prevention.