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
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.
Genetic fingerprinting

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

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

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


Phylogenetic tree from Trevor Bedford Group: [https://docs.nextstrain.org](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**

VirusA: CGTTGCTGAAAT
VirusB: CGTTGATGAGAA
VirusC: GGTAGATGAAACG
VirusD: GGCTGAAGATCT

Image adapted from Nathan Grubaugh
Basic unit of difference: Single nucleotide polymorphisms

- SNP = Single Nucleotide Polymorphism
  - ATGTT\textcolor{red}{C}CTC sequence
  - ATGTT\textcolor{green}{G}CTC reference

- SNPs occur across the full genome, with varied frequency:

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

Multiple sequence alignment

**Genetic relatedness**
(SNP differences)

<table>
<thead>
<tr>
<th></th>
<th>2016K-0438</th>
<th>PNUSAS010651</th>
<th>PNUSAS011968</th>
<th>PNUSAS013510</th>
<th>PNUSAS013903</th>
<th>PNUSAS013904</th>
<th>PNUSAS013905</th>
</tr>
</thead>
<tbody>
<tr>
<td>2016K-0438</td>
<td>0</td>
<td>35</td>
<td>37</td>
<td>36</td>
<td>25</td>
<td>28</td>
<td>26</td>
</tr>
<tr>
<td>PNUSAS010651</td>
<td>35</td>
<td>0</td>
<td>2</td>
<td>41</td>
<td>31</td>
<td>35</td>
<td>34</td>
</tr>
<tr>
<td>PNUSAS011968</td>
<td>37</td>
<td>2</td>
<td>0</td>
<td>41</td>
<td>30</td>
<td>34</td>
<td>33</td>
</tr>
<tr>
<td>PNUSAS013510</td>
<td>36</td>
<td>41</td>
<td>41</td>
<td>0</td>
<td>28</td>
<td>33</td>
<td>30</td>
</tr>
<tr>
<td>PNUSAS013903</td>
<td>25</td>
<td>31</td>
<td>30</td>
<td>28</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>PNUSAS013904</td>
<td>28</td>
<td>35</td>
<td>34</td>
<td>33</td>
<td>1</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>PNUSAS013905</td>
<td>26</td>
<td>34</td>
<td>33</td>
<td>30</td>
<td>1</td>
<td>2</td>
<td>0</td>
</tr>
</tbody>
</table>

**Phylogenetic trees**

Tree image from Trevor Bedford Group: [https://docs.nextstrain.org](https://docs.nextstrain.org)

**Fingerprint**
## Growing trees from MSA

<table>
<thead>
<tr>
<th>Isolate</th>
<th>Fingerprint</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ancestor</td>
<td><strong>ACTGAATTA</strong></td>
</tr>
<tr>
<td>A</td>
<td><strong>GGAGAGTTTA</strong></td>
</tr>
<tr>
<td>B</td>
<td><strong>GGATCCCCC</strong></td>
</tr>
<tr>
<td>C</td>
<td><strong>GGATTATTA</strong></td>
</tr>
<tr>
<td>D</td>
<td><strong>ACTGCCGGT</strong></td>
</tr>
</tbody>
</table>
# Growing trees from MSA

<table>
<thead>
<tr>
<th>Isolate</th>
<th>Fingerprint</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ancestor</td>
<td>ACTGAATTA</td>
</tr>
<tr>
<td>A</td>
<td>GGAGAGTTA</td>
</tr>
<tr>
<td>B</td>
<td>GGATCCCCC</td>
</tr>
<tr>
<td>C</td>
<td>GGATTATTA</td>
</tr>
<tr>
<td>D</td>
<td>ACTGCCGGT</td>
</tr>
</tbody>
</table>

The table shows the genetic fingerprints of different isolates, with the Ancestor being the starting point. The diagram illustrates the genetic changes from the Ancestor isolate to the other isolates, showing how genetic changes have occurred over time.
Anatomy of a phylogenetic tree

Leaf node: sequenced samples

Outgroup: less related sequence, not part of outbreak cluster
Anatomy of a phylogenetic tree

Genetic distance

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>7</td>
<td>3</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>7</td>
<td>6</td>
<td>14</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>3</td>
<td>6</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>9</td>
<td>14</td>
<td>10</td>
<td></td>
</tr>
</tbody>
</table>

Leaf node: sequenced samples

Outgroup: less related sequence, not part of outbreak cluster

Branch: genetic distance
Anatomy of a phylogenetic tree

**Internal node:** common ancestor to isolates B and C

**Leaf node:** sequenced samples

**Root:** common ancestor to whole tree

**Outgroup:** less related sequence, not part of outbreak cluster

**Branch:** genetic distance
Anatomy of a phylogenetic tree

**Leaf node:** sequenced samples

**Internal node:** common ancestor to isolates B and C

**Root:** common ancestor to whole tree

**Clade:** group of closely related sequences

**Outgroup:** less related sequence, not part of outbreak cluster

**Branch:** genetic distance
Branch rotations don’t change the tree

This axis means nothing!

Only this axis matters!

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
| A: These isolates have identical sequences |
| B: So do these |

Mutations
Mutations vs Collection date

Trees from Hayley Yaglom
Visualizing trees: other tools

- FigTree (download, free): [http://tree.bio.ed.ac.uk/software/figtree/](http://tree.bio.ed.ac.uk/software/figtree/)
- Geneious (download, $$): [https://www.geneious.com/](https://www.geneious.com/)
- iTOL (online, free or $$): [https://itol.embl.de/](https://itol.embl.de/)

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:
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)

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.