Advanced Molecular Detection
Southeast Region Bioinformatics

AMD Southeast Region Genomic Epidemiology Training
03/04/2024

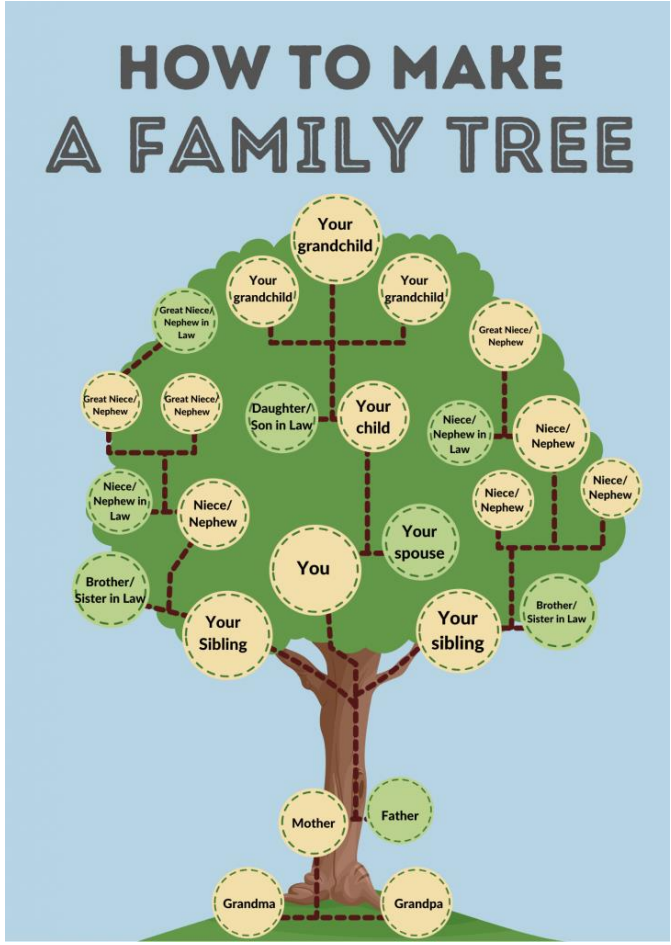
Outline: Phylogenetic Trees

- Components and Vocabulary
- Available Tools and Resources
- Communicating Results
- SNP Matrices

Components and Vocabulary

What is a phylogenetic tree?

- A diagram that shows the evolutionary relationships among species, organisms, or whatever you're interested in
- Aka phylogeny
- Depicts common ancestors – shared ancestral organism by at least two lineages

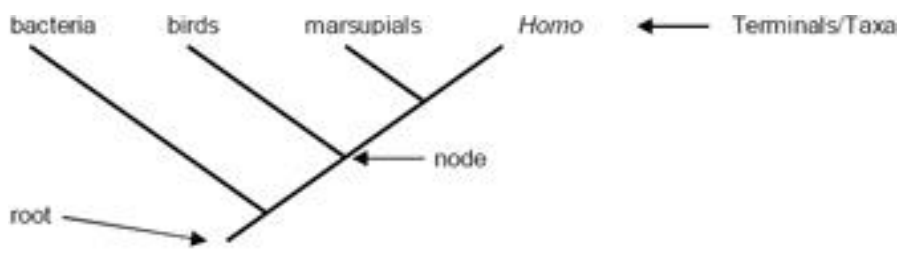


Advanced Molecular Detection
Southeast Region Bioinformatics

https://www.khanacademy.org/science/high-school-biology/hs-evolution/hs-phylogeny/a/phylogenetic-trees

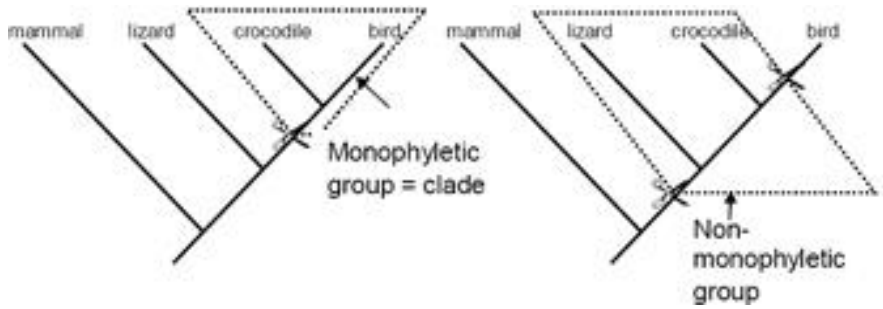
David Baum, Ph.D. (*Dept. of Botany, University of Wisconsin, 430 Lincoln Ave., Madison, WI*) © 2008 Nature Education
Citation: Baum, D. (2008) Reading a Phylogenetic Tree: The Meaning of Monophyletic Groups. *Nature Education* 1(1):190

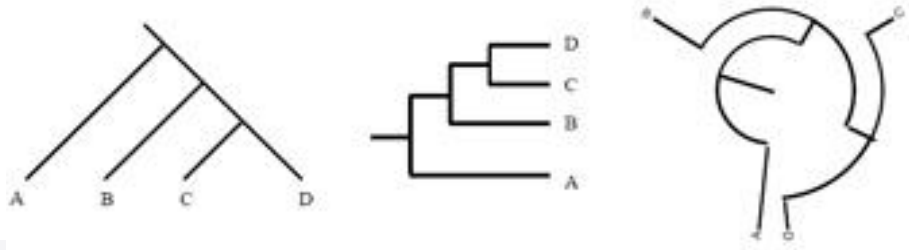
Parts of a Tree



1. Root: Common ancestor of all the organisms on the tree
2. Node: branching point marking change between two lineages
3. Terminal: End of the branch, usually the most recent sample

1. Clade: aka monophyletic group, a part of the bigger phylogeny that shares a common ancestor





1. Left: Rooted
2. Center: Rooted and the most common
3. Right: Unrooted, show relatedness without assuming ancestry

Advanced Molecular Detection
Southeast Region Bioinformatics

Parts of a Tree: Time vs Mutation

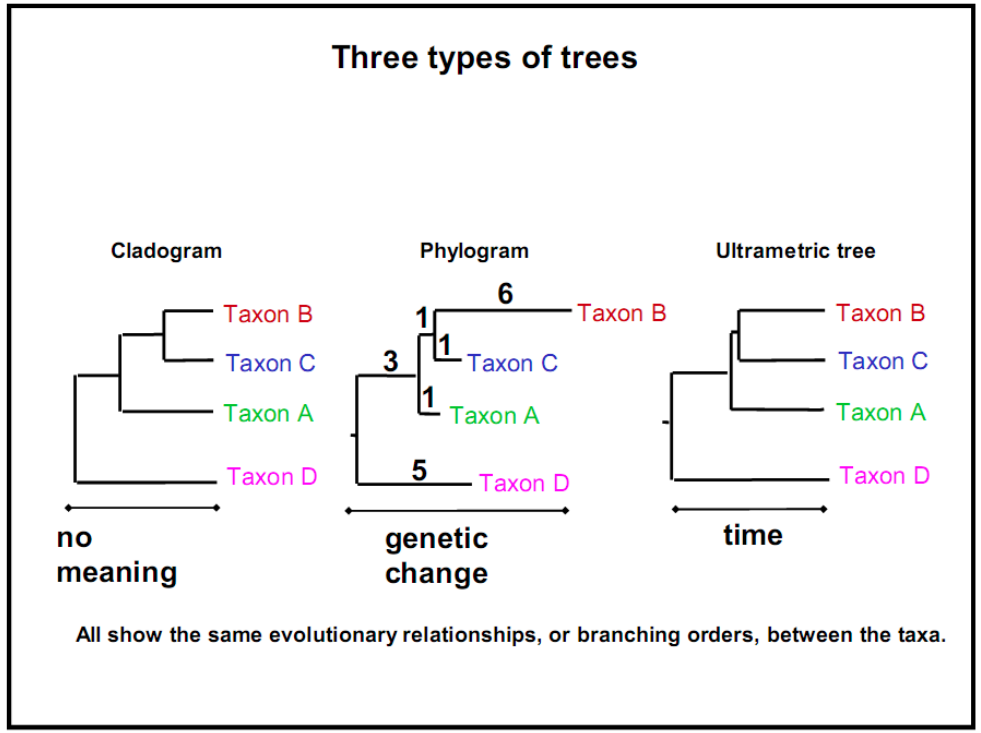


Figure 26.3: Three types of trees.

Branch length can be used to show time, mutation rate, or nothing at all

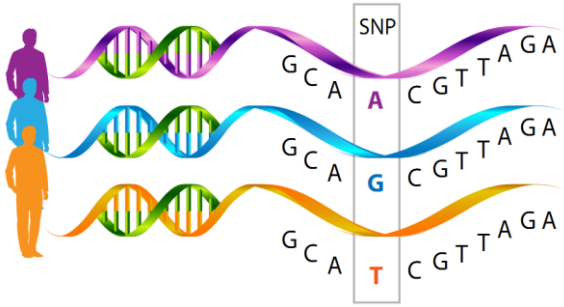
Important to check branch length meaning when reading a tree

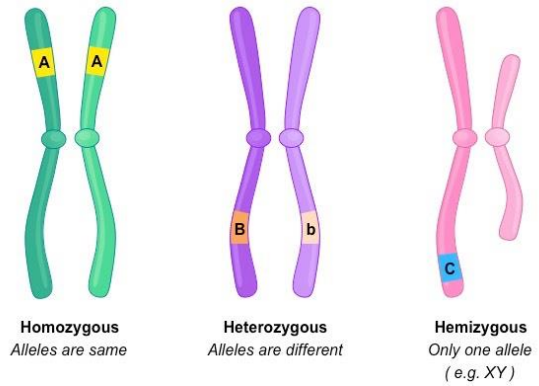
Advanced Molecular Detection
Southeast Region Bioinformatics

Kellis et al. Computational Biology-Genomes, Networks, and Evolution. Chapter 26: Molecular Evolution. Section 26.2 Basic of Phylogeny

Components and Vocabulary

- SNP- Single Nucleotide Polymorphism
- MSA- Multiple Sequence Alignment
  - Combined SNP profiles of many people
  - Used to measure relatedness and create trees
- Indels-Insertion/Deletion
- Allele-one of two versions of a DNA sequence at specific location

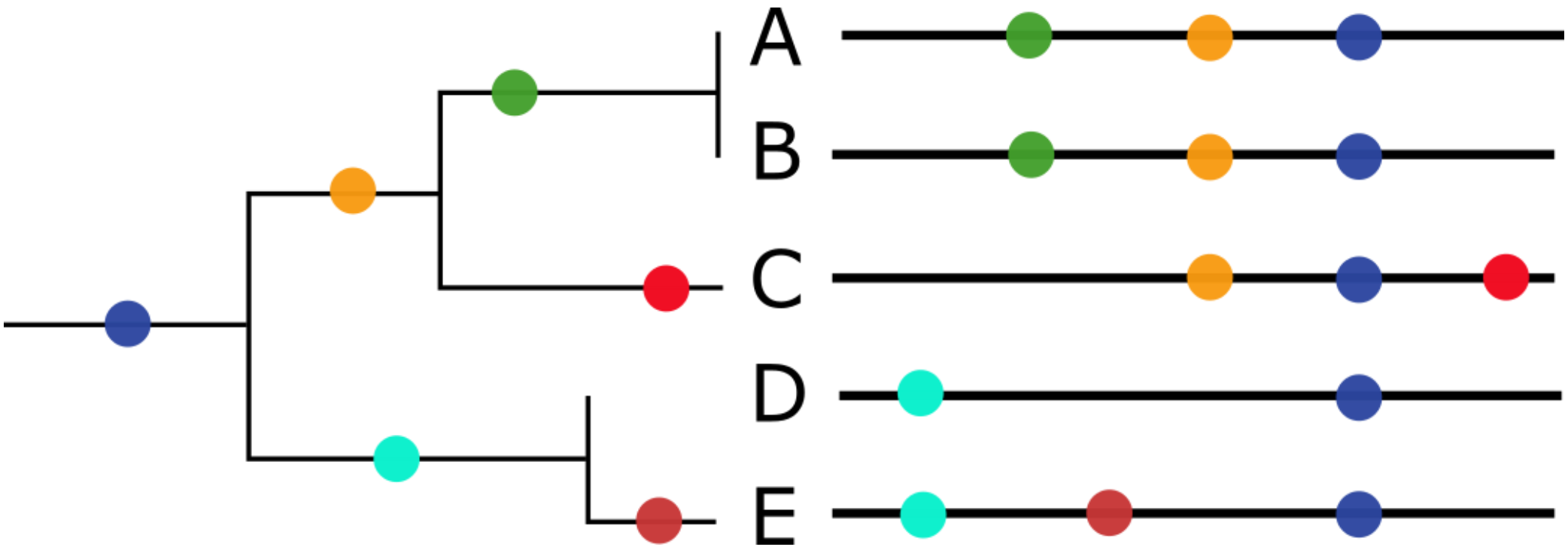




Advanced Molecular Detection
Southeast Region Bioinformatics

SNP Image: https://www.nutrigeneticsspecialists.com/single-post/2017/03/27/what-is-a-snp
Allele Image: https://ib.bioninja.com.au/standard-level/topic-3-genetics/34-inheritance/types-of-zygosity.html

Mutation Accumulation



Advanced Molecular Detection
Southeast Region Bioinformatics

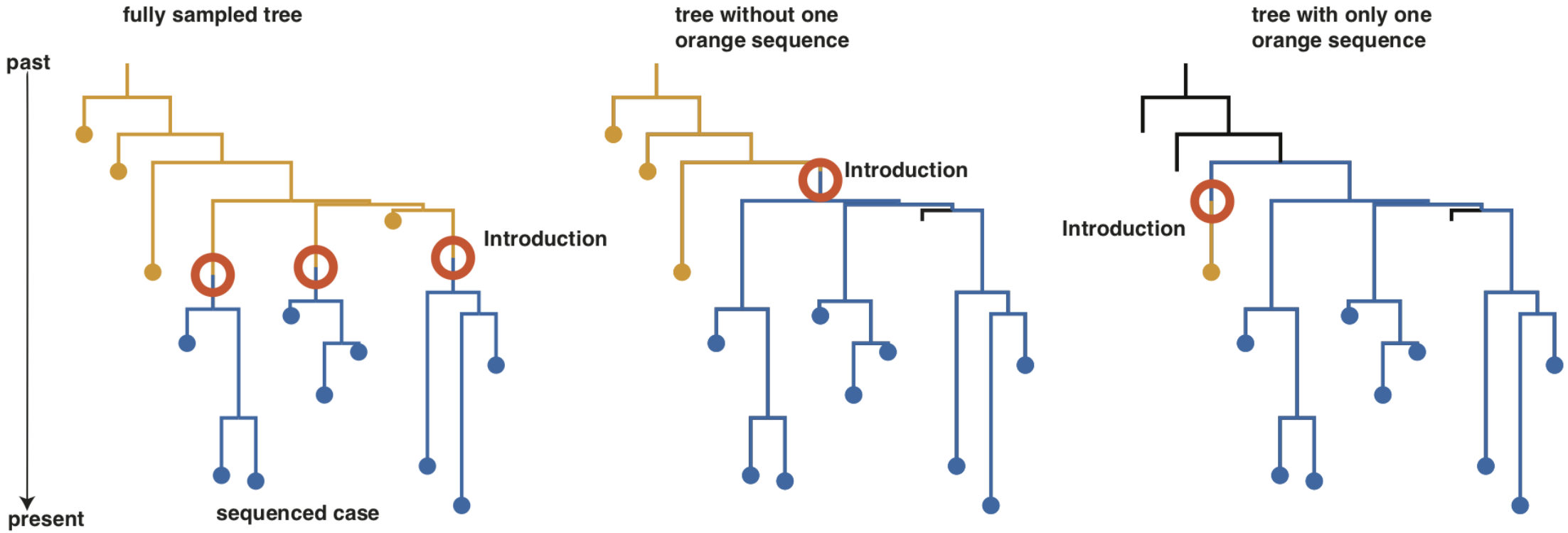
https://nextstrain.org/narratives/trees-background?n=3

Components and Vocabulary

- Genetic Relatedness-probability that one allele from one organism is present in another organism
- Convenience Sampling – sample whoever is available and willing to be sampled
  - Most common sampling method for DNA Sequencing
- Glossary of Bioinformatics and Gen Epi Terms
  - https://docs.nextstrain.org/en/latest/reference/glossary.html

Advanced Molecular Detection
Southeast Region Bioinformatics

Tree Interpretation

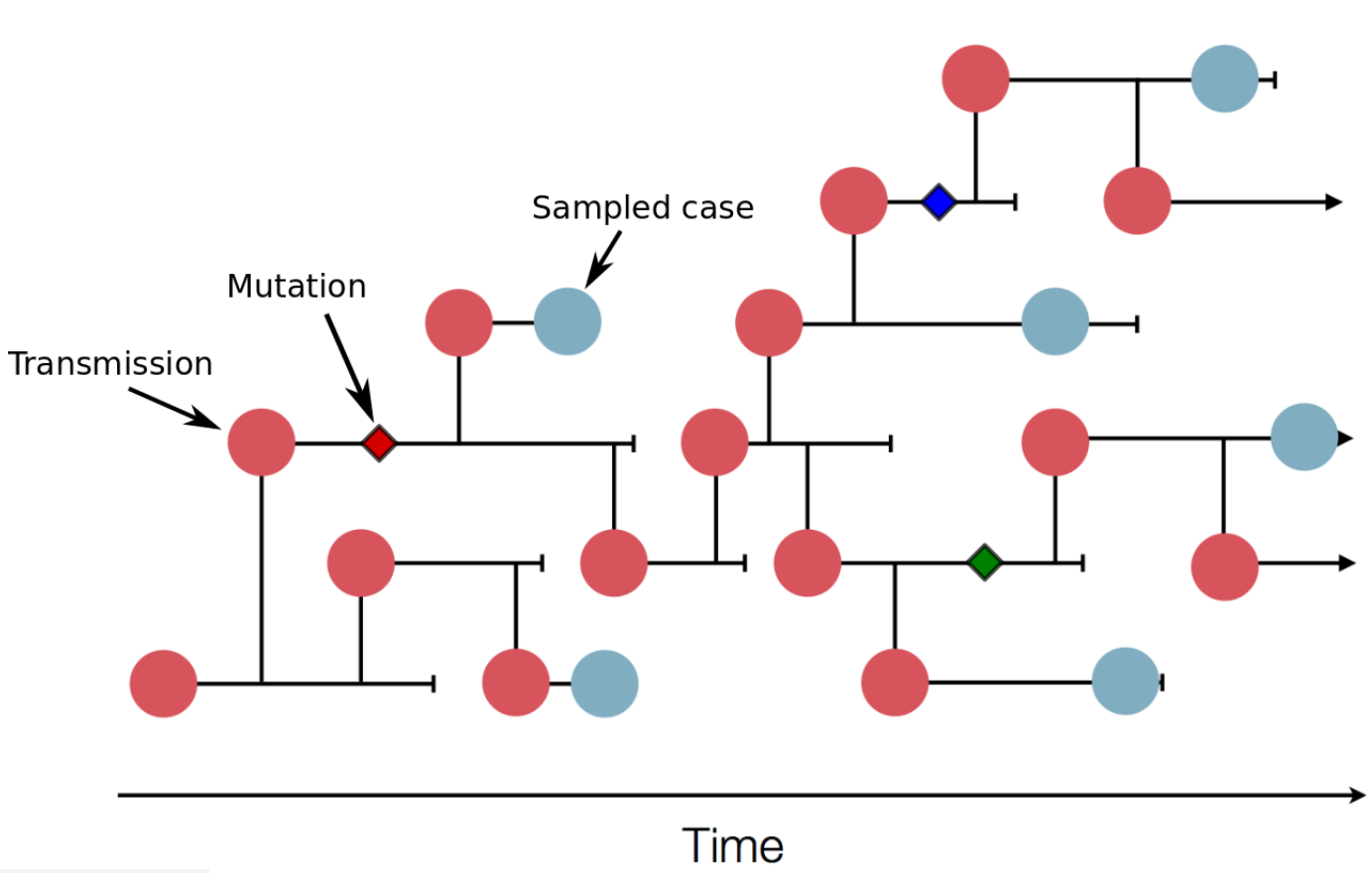


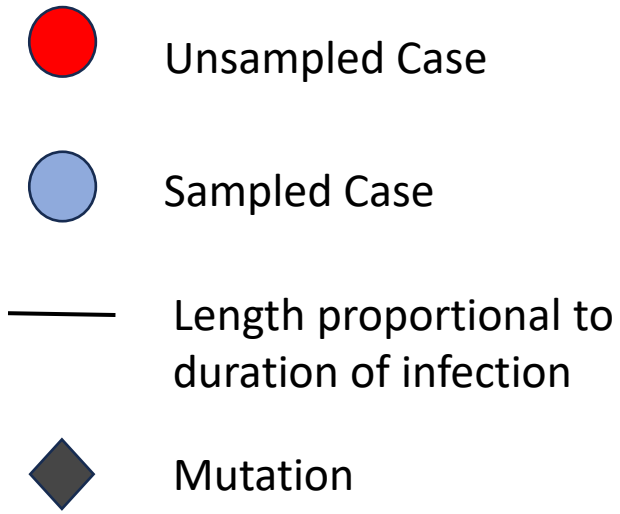
- Orange and Blue represent two different locations
- Changes in sampling drastically affect the trees generated and conclusions drawn
- Moral of the story: use caution when interpreting trees

Advanced Molecular Detection
Southeast Region Bioinformatics

https://nextstrain.org/narratives/trees-background?n=8

Other Trees: Transmission Tree





Advanced Molecular Detection
Southeast Region Bioinformatics

https://nextstrain.org/narratives/trees-background?n=2

Additional Learning Resources

- CDC Gen Epi Tool Kit
  - https://www.cdc.gov/amd/training/covid-19-gen-epi-toolkit.html
- NextStrain: How to Interpret Phylogenetic Trees
  - https://nextstrain.org/narratives/trees-background
- ARTIC Network
  - https://artic.network/how-to-read-a-tree.html
- Khan Academy
  - https://www.khanacademy.org/science/high-school-biology/hs-evolution/hs-phylogeny/a/phylogenetic-trees

Advanced Molecular Detection
Southeast Region Bioinformatics

Available Tools

1. NextStrain
    1. Web or CLI (Command Line Interface)
2. IqTree with ggtree
    1. CLI and the R
3. iTOL (interactive tree of life)
    1. CLI or Web

*It's not important to know fine details, just want to introduce them as options

Advanced Molecular Detection
Southeast Region Bioinformatics

If you're feeling adventurous

Untitled1*

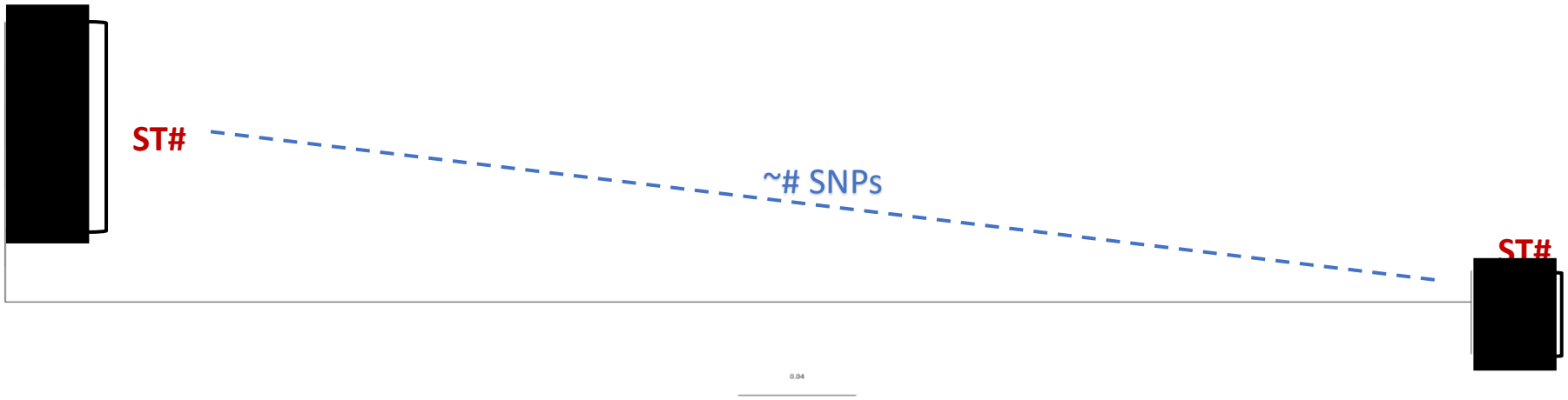
Source on Save | Run | Source

```r
library(readr)
SNPs_boot <- read_csv("path/to/SNPs_boot.treefile")
View(SNPs_boot)

library(ggtree)
library(ggplot2)
library(tidyverse)

#Read in tree file with ggtree"
tree <- read.tree("path/to/SNPs_boot.treefile")

ggtree(tree, right=TRUE) + geom_treescale() + geom_tiplab(size=6)

#Save plot as image
ggsave("SNPs_boot_tree.tiff", width = 85, height = 25, units = "cm")
```

15:23 (Top Level) R Script

IqTree and ggtree

- Example of ggtree output
- Adding ST values and numbers of SNPs between clades can help with drawing conclusions



SNP Matrices

- Matrix depicting the number of SNPs between two samples
- More SNPs means more genetic distance between two samples
- The larger the SNP number, the less likely two samples are related
- Is used in conjunction with phylogenetic trees to determine relatedness between two samples (usually during an outbreak)

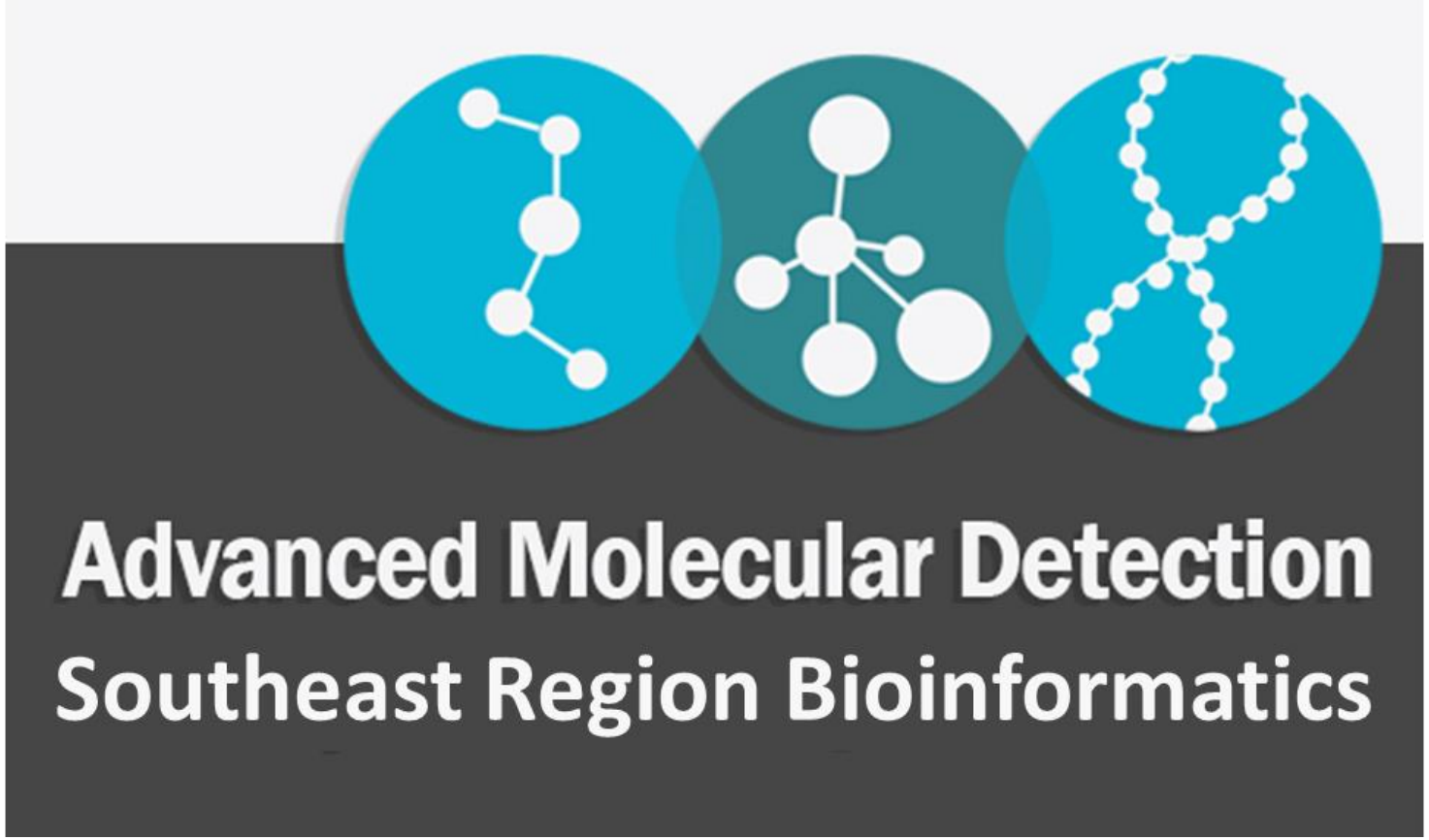
snp-dists 0.6.2	Strain 1	Strain 2	Strain 3
Strain 1	0	22	13
Strain 2	22	0	17
Strain 3	13	17	0

Advanced Molecular Detection
Southeast Region Bioinformatics

Communicating Results

- Sequencing information can influence a variety of public health decisions
  - Masking policies
  - Targeted vaccine campaigns
  - Nursing home lockdowns
  - Travel advisories
  - Determining cases in an outbreak

Advanced Molecular Detection
Southeast Region Bioinformatics



Questions?

Bphl-sebioinformatics@flhealth.gov

TBD
Lead Bioinformatician
TBD@flhealth.gov

Molly Mitchell, PhD
Bioinformatician
Molly.Mitchell@flhealth.gov

Lakshmi Thsaliki, MS
Bioinformatician
Lakshmi.Thsaliki@flhealth.gov

Sam Marcellus, MPH
Bioinformatician
Samantha.marcellus@flhealth.gov