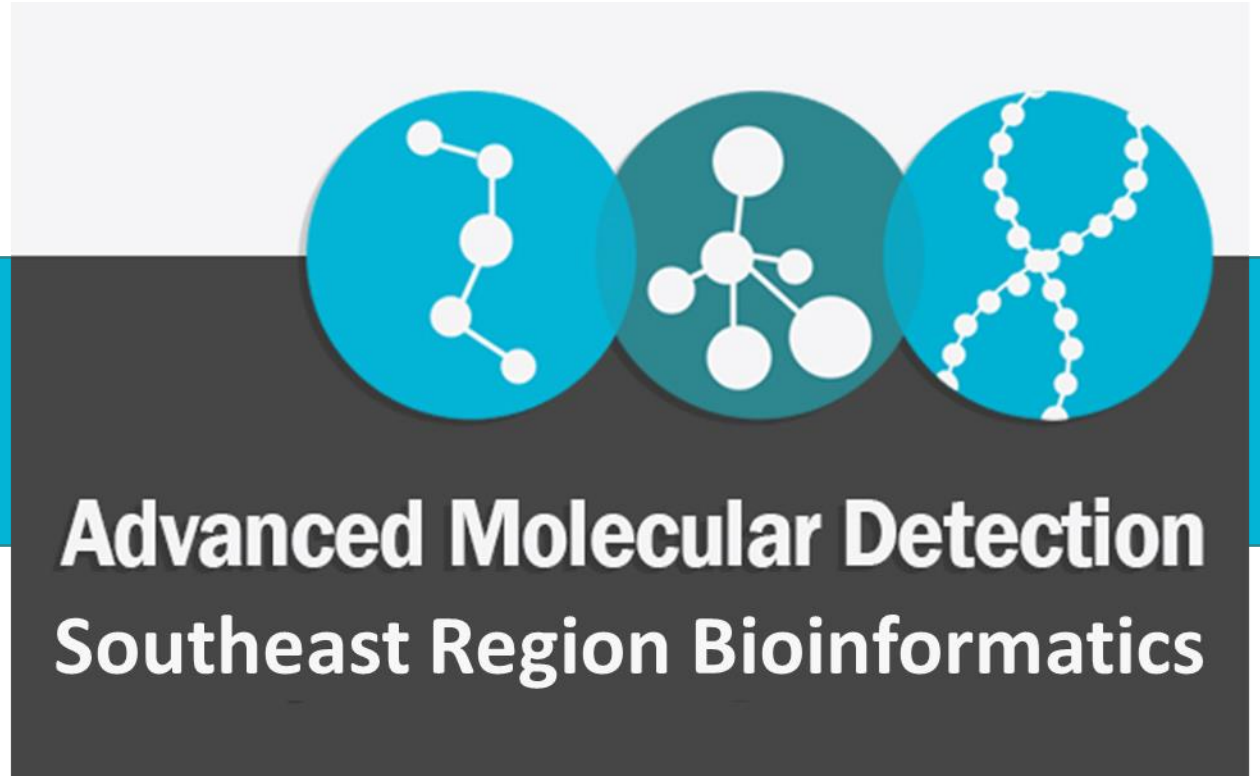


AMD Southeast Region Genomic Epidemiology Training

Gen Epi Tools

3/18/2024

Genomic Epidemiology Visualization Tools

1. IQ-Tree
2. NextStrain
3. MicrobeTrace
4. Microreact
5. CZ Gen Epi
6. General Recommendations

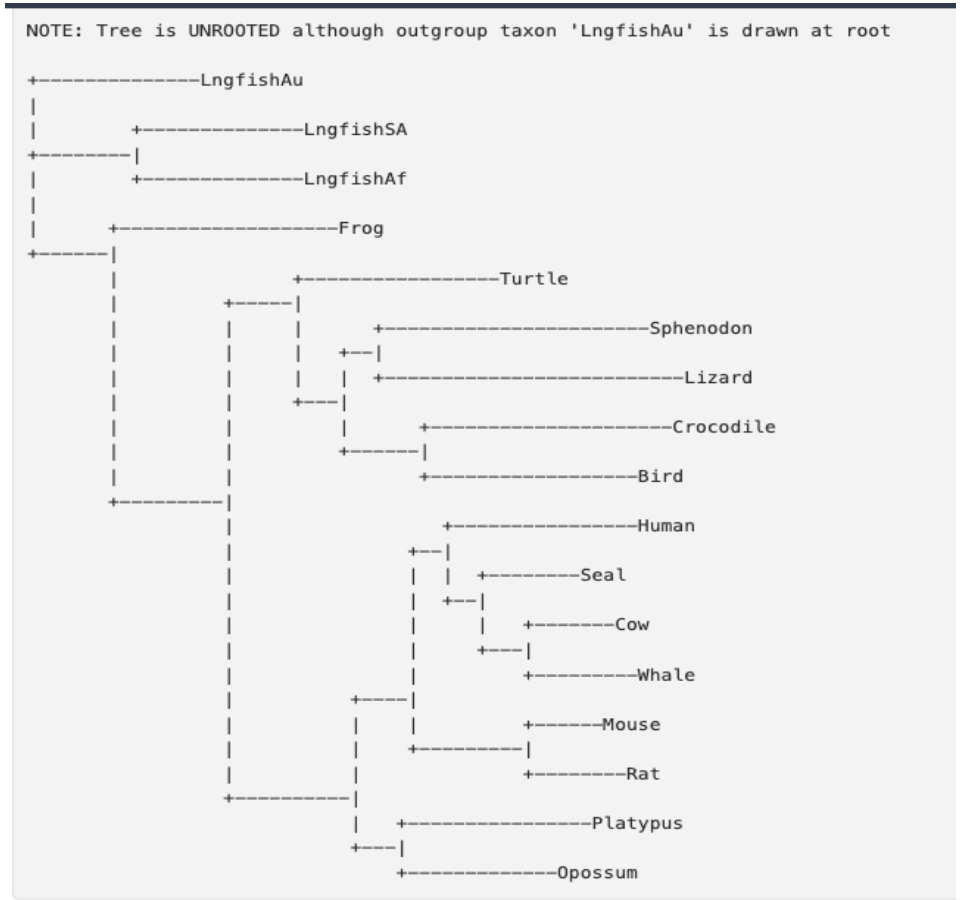
IQ-Tree

- Creates trees by Maximum Likelihood method
  - Popular method
  - Can be slower than other Maximum Parsimony
    - Faster than RAxML and PhyML
  - Looks for the highest probability of relationship between samples
  - Estimates unknown parameters of a probability model
    - Parameters like rate of transmission, rate of mutation, tree construction
- Included in the Cecret Pipeline (for SARS-CoV-2)

Advanced Molecular Detection
Southeast Region Bioinformatics

IQ-Tree

- Open-Source
- Requires access to command line
  - Like AWS or a Linux environment
- Documentation at http://www.iqtree.org/
  - Beginner's Tutorial http://www.iqtree.org/doc/Tutorial
- Not visually appealing so we output to ggtree2 in R



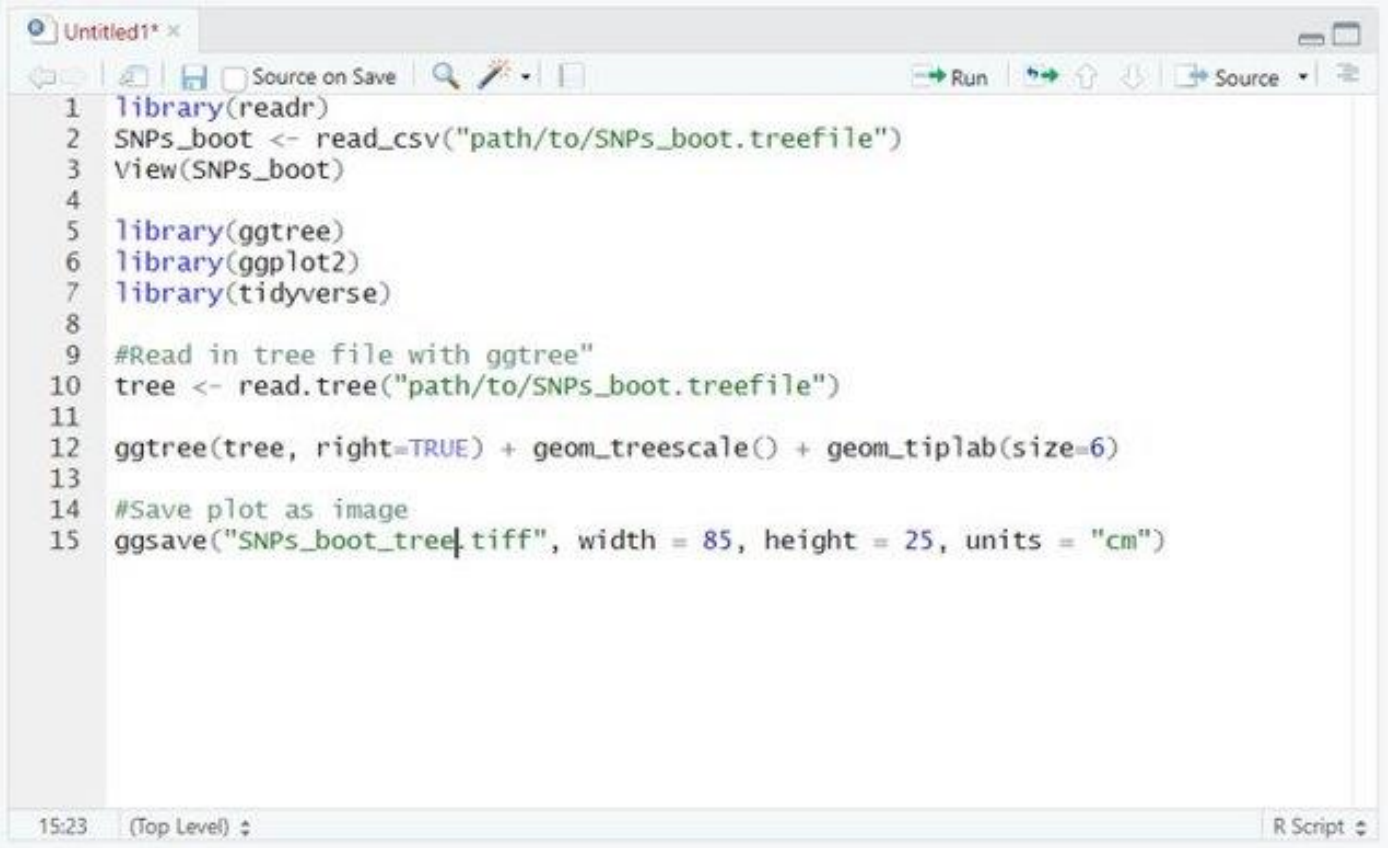
Advanced Molecular Detection
Southeast Region Bioinformatics

IQ-Tree Beginner's Tutorial

- You can use your own files or copy the small files provided
  - If you use your own files they need to be aligned (I suggest MAFFT)
  - If you do not have access to a Linux environment you can use a Jupyter Notebook.
    - https://jupyter.org/try-jupyter/notebooks/?path=notebooks/Intro.ipynb
    - Note: don't use PHI in the online version

Advanced Molecular Detection
Southeast Region Bioinformatics

IQ-Tree



If you have a Newick (tree) file and want to change the branch length to time, use the following code

iqtree -s ALN_FILE --date DATE_FILE -te TREE_FILE

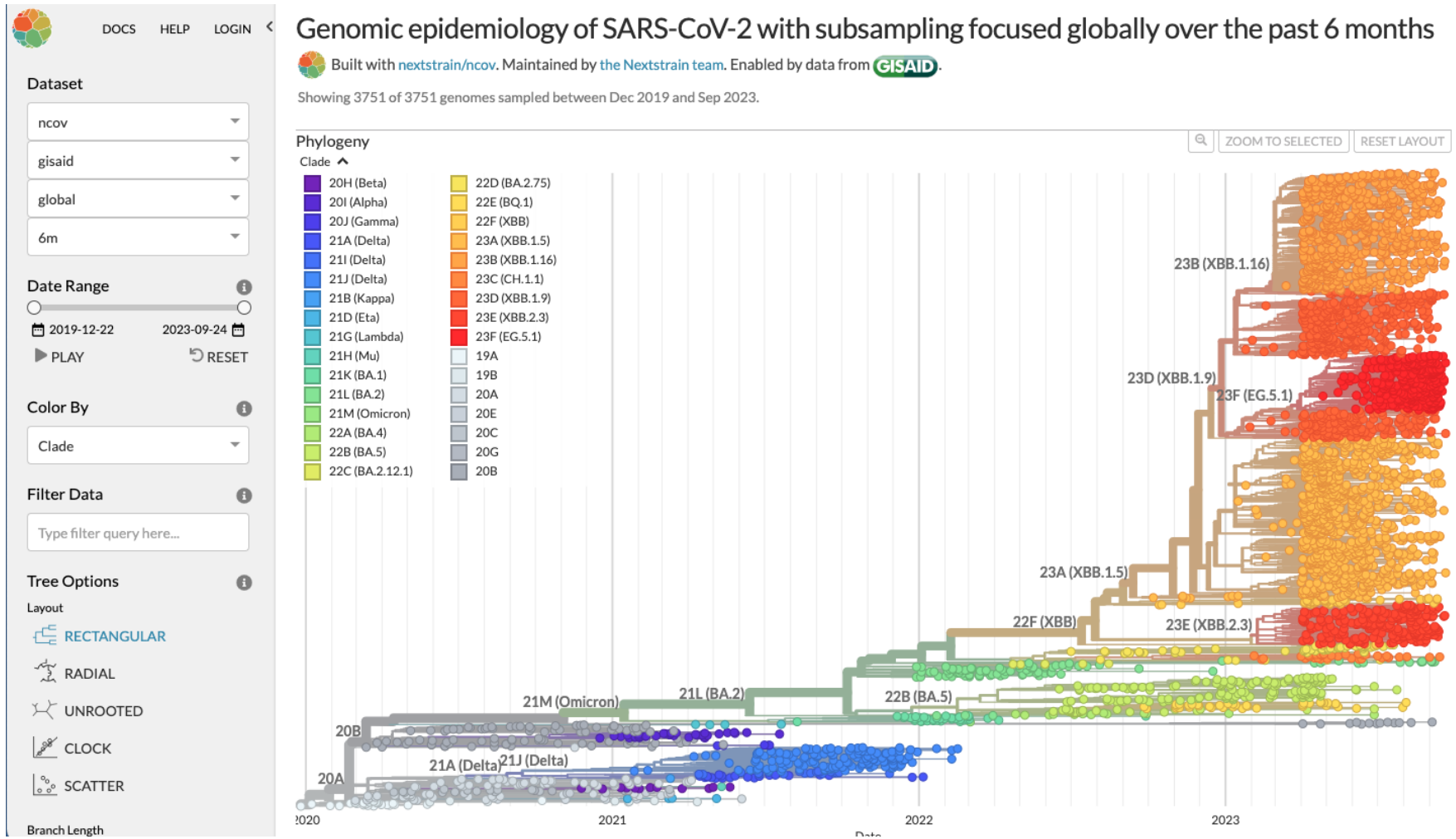
Advanced Molecular Detection
Southeast Region Bioinformatics

Nextstrain

- Pulls sample data from GISAID (worldwide repository for Covid and Flu data) to create worldwide phylogenetic trees
- Available for Influenza, Covid, MPox, Ebola, Enterovirus D68, Measels, Mumps, RSV, TB, West Nile, and Zika
- Options to create trees of various sized data sets and regions
  - Color coding options
  - Tree layout options
  - Many customization choices!
- Web-based, no advanced computing required (nextstrain.org)

Advanced Molecular Detection
Southeast Region Bioinformatics

Nextstrain



Advanced Molecular Detection
Southeast Region Bioinformatics

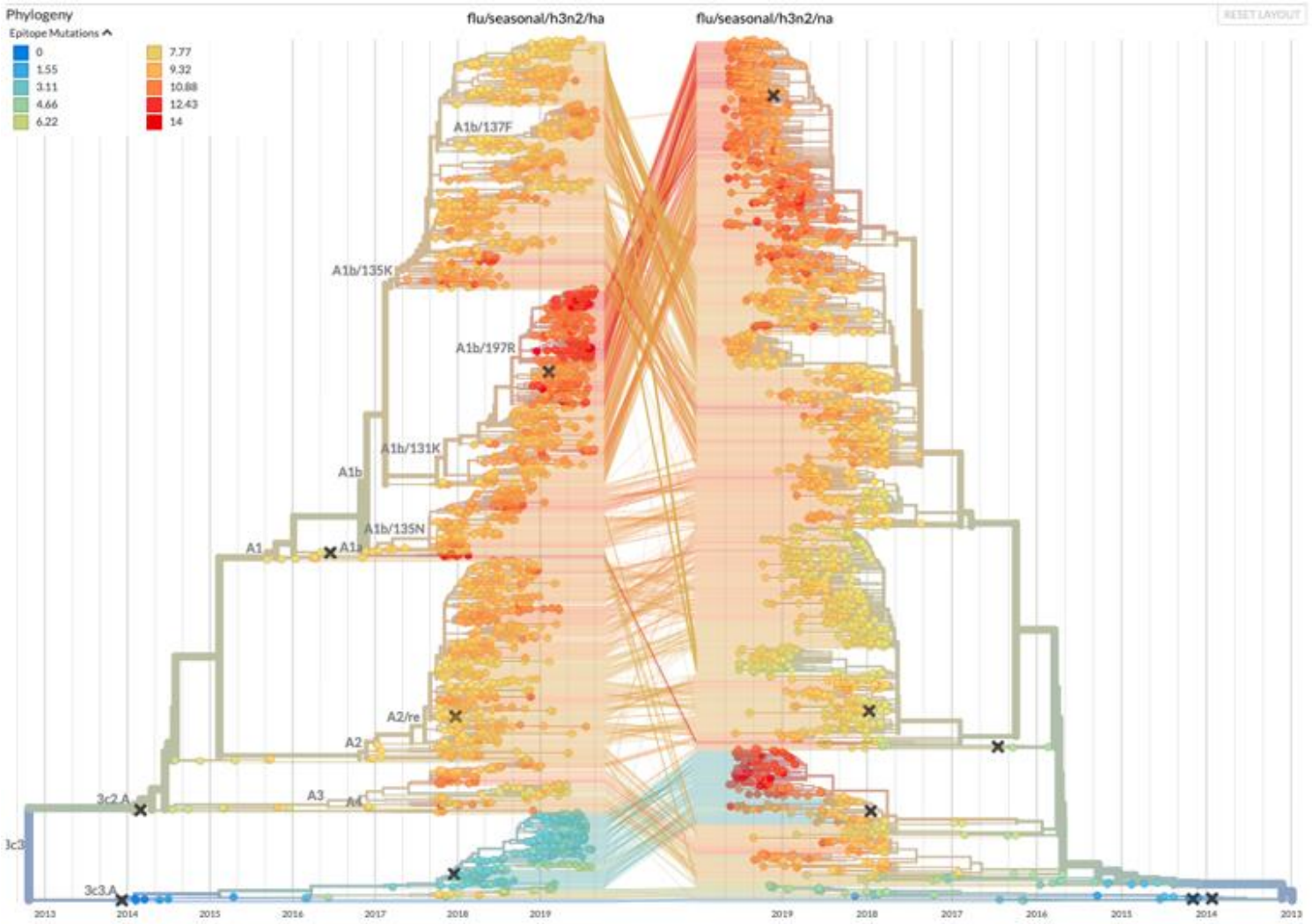
Nextstrain

- Using Augur and Auspice, you can create local builds (tree and everything that goes with it) using your own computing resources
  - Private
  - Can incorporate PHI/PII
  - Can incorporate additional metadata
- Local builds require open-source access and a very large amount of computing resources (larger than would normally be available)
  - Subsampling reduces the computing resources required
- Produces a shareable .JSON file

Advanced Molecular Detection
Southeast Region Bioinformatics

Nextstrain

- With Auspice you can compare different trees directly
- Figure right, compares HA and NA mutation in H3N2 in the same flu season



Advanced Molecular Detection
Southeast Region Bioinformatics

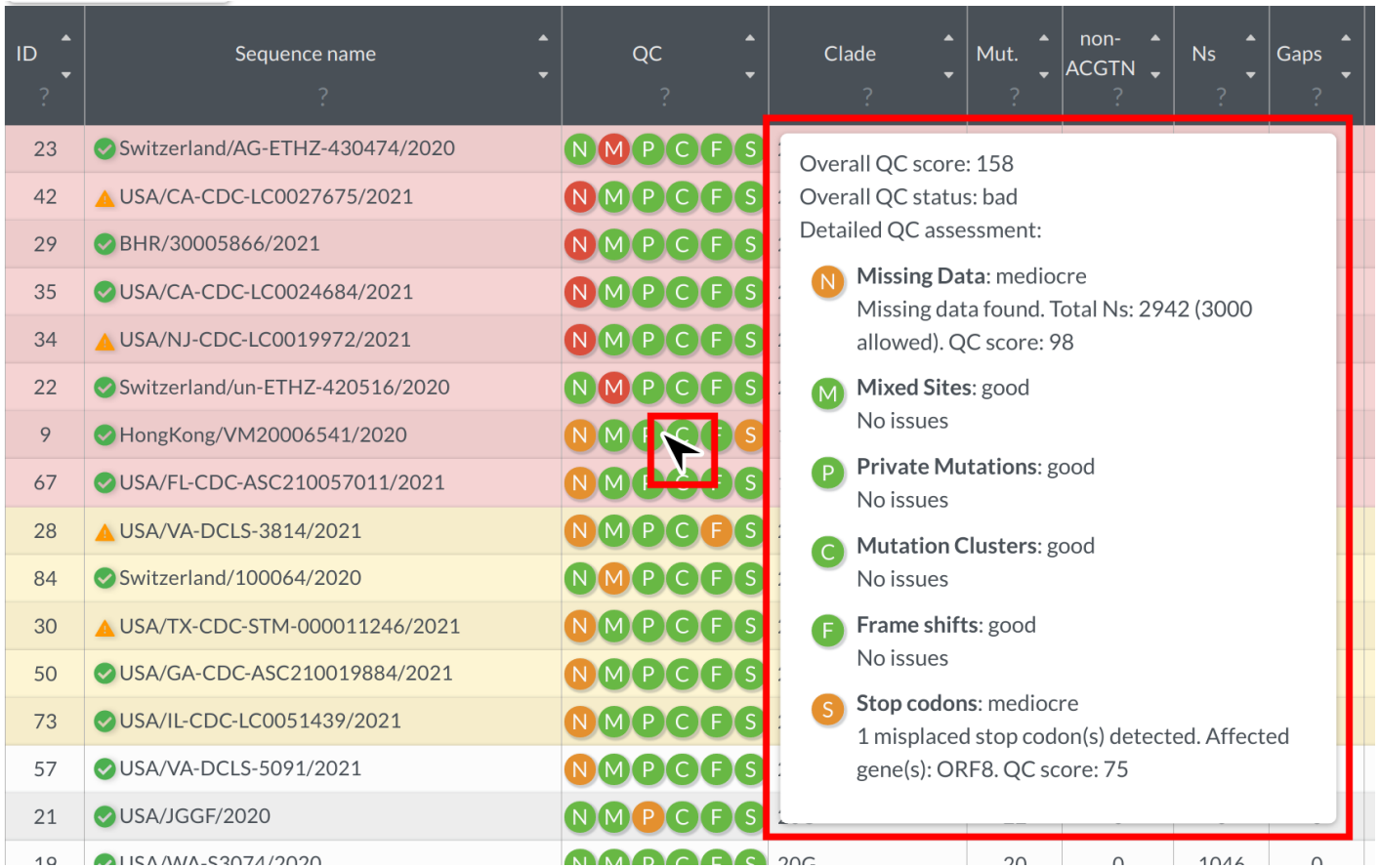
Nextclade

- Function within NextStrain
  - Available web-based or in CLI (docker container)
    - Web-Based: clades.nextstrain.org
    - CLI: https://github.com/nextstrain/nextclade
  - Start with web-based, very user friendly
  - CLI allows for more advanced options
- Web-based runs on your computer but requires access to the internet
- Built to be quicker than NextStrain, great for small data sets
- Limitation: Samples are placed one by one on a phylogenetic tree, no internal nodes (last common ancestor) will be detected

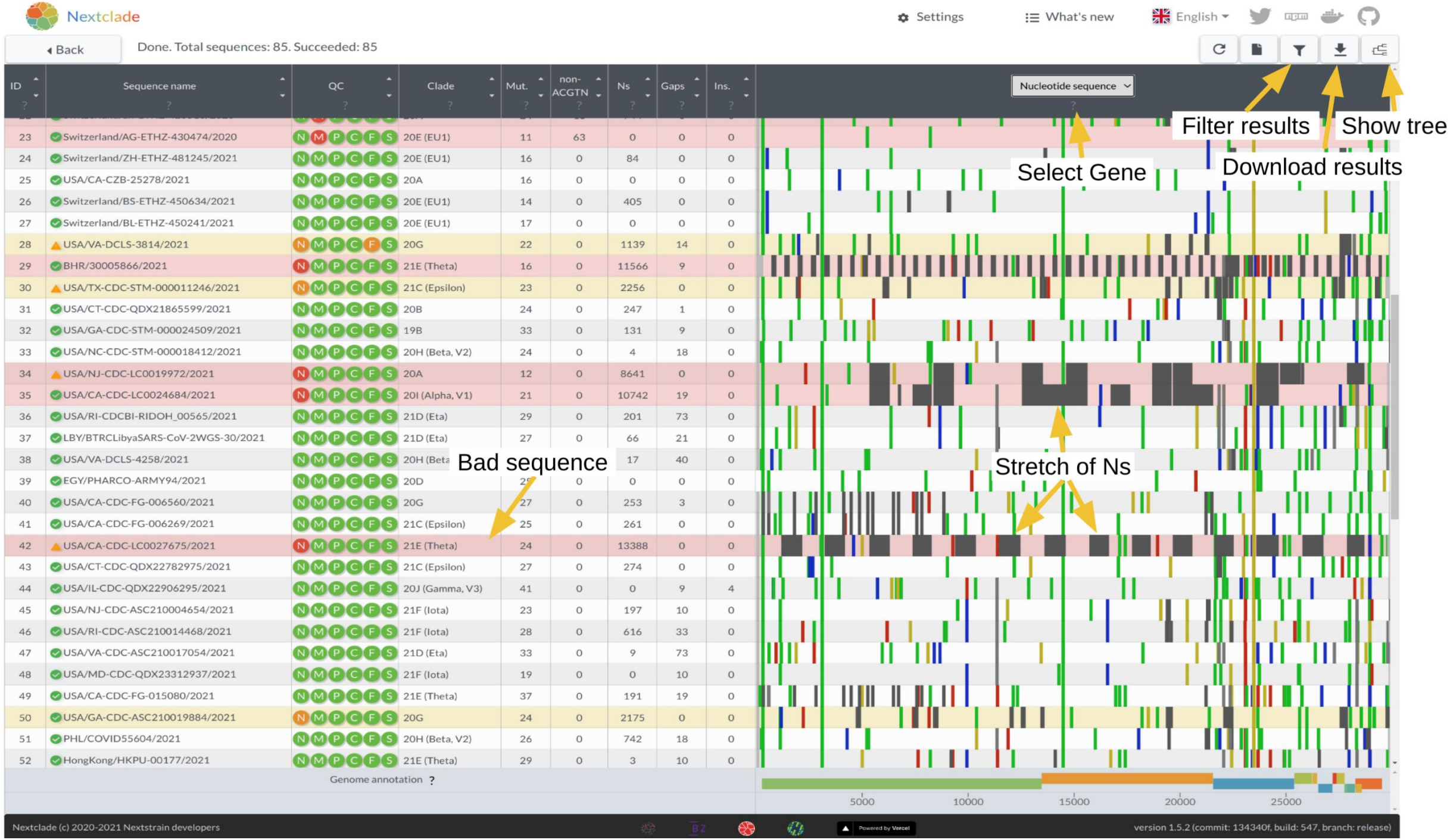
Advanced Molecular Detection
Southeast Region Bioinformatics

Nextclade

- Takes .fasta from raw, through alignment, calls mutations, and determines clade
- Results available as table and phylogenetic tree



Advanced Molecular Detection
Southeast Region Bioinformatics



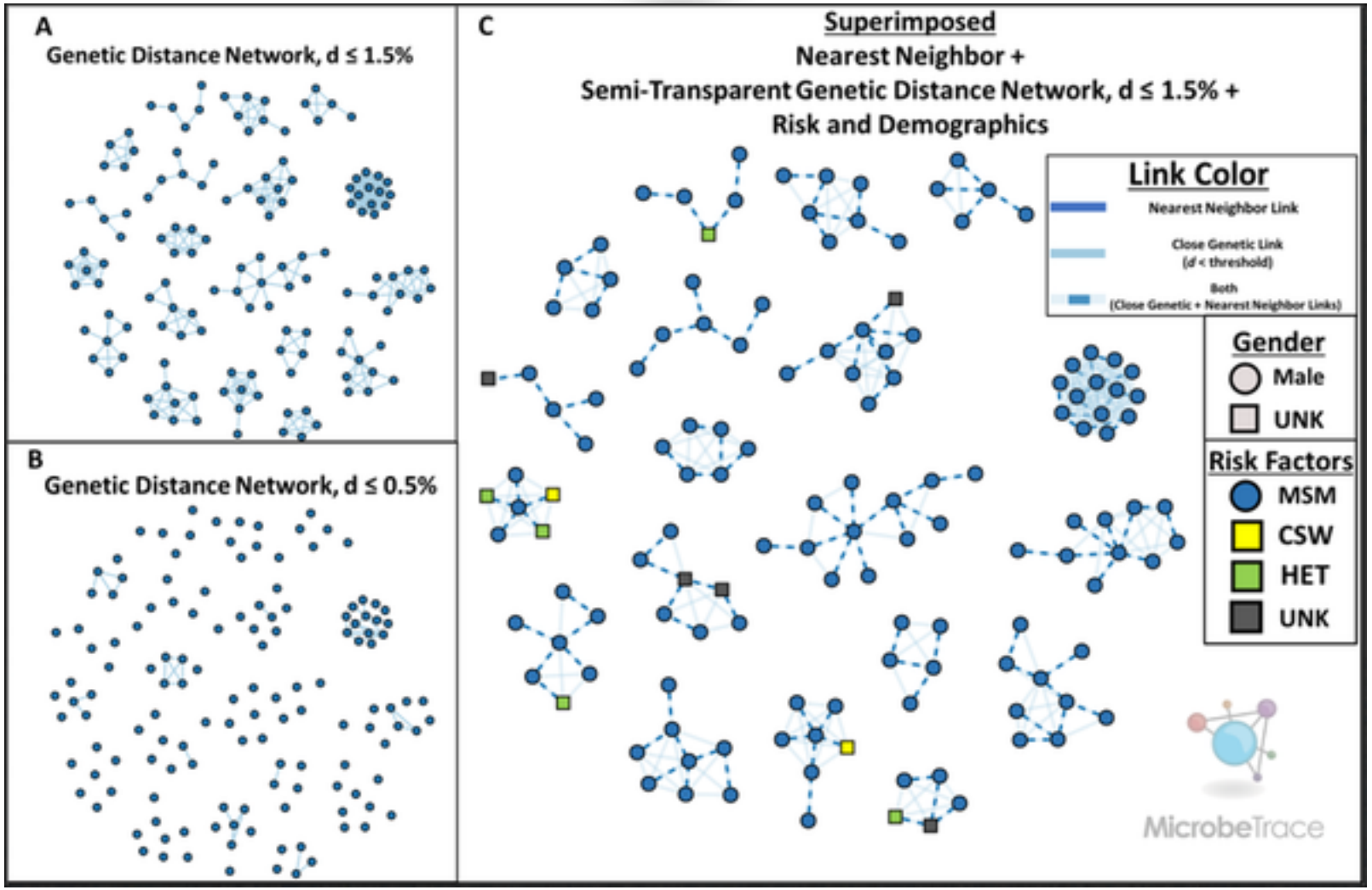
ID	Sequence name	QC	Clade	Mut.	non-ACGTN	Ns	Gaps	Ins.
23	Switzerland/AG-ETHZ-430474/2020	N M P C F S	20E (EU1)	11	63	0	0	0
24	Switzerland/ZH-ETHZ-481245/2021	N M P C F S	20E (EU1)	16	0	84	0	0
25	USA/CA-CZB-25278/2021	N M P C F S	20A	16	0	0	0	0
26	Switzerland/BS-ETHZ-450634/2021	N M P C F S	20E (EU1)	14	0	405	0	0
27	Switzerland/BL-ETHZ-450241/2021	N M P C F S	20E (EU1)	17	0	0	0	0
28	USA/VA-DCLS-3814/2021	N M P C F S	20G	22	0	1139	14	0
29	BHR/30005866/2021	N M P C F S	21E (Theta)	16	0	11566	9	0
30	USA/TX-CDC-STM-000011246/2021	N M P C F S	21C (Epsilon)	23	0	2256	0	0
31	USA/CT-CDC-QDX21865599/2021	N M P C F S	20B	24	0	247	1	0
32	USA/GA-CDC-STM-000024509/2021	N M P C F S	19B	33	0	131	9	0
33	USA/NC-CDC-STM-000018412/2021	N M P C F S	20H (Beta, V2)	24	0	4	18	0
34	USA/NJ-CDC-LC0019972/2021	N M P C F S	20A	12	0	8641	0	0
35	USA/CA-CDC-LC0024684/2021	N M P C F S	20I (Alpha, V1)	21	0	10742	19	0
36	USA/RI-CDCBI-RIDOH_00565/2021	N M P C F S	21D (Eta)	29	0	201	73	0
37	LBY/BTRCLibyaSARS-CoV-2WGS-30/2021	N M P C F S	21D (Eta)	27	0	66	21	0
38	USA/VA-DCLS-4258/2021	N M P C F S	20H (Beta	17			40	0
39	EGY/PHARCO-ARMY94/2021	N M P C F S	20D	25	0	0	0	0
40	USA/CA-CDC-FG-006560/2021	N M P C F S	20G	27	0	253	3	0
41	USA/CA-CDC-FG-006269/2021	N M P C F S	21C (Epsilon)	25	0	261	0	0
42	USA/CA-CDC-LC0027675/2021	N M P C F S	21E (Theta)	24	0	13388	0	0
43	USA/CT-CDC-QDX22782975/2021	N M P C F S	21C (Epsilon)	27	0	274	0	0
44	USA/IL-CDC-QDX22906295/2021	N M P C F S	20J (Gamma, V3)	41	0	0	9	4
45	USA/NJ-CDC-ASC210004654/2021	N M P C F S	21F (Iota)	23	0	197	10	0
46	USA/RI-CDC-ASC210014468/2021	N M P C F S	21F (Iota)	28	0	616	33	0
47	USA/VA-CDC-ASC210017054/2021	N M P C F S	21D (Eta)	33	0	9	73	0
48	USA/MD-CDC-QDX23312937/2021	N M P C F S	21F (Iota)	19	0	0	10	0
49	USA/CA-CDC-FG-015080/2021	N M P C F S	21E (Theta)	37	0	191	19	0
50	USA/GA-CDC-ASC210019884/2021	N M P C F S	20G	24	0	2175	0	0
51	PHL/COVID55604/2021	N M P C F S	20H (Beta, V2)	26	0	742	18	0
52	HongKong/HKPU-00177/2021	N M P C F S	21E (Theta)	29	0	3	10	0

MicrobeTrace

- Web-browser based, but runs locally on your computer
  - Can turn internet off and it will still run, the data stays on your computer
- Creates a network (web) of cases using a combination of pathogen sequences and epi metadata
- Can help identify pathogen and transmission hotspots
- Originally made for HIV transmission but has been used for other diseases
- https://microbetrace.cdc.gov/MicrobeTrace/
- Tends to be buggy

Advanced Molecular Detection
Southeast Region Bioinformatics

MicrobeTrace



Campbell EM, Boyles A, Shankar A, Kim J, Knyazev S, et al. (2021) MicrobeTrace: Retooling molecular epidemiology for rapid public health response. PLOS Computational Biology 17(9): e1009300. https://doi.org/10.1371/journal.pcbi.1009300

Advanced Molecular Detection
Southeast Region Bioinformatics

MicrobeTrace

MicrobeTrace™

Common Data Types and Combinations

- DNA Only (.fasta)
- Edge List Only (.csv)
- Distance Matrix from Genetic Data (.csv)
- DNA and Node Attribute Table (.fasta and .csv)
- Edge List and Node Attribute Table (.csv)
- Tree File (.nwk)

(Data is linked on the ID column)

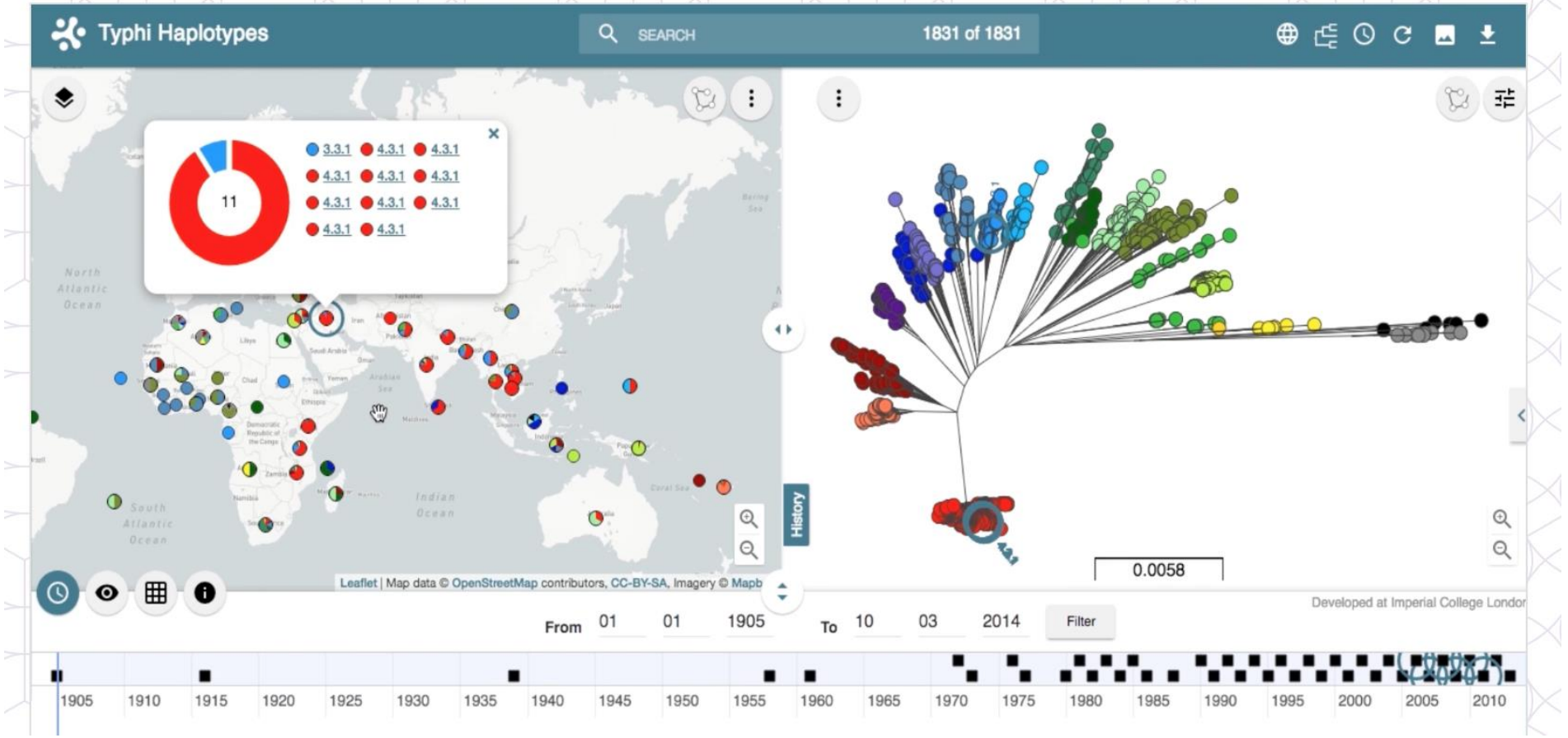
Advanced Molecular Detection
Southeast Region Bioinformatics

Microreact

- Web-based application for the interactive visualization of genetic clustering (via trees), geographic maps, and time
- Additional metadata can be attached as a table
- A few states have a desktop version of Microreact so that PHI can be included in visualizations, crazy expensive
  - Florida has funding to purchase
- Has wide customization options to fit a variety of needs

Advanced Molecular Detection
Southeast Region Bioinformatics

Microreact



Advanced Molecular Detection
Southeast Region Bioinformatics

CZ GEN EPI

- Web-based and open source application
  - There is a waitlist to join, still usually quick acceptance
  - Only for SAR—CoV-2 and Mpox
- Combines NCBI Virus, NextStrain, NextClade, Pangolin, and UShER into one pipeline
- Not extensively used
- Does have a good library of Gen Epi learning resources
  - https://help.czgenepi.org/hc/en-us/categories/6217716150804-Genomic-Epidemiology-Learning-Center

Advanced Molecular Detection
Southeast Region Bioinformatics

CZ GEN EPI

Feature	UShER	Nextstrain
Phylogenetic placement approach	✓	
Ultrafast (done within minutes)	✓	
Provides placement confidence metrics	✓	
Focuses on subtrees with closely related samples	✓	
Phylogenetic tree building (from scratch)		✓
Takes up to 12 hours		✓
Tree displays samples of interest and contextual data		✓
Tree displays samples over time		✓
Tree can be viewed in Nextstrain (interactive visualization program)	✓	✓
Nextstrain visualization incorporates genetic, temporal, and spatial data		✓

Upload consensus genomes

Easily upload sample data into a secure, cloud-based platform.

Lineage assignments

Automated assignment of virus lineages through Pangolin or Nextclade as soon as you upload data.

Select samples for tree builds

Point-and-click, no-code interface allows you to sort, filter, search, and select samples to build phylogenetic trees.

Generate phylogenetic trees

Generate Nextstrain trees and run phylogenetic placements with UShER with a few clicks.

Overlay metadata on your trees

Explore interactive trees using Nextstrain and overlay metadata for a comprehensive view of where and how transmission is occurring.

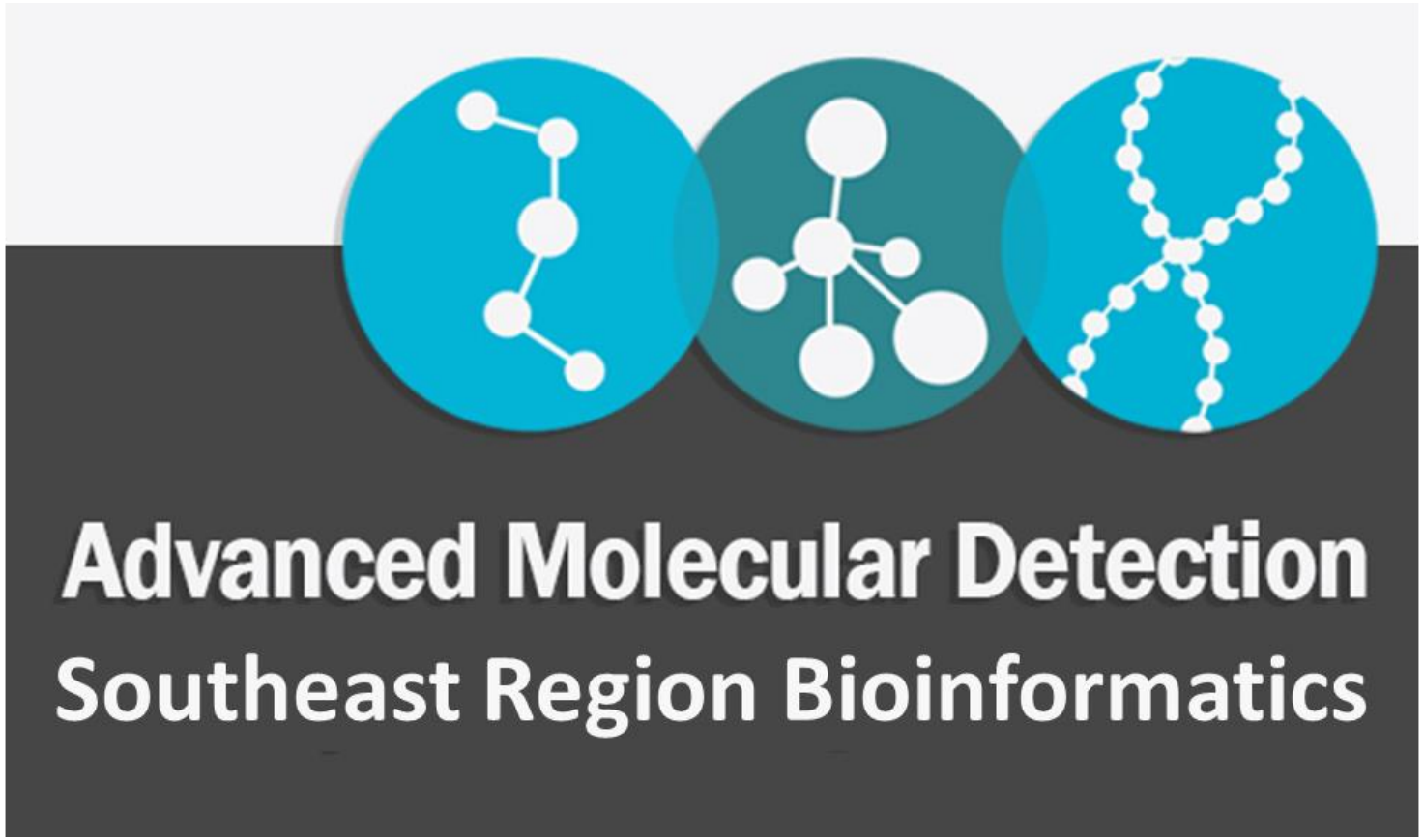
Advanced Molecular Detection
Southeast Region Bioinformatics

https://help.czgenepi.org/hc/en-us/sections/5686512995988-Get-Started

General Recommendations

1. Use large graphs
2. Consider color blind-friendly color palettes
3. Don’t be afraid to ask your bioinformaticians for help

Advanced Molecular Detection
Southeast Region Bioinformatics



Questions?

Bphl-sebioinformatics@flhealth.gov

TBD
Lead Bioinformatician & Supervisor
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