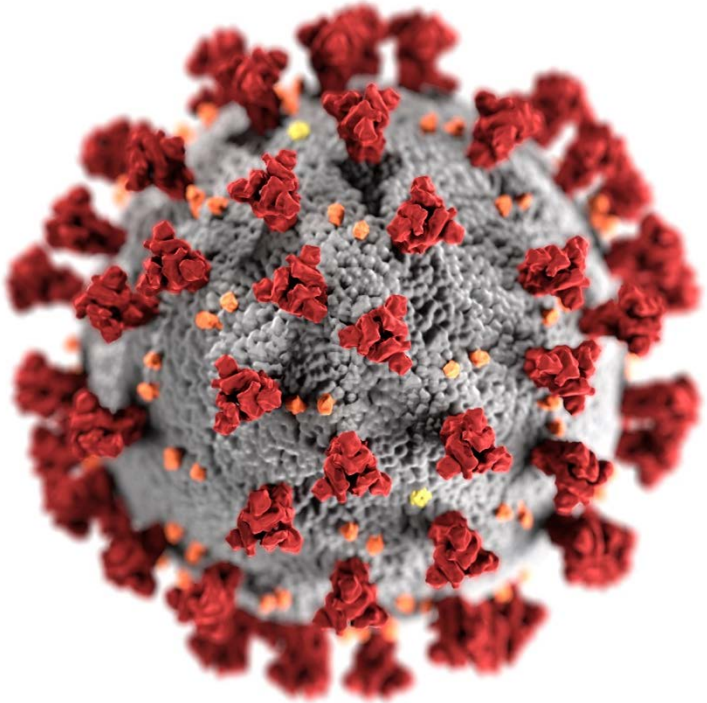
Real-time phylogenetics with UShER

COVID-19 Genomic Epidemiology Toolkit: Module 3.3



Russ Corbett-Detig, PhD
Assistant Professor
Department of Biomolecular Engineering
University of California, Santa Cruz

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

1.4 Emerging variants of SARS-CoV-2

Part 2: Case Studies

2.1 SARS-CoV-2 sequencing in Arizona

2.2 Healthcare cluster transmission

2.3 Community transmission

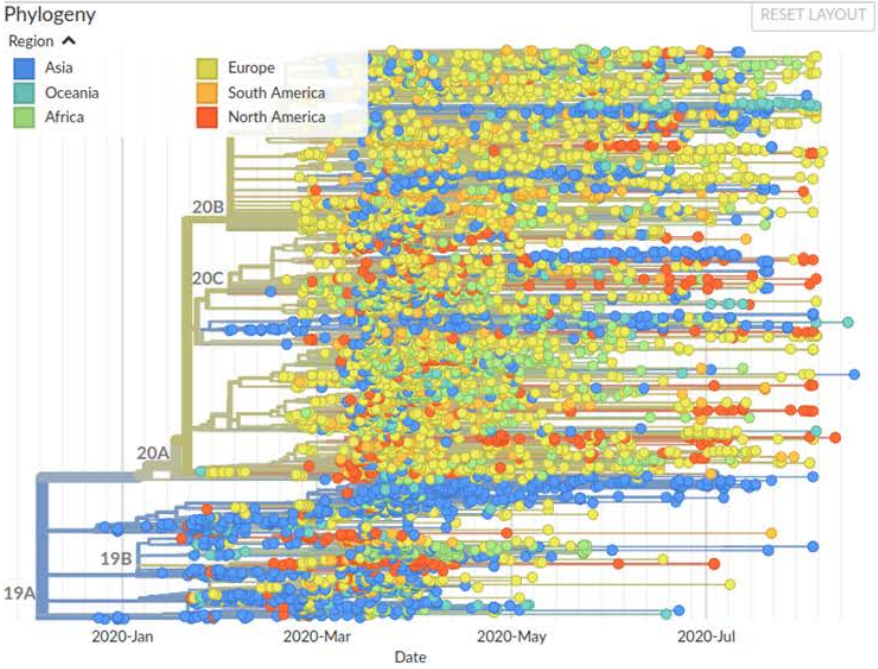
Part 3: Implementation

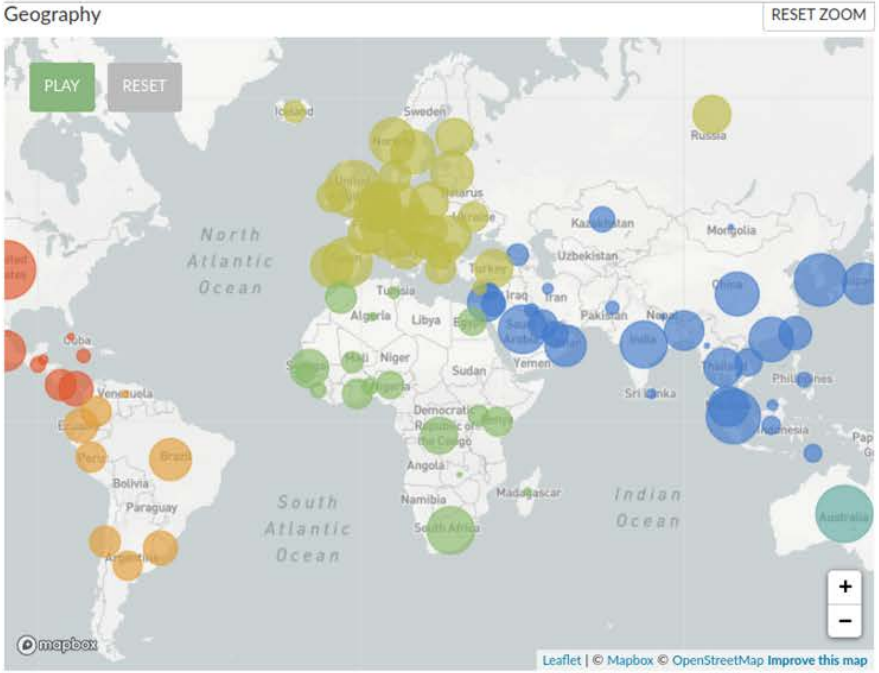
3.1 Getting started with Nextstrain

3.2 Getting started with MicrobeTrace

3.3 Linking epidemiologic data

Tracking viral evolution





Images from nextstrain.org

UShER: Real-time phylogenetic placement

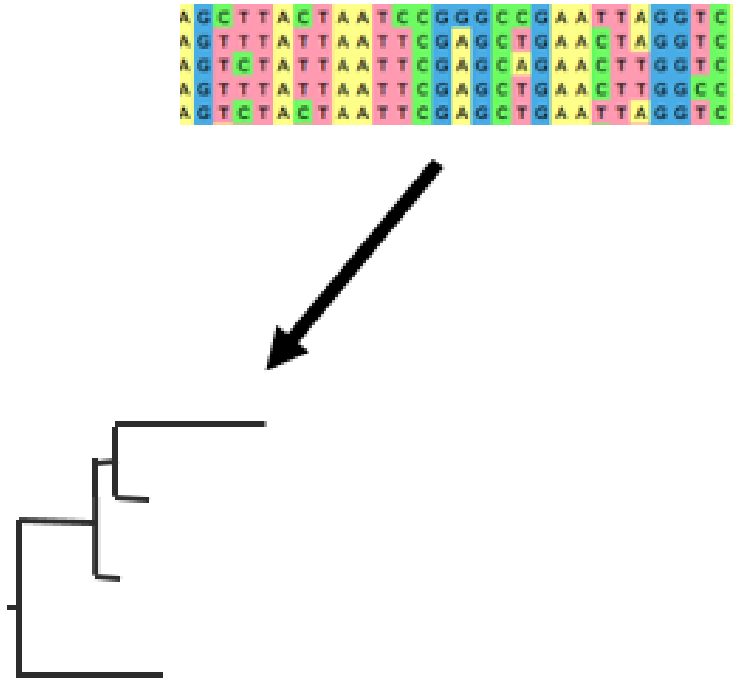
- **U**ltrafast **S**ample placement on **E**xisting t**R**ees

- Designed to take user sequences and
  1. Accurately place them onto global phylogeny
  2. Construct new subtrees
  3. Enable easy visualization

- Runs quickly (<1 second) to facilitate genomic epidemiology

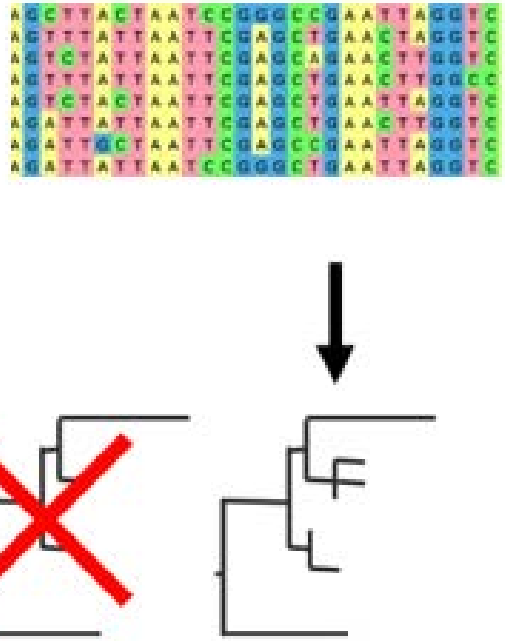
Constant flow and huge datasets overwhelm typical phylogenetics approaches

- Typical phylogenetic workflow:
  1. Gather data
  2. Calculate tree



Constant flow and huge datasets overwhelm typical phylogenetics approaches

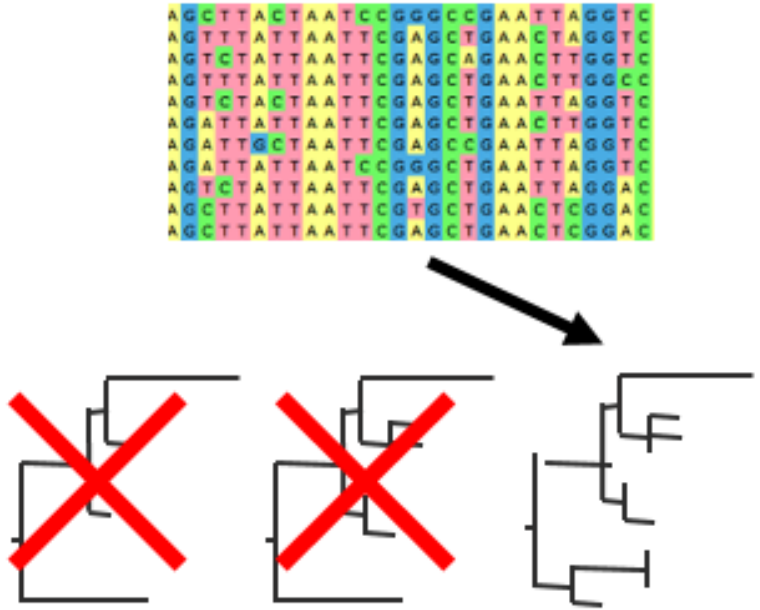
- Typical phylogenetic workflow:
  1. Gather data
  2. Calculate tree
  3. **More data!**
  4. **Recalculate tree?**



Constant flow and huge datasets overwhelm typical phylogenetics approaches

- Typical phylogenetic workflow:
  1. Gather data
  2. Calculate tree
  3. More data
  4. Recalculate tree
  5. **More data!**
  6. **Recalculate tree?**

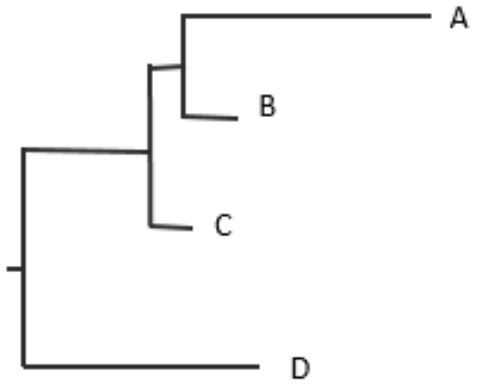
  Repeat… forever



The UShER approach for phylogenetics

UShER takes as input:

1. phylogenetic tree
2. list of mutations in each sample

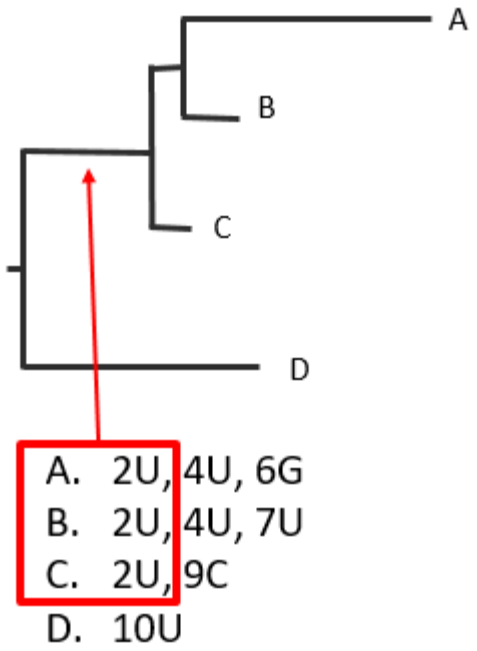


Sample A	A	U	C	U	U	G	A	G	U	C
Sample B	A	U	C	U	U	A	U	G	U	C
Sample C	A	U	C	G	U	A	A	G	C	C
Sample D	A	C	C	G	U	A	A	G	U	U

A. 2U, 4U, 6G
B. 2U, 4U, 7U
C. 2U, 9C
D. 10U

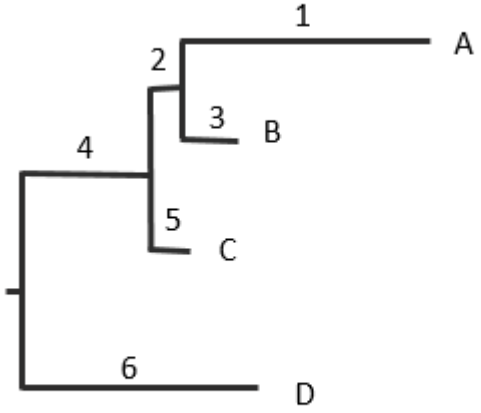
The UShER approach for phylogenetics

Using parsimony, UShER maps mutations onto the existing tree.



The UShER approach for phylogenetics

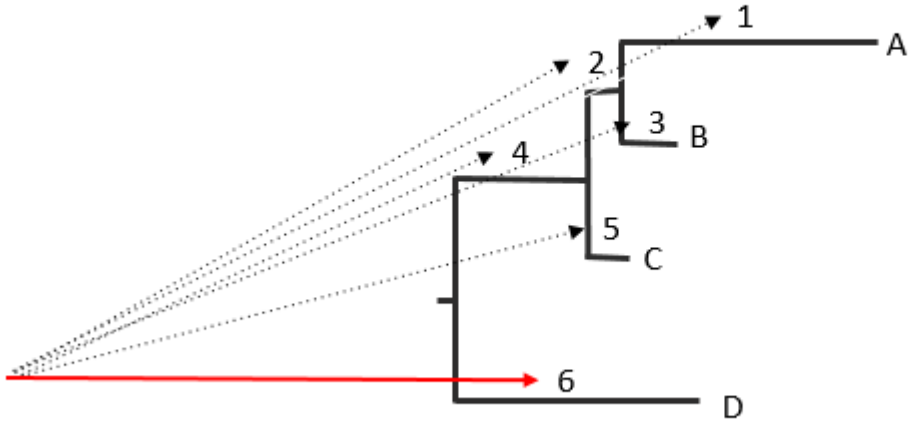
UShER stores this **mutation annotated tree**.



1. 6G
2. 4U
3. 7U
4. 2U
5. 9C
6. 10U

The UShER approach for phylogenetics

New samples are added using maximum parsimony by checking every possible placement.



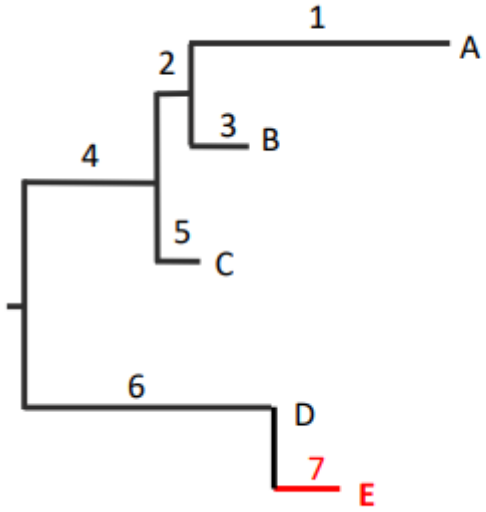
Sample E: 10U, 14G

1. 6G
2. 4U
3. 7U
4. 2U
5. 9C
6. 10U
7. 14G

UShER placements are highly accurate

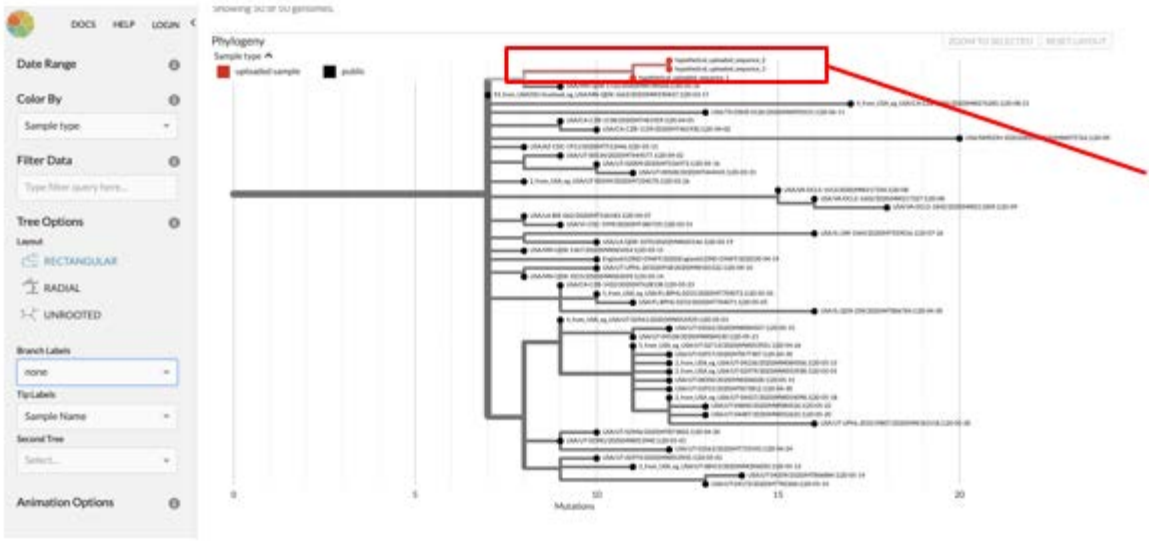
UShER finds the correct placement in 97% of the cases.

When incorrect, placements are still usually very close to the true site.



1. 6G
2. 4U
3. 7U
4. 2U
5. 9C
6. 10U
7. 14G

UShER output



UShER outputs a subtree of 50 most closely related samples to a user's sample.

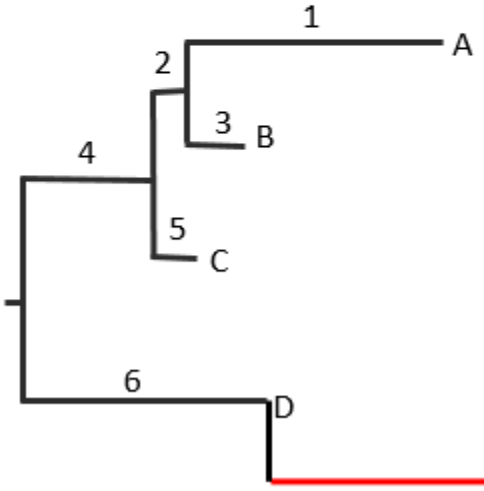
User's sample in red

This subtree can be visualized and explored using the Nextstrain platform.

UShER's quality control metrics

Fasta Sequence	Size (?)	#Ns (?)	#Mixed (?)	Bases aligned (?)	Insertions (?)	Deletions (?)	#SNVs used for placement (?)	#Masked SNVs (?)	Neighboring sample in tree (?)	Lineage of neighbor (?)	#Imputed values for mixed bases (?)	#Maximally parsimonious placements (?)	Parsimony score (?)	Subtree number (?)
hypothetical_uploaded_sequence_1	29903	0	0	29903 (?)	0	0	37 (?)	2 (?)	England/CAMC-AEAAD7/2020 | 20-10-26	B.1.5	0	2	32	1 (view in Nextstrain)

The parsimony score - Number of mutations unique to a user's sample branch.

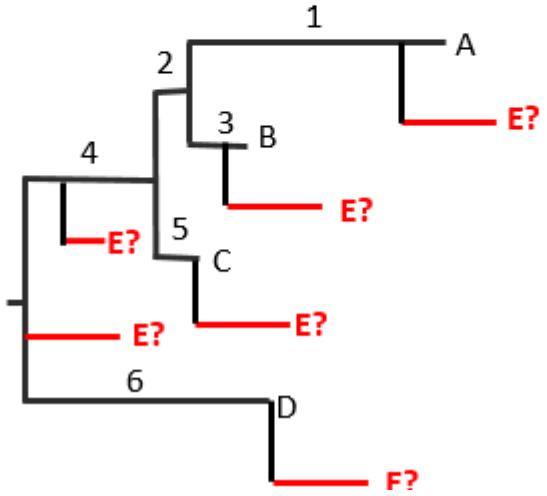


1. 6G
2. 4U
3. 7U
4. 2U
5. 9C
6. 10U
7. 14G, 22U, 24G, 33U, 38C, 49U, 51G, 55A, 56G…

UShER's quality control metrics

Fasta Sequence	Size (?)	#Ns (?)	#Mixed (?)	Bases aligned (?)	Insertions (?)	Deletions (?)	#SNVs used for placement (?)	#Masked SNVs (?)	Neighboring sample in tree (?)	Lineage of neighbor (?)	#Imputed values for mixed bases (?)	#Maximally parsimonious placements (?)	Parsimony score (?)	Subtree number (?)
hypothetical_uploaded_sequence_1	29903	0	0	29903 (?)	0	0	37 (?)	2 (?)	England/CAMC-AEAAD7/2020 | 20-10-26	B.1.5	0	2	32	1 (view in Nextstrain)

The number of equally parsimonious placements for an added sample.



1. 6G
2. 4U
3. 7U
4. 2U
5. 9C
6. 10U

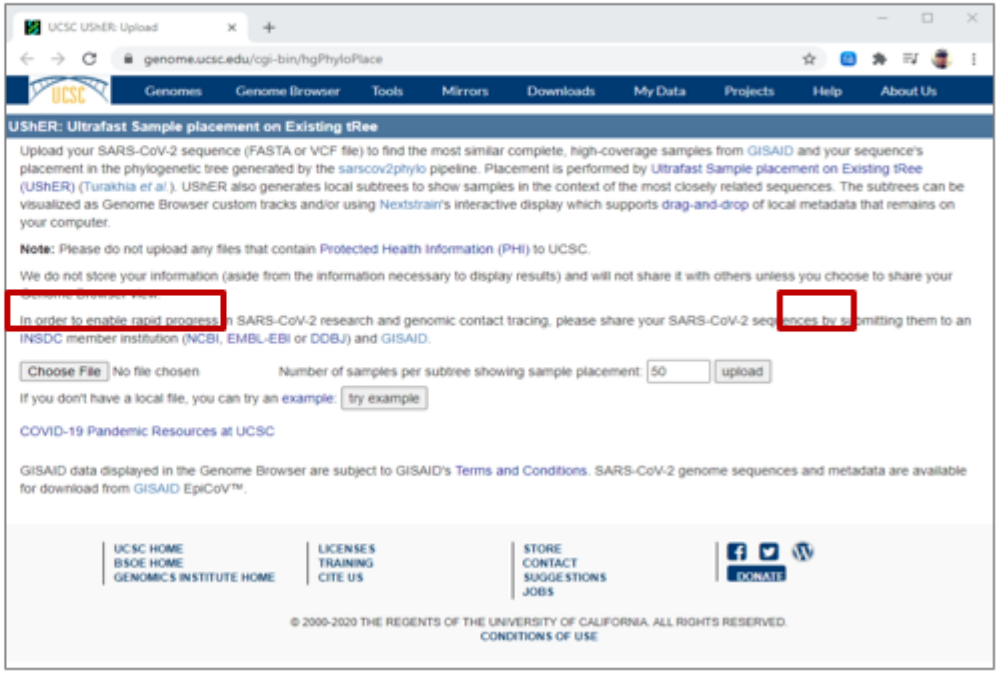
Sample E: 2N, 4N, 6N, 7N, 9N, 10N…

Uploading data

https://genome.ucsc.edu/cgi-bin/hgPhyloPlace

sequences.fasta

>genome_01
AUGAUGCAUGCUGCAUGAUG
CGUCAUGACACUGAUCG
>genome_02
AUGAUGCAUGCUGCAUGAUG
CGUCAUGACACUGAUCG
…



Summary

- UShER places samples onto a global phylogeny of SARS-CoV-2 genomes.
  - Learning about relationships among user samples, e.g., the number of unique introductions in an area.
  - Rapid sequence quality control.

- UShER resources:
  - Hands-on example data: https://github.com/russcd/USHER_DEMO
  - The UShER source code: https://github.com/yatisht/usher
  - Manuscript: https://www.biorxiv.org/content/10.1101/2020.09.26.314971v1
  - UShER's web resource: https://genome.ucsc.edu/cgi-bin/hgPhyloPlace

Acknowledgements - The UShER Team and Funding

Yatish Turakhia, UCSC

Nicola DeMaio, EBI

Angie Hinrichs, UCSC

Landen Gozashti, UCSC

Bryan Thornlow, UCSC

Ron Lanfear, ANU

David Haussler, UCSC

UNIVERSITY OF CALIFORNIA Office of the President

SCHMIDT FUTURES

UNIVERSITY OF CALIFORNIA SANTA CRUZ Genomics Institute

Alfred P. Sloan FOUNDATION

Pat & Rowland Rebele

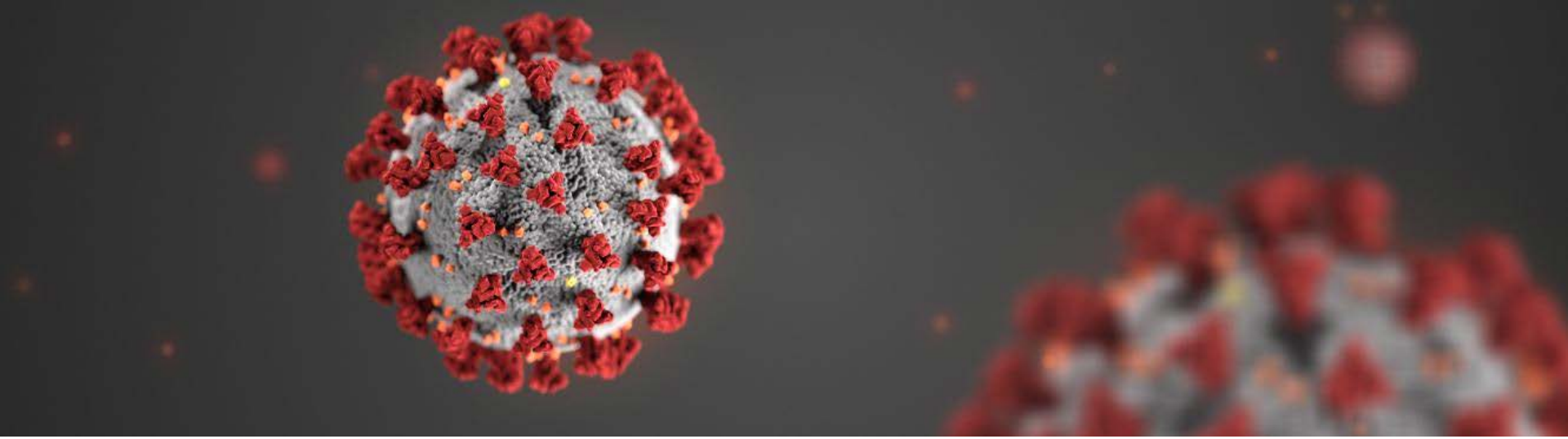
NIH

CITRIS AND THE BANATAO INSTITUTE

Learn More

- Other modules
  - Getting started with Nextstrain – Module 3.1
  - Getting started with MicrobeTrace – Module 3.2

- 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.