Pandemic-scale Phylogenetics

Yatish Turakhia

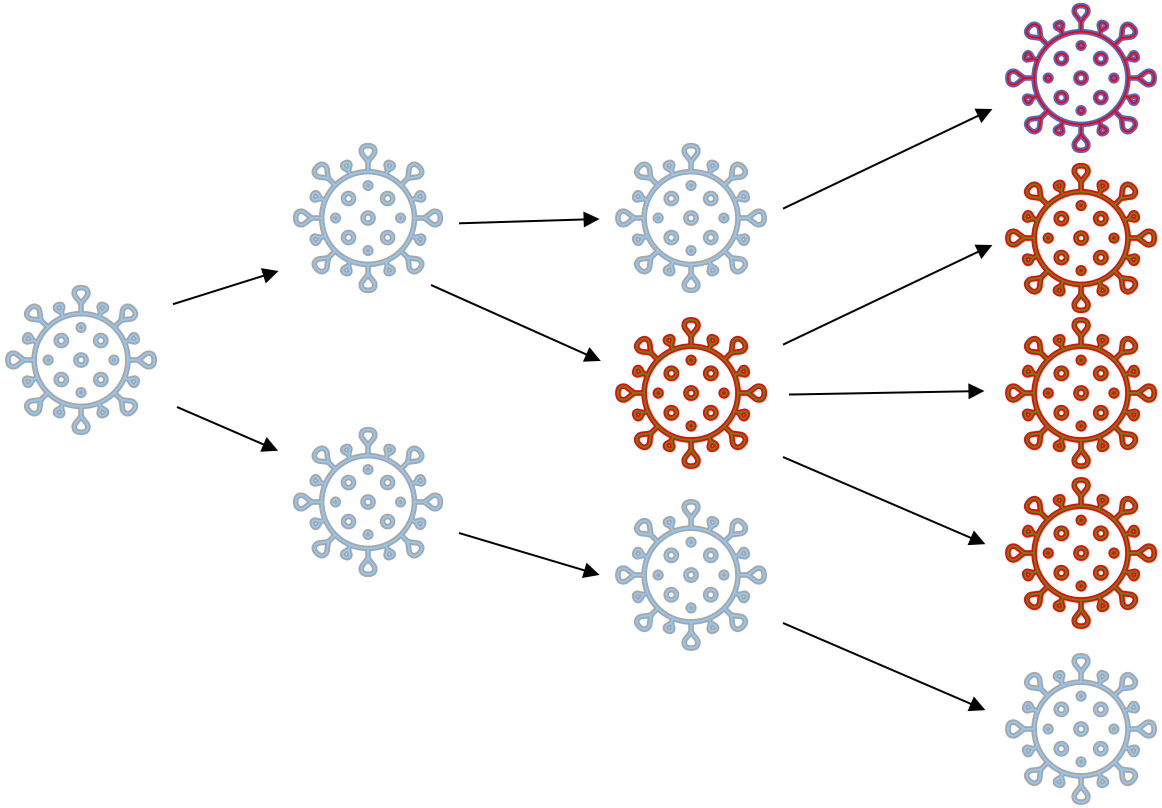
Assistant Professor, UC San Diego

COVID-19 virus (SARS-CoV-2) is constantly mutating

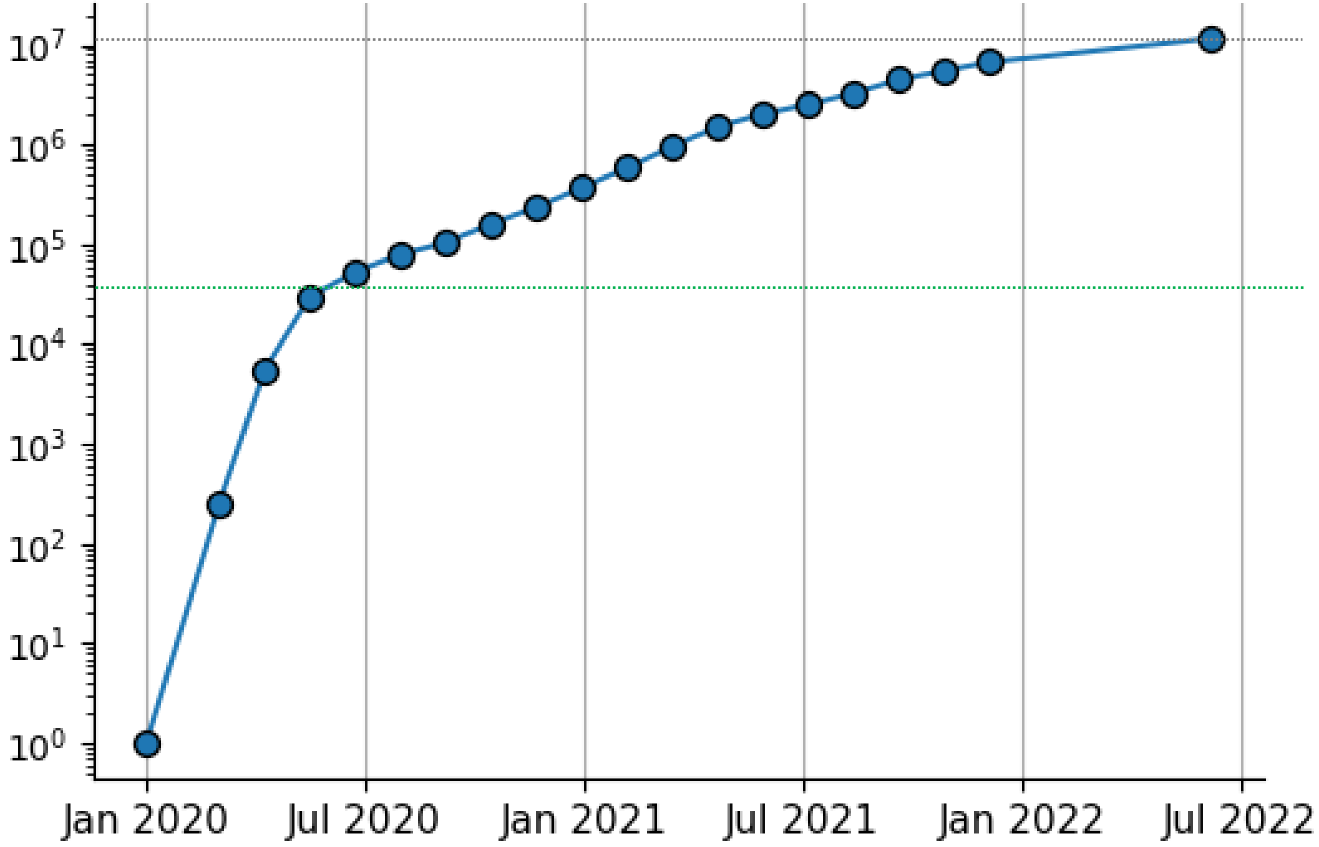
- As the COVID-19 virus (SARS-COV-2) spreads, it **mutates**

- Certain mutations render the virus more **contagious**, **virulent** or capable of **evading** the vaccines and antibody-based therapies

- Genome sequencing helps **monitor** the **viral mutations** and the **evolutionary dynamics**



Number of Global SARS-CoV-2 Genome Sequences



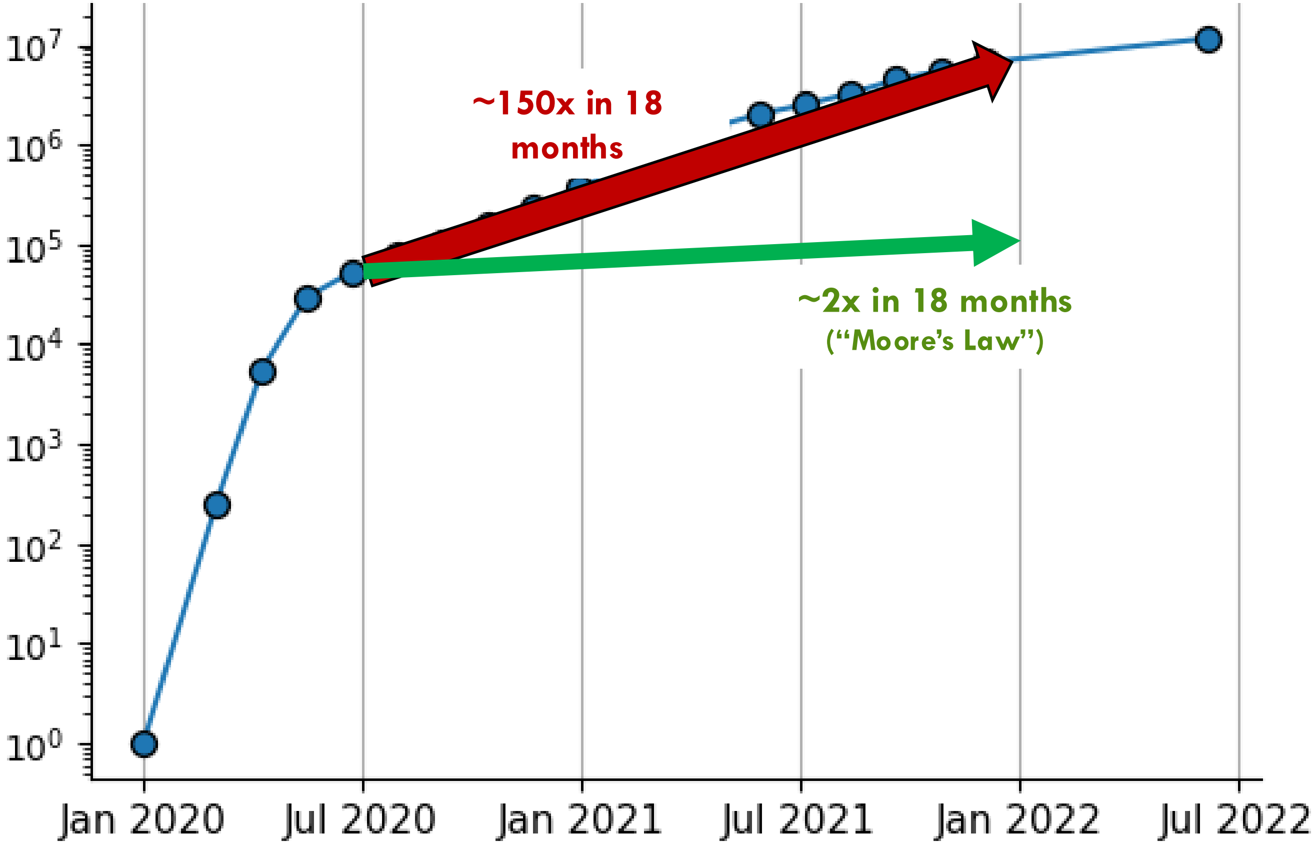
~11.6M sequences

>500x

HIV database
(~23k sequences)

(GISAID.org)

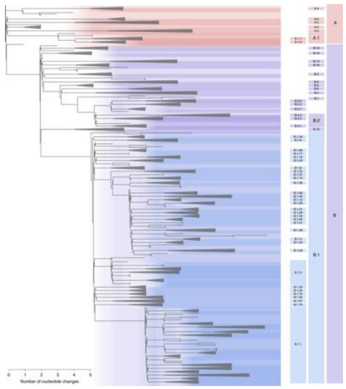
Number of Global SARS-CoV-2 Genome Sequences



~75x higher semi-logarithmic slope compared to Moore's Law

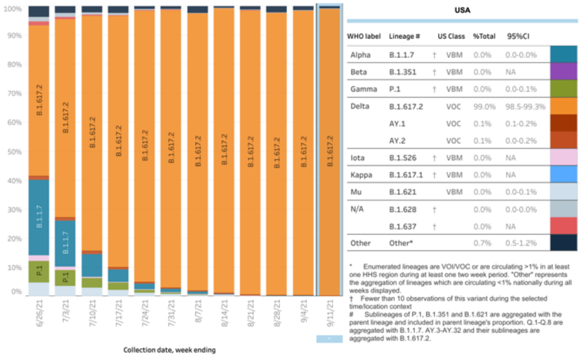
(GISAID.org)

Naming lineages



(Rambaut et al., Nat. Microbiol. 2020)

Monitoring circulating lineages



USA

WHO label	Lineage #		US Class	%Total	95%CI	
Alpha	B.1.1.7	†	VBM	0.0%	0.0-0.0%	🟦
Beta	B.1.351	†	VBM	0.0%	NA	🟪
Gamma	P.1	†	VBM	0.0%	0.0-0.1%	🟩
Delta	B.1.617.2		VOC	99.0%	98.5-99.3%	🟧
	AY.1		VOC	0.1%	0.1-0.2%	🟫
	AY.2		VOC	0.1%	0.0-0.2%	🟥
Iota	B.1.526	†	VBM	0.0%	NA	🟪
Kappa	B.1.617.1	†	VBM	0.0%	NA	🟦
Mu	B.1.621		VBM	0.0%	0.0-0.1%	⬜
N/A	B.1.628	†		0.0%	0.0-0.0%	⬜
	B.1.637	†		0.0%	NA	🟥
Other	Other*			0.7%	0.5-1.2%	⬛

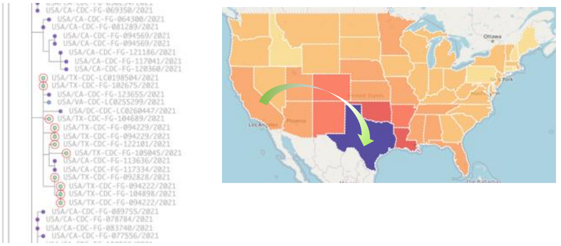
* Enumerated lineages are VOI/VOC or are circulating >1% in at least one HHS region during at least one two week period. "Other" represents the aggregation of lineages which are circulating <1% nationally during all weeks displayed.

† Fewer than 10 observations of this variant during the selected time/location context

Sublineages of P.1, B.1.351 and B.1.621 are aggregated with the parent lineage and included in parent lineage's proportion. Q.1-Q.8 are aggregated with B.1.1.7. AY.3-AY.32 and their sublineages are aggregated with B.1.617.2.

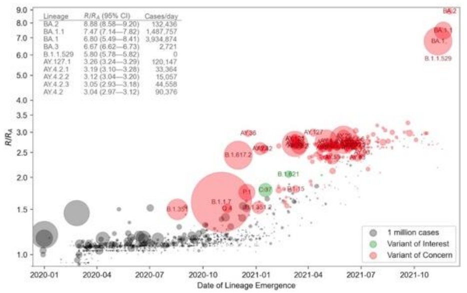
(CDC.gov dashboard 2021)

Identify newly-introduced strains



(McBroome et al., Virus Evol. 2022)

Predicting fitness of a new strain

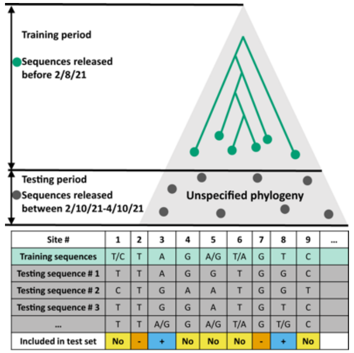


Lineage	R/R_A (95% CI)	Cases/day
BA.2	8.88 (8.58—9.20)	132,436
BA.1.1	7.47 (7.14—7.82)	1,487,757
BA.1	6.80 (5.49—8.41)	3,934,874
BA.3	6.67 (6.62—6.73)	2,721
B.1.1.529	5.80 (5.78—5.82)	0
AY.127.1	3.26 (3.24—3.29)	120,147
AY.4.2.1	3.19 (3.10—3.28)	33,364
AY.4.2.2	3.12 (3.04—3.20)	15,057
AY.4.2.3	3.05 (2.93—3.18)	44,558
AY.4.2	3.04 (2.97—3.12)	90,376

(Obermeyer et al., Science 2022)

UShER

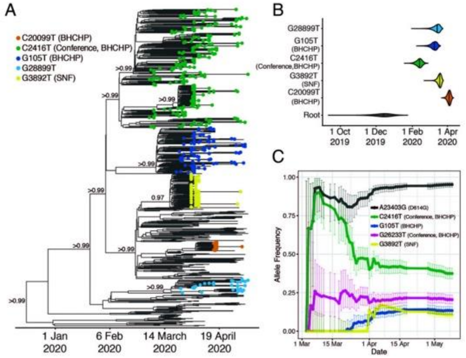
Predicting the next mutation


Training period
Sequences released before 2/8/21
Testing period
Sequences released between 2/10/21-4/10/21
Unspecified phylogeny

Site #	1	2	3	4	5	6	7	8	9	...
Training sequences	T/C	T	A	G	A/G	T/A	G	T	C	
Testing sequence # 1	T	T	A	G	G	T	G	G	C	
Testing sequence # 2	C	T	G	A	A	T	G	G	T	
Testing sequence # 3	T	T	G	G	A	T	G	T	C	
...	T	T	A/G	G	A/G	T/A	G	T/G	C	
Included in test set	No	-	+	No	No	No	-	+	No	

(Hallak et al., Nat. Comm Biol. 2022)

Analyze outbreaks and superspreader events



(Lemieux et al., Science 2021)

Wastewater surveillance

(Karthikeyan et al., Nature 2022)

Overview of the UShER Package

- **UShER:** Phylogenetic placement
  - *Turakhia et al.*, **Nature Genetics** 2021

- **matOptimize:** Phylogenetic tree optimization
  - *Ye et al.*, **Bioinformatics** 2022

- **RIPPLES:** Find recombinant sequences using a phylogenomic approach
  - *Turakhia et al.*, bioRxiv 2021 (under revision, **Nature**)

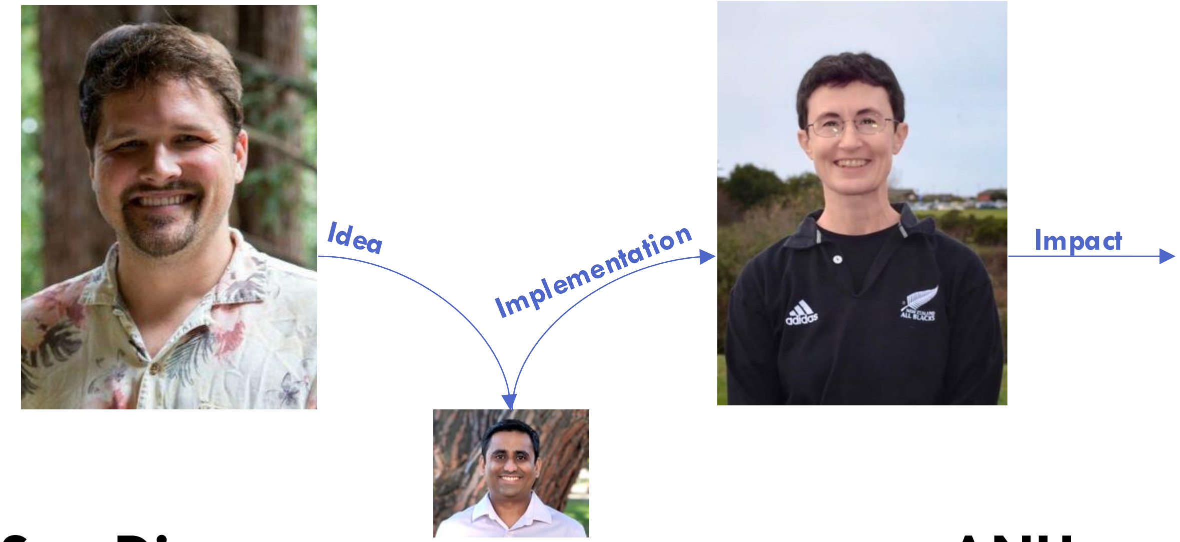
- **matUtils:** Command-line tools for rapidly analyzing and interpreting SARS-CoV-2 mutation-annotated phylogenetic trees
  - *McBroome et al.*, Molecular Biology and Evolution (**MBE**) 2021

https://usher-wiki.readthedocs.io/

Acknowledgments

Russell Corbett-Detig (UCSC)

Angie Hinrichs (UCSC)



UC Santa Cruz

- Bryan Thornlow
- Jakob McBroome
- Alexander Kramer
- Landen Gozashti
- Adriano Schneider
- Cade Mirchandani
- David Haussler

UC San Diego

- Cheng Ye
- Sumit Walia
- Alireza M.
- Kyle Smith
- Kevin Liu
- Xuan Wang
- Carol Wang
- Devika Torvi
- Shoh Mollenkamp
- Arthur Lu

ANU

- Robert Lanfear

EBI/EMBL

- Nicola DeMaio
- Nick Goldman

Funding

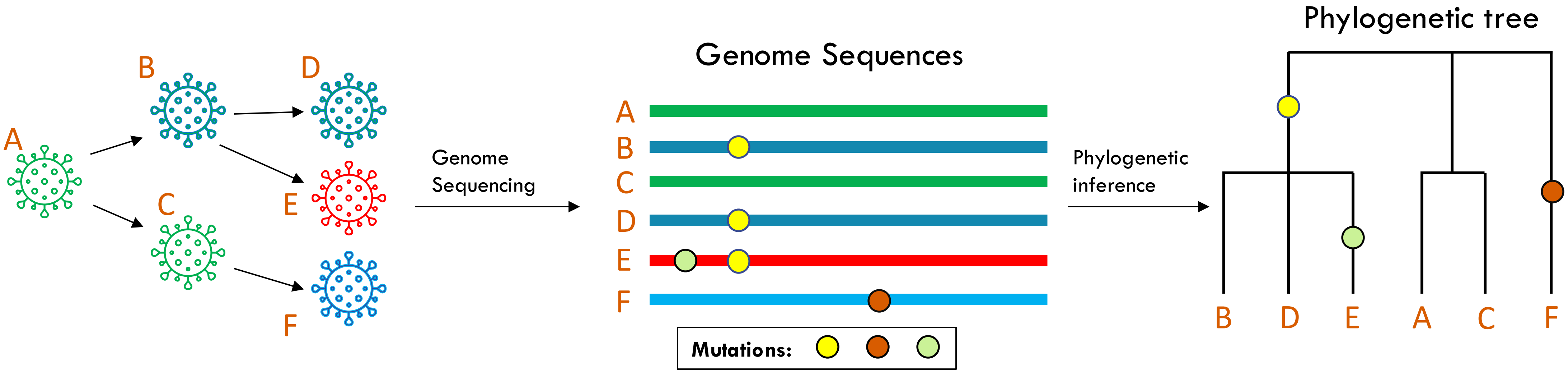
- CDC
- NIH
- Schmidt Foundation
- UCOP Seed Funding for COVID-19

Overview of the UShER Package

- **UShER: Phylogenetic placement**
- **matOptimize:** Phylogenetic tree optimization
- **RIPPLES:** Find recombinant sequences using a phylogenomic approach
- **matUtils:** Command-line tools for rapidly analyzing and interpreting SARS-CoV-2 mutation-annotated phylogenetic trees

https://usher-wiki.readthedocs.io/

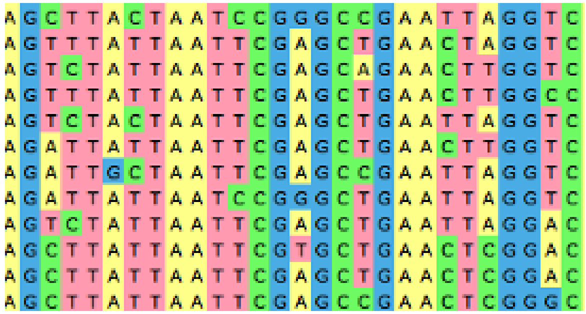
Phylogenetic analysis using genome sequence data



SARS-CoV-2 phylogenetics with <u>new</u> sequences

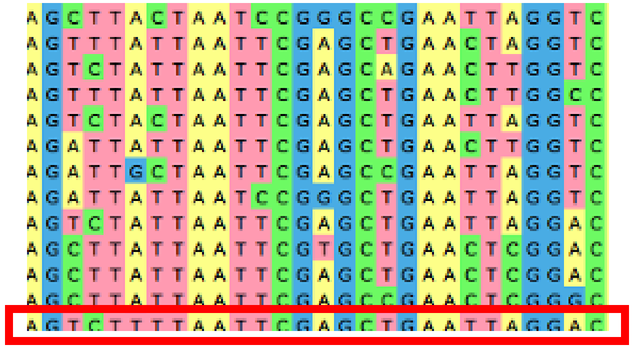
Approach 1: **Re-infer** global phylogeny including the new sequences

1. Gather global sequences (GISAID etc.)



```
AGCTTACTAATCCGGGCCGAATTAGGTC
AGTTTATTAATTCGAGCTGAACTAGGTC
AGTCTATTAATTCGAGCAGAACTTGGTC
AGTTTATTAATTCGAGCTGAACTTGGCC
AGTCTACTAATTCGAGCTGAATTAGGTC
AGATTATTAATTCGAGCTGAACTTGGTC
AGATTGCTAATTCGAGCCGAATTAGGTC
AGATTATTAATCCGGGCTGAATTAGGTC
AGTCTATTAATTCGAGCTGAATTAGGAC
AGCTTATTAATTCGTGCTGAACTCGGAC
AGCTTATTAATTCGAGCTGAACTCGGAC
AGCTTATTAATTCGAGCCGAACTCGGGC
```

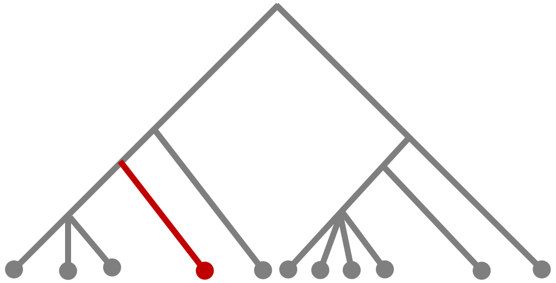
2. Add new sequences to MSA (MAFFT etc.)



```
AGCTTACTAATCCGGGCCGAATTAGGTC
AGTTTATTAATTCGAGCTGAACTAGGTC
AGTCTATTAATTCGAGCAGAACTTGGTC
AGTTTATTAATTCGAGCTGAACTTGGCC
AGTCTACTAATTCGAGCTGAATTAGGTC
AGATTATTAATTCGAGCTGAACTTGGTC
AGATTGCTAATTCGAGCCGAATTAGGTC
AGATTATTAATCCGGGCTGAATTAGGTC
AGTCTATTAATTCGAGCTGAATTAGGAC
AGCTTATTAATTCGTGCTGAACTCGGAC
AGCTTATTAATTCGAGCTGAACTCGGAC
AGCTTATTAATTCGAGCCGAACTCGGGC
AGTCTTTTAATTCGAGCTGAATTAGGAC
```

Intractable for SARS-CoV-2 scale!

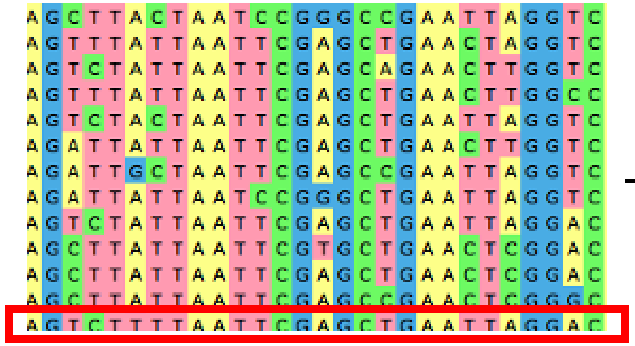
3. De novo inference of phylogeny (IQ-Tree etc.)



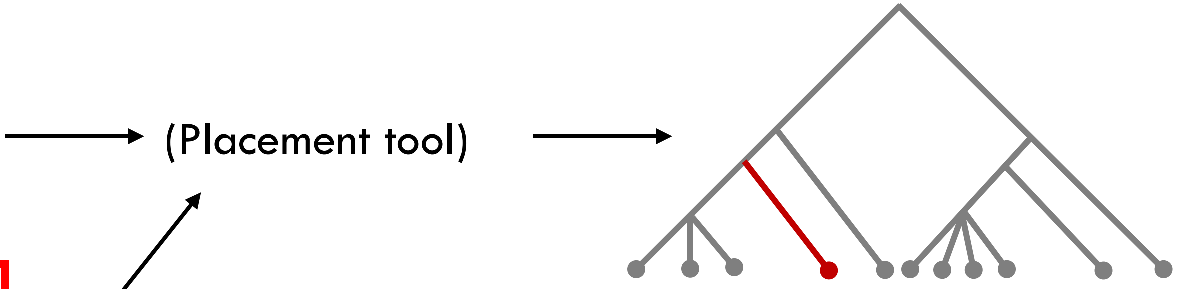
SARS-CoV-2 phylogenetics with new sequences

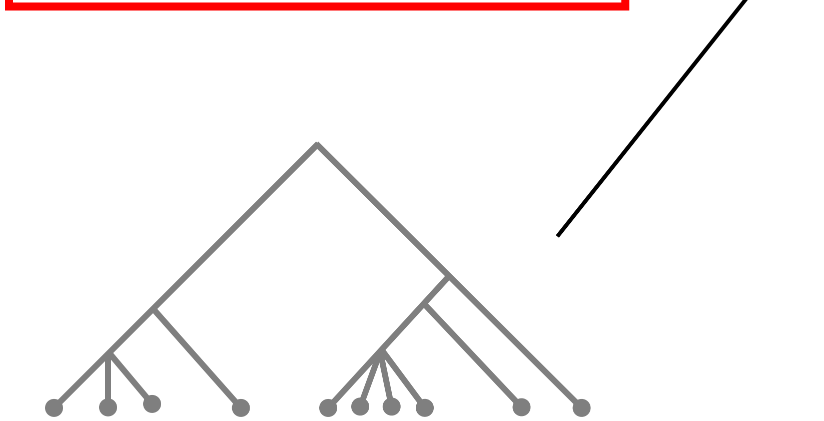
~~**Approach 1: Re-infer** global phylogeny including the new sequences~~

Approach 2: Place new sequences on an existing phylogeny



```
AGCTTACTAATCCGGGCCGAATTAGGTC
AGTTTATTAATTCGAGCTGAACTAGGTC
AGTCTATTAATTCGAGCAGAACTTGGTC
AGTTTATTAATTCGAGCTGAACTTGGCC
AGTCTACTAATTCGAGCTGAATTAGGTC
AGATTATTAATTCGAGCTGAACTTGGTC
AGATTGCTAATTCGAGCCGAATTAGGTC
AGATTATTAATCCGGGCTGAATTAGGTC
AGTCTATTAATTCGAGCTGAATTAGGAC
AGCTTATTAATTCGTGCTGAACTCGGAC
AGCTTATTAATTCGAGCTGAACTCGGAC
AGCTTATTAATTCGAGCCGAACTCGGGC
AGTCTTTTAATTCGAGCTGAATTAGGAC
```





Tool	Time to place 1K samples on 100K tree	Memory required
IQ-TREE 2	6h 9m	120.2GB

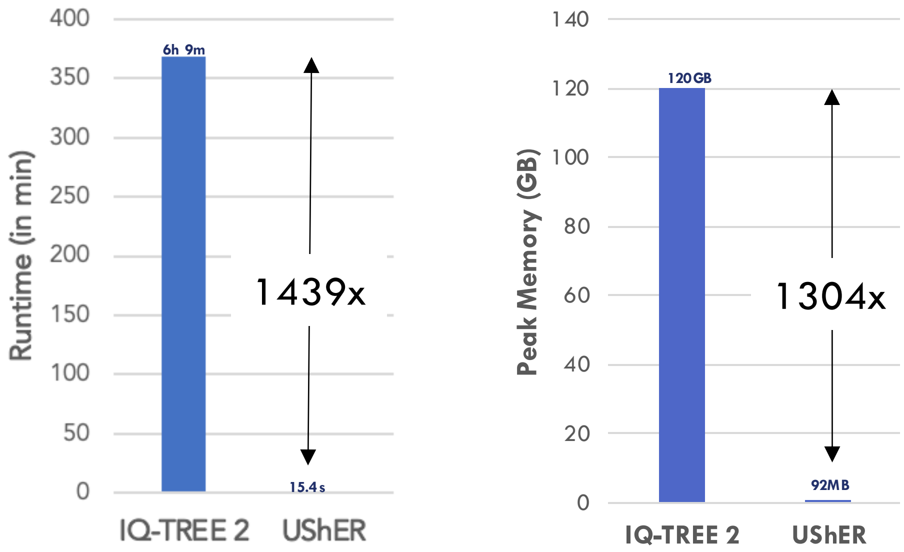
(using e2-highmem-16 instance)

UShER

Ultrafast Sample Placement on Existing Trees

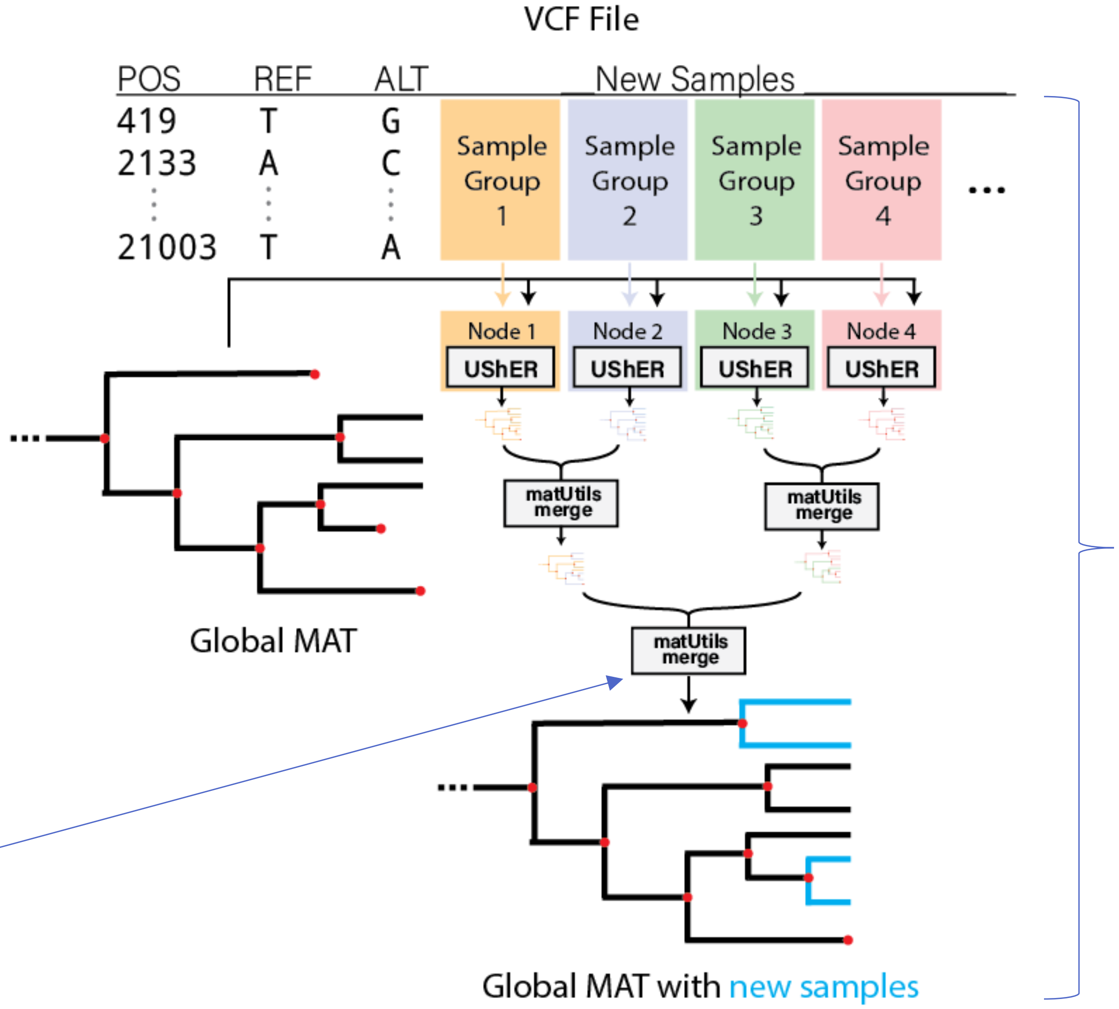
https://github.com/yatisht/usher

Placing 1K samples on 100K tree



(Turakhia et al., Nature Genetics 2021)

Parallelizing over multiple CPU instances



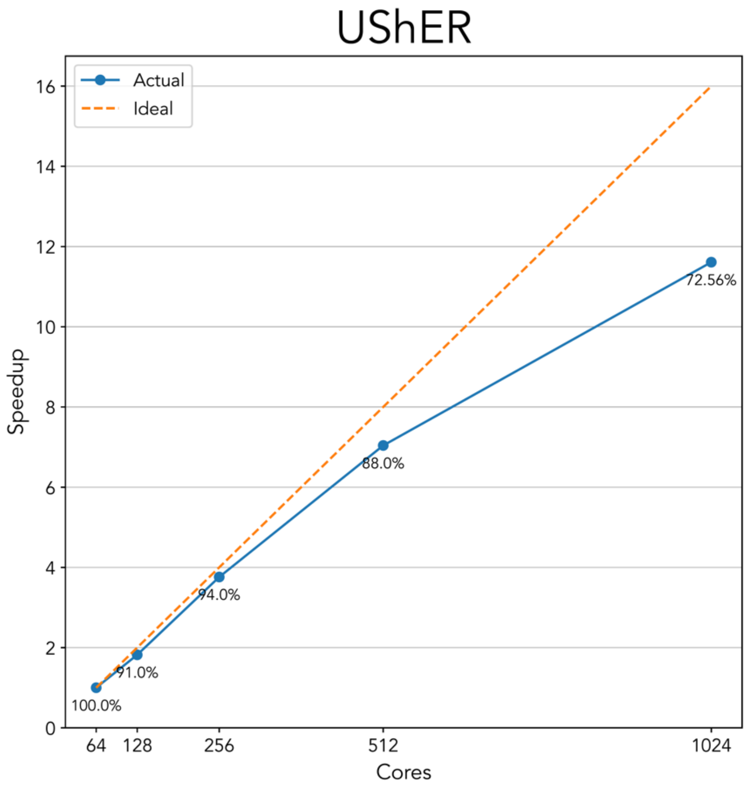
Google Cloud Platform

Devika Torvi, UCSD **Bioinformatics** undergrad

Kyle Smith, UCSD **Bioinformatics** undergrad

Scaling analysis for placing 100K samples on a 1M-sample tree

Strong Scaling



Weak Scaling

UShER		
vCPU	**Samples placed**	**Time**
64	6.25K	26m 48s
128	12.5K	28m 22s
256	25K	30m 41s
512	50K	33m 36s
1024	100K	37m 07s

Why is UShER so fast?

For SARS-CoV-2 relative to the highly popular software, **IQ-TREE,** that has amassed **>10K citations**

What makes UShER so fast?

1. Choice of **algorithm**: maximum parsimony over maximum likelihood

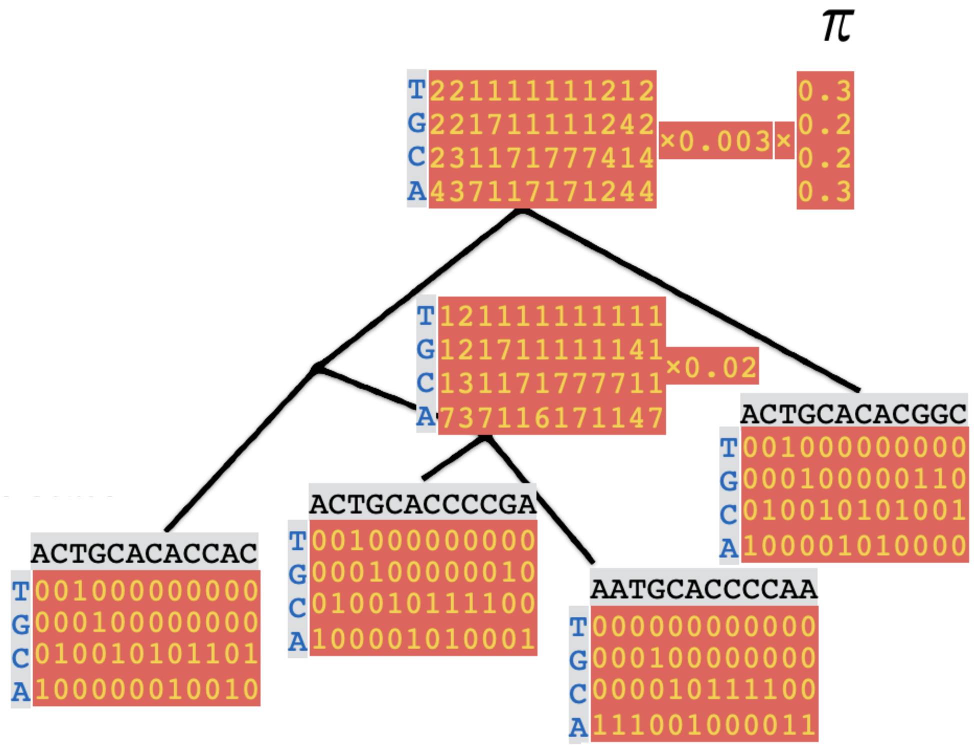
2. Efficient **data structure**: mutation-annotated tree (MAT)

3. **Pre-processing** for sequential placement

4. Efficiently **parallelizing** the placement step

1. Choice of Algorithm: MP over ML (10-100x speedup)

Maximum Likelihood (ML)



T 221111111212
G 221711111242
C 231171777414
A 437117171244

0.3
0.2
0.2
0.3

T 121111111111
G 121711111141
C 131171777711
A 737116171147

T 001000000000
G 000100000110
C 010010101001
A 100001010000

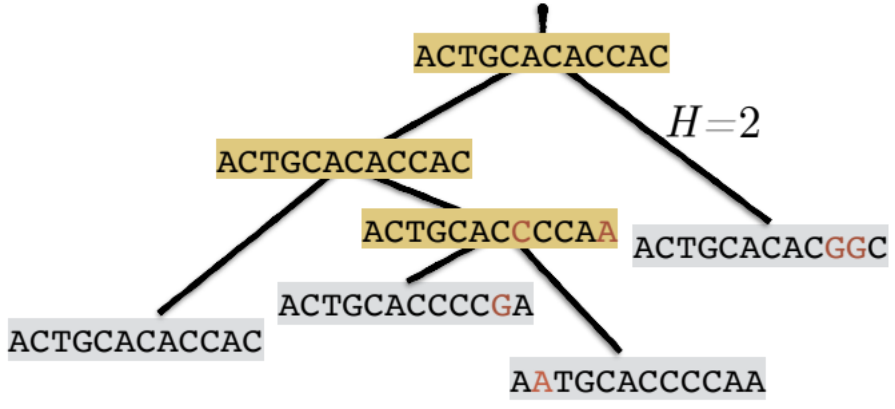
T 001000000000
G 000100000010
C 010010111100
A 100001010001

T 001000000000
G 000100000000
C 010010101101
A 100000010010

T 000000000000
G 000100000000
C 000010111100
A 111001000011

$$L(\theta; D) = \prod_{i=1}^{L} \sum_{x \in X(D)} \pi_{x_r^i} \prod_{(p,c) \in E} P(x_c^i | x_p^i, t_e, \theta) \quad P(D|\theta) = \prod_{1}^{L} P(D^i|\theta)$$

Maximum Parsimony (MP)

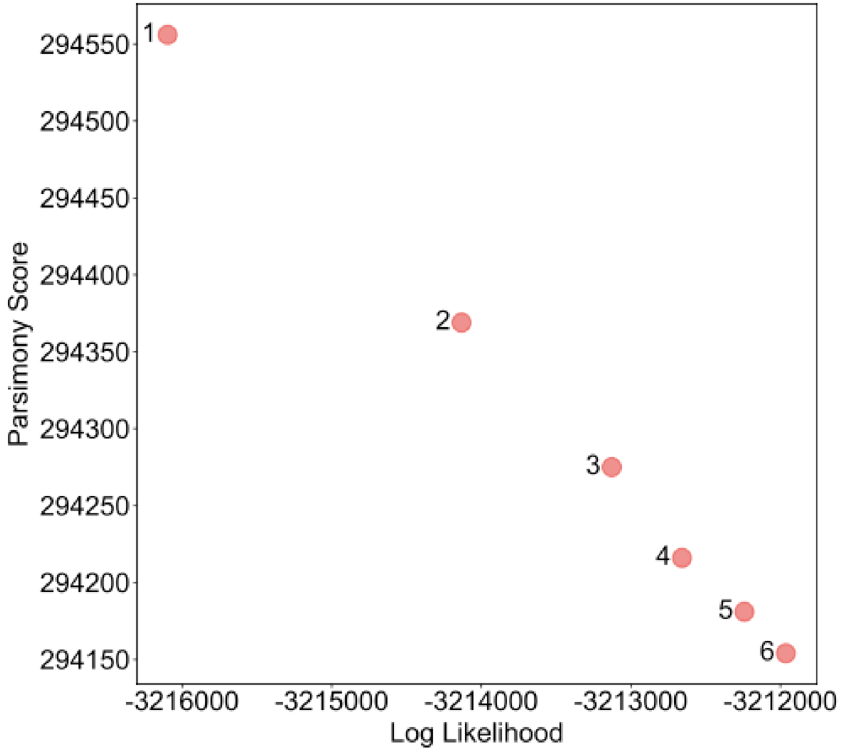


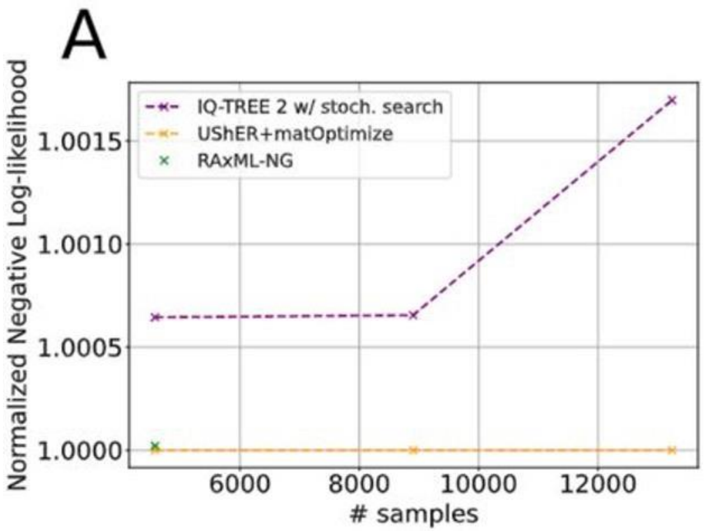
Internal node u with children v and w:

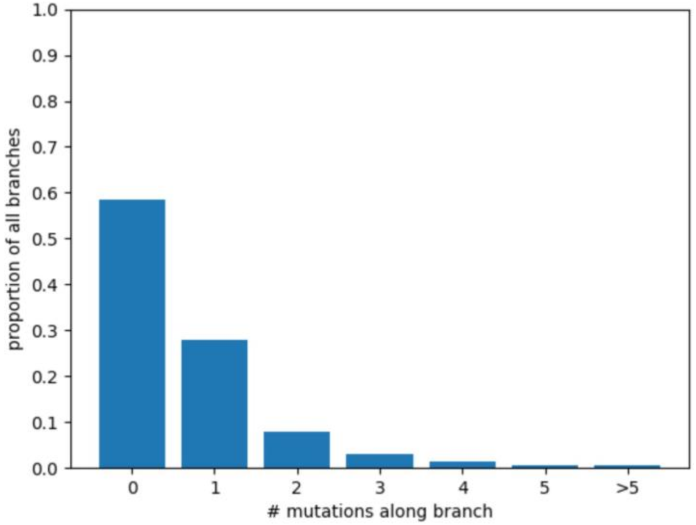
$$S_u(x) = \min_y(S_v(y) + W_{xy}) + \min_y(S_w(y) + W_{xy})$$

(S. Mirarab, UCSD ECE 208)

MP and ML trees practically the same for SARS-CoV-2







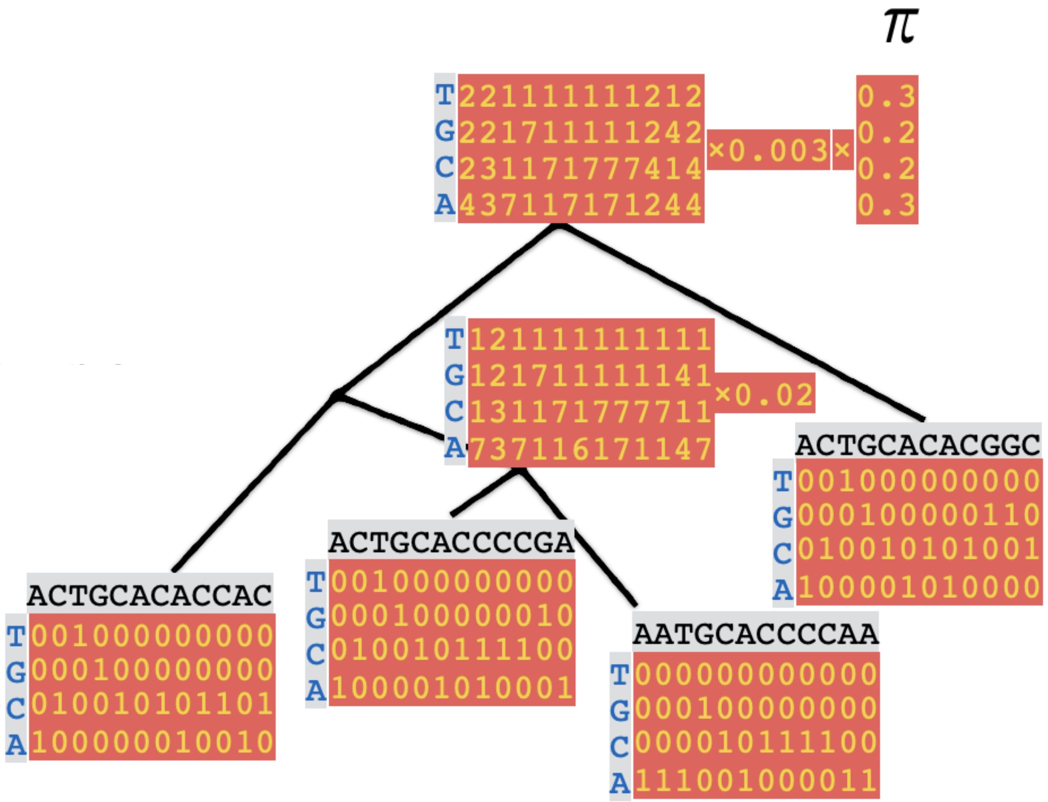
Bryan Thornlow,

UCSC -> ROME Therapeutics

(Thornlow et al., bioRxiv 2022) | https://github.com/bpt26/parsimony

MP trees are also easier to analyze & interpret!

Maximum Likelihood



T 221111111212
G 221711111242
C 231171777414
A 437117171244

0.3
0.2
0.2
0.3

T 121111111111
G 121711111141
C 131171777711
A 737116171147

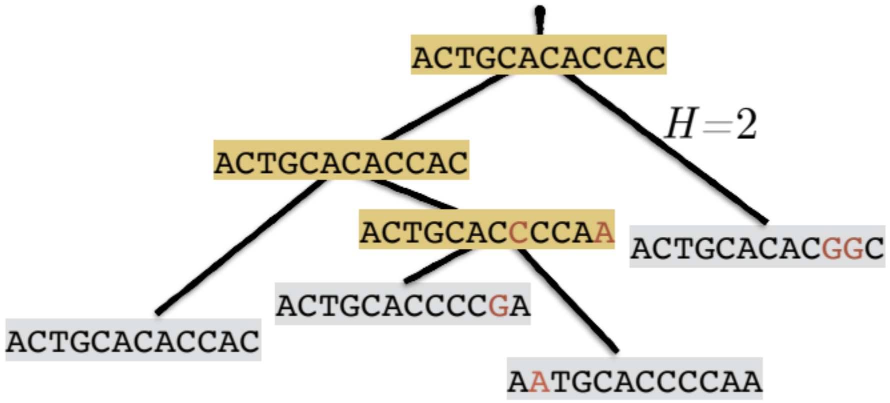
ACTGCACACGGC
T 001000000000
G 000100000110
C 010010101001
A 100001010000

ACTGCACCCCGA
T 001000000000
G 000100000010
C 010010111100
A 100001010001

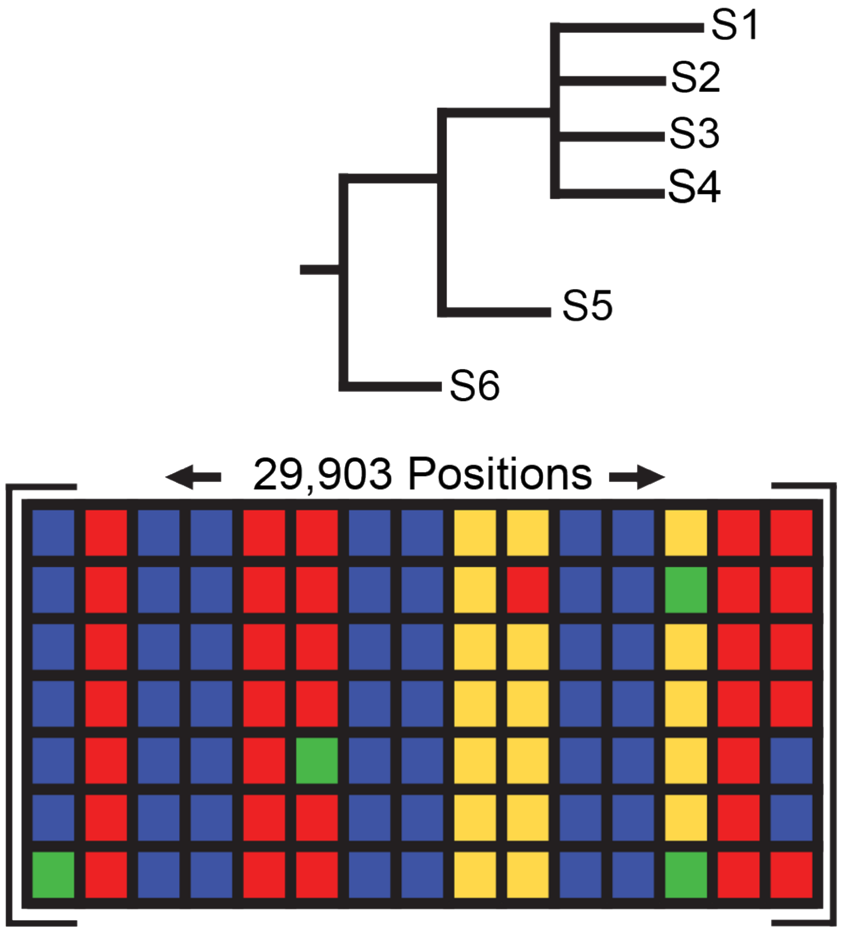
ACTGCACACCAC
T 001000000000
G 000100000000
C 010010101101
A 100000010010

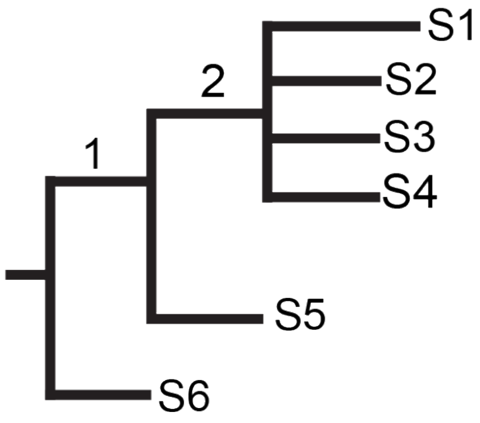
AATGCACCCCAA
T 000000000000
G 000100000000
C 000010111100
A 111001000011

Maximum Parsimony



2. Efficient data structure: Mutation-annotated tree (MAT)



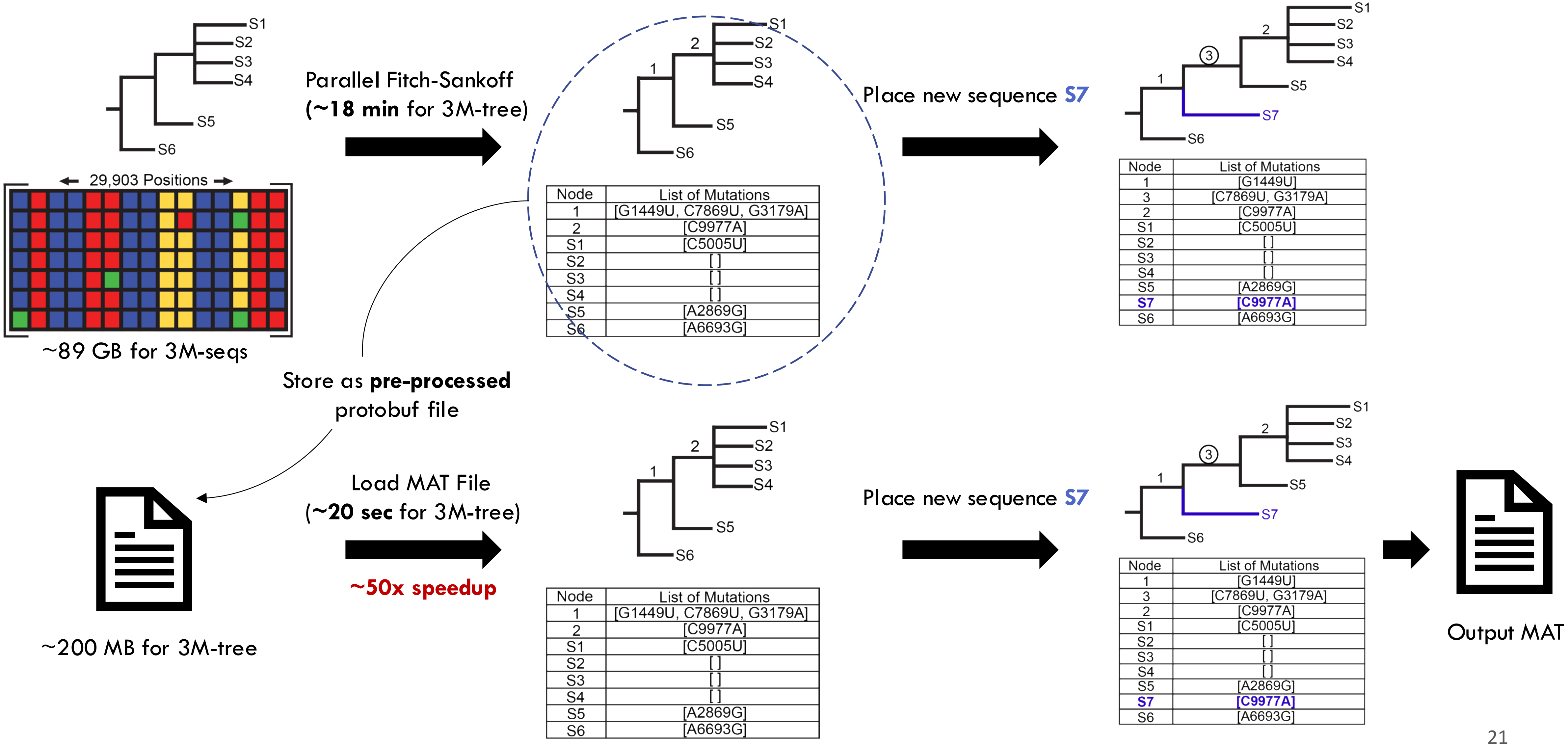


Node	List of Mutations
1	[G1449U, C7869U, G3179A]
2	[C9977A]
S1	[C5005U]
S2	[ ]
S3	[ ]
S4	[ ]
S5	[A2869G]
S6	[A6693G]

174,000 bytes → 186 bytes

"evolutionary compression"

3. Pre-processing for sequential placement: ~50x speedup



~89 GB for 3M-seqs

Parallel Fitch-Sankoff
(**~18 min** for 3M-tree)

Node	List of Mutations
1	[G1449U, C7869U, G3179A]
2	[C9977A]
S1	[C5005U]
S2	[ ]
S3	[ ]
S4	[ ]
S5	[A2869G]
S6	[A6693G]

Place new sequence **S7**

Node	List of Mutations
1	[G1449U]
3	[C7869U, G3179A]
2	[C9977A]
S1	[C5005U]
S2	[ ]
S3	[ ]
S4	[ ]
S5	[A2869G]
S7	**[C9977A]**
S6	[A6693G]

Store as **pre-processed** protobuf file

~200 MB for 3M-tree

Load MAT File
(**~20 sec** for 3M-tree)

~50x speedup

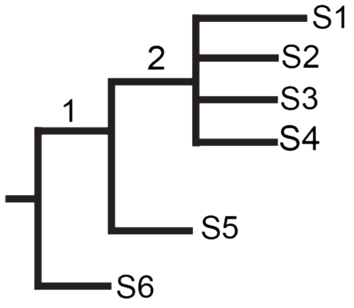
Node	List of Mutations
1	[G1449U, C7869U, G3179A]
2	[C9977A]
S1	[C5005U]
S2	[ ]
S3	[ ]
S4	[ ]
S5	[A2869G]
S6	[A6693G]

Place new sequence **S7**

Node	List of Mutations
1	[G1449U]
3	[C7869U, G3179A]
2	[C9977A]
S1	[C5005U]
S2	[ ]
S3	[ ]
S4	[ ]
S5	[A2869G]
S7	**[C9977A]**
S6	[A6693G]

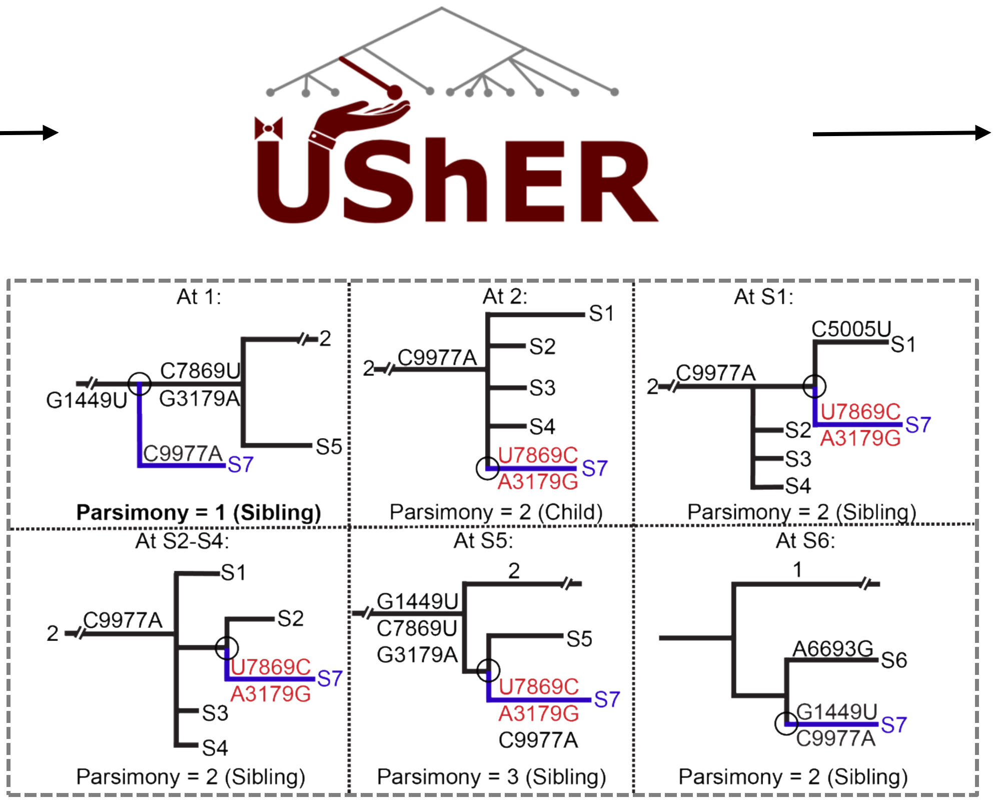
Output MAT

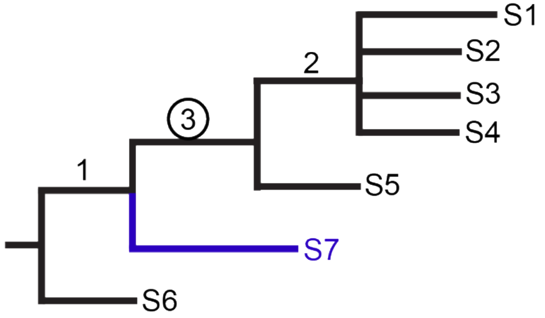
4. Efficiently parallelizing the placement step



Node	List of Mutations
1	[G1449U, C7869U, G3179A]
2	[C9977A]
S1	[C5005U]
S2	[ ]
S3	[ ]
S4	[ ]
S5	[A2869G]
S6	[A6693G]

S7 [G1449U, C9977A]





Node	List of Mutations
1	[G1449U]
3	[C7869U, G3179A]
2	[C9977A]
S1	[C5005U]
S2	[ ]
S3	[ ]
S4	[ ]
S5	[A2869G]
S7	**[C9977A]**
S6	[A6693G]

But … (greedy) sequential placement can lead to suboptimal trees occasionally

And suboptimalities could accumulate!

Overview of the UShER Package

- UShER: Phylogenetic placement
- **matOptimize: Phylogenetic tree optimization**
- **RIPPLES:** Find recombinant sequences using a phylogenomic approach
- **matUtils:** Command-line tools for rapidly analyzing and interpreting SARS-CoV-2 mutation-annotated phylogenetic trees

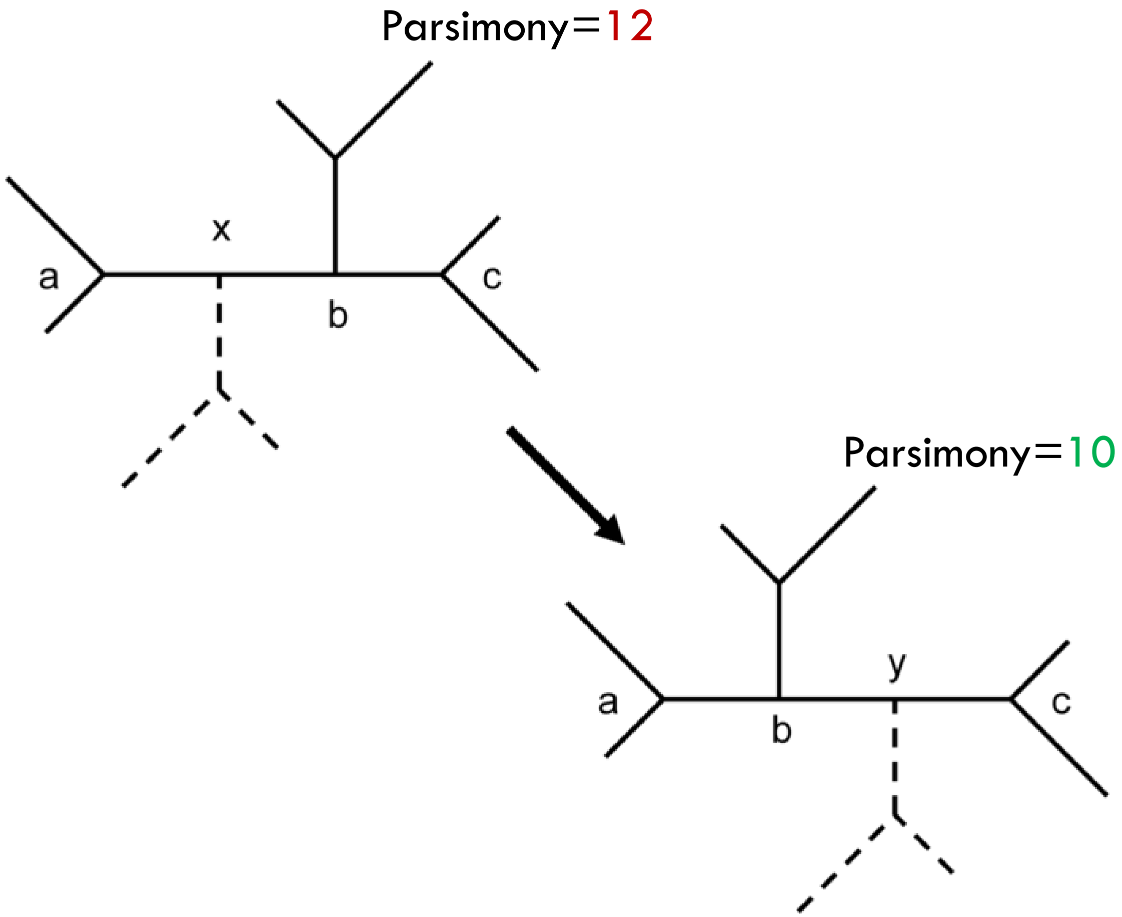
https://usher-wiki.readthedocs.io/

Tree optimization programs can help ameliorate suboptimal placements

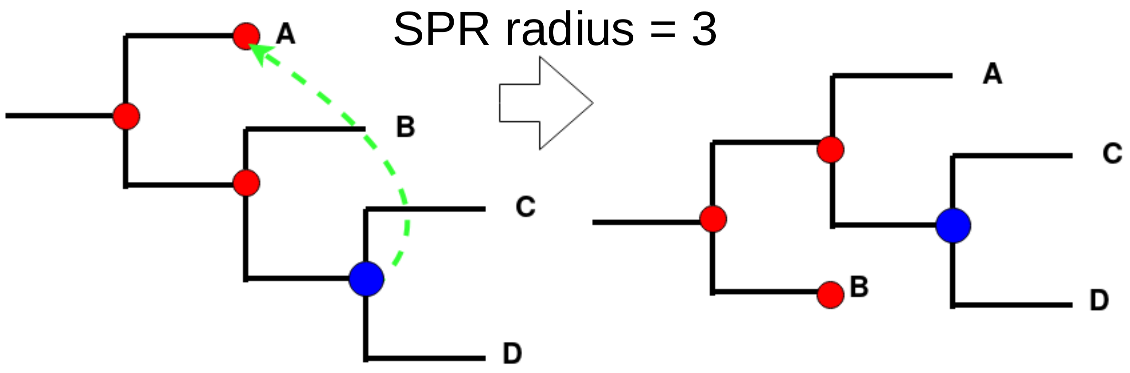
- Use tree re-arrangement (NNI, SPR, TBR etc.)

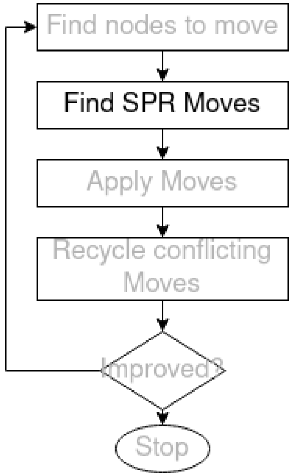
- Three step process:
  - (Split tree into subtrees/sectors)
  - Identify **profitable** tree rearrangement moves in each sector
  - Apply **non-conflicting** profitable moves in each sector, with a tie-breaking strategy for conflicts
  - (Merge sectors to optimized full tree)

- Some programs maintain several tree copies (applying different moves in each copy)



matOptimize (ours) outperforms TNT for SARS-CoV-2





Find nodes to move

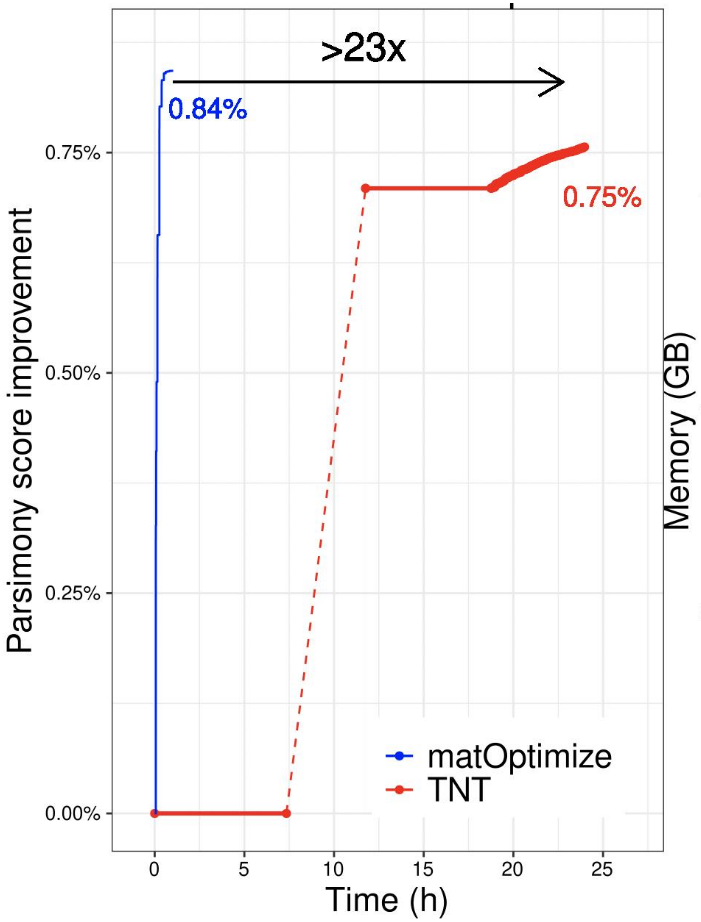
Find SPR Moves

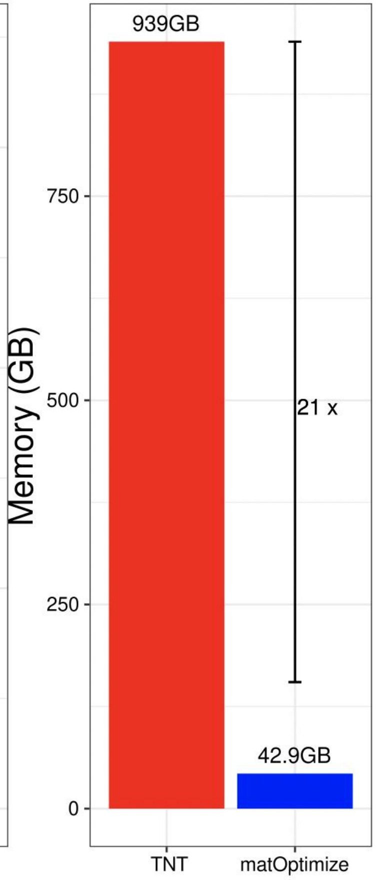
Apply Moves

Recycle conflicting Moves

Improved?

Stop

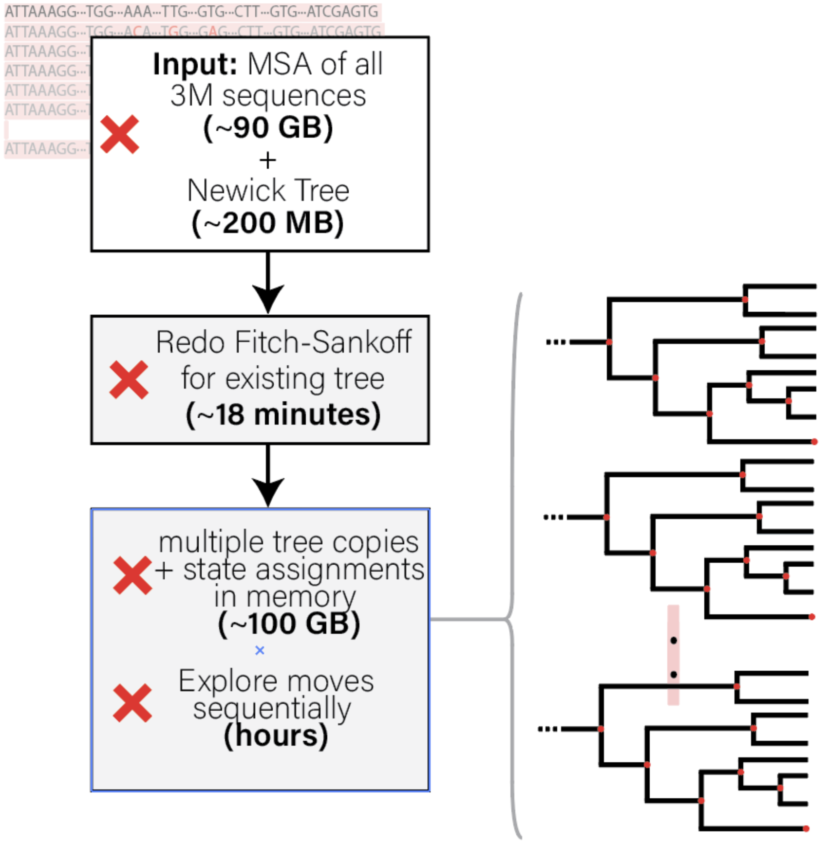




(Cheng Ye, UCSD ECE undergrad)

Innovative optimizations in matOptimize

Previous approaches



☒ **Input:** MSA of all 3M sequences **(~90 GB)** + Newick Tree **(~200 MB)**

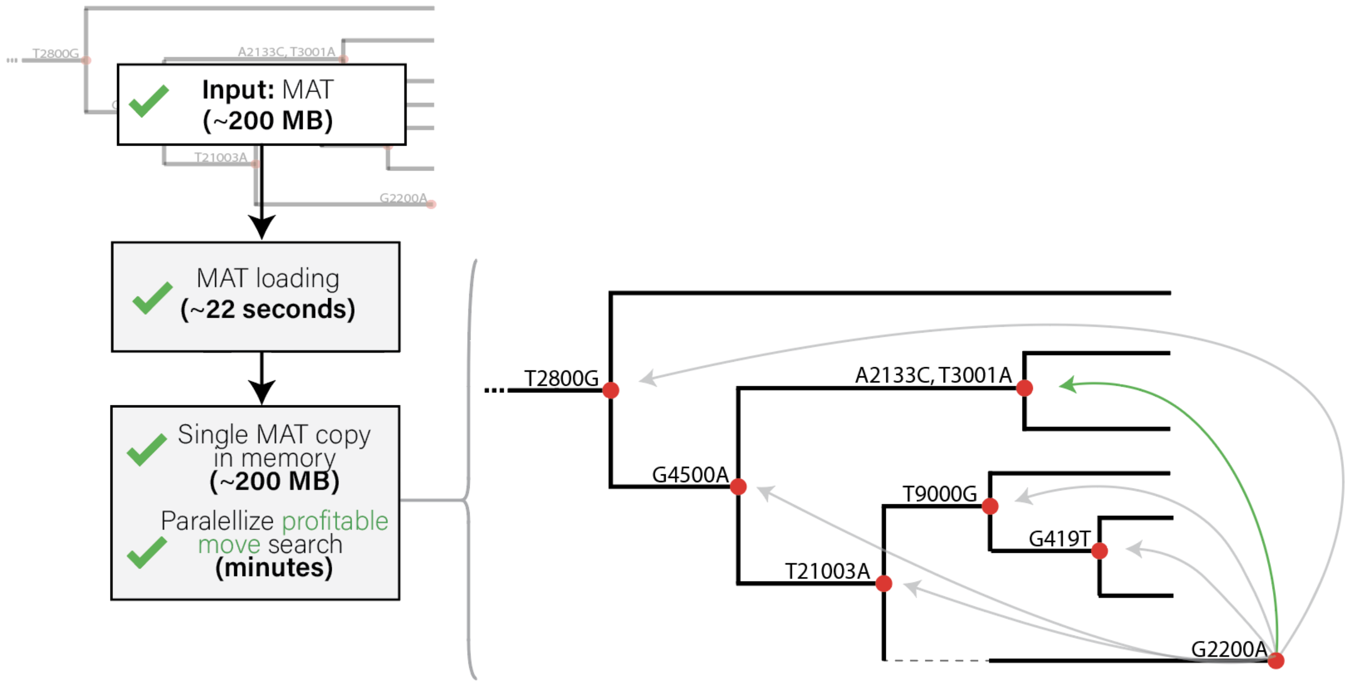
☒ Redo Fitch-Sankoff for existing tree **(~18 minutes)**

☒ multiple tree copies + state assignments in memory **(~100 GB)**

☒ Explore moves sequentially **(hours)**

ATTAAAGG···TGG···AAA···TTG···GTG···CTT···GTG···ATCGAGTG
ATTAAAGG···TGG···ACA···TGG···GAG···CTT···GTG···ATCGAGTG
ATTAAAGG···
ATTAAAGG···
ATTAAAGG···
ATTAAAGG···
ATTAAAGG···

Ours (matOptimize)



✓ **Input:** MAT **(~200 MB)**

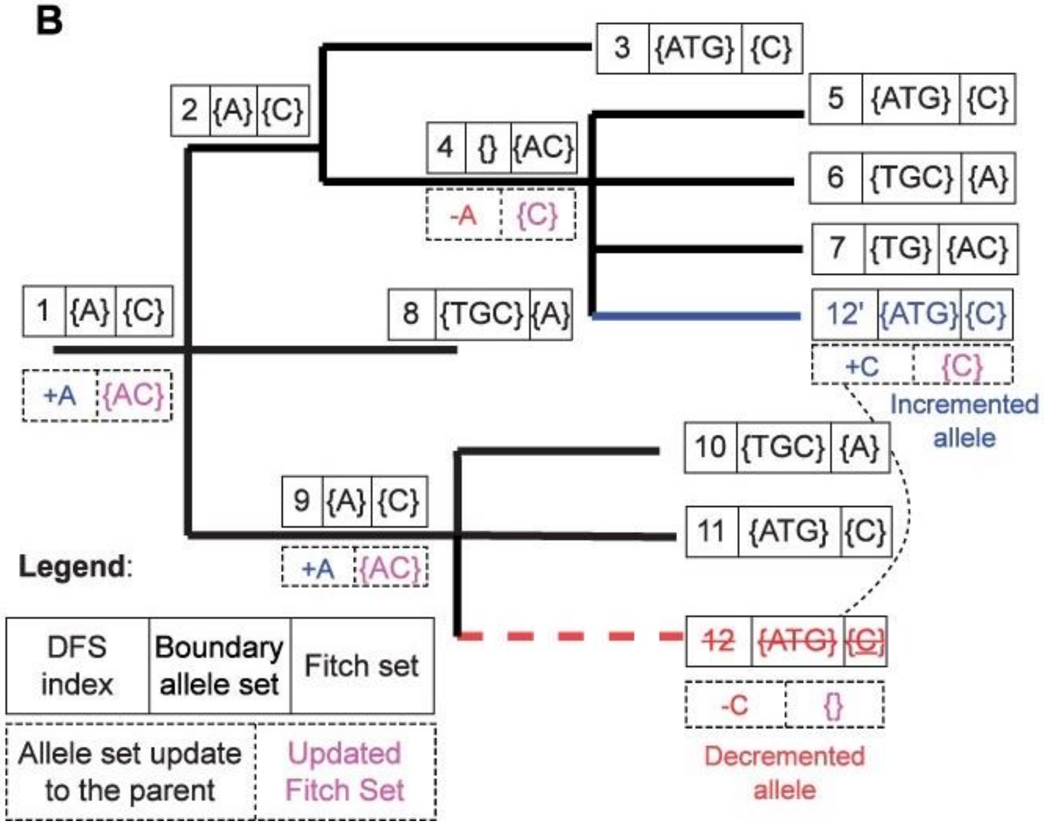
✓ MAT loading **(~22 seconds)**

✓ Single MAT copy in memory **(~200 MB)**

✓ Paralellize profitable move search **(minutes)**

Innovative optimizations in matOptimize

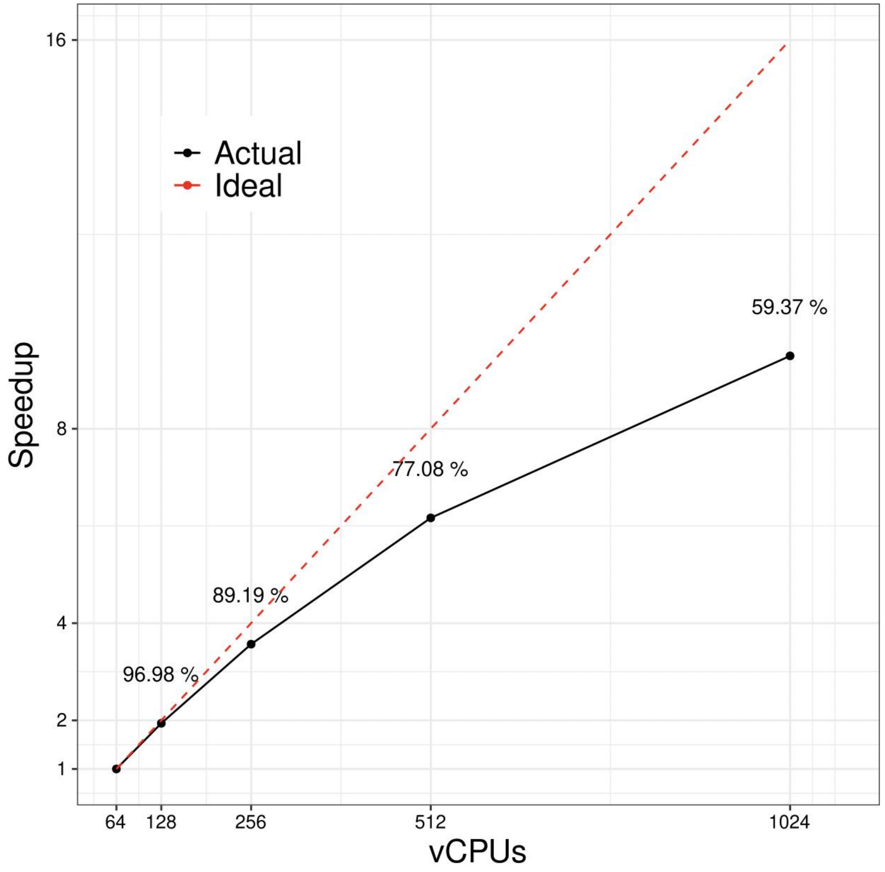
- More space-efficient and optimization-friendly **MAT format**
- Separate profitable move search and application phase to achieve **high parallelism**
  - Supports multi-node parallelism with MPI
- Modified **Gladstein's incremental update** method to calculate change in parsimony score resulting from a move
- Novel **search space pruning**



(Ye et al. Bioinformatics 2022)

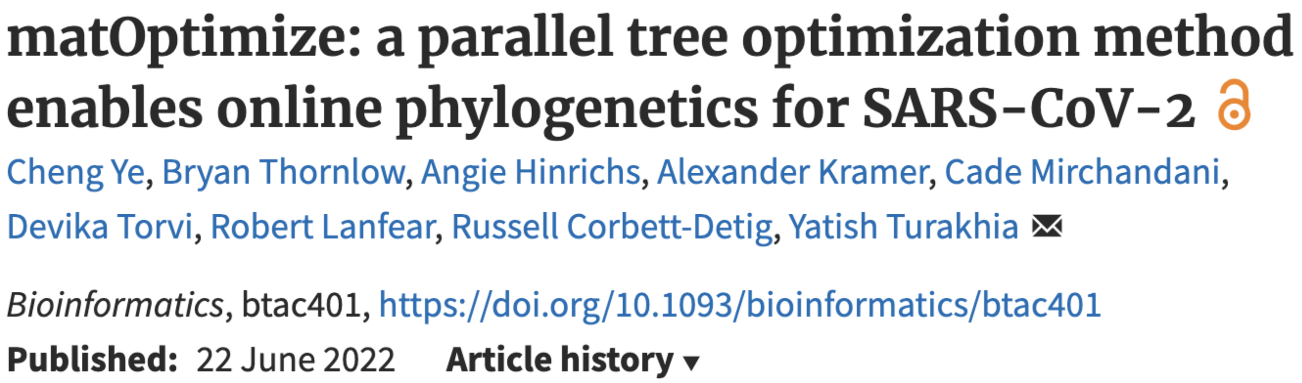
Multi-node scaling of matOptimize

Strong Scaling



Weak Scaling

vCPU	Source nodes explored	Time
64	39789	10m 45s
128	79577	11m 54s
256	159154	11m 51s
512	318308	11m 58s
1024	636616	11m 30s



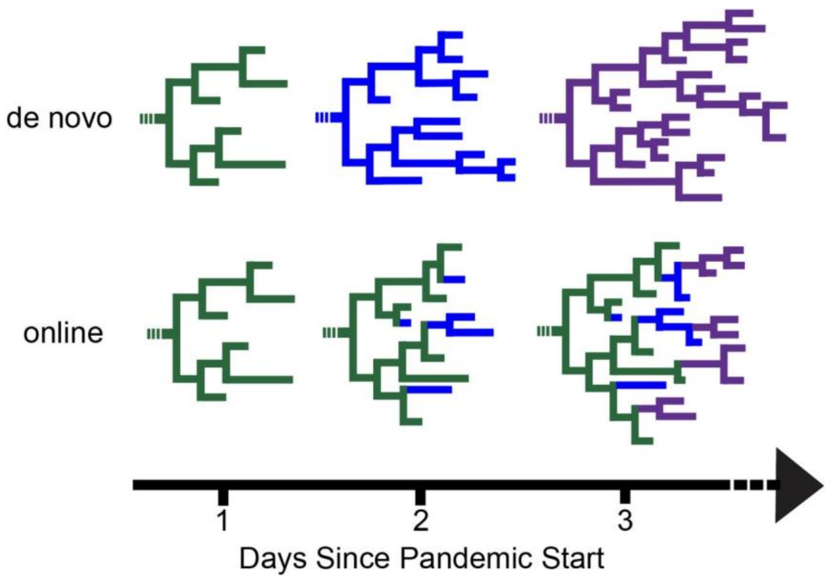
matOptimize: a parallel tree optimization method enables online phylogenetics for SARS-CoV-2

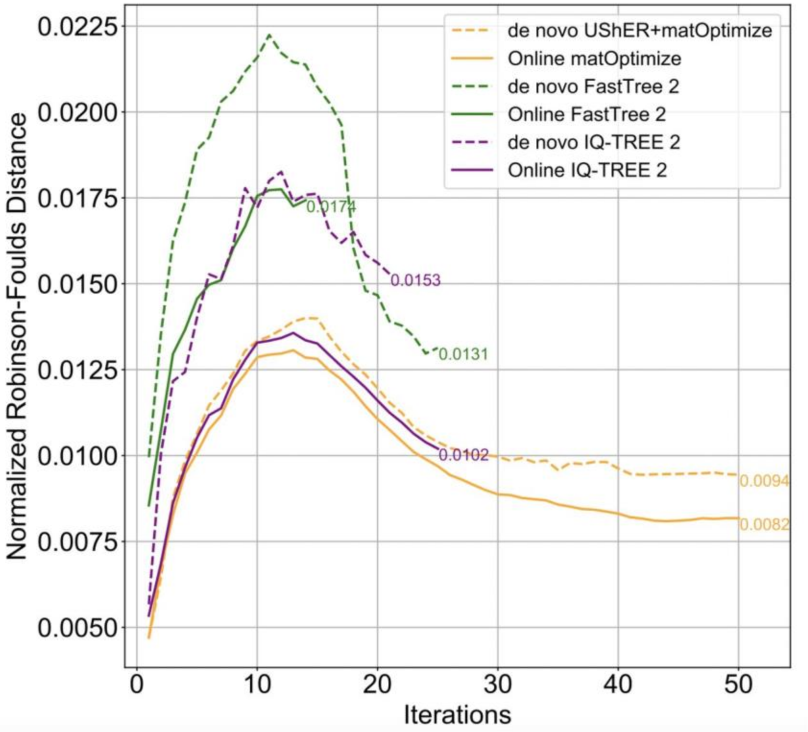
Cheng Ye, Bryan Thornlow, Angie Hinrichs, Alexander Kramer, Cade Mirchandani, Devika Torvi, Robert Lanfear, Russell Corbett-Detig, Yatish Turakhia ✉

Bioinformatics, btac401, https://doi.org/10.1093/bioinformatics/btac401

Published: 22 June 2022 **Article history** ▾

Online matOptimize produces phylogenies most similar to ground truth on simulated SARS2 data





Bryan Thornlow

(Thornlow et al., bioRxiv 2022)

There is another spooky mechanism through which the virus evolves …

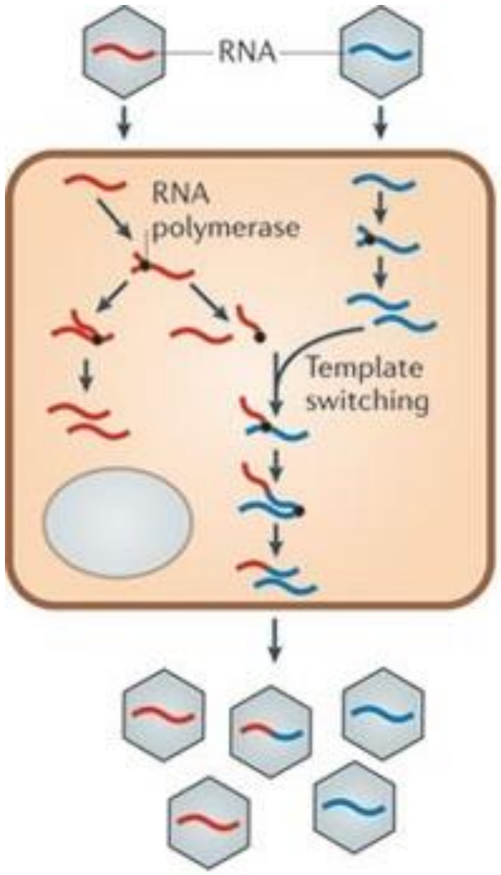
It’s called Recombination

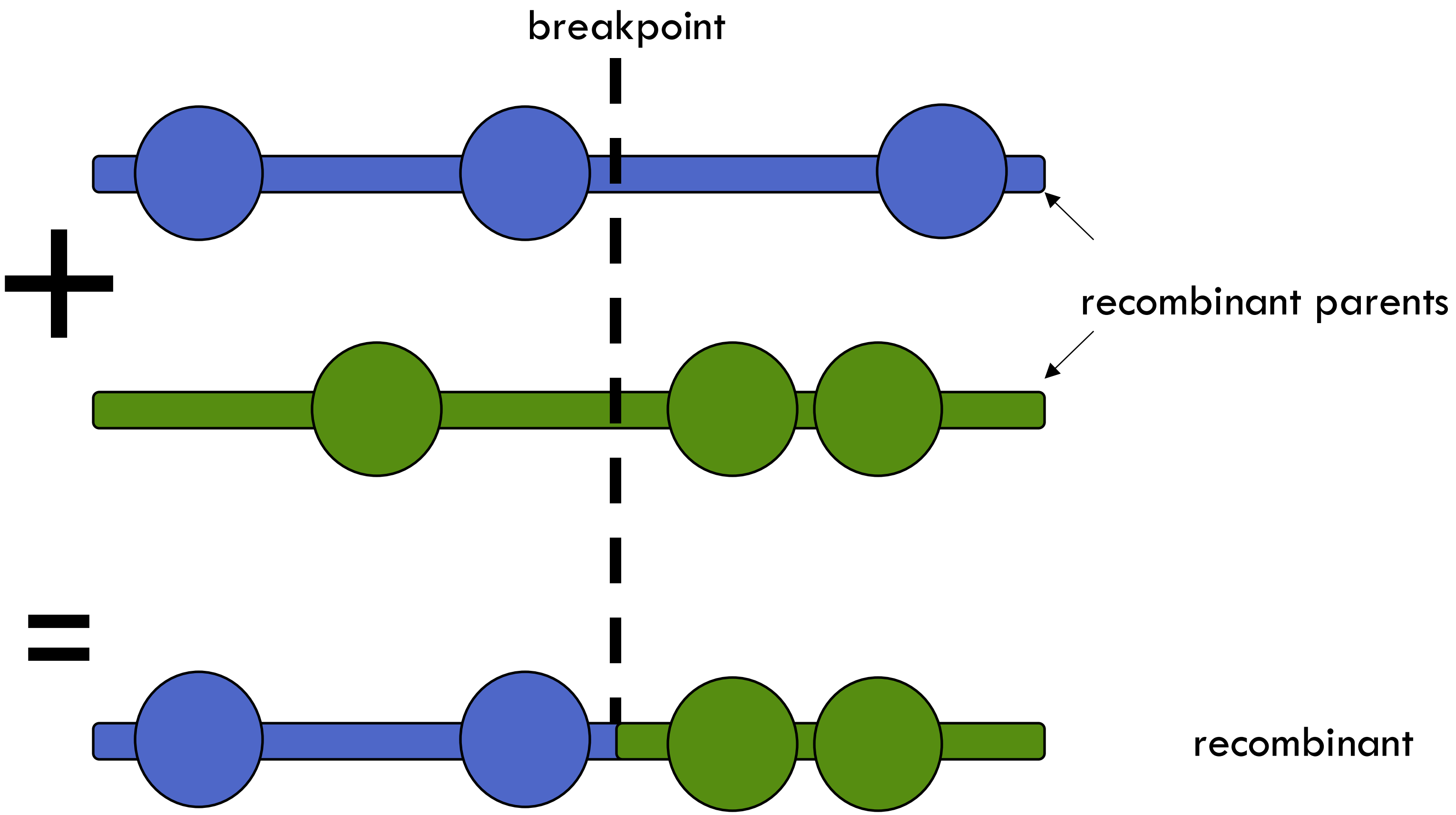
Overview of the UShER Package

- UShER: Phylogenetic placement
- matOptimize: Phylogenetic tree optimization
- **RIPPLES: Find recombinant sequences using a phylogenomic approach**
- **matUtils:** Command-line tools for rapidly analyzing and interpreting SARS-CoV-2 mutation-annotated phylogenetic trees

https://usher-wiki.readthedocs.io/

Viral Recombination

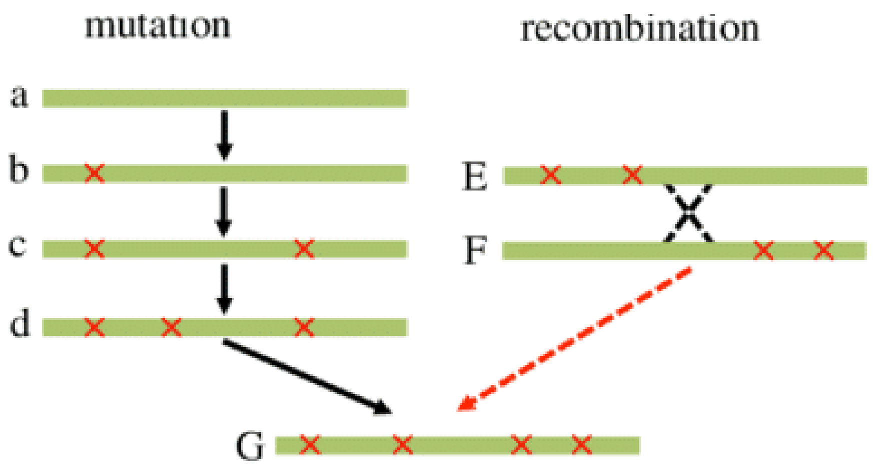




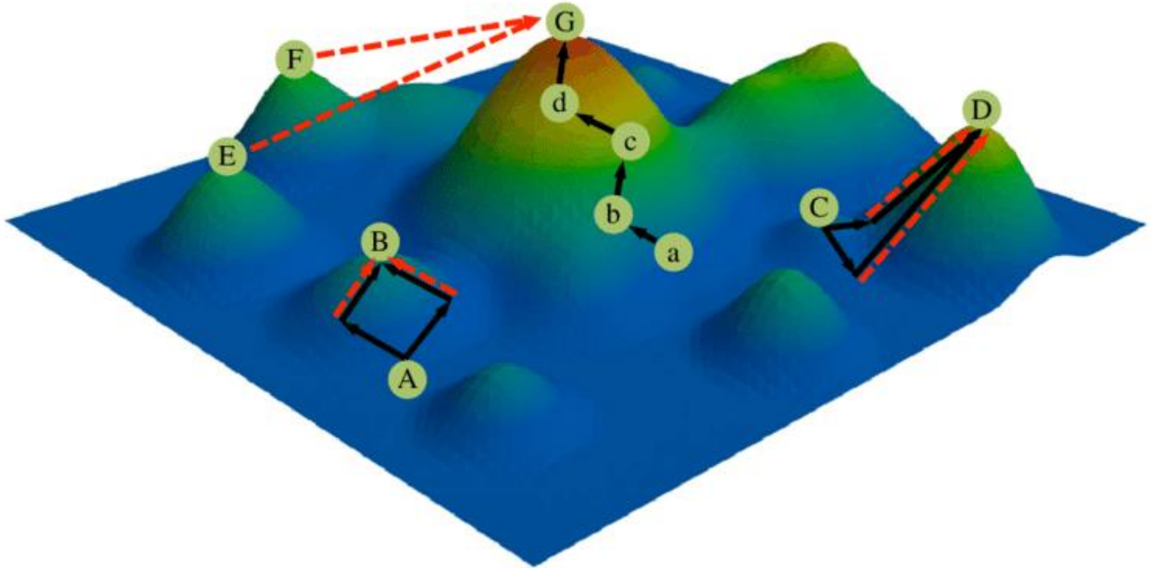
(Simon-Loriere et al., Nature Microbiology 2011)

Recombination may lead to drastic jumps in fitness!

Sequence evolution by single mutations v/s recombination

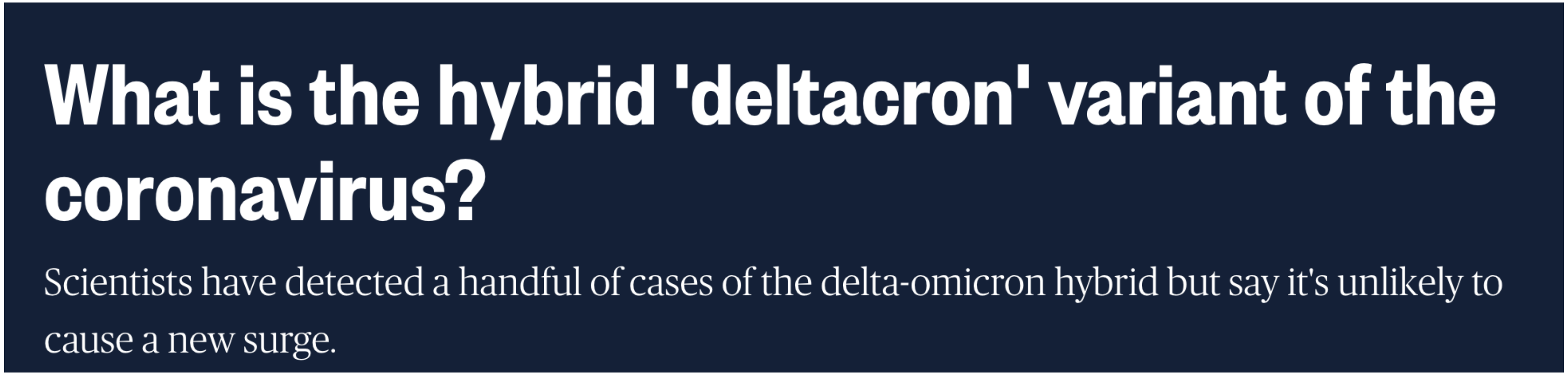


Fitness landscape



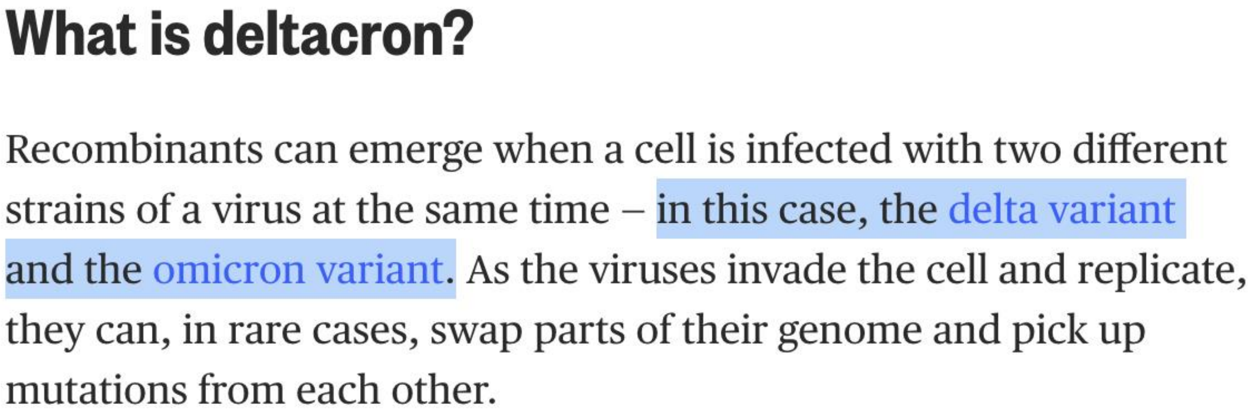
(Pesce et al., RSTB 2016)

Deltacron & some other recombinants are real!



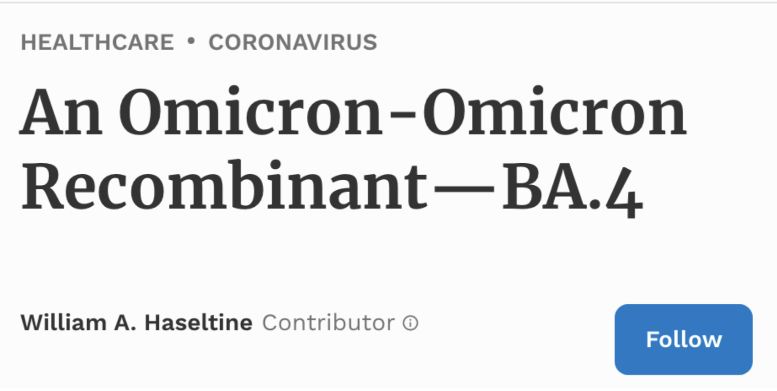
What is the hybrid 'deltacron' variant of the coronavirus?

Scientists have detected a handful of cases of the delta-omicron hybrid but say it's unlikely to cause a new surge.



What is deltacron?

Recombinants can emerge when a cell is infected with two different strains of a virus at the same time – in this case, the delta variant and the omicron variant. As the viruses invade the cell and replicate, they can, in rare cases, swap parts of their genome and pick up mutations from each other.



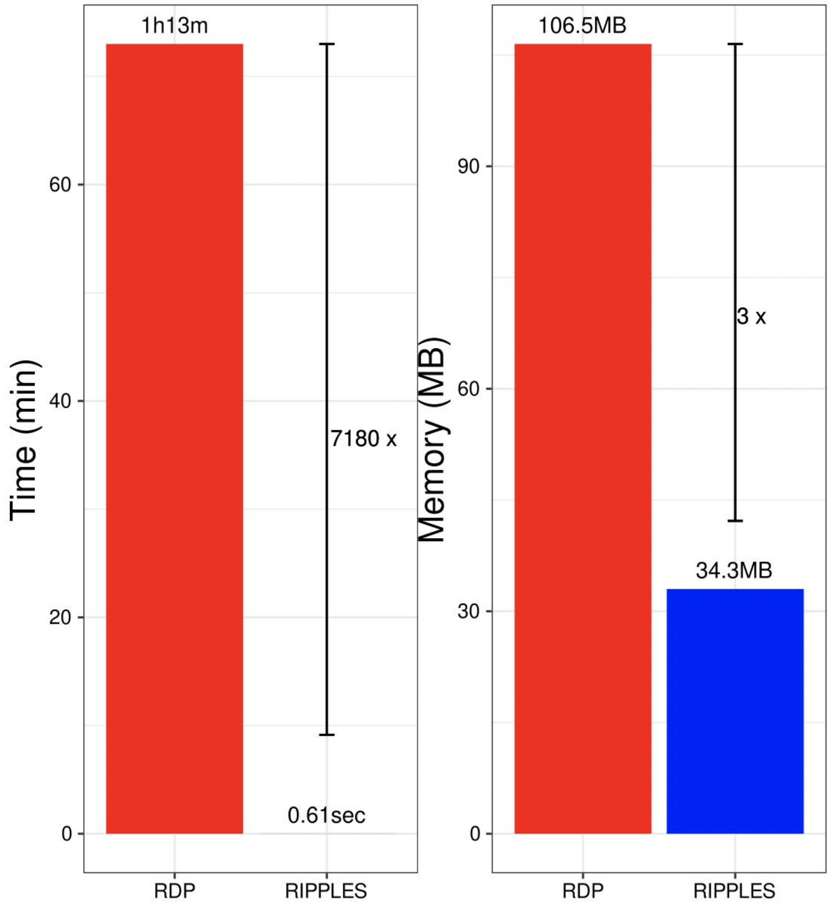
HEALTHCARE • CORONAVIRUS

An Omicron-Omicron Recombinant—BA.4

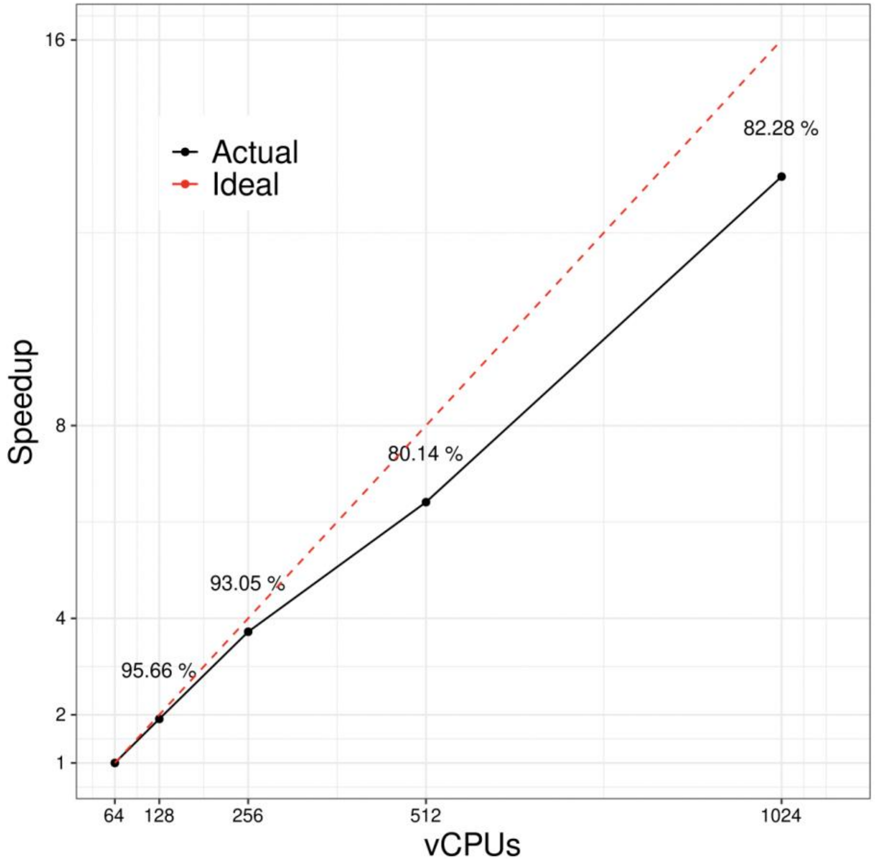
William A. Haseltine Contributor

https://www.nbcnews.com/science/science-news/deltacron-hybrid-variant-coronavirus-rcna20103

RIPPLES orders of magnitude faster using phylogenomic insights



Analyzed on 1K SARS-CoV-2 seqs

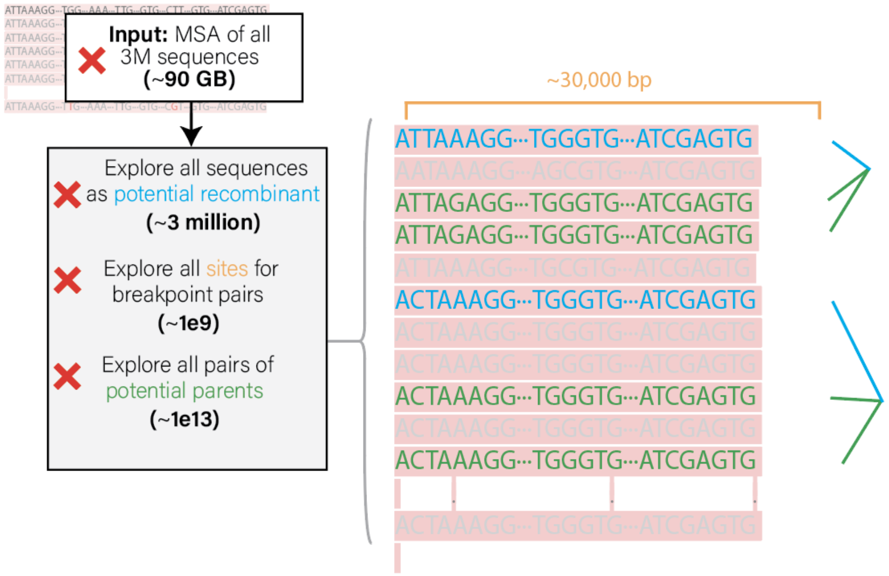


~50 min for 1M-sample tree

(Turakhia et al., bioRxiv 2021)

Innovative optimizations in RIPPLES

Previous approaches



Input: MSA of all 3M sequences **(~90 GB)**

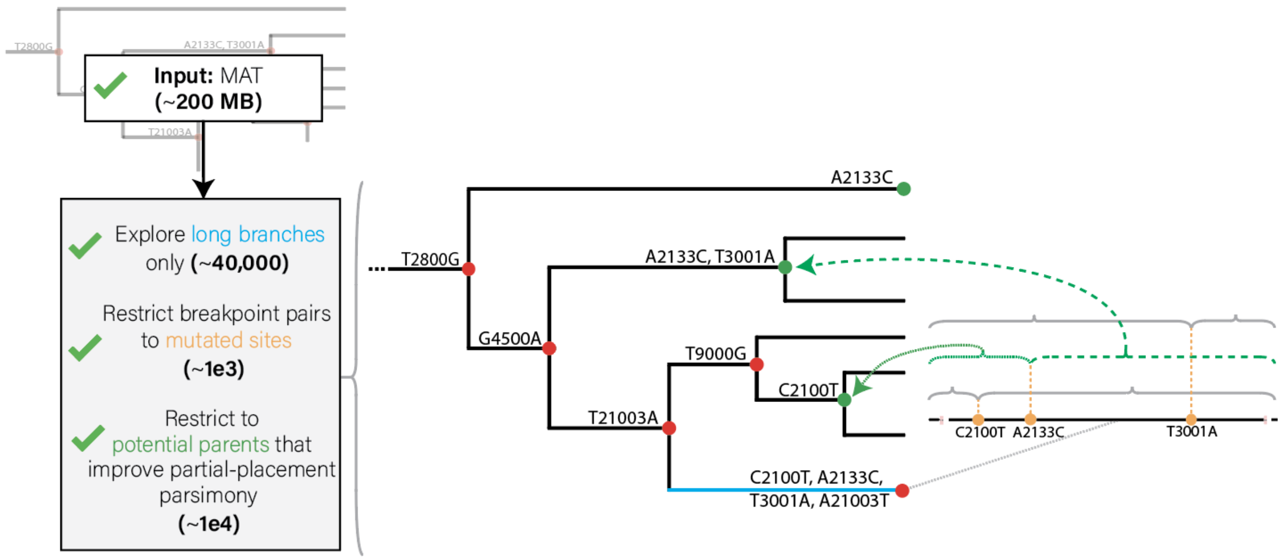
Explore all sequences as potential recombinant **(~3 million)**

Explore all sites for breakpoint pairs **(~1e9)**

Explore all pairs of potential parents **(~1e13)**

ATTAAAGG···TGG···AAA···TTG···GTG···CTT···GTG···ATCGAGTG
ATTAAAGG···TTG···AAA···TTG···GTG···CGT···GTG···ATCGAGTG
~30,000 bp
ATTAAAGG···TGGGTG···ATCGAGTG
AATAAAGG···AGCGTG···ATCGAGTG
ATTAGAGG···TGGGTG···ATCGAGTG
ATTAGAGG···TGGGTG···ATCGAGTG
ATTAAAGG···TGCGTG···ATCGAGTG
ACTAAAGG···TGGGTG···ATCGAGTG
ACTAAAGG···TGGGTG···ATCGAGTG
ACTAAAGG···TGGGTG···ATCGAGTG
ACTAAAGG···TGGGTG···ATCGAGTG
ACTAAAGG···TGGGTG···ATCGAGTG
ACTAAAGG···TGGGTG···ATCGAGTG
ACTAAAGG···TGGGTG···ATCGAGTG

Ours (RIPPLES)



Input: MAT **(~200 MB)**

Explore long branches only **(~40,000)**

Restrict breakpoint pairs to mutated sites **(~1e3)**

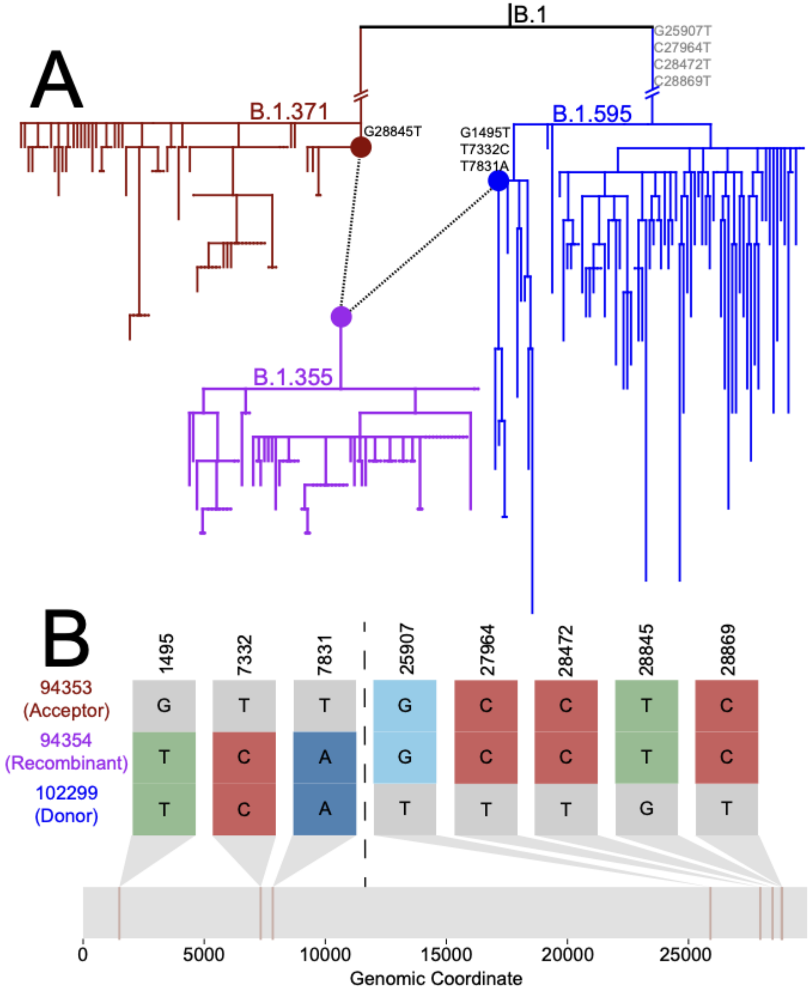
Restrict to potential parents that improve partial-placement parsimony **(~1e4)**

RIPPLES discovers >600 recombination events!

- **>600 unique** SARS-CoV-2 recombination events discovered from a 1.6M sample phylogeny!
  - ~2.7% sequences have a **detectable** recombinant ancestry

- This is the **largest recombination study** to our knowledge

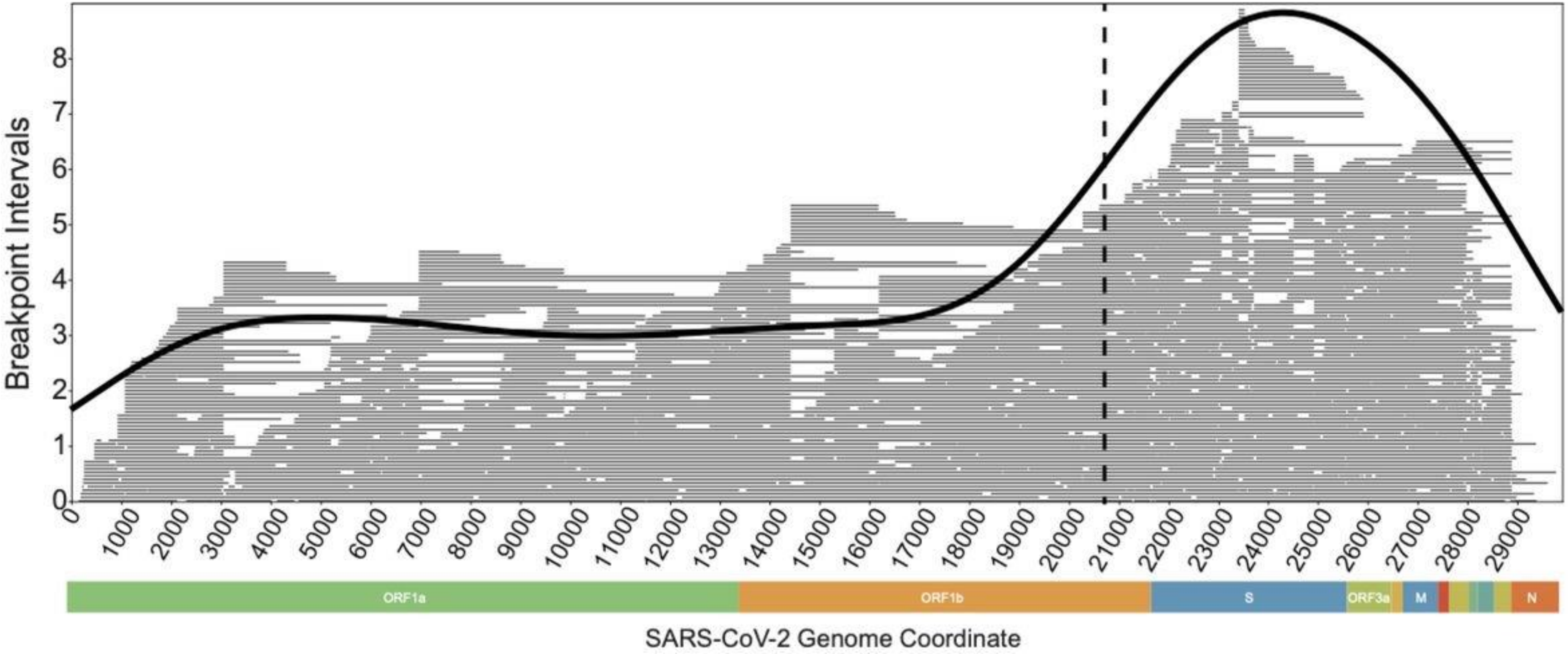
- With our latest RIPPLES software (`ripples-fast`), we can infer recombinants from a **~10M SARS-CoV-2 mutation-annotated tree** in **~2 hours!**



Cheng Ye

(Turakhia et al., bioRxiv 2021)

Recombination breakpoints are elevated in the SARS-CoV-2 Spike protein region



(Turakhia et al., bioRxiv 2021)

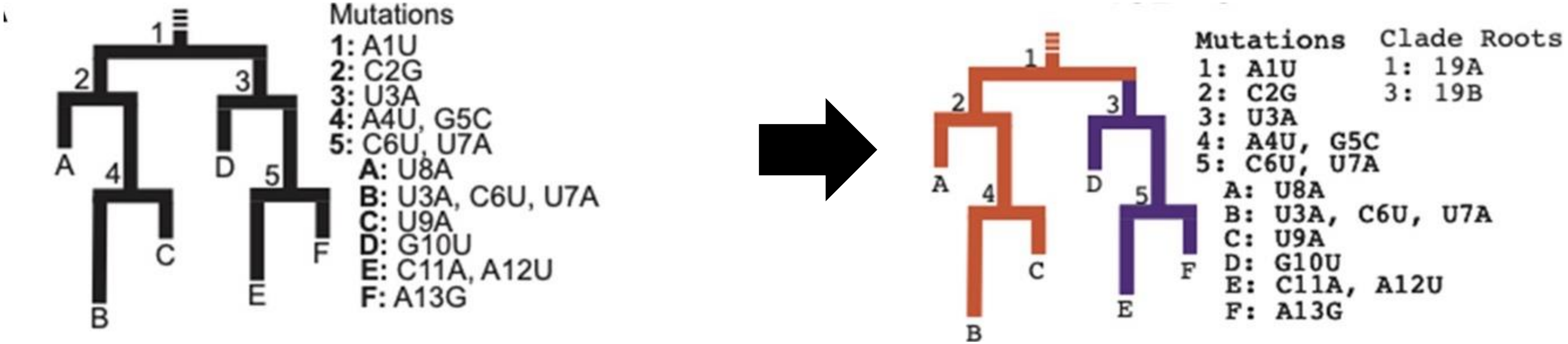
Overview of the UShER Package

- UShER: Phylogenetic placement
- matOptimize: Phylogenetic tree optimization
- RIPPLES: Find recombinant sequences using a phylogenomic approach
- **matUtils: Command-line tools for rapidly analyzing and interpreting SARS-CoV-2 mutation-annotated phylogenetic trees**

https://usher-wiki.readthedocs.io/

matUtils: Toolkit to rapidly query and interpret MATs

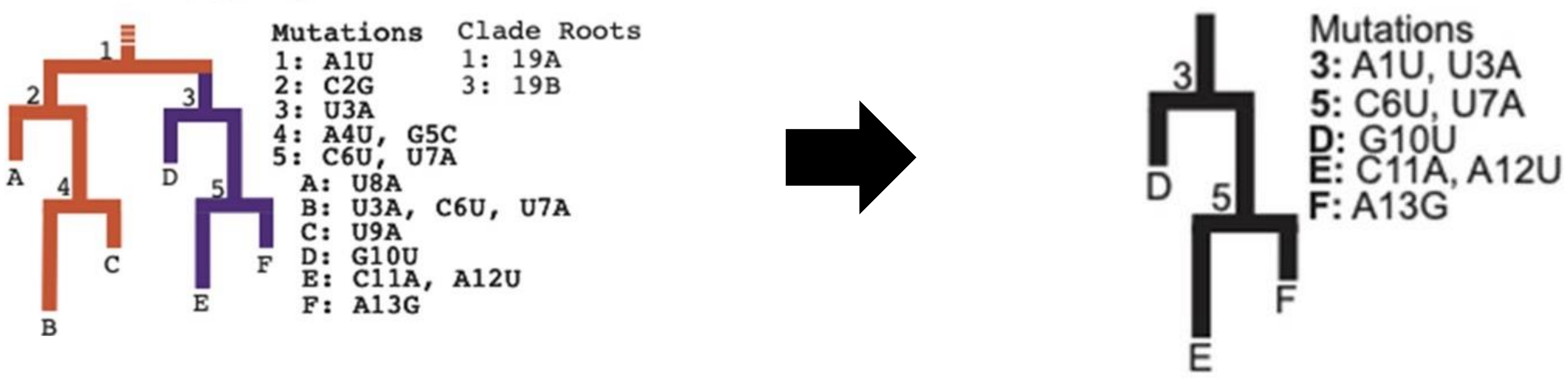
Subcommand	Arguments	Description
annotate	**--clade-to-nid (1: 19A, 3: 19B)**	**Assigns clades to nodes**
extract	--clade 19B	Extract subtree based on clade, mutation, branch length and other conditions
uncertainty	--find-epps	Output sample placement uncertainty metrics, for e.g., number of equally parsimonious placements
introduce	--population-samples (D: USA, E: USA, F:USA)	Identify internal nodes corresponding to introduction of one or more infection clusters in a geographic region of interest
summary	N/A	Output basic statistics of a MAT



(McBroome et al., MBE 2021)

matUtils: Toolkit to rapidly query and interpret MATs

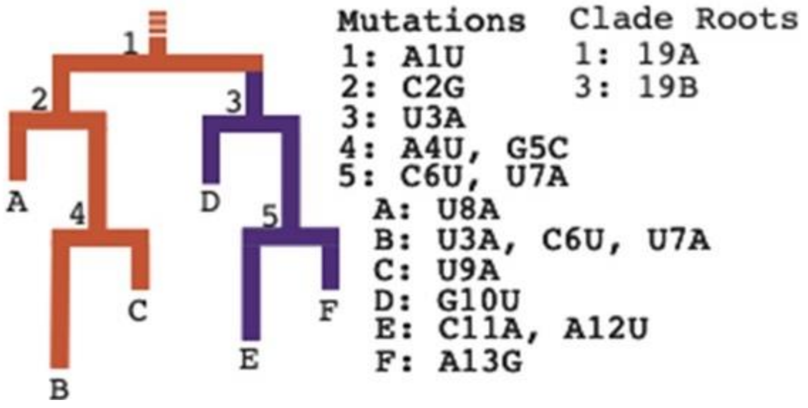
Subcommand	Arguments	Description
`annotate`	`--clade-to-nid (1: 19A, 3: 19B)`	Assigns clades to nodes
`extract`	**`--clade 19B`**	**Extract subtree based on clade, mutation, branch length and other conditions**
`uncertainty`	`--find-epps`	Output sample placement uncertainty metrics, for e.g., number of equally parsimonious placements
`introduce`	`--population-samples (D: USA, E: USA, F:USA)`	Identify internal nodes corresponding to introduction of one or more infection clusters in a geographic region of interest
`summary`	**`N/A`**	**Output basic statistics of a** MAT

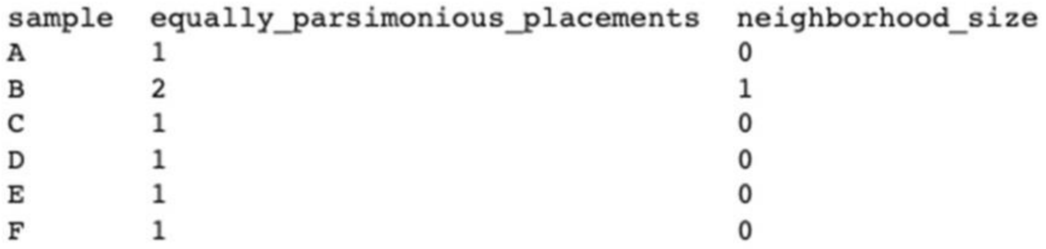


(McBroome et al., MBE 2021)

matUtils: Toolkit to rapidly query and interpret MATs

Subcommand	Arguments	Description
`annotate`	`--clade-to-nid (1: 19A, 3: 19B)`	Assigns clades to nodes
`extract`	`--clade 19B`	Extract subtree based on clade, mutation, branch length and other conditions
`uncertainty`	**`--find-epps`**	**Output sample placement uncertainty metrics, for e.g., number of equally parsimonious placements**
`introduce`	`--population-samples (D: USA, E: USA, F:USA)`	Identify internal nodes corresponding to introduction of one or more infection clusters in a geographic region of interest
`summary`	N/A	Output basic statistics of a MAT



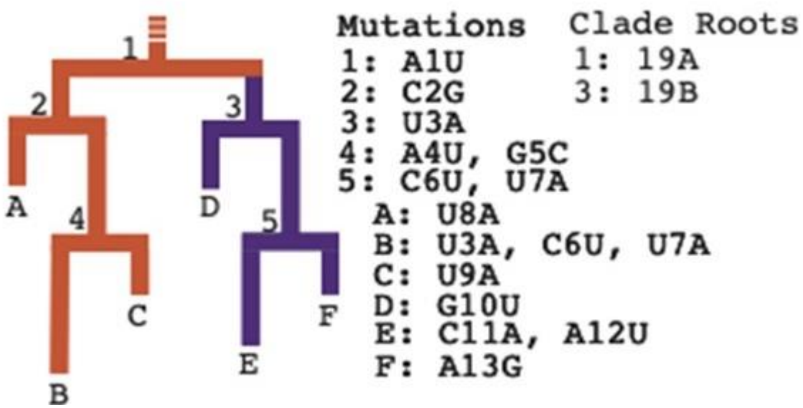


```
sample  equally_parsimonious_placements  neighborhood_size
A       1                                0
B       2                                1
C       1                                0
D       1                                0
E       1                                0
F       1                                0
```

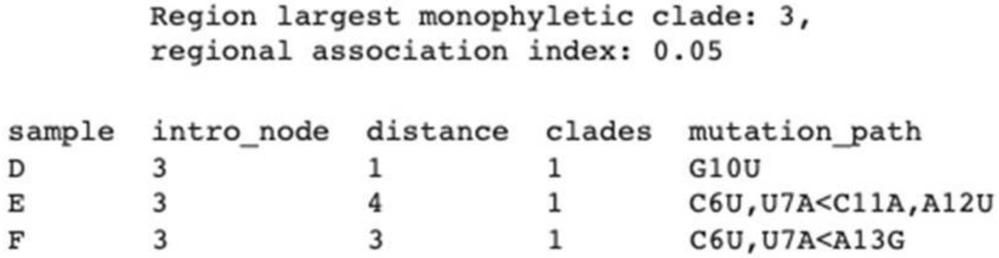
(McBroome et al., MBE 2021)

matUtils: Toolkit to rapidly query and interpret MATs

Subcommand	Arguments	Description
`annotate`	`--clade-to-nid (1: 19A, 3: 19B)`	Assigns clades to nodes
`extract`	`--clade 19B`	Extract subtree based on clade, mutation, branch length and other conditions
`uncertainty`	`--find-epps`	Output sample placement uncertainty metrics, for e.g., number of equally parsimonious placements
`introduce`	**`--population-samples (D: USA, E: USA, F:USA)`**	**Identify internal nodes corresponding to introduction of one or more infection clusters in a geographic region of interest**
`summary`	`N/A`	Output basic statistics of a MAT



```
Mutations     Clade Roots
1: A1U        1: 19A
2: C2G        3: 19B
3: U3A
4: A4U, G5C
5: C6U, U7A
  A: U8A
  B: U3A, C6U, U7A
  C: U9A
  D: G10U
  E: C11A, A12U
  F: A13G
```



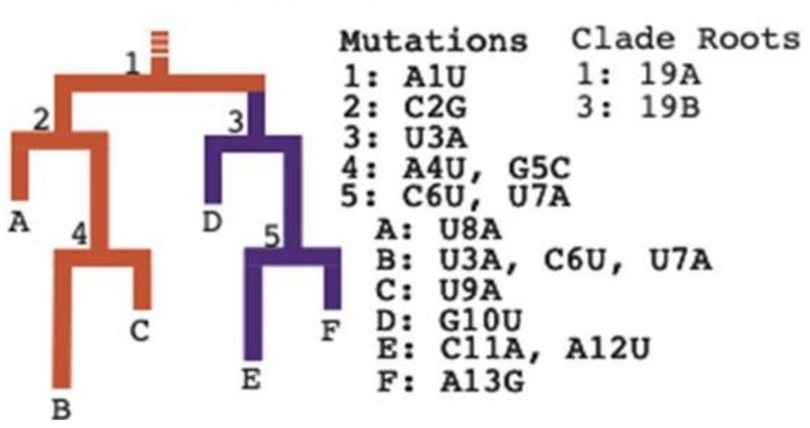
```
        Region largest monophyletic clade: 3,
        regional association index: 0.05

sample  intro_node  distance  clades  mutation_path
D       3           1         1       G10U
E       3           4         1       C6U,U7A<C11A,A12U
F       3           3         1       C6U,U7A<A13G
```

(McBroome et al., MBE 2021)

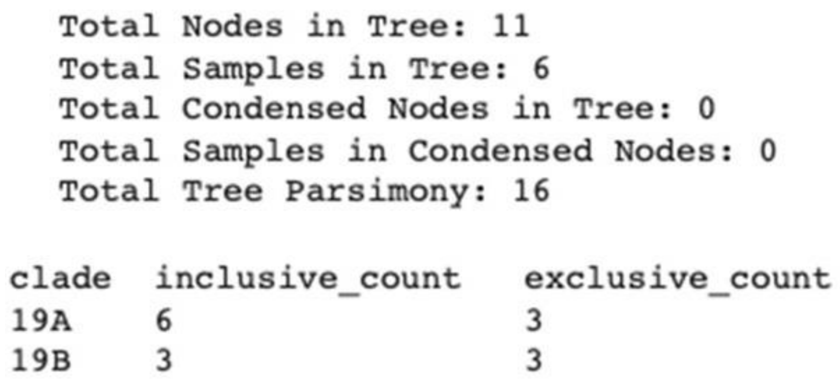
matUtils: Toolkit to rapidly query and interpret MATs

Subcommand	Arguments	Description
`annotate`	`--clade-to-nid (1: 19A, 3: 19B)`	Assigns clades to nodes
`extract`	`--clade 19B`	Extract subtree based on clade, mutation, branch length and other conditions
`uncertainty`	`--find-epps`	Output sample placement uncertainty metrics, for e.g., number of equally parsimonious placements
`introduce`	`--population-samples (D: USA, E: USA, F:USA)`	Identify internal nodes corresponding to introduction of one or more infection clusters in a geographic region of interest
`summary`	**N/A**	**Output basic statistics of a MAT**



Mutations
1: A1U
2: C2G
3: U3A
4: A4U, G5C
5: C6U, U7A
A: U8A
B: U3A, C6U, U7A
C: U9A
D: G10U
E: C11A, A12U
F: A13G

Clade Roots
1: 19A
3: 19B



```
    Total Nodes in Tree: 11
    Total Samples in Tree: 6
    Total Condensed Nodes in Tree: 0
    Total Samples in Condensed Nodes: 0
    Total Tree Parsimony: 16

clade   inclusive_count     exclusive_count
19A     6                   3
19B     3                   3
```

(McBroome et al., MBE 2021)

matUtils is fast!

On **~1M-sample** SARS-CoV-2 phylogeny:

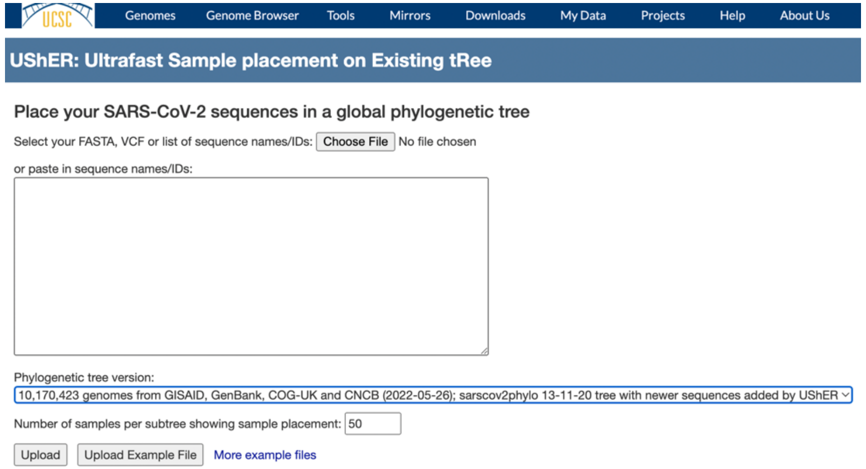
- ~5 sec to compute **summary** statistics
- ~5 sec to **extract** a subtree of specified samples
- ~15 sec to **extract** mutation paths from root to every sample in the tree
- ~9 sec to **resolve** all polytomies
  - Takes ~37 min using `ape`
- ~1 min to identify **introductions**

Jacob McBroome, UCSC

(McBroome et al., MBE 2021)

Real-world impact

UShER added to UCSC Genome Browser



UShER: Ultrafast Sample placement on Existing tRee

Place your SARS-CoV-2 sequences in a global phylogenetic tree

Select your FASTA, VCF or list of sequence names/IDs: Choose File No file chosen

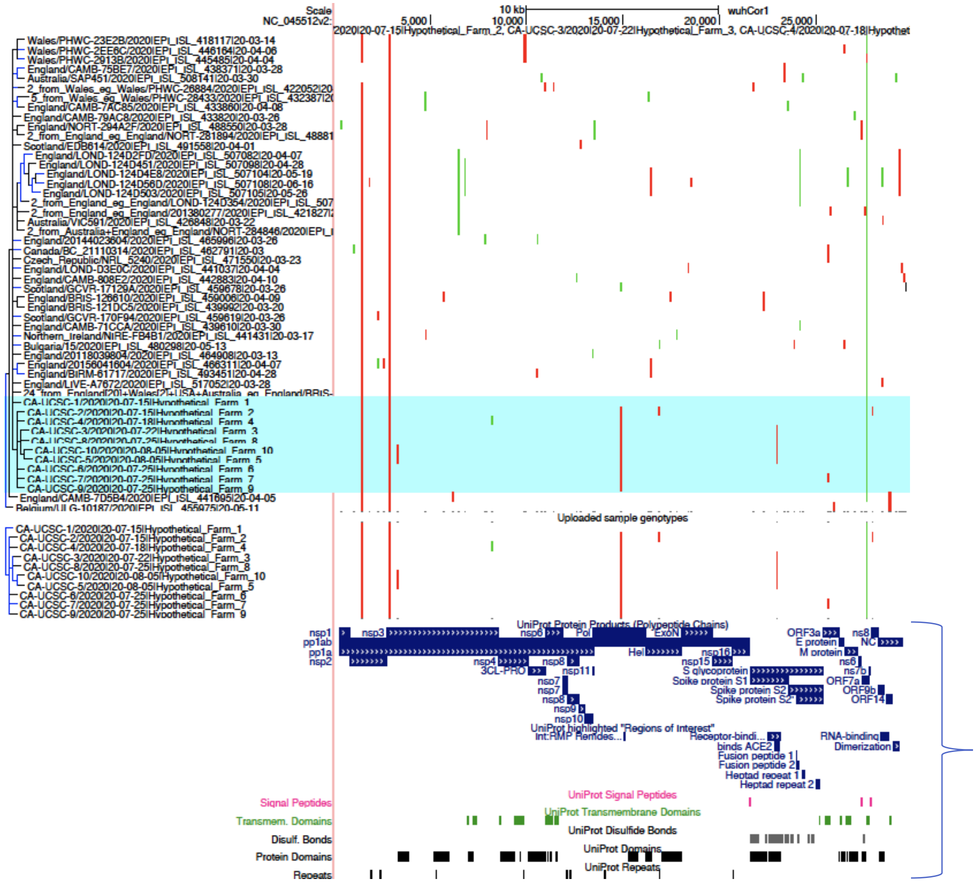
or paste in sequence names/IDs:

Phylogenetic tree version:

10,170,423 genomes from GISAID, GenBank, COG-UK and CNCB (2022-05-26); sarscov2phylo 13-11-20 tree with newer sequences added by UShER

Number of samples per subtree showing sample placement: 50

Upload | Upload Example File | More example files

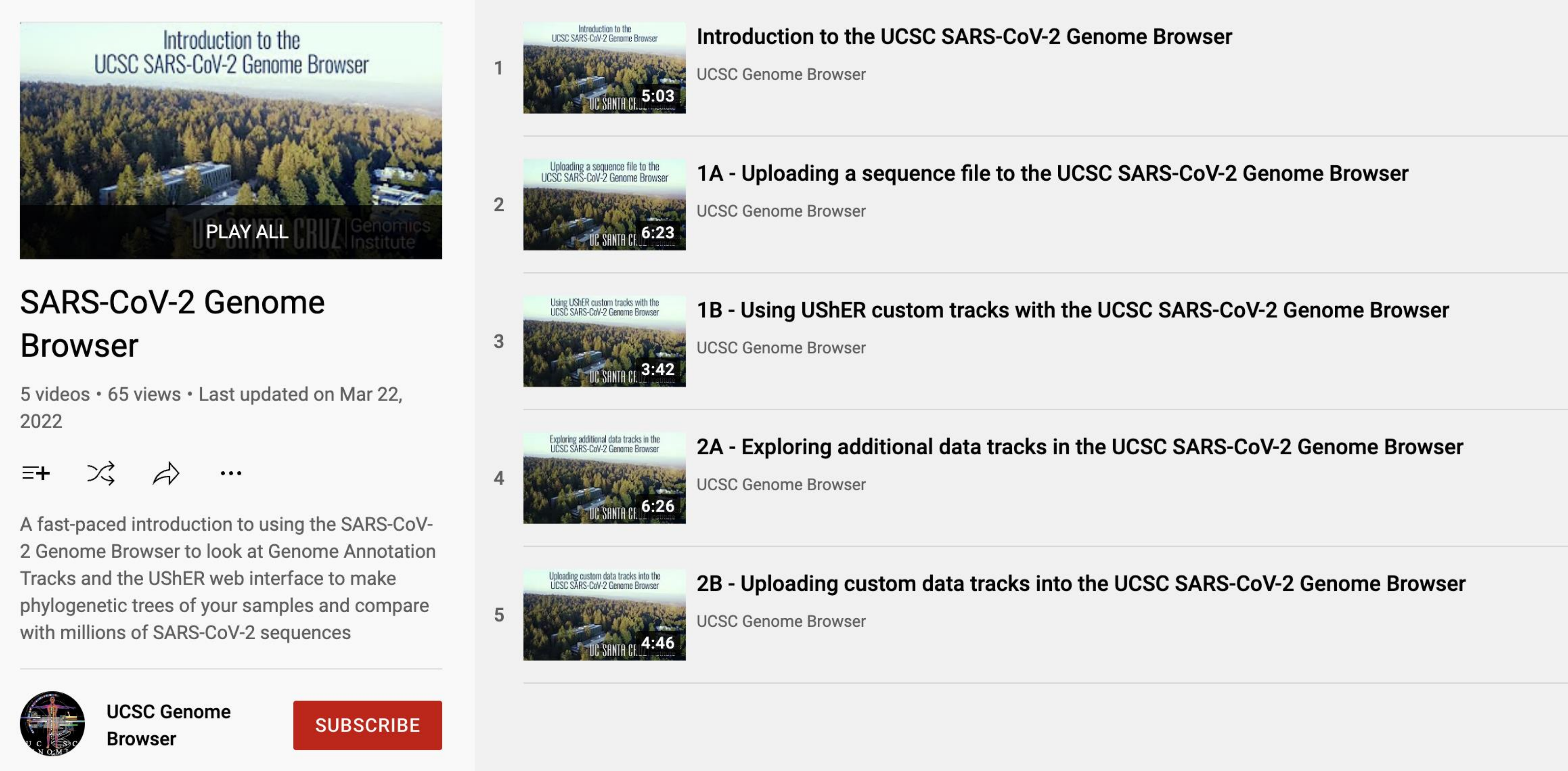

Scale
10 kb
wuhCor1
NC_045512v2:
5,000
10,000
15,000
20,000
25,000
Uploaded sample genotypes
CA-UCSC-1/2020|20-07-15|Hypothetical_Farm_1
CA-UCSC-2/2020|20-07-15|Hypothetical_Farm_2
CA-UCSC-4/2020|20-07-18|Hypothetical_Farm_4
CA-UCSC-3/2020|20-07-22|Hypothetical_Farm_3
CA-UCSC-8/2020|20-07-25|Hypothetical_Farm_8
CA-UCSC-10/2020|20-08-05|Hypothetical_Farm_10
CA-UCSC-5/2020|20-08-05|Hypothetical_Farm_5
CA-UCSC-6/2020|20-07-25|Hypothetical_Farm_6
CA-UCSC-7/2020|20-07-25|Hypothetical_Farm_7
CA-UCSC-9/2020|20-07-25|Hypothetical_Farm_9
UniProt Protein Products (Polypeptide Chains)
nsp1
pp1ab
pp1a
nsp2
nsp3
Pol
ExoN
Hel
nsp16
3CL-PRO
S glycoprotein
Spike protein S1
Spike protein S2
ORF3a
E protein
M protein
ns6
ns7b
ORF7a
ORF9b
ORF14
ns8
NC
UniProt highlighted "Regions of Interest"
Receptor-bindi...
binds ACE2
Fusion peptide 1
Fusion peptide 2
Heptad repeat 1
Heptad repeat 2
RNA-binding
Dimerization
UniProt Signal Peptides
UniProt Transmembrane Domains
UniProt Disulfide Bonds
UniProt Domains
UniProt Repeats
Signal Peptides
Transmem. Domains
Disulf. Bonds
Protein Domains
Repeats

Rapid cross-referencing with 50+ molecular and structural biology tracks

Angie Hinrichs, UCSC

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

YouTube Tutorials on the UCSC SARS-CoV-2 Genome Browser



Introduction to the UCSC SARS-CoV-2 Genome Browser
UC SANTA CRUZ Genomics Institute
PLAY ALL

SARS-CoV-2 Genome Browser

5 videos • 65 views • Last updated on Mar 22, 2022

A fast-paced introduction to using the SARS-CoV-2 Genome Browser to look at Genome Annotation Tracks and the UShER web interface to make phylogenetic trees of your samples and compare with millions of SARS-CoV-2 sequences

UCSC Genome Browser

SUBSCRIBE

1 5:03 **Introduction to the UCSC SARS-CoV-2 Genome Browser**
UCSC Genome Browser

2 6:23 **1A - Uploading a sequence file to the UCSC SARS-CoV-2 Genome Browser**
UCSC Genome Browser

3 3:42 **1B - Using UShER custom tracks with the UCSC SARS-CoV-2 Genome Browser**
UCSC Genome Browser

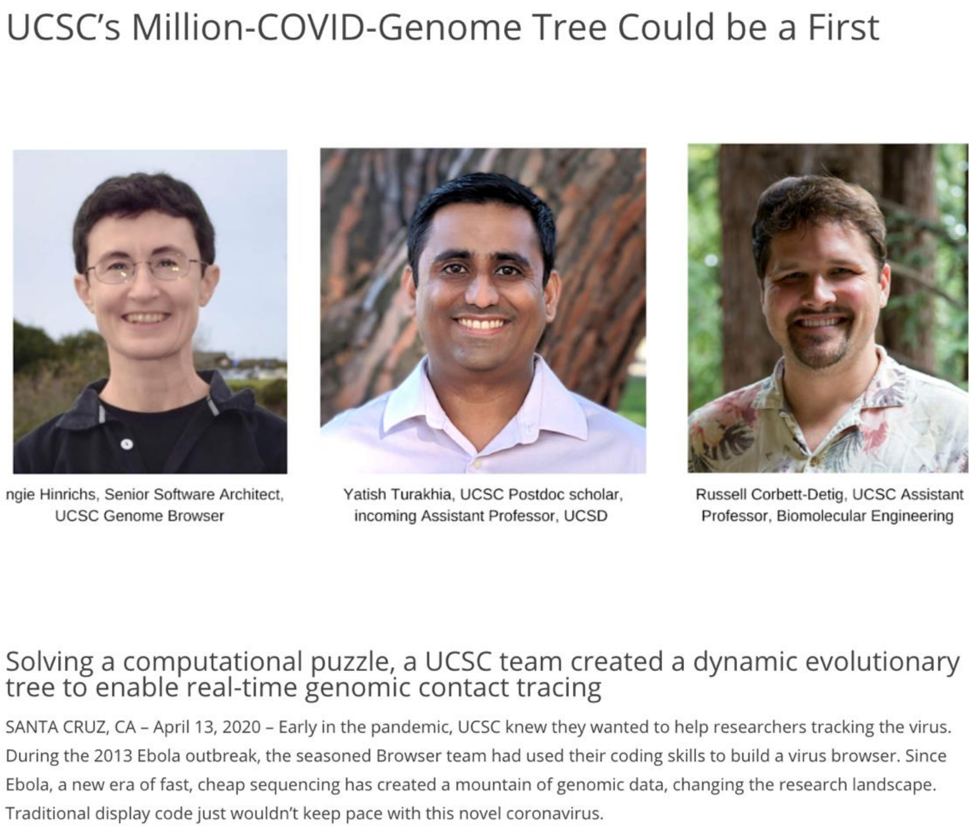
4 6:26 **2A - Exploring additional data tracks in the UCSC SARS-CoV-2 Genome Browser**
UCSC Genome Browser

5 4:46 **2B - Uploading custom data tracks into the UCSC SARS-CoV-2 Genome Browser**
UCSC Genome Browser

https://www.youtube.com/playlist?list=PL5G1tMPaNuswZM4zAkmS6F7q1G-7D-PI-

Largest SARS-CoV-2 phylogenies

Covered by UCSC news and
Santa Cruz Tech Beat



UCSC's Million-COVID-Genome Tree Could be a First

ngie Hinrichs, Senior Software Architect, UCSC Genome Browser

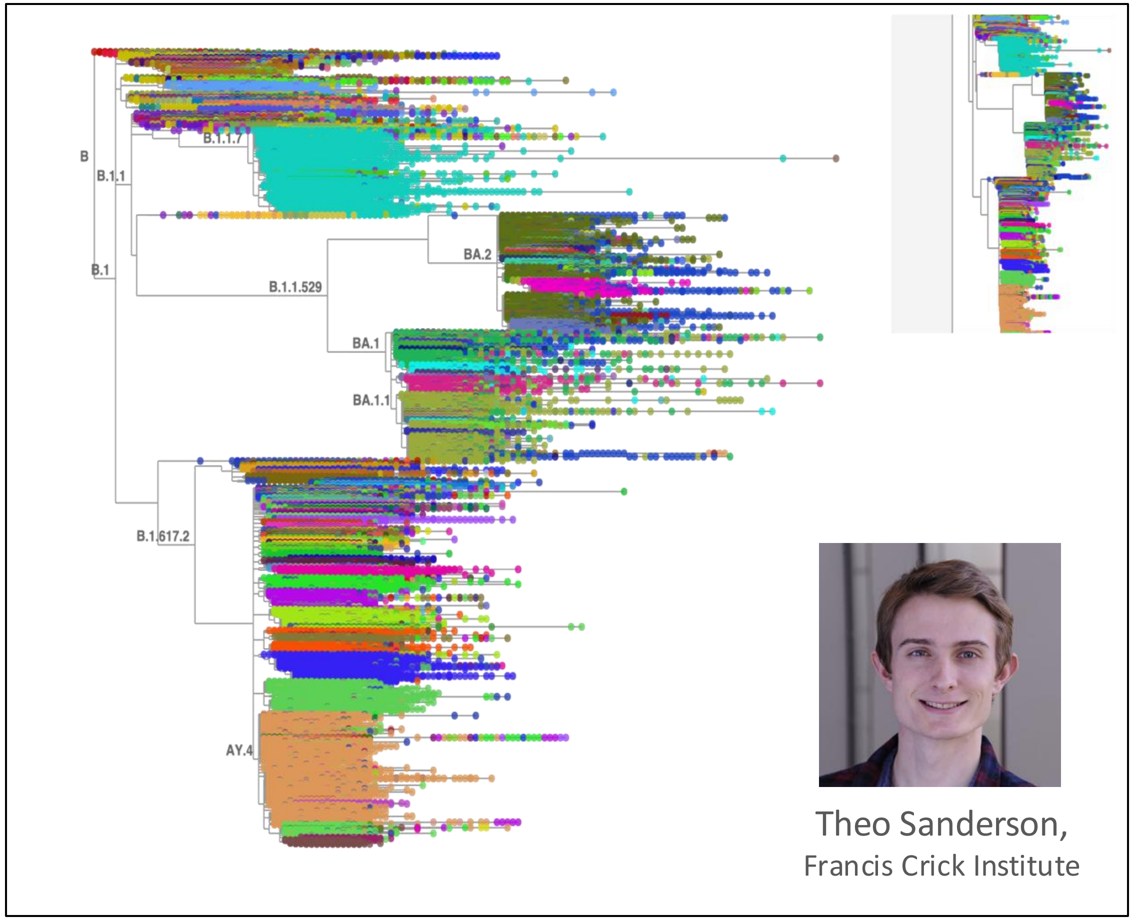
Yatish Turakhia, UCSC Postdoc scholar, incoming Assistant Professor, UCSD

Russell Corbett-Detig, UCSC Assistant Professor, Biomolecular Engineering

Solving a computational puzzle, a UCSC team created a dynamic evolutionary tree to enable real-time genomic contact tracing

SANTA CRUZ, CA – April 13, 2020 – Early in the pandemic, UCSC knew they wanted to help researchers tracking the virus. During the 2013 Ebola outbreak, the seasoned Browser team had used their coding skills to build a virus browser. Since Ebola, a new era of fast, cheap sequencing has created a mountain of genomic data, changing the research landscape. Traditional display code just wouldn't keep pace with this novel coronavirus.

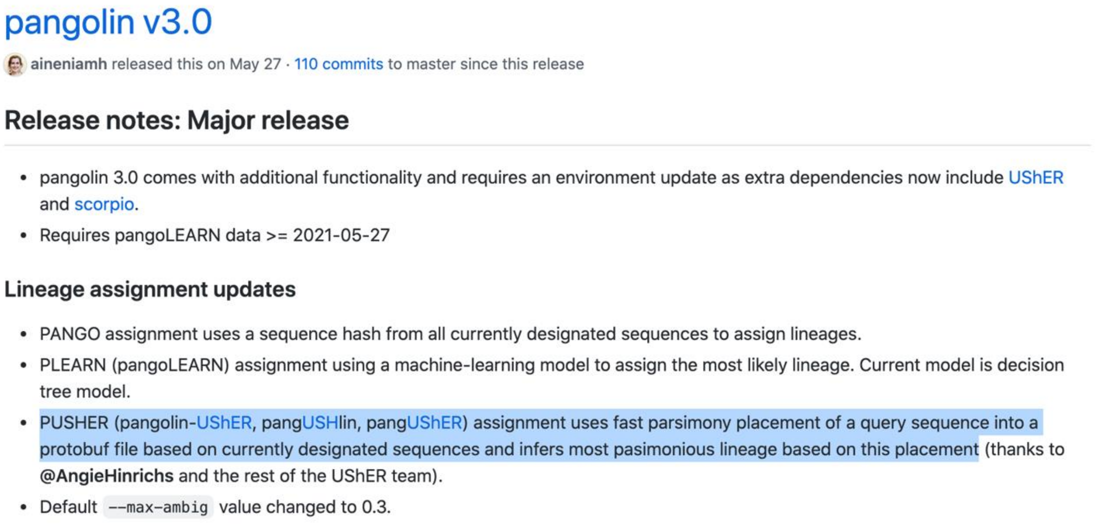
Taxonium view of >10M SARS-CoV-2 sample phylogeny



Theo Sanderson,
Francis Crick Institute

Public version: http://hgdownload.soe.ucsc.edu/goldenPath/wuhCor1/UShER_SARS-CoV-2/

UShER default engine in Pangolin SARS-CoV-2 lineage assigner



pangolin v3.0

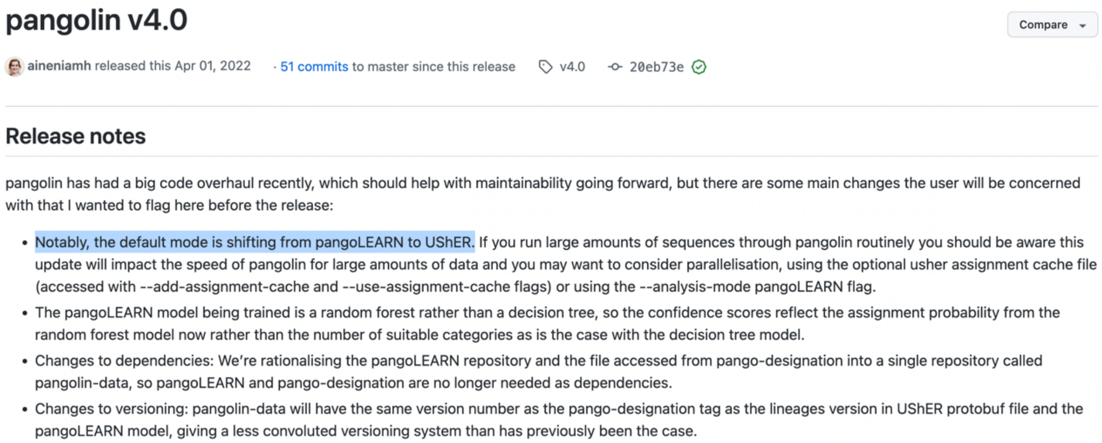
aineniamh released this on May 27 · 110 commits to master since this release

Release notes: Major release

- pangolin 3.0 comes with additional functionality and requires an environment update as extra dependencies now include UShER and scorpio.
- Requires pangoLEARN data >= 2021-05-27

Lineage assignment updates

- PANGO assignment uses a sequence hash from all currently designated sequences to assign lineages.
- PLEARN (pangoLEARN) assignment using a machine-learning model to assign the most likely lineage. Current model is decision tree model.
- PUSHER (pangolin-UShER, pangUSHlin, pangUShER) assignment uses fast parsimony placement of a query sequence into a protobuf file based on currently designated sequences and infers most pasimonious lineage based on this placement (thanks to **@AngieHinrichs** and the rest of the UShER team).
- Default `--max-ambig` value changed to 0.3.



pangolin v4.0

aineniamh released this Apr 01, 2022 · 51 commits to master since this release v4.0 20eb73e

Release notes

pangolin has had a big code overhaul recently, which should help with maintainability going forward, but there are some main changes the user will be concerned with that I wanted to flag here before the release:

- Notably, the default mode is shifting from pangoLEARN to UShER. If you run large amounts of sequences through pangolin routinely you should be aware this update will impact the speed of pangolin for large amounts of data and you may want to consider parallelisation, using the optional usher assignment cache file (accessed with --add-assignment-cache and --use-assignment-cache flags) or using the --analysis-mode pangoLEARN flag.
- The pangoLEARN model being trained is a random forest rather than a decision tree, so the confidence scores reflect the assignment probability from the random forest model now rather than the number of suitable categories as is the case with the decision tree model.
- Changes to dependencies: We're rationalising the pangoLEARN repository and the file accessed from pango-designation into a single repository called pangolin-data, so pangoLEARN and pango-designation are no longer needed as dependencies.
- Changes to versioning: pangolin-data will have the same version number as the pango-designation tag as the lineages version in UShER protobuf file and the pangoLEARN model, giving a less convoluted versioning system than has previously been the case.

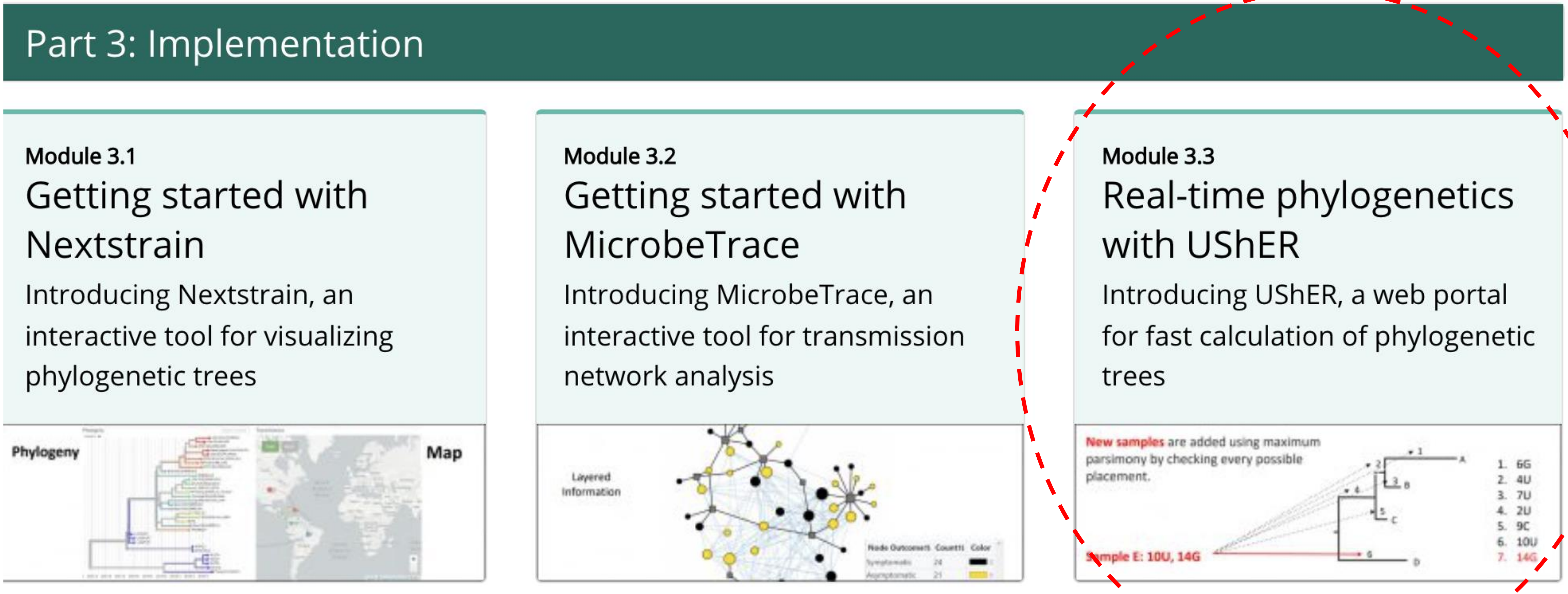
Angie Hinrichs, UCSC

Thanks to:

- Áine O'Toole
- Emily Scher
- Rachel Colquhoun
- Andrew Rambaut

https://github.com/cov-lineages/pangolin/

UShER included in the CDC COVID-19 Genomic Epidemiology Toolkit



Part 3: Implementation

Module 3.1

Getting started with Nextstrain

Introducing Nextstrain, an interactive tool for visualizing phylogenetic trees

Module 3.2

Getting started with MicrobeTrace

Introducing MicrobeTrace, an interactive tool for transmission network analysis

Module 3.3

Real-time phylogenetics with UShER

Introducing UShER, a web portal for fast calculation of phylogenetic trees

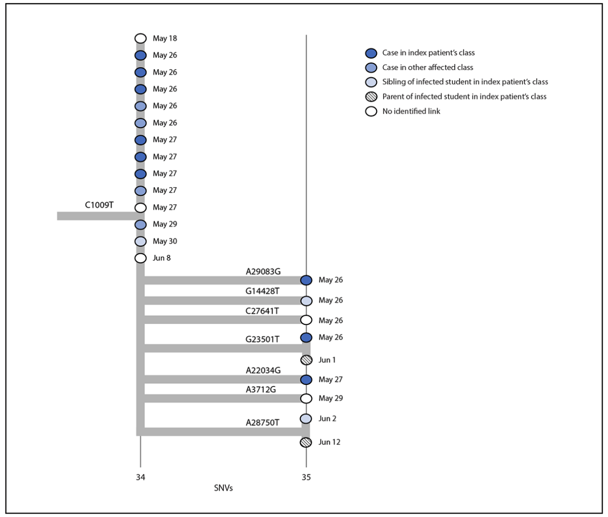
https://www.cdc.gov/amd/training/covid-19-gen-epi-toolkit.html

Delta outbreak analysis at a Calif. elementary school used UShER

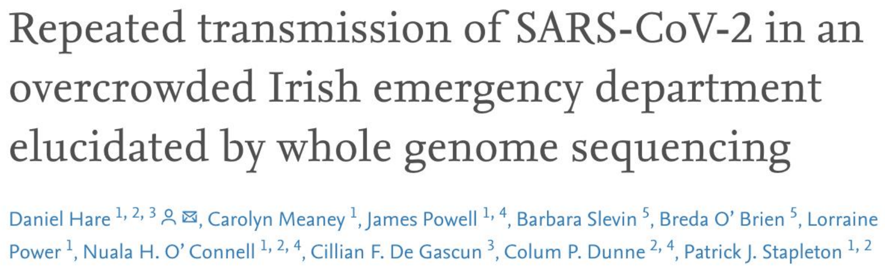
The New York Times

How the Delta Variant Infiltrated an Elementary School Classroom

A detailed study in California found that the variant easily spread from an unvaccinated teacher to children and, in a few cases,

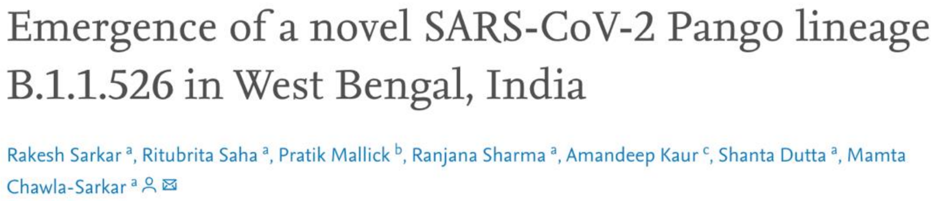


UShER has been used for outbreak analysis in several parts of the world (including LMIC)



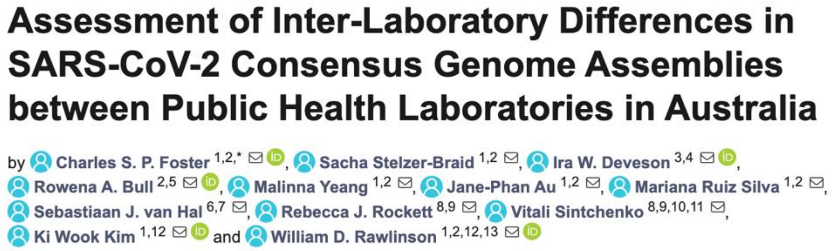
Repeated transmission of SARS-CoV-2 in an overcrowded Irish emergency department elucidated by whole genome sequencing

Daniel Hare [1, 2, 3], Carolyn Meaney [1], James Powell [1, 4], Barbara Slevin [5], Breda O' Brien [5], Lorraine Power [1], Nuala H. O' Connell [1, 2, 4], Cillian F. De Gascun [3], Colum P. Dunne [2, 4], Patrick J. Stapleton [1, 2]



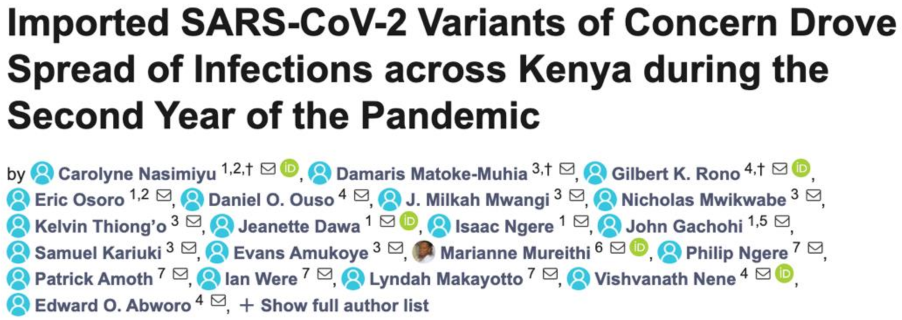
Emergence of a novel SARS-CoV-2 Pango lineage B.1.1.526 in West Bengal, India

Rakesh Sarkar [a], Ritubrita Saha [a], Pratik Mallick [b], Ranjana Sharma [a], Amandeep Kaur [c], Shanta Dutta [a], Mamta Chawla-Sarkar [a]



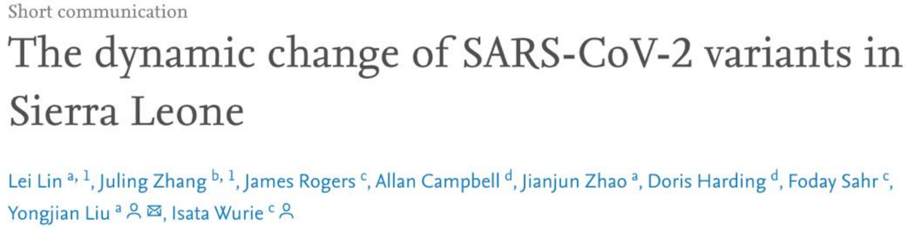
Assessment of Inter-Laboratory Differences in SARS-CoV-2 Consensus Genome Assemblies between Public Health Laboratories in Australia

by Charles S. P. Foster [1,2,*], Sacha Stelzer-Braid [1,2], Ira W. Deveson [3,4], Rowena A. Bull [2,5], Malinna Yeang [1,2], Jane-Phan Au [1,2], Mariana Ruiz Silva [1,2], Sebastiaan J. van Hal [6,7], Rebecca J. Rockett [8,9], Vitali Sintchenko [8,9,10,11], Ki Wook Kim [1,12] and William D. Rawlinson [1,2,12,13]



Imported SARS-CoV-2 Variants of Concern Drove Spread of Infections across Kenya during the Second Year of the Pandemic

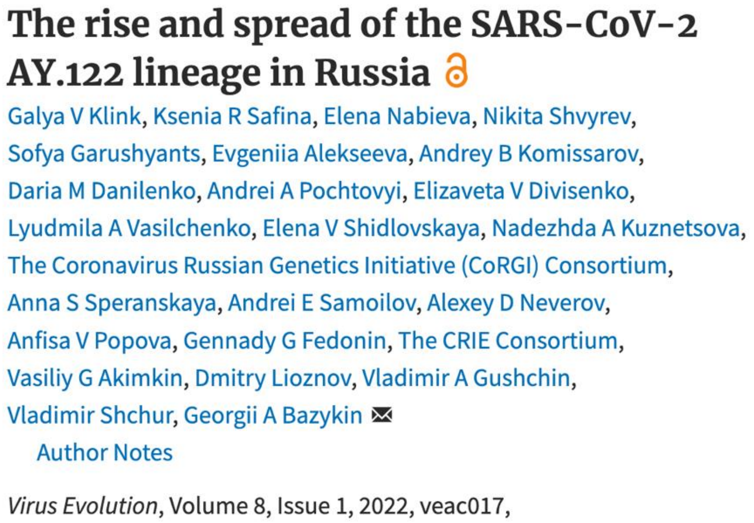
by Carolyne Nasimiyu [1,2,†], Damaris Matoke-Muhia [3,†], Gilbert K. Rono [4,†], Eric Osoro [1,2], Daniel O. Ouso [4], J. Milkah Mwangi [3], Nicholas Mwikwabe [3], Kelvin Thiong'o [3], Jeanette Dawa [1], Isaac Ngere [1], John Gachohi [1,5], Samuel Kariuki [3], Evans Amukoye [3], Marianne Mureithi [6], Philip Ngere [7], Patrick Amoth [7], Ian Were [7], Lyndah Makayotto [7], Vishvanath Nene [4], Edward O. Abworo [4], + Show full author list



Short communication

The dynamic change of SARS-CoV-2 variants in Sierra Leone

Lei Lin [a, 1], Juling Zhang [b, 1], James Rogers [c], Allan Campbell [d], Jianjun Zhao [a], Doris Harding [d], Foday Sahr [c], Yongjian Liu [a], Isata Wurie [c]

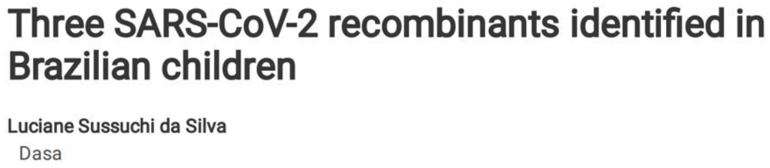


The rise and spread of the SARS-CoV-2 AY.122 lineage in Russia

Galya V Klink, Ksenia R Safina, Elena Nabieva, Nikita Shvyrev, Sofya Garushyants, Evgeniia Alekseeva, Andrey B Komissarov, Daria M Danilenko, Andrei A Pochtovyi, Elizaveta V Divisenko, Lyudmila A Vasilchenko, Elena V Shidlovskaya, Nadezhda A Kuznetsova, The Coronavirus Russian Genetics Initiative (CoRGI) Consortium, Anna S Speranskaya, Andrei E Samoilov, Alexey D Neverov, Anfisa V Popova, Gennady G Fedonin, The CRIE Consortium, Vasiliy G Akimkin, Dmitry Lioznov, Vladimir A Gushchin, Vladimir Shchur, Georgii A Bazykin

Author Notes

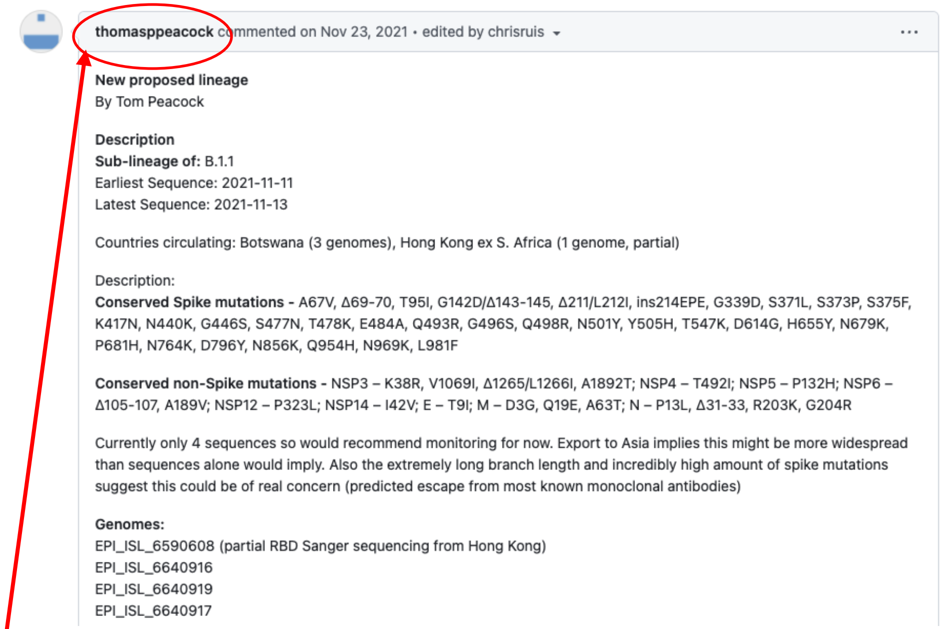
Virus Evolution, Volume 8, Issue 1, 2022, veac017,



Three SARS-CoV-2 recombinants identified in Brazilian children

Luciane Sussuchi da Silva
Dasa

First flagging of Omicron (B.1.1.529) variant



thomasppeacock commented on Nov 23, 2021 • edited by chrisruis

New proposed lineage
By Tom Peacock

Description
Sub-lineage of: B.1.1
Earliest Sequence: 2021-11-11
Latest Sequence: 2021-11-13

Countries circulating: Botswana (3 genomes), Hong Kong ex S. Africa (1 genome, partial)

Description:
Conserved Spike mutations - A67V, Δ69-70, T95I, G142D/Δ143-145, Δ211/L212I, ins214EPE, G339D, S371L, S373P, S375F, K417N, N440K, G446S, S477N, T478K, E484A, Q493R, G496S, Q498R, N501Y, Y505H, T547K, D614G, H655Y, N679K, P681H, N764K, D796Y, N856K, Q954H, N969K, L981F

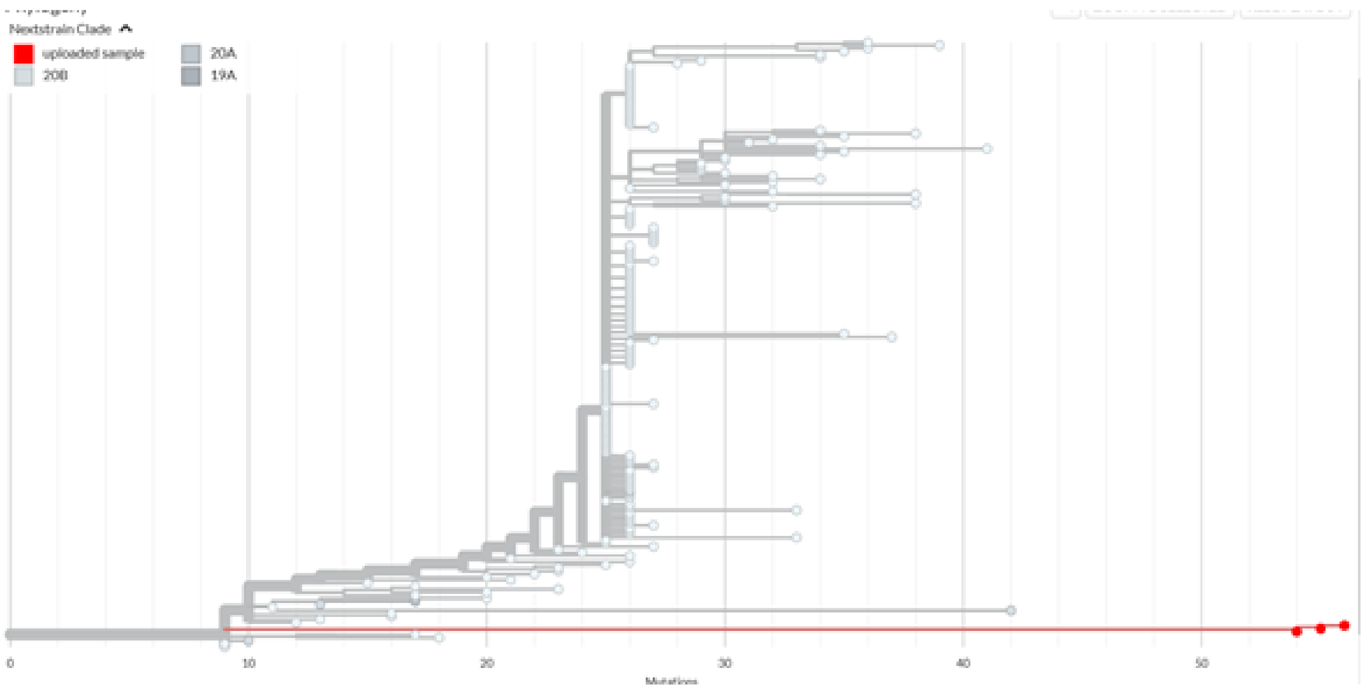
Conserved non-Spike mutations - NSP3 – K38R, V1069I, Δ1265/L1266I, A1892T; NSP4 – T492I; NSP5 – P132H; NSP6 – Δ105-107, A189V; NSP12 – P323L; NSP14 – I42V; E – T9I; M – D3G, Q19E, A63T; N – P13L, Δ31-33, R203K, G204R

Currently only 4 sequences so would recommend monitoring for now. Export to Asia implies this might be more widespread than sequences alone would imply. Also the extremely long branch length and incredibly high amount of spike mutations suggest this could be of real concern (predicted escape from most known monoclonal antibodies)

Genomes:
EPI_ISL_6590608 (partial RBD Sanger sequencing from Hong Kong)
EPI_ISL_6640916
EPI_ISL_6640919
EPI_ISL_6640917

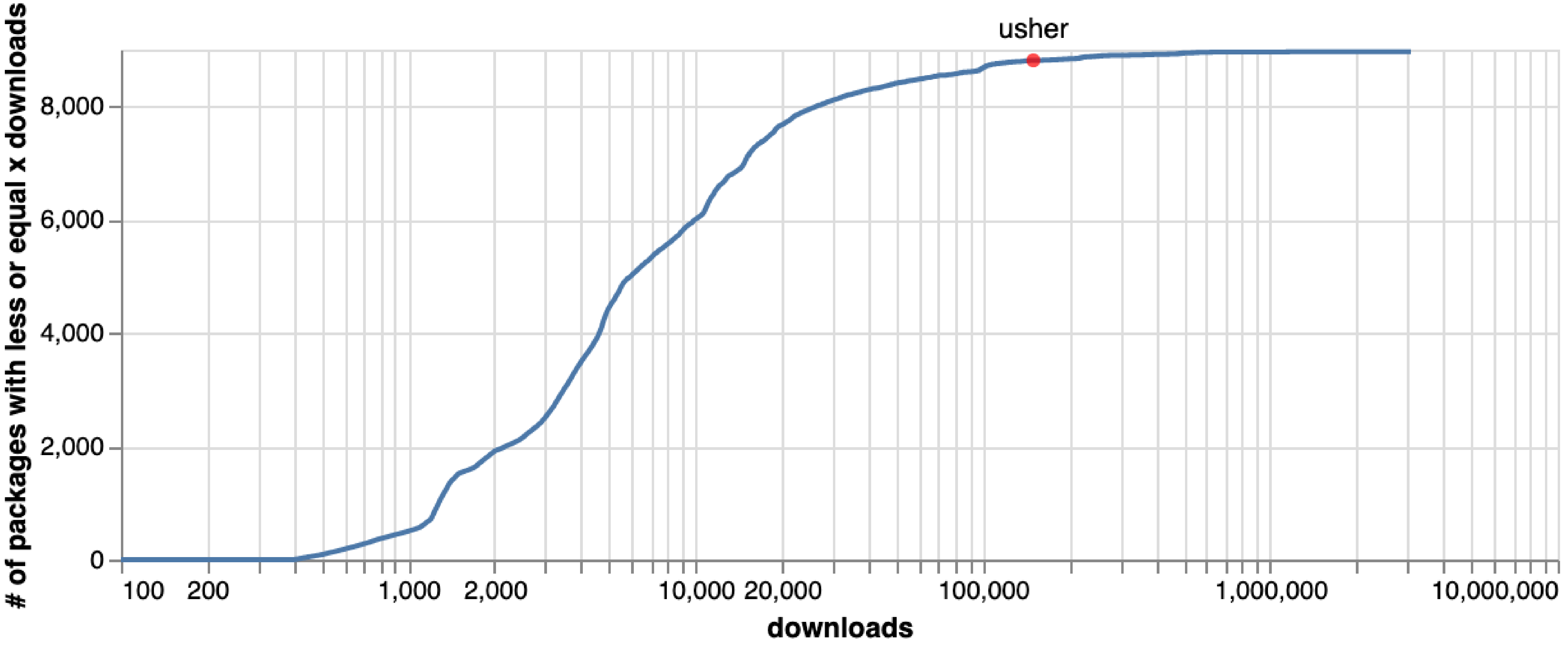
Tom Peacock, Imperial College

UShER-based phylogenetic analysis with the first Omicron sequences in red



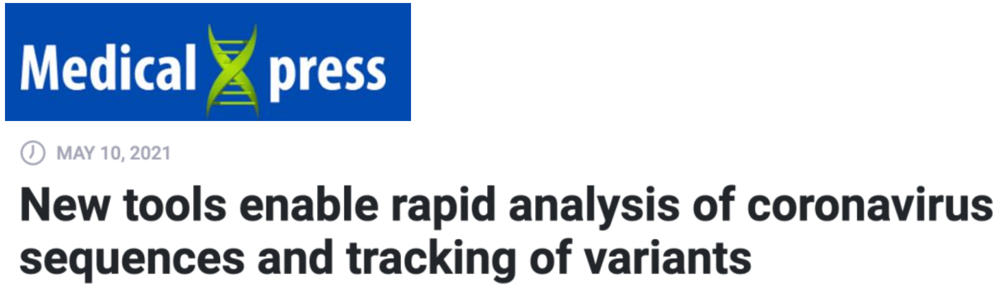
https://github.com/cov-lineages/pango-designation/issues/343

150K+ downloads on bioconda



https://bioconda.github.io/recipes/usher/README.html

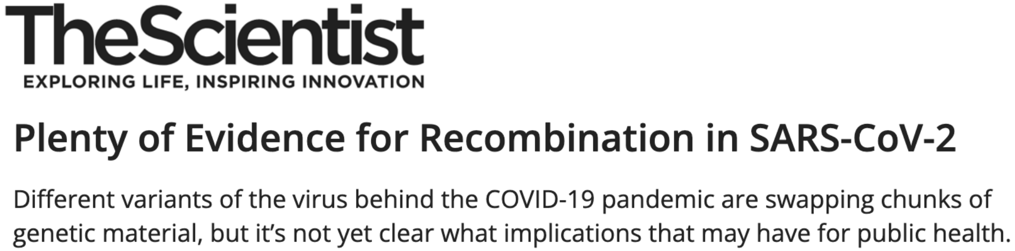
Press coverage



Medical press

MAY 10, 2021

New tools enable rapid analysis of coronavirus sequences and tracking of variants

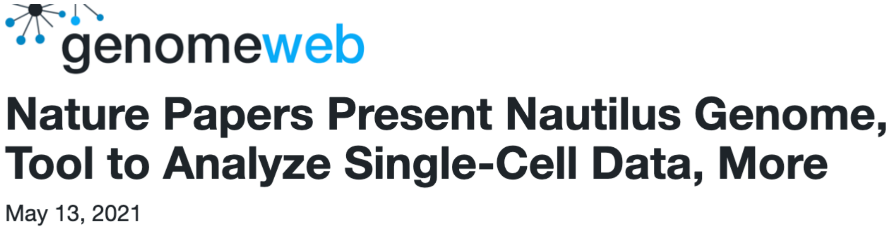


TheScientist

EXPLORING LIFE, INSPIRING INNOVATION

Plenty of Evidence for Recombination in SARS-CoV-2

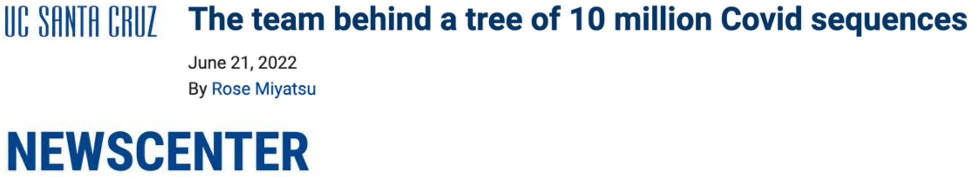
Different variants of the virus behind the COVID-19 pandemic are swapping chunks of genetic material, but it's not yet clear what implications that may have for public health.



genomeweb

Nature Papers Present Nautilus Genome, Tool to Analyze Single-Cell Data, More

May 13, 2021

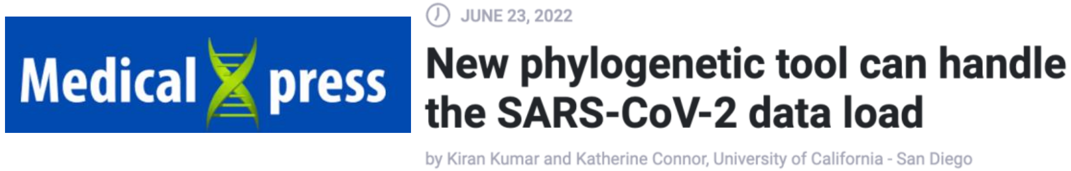


UC SANTA CRUZ

The team behind a tree of 10 million Covid sequences

June 21, 2022
By Rose Miyatsu

NEWSCENTER

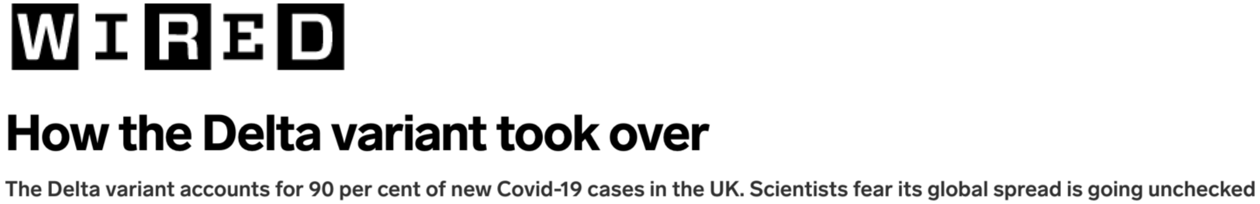


Medical press

JUNE 23, 2022

New phylogenetic tool can handle the SARS-CoV-2 data load

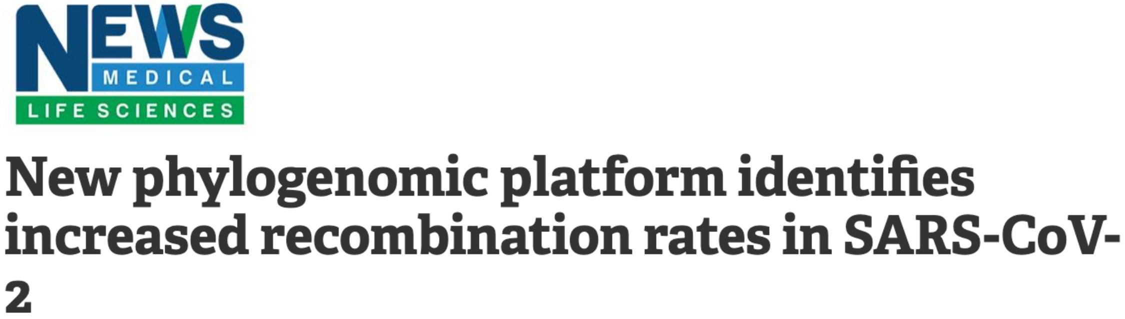
by Kiran Kumar and Katherine Connor, University of California - San Diego



WIRED

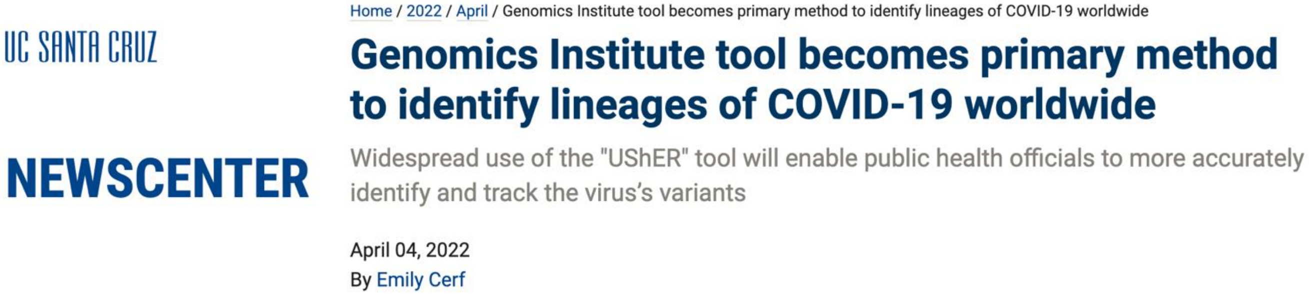
How the Delta variant took over

The Delta variant accounts for 90 per cent of new Covid-19 cases in the UK. Scientists fear its global spread is going unchecked



NEWS MEDICAL LIFE SCIENCES

New phylogenomic platform identifies increased recombination rates in SARS-CoV-2



UC SANTA CRUZ

NEWSCENTER

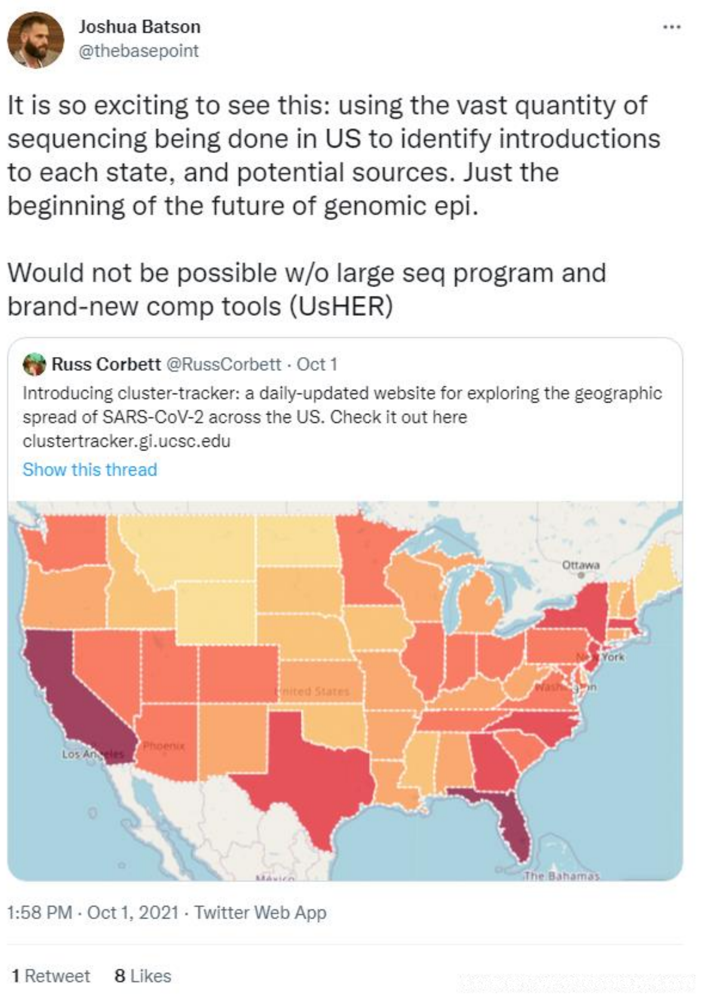
Home / 2022 / April / Genomics Institute tool becomes primary method to identify lineages of COVID-19 worldwide

Genomics Institute tool becomes primary method to identify lineages of COVID-19 worldwide

Widespread use of the "UShER" tool will enable public health officials to more accurately identify and track the virus's variants

April 04, 2022
By Emily Cerf

Community response



Joshua Batson
@thebasepoint

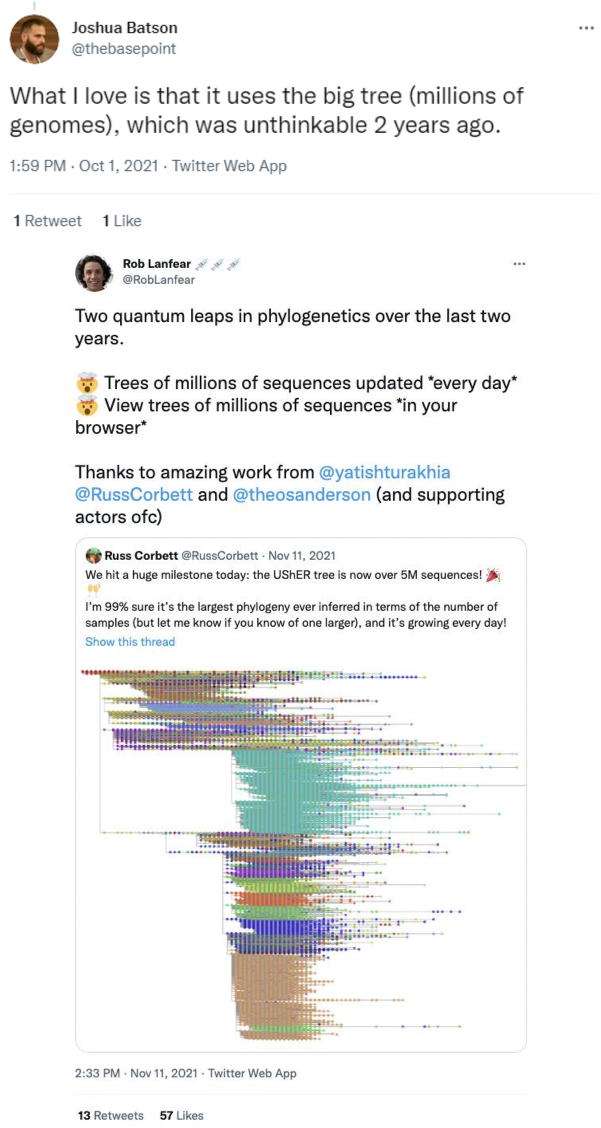
It is so exciting to see this: using the vast quantity of sequencing being done in US to identify introductions to each state, and potential sources. Just the beginning of the future of genomic epi.

Would not be possible w/o large seq program and brand-new comp tools (UsHER)

Russ Corbett @RussCorbett · Oct 1
Introducing cluster-tracker: a daily-updated website for exploring the geographic spread of SARS-CoV-2 across the US. Check it out here
clustertracker.gi.ucsc.edu
Show this thread

1:58 PM · Oct 1, 2021 · Twitter Web App

1 Retweet 8 Likes



Joshua Batson
@thebasepoint

What I love is that it uses the big tree (millions of genomes), which was unthinkable 2 years ago.

1:59 PM · Oct 1, 2021 · Twitter Web App

1 Retweet 1 Like

Rob Lanfear 💉💉💉
@RobLanfear

Two quantum leaps in phylogenetics over the last two years.

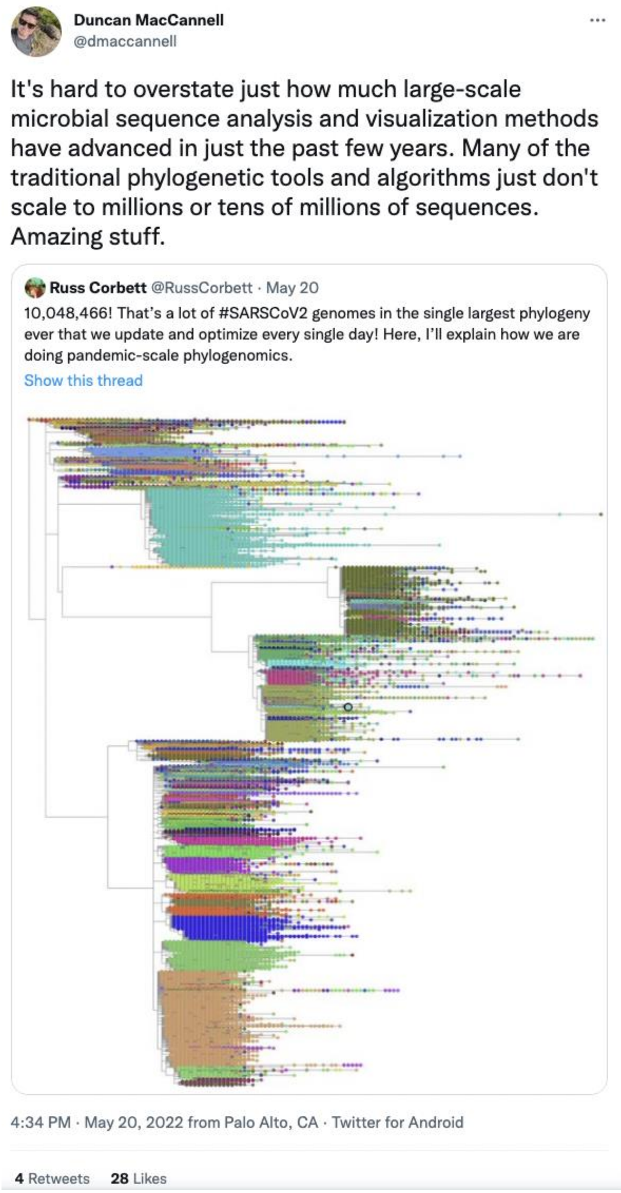
🤯 Trees of millions of sequences updated *every day*
🤯 View trees of millions of sequences *in your browser*

Thanks to amazing work from @yatishturakhia @RussCorbett and @theosanderson (and supporting actors ofc)

Russ Corbett @RussCorbett · Nov 11, 2021
We hit a huge milestone today: the UShER tree is now over 5M sequences! 🎉 🥂
I'm 99% sure it's the largest phylogeny ever inferred in terms of the number of samples (but let me know if you know of one larger), and it's growing every day!
Show this thread

2:33 PM · Nov 11, 2021 · Twitter Web App

13 Retweets 57 Likes



Duncan MacCannell
@dmaccannell

It's hard to overstate just how much large-scale microbial sequence analysis and visualization methods have advanced in just the past few years. Many of the traditional phylogenetic tools and algorithms just don't scale to millions or tens of millions of sequences. Amazing stuff.

Russ Corbett @RussCorbett · May 20
10,048,466! That's a lot of #SARSCoV2 genomes in the single largest phylogeny ever that we update and optimize every single day! Here, I'll explain how we are doing pandemic-scale phylogenomics.
Show this thread

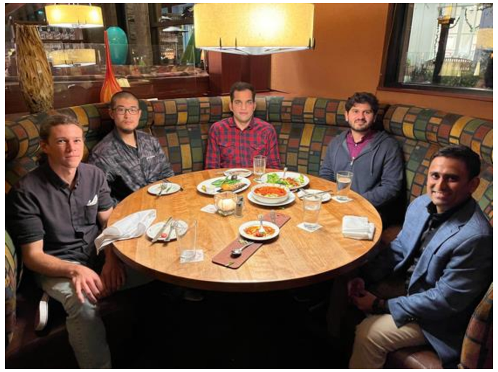
4:34 PM · May 20, 2022 from Palo Alto, CA · Twitter for Android

4 Retweets 28 Likes

TURAKHIA LAB

We are hiring at all levels!

- Computational Genomics / Bioinformatics
- High-performance computing (HPC)
- GPU / FPGA computing
- Computer Architecture
- Hardware-Software Co-design
- VLSI Design



Webpage: https://turakhia.eng.ucsd.edu/
Email: yturakhia@ucsd.edu