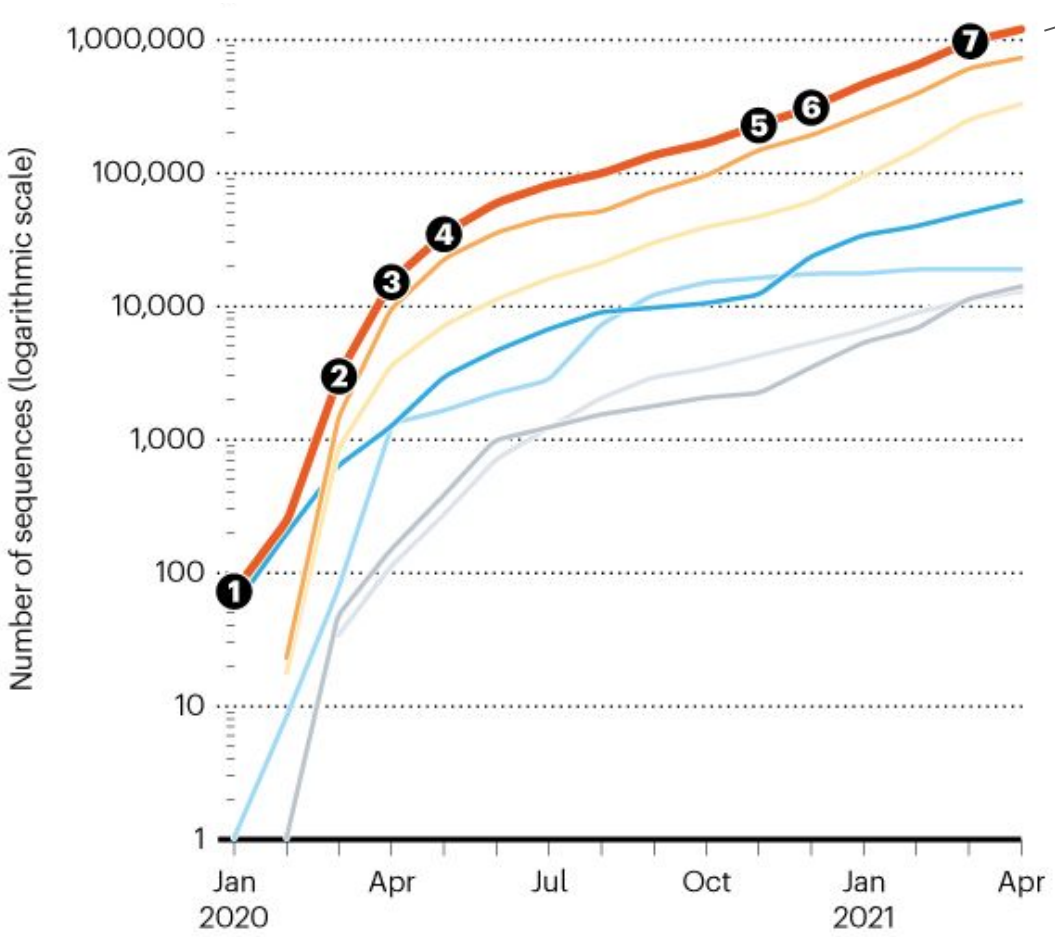
Accelerating phylogenetic tree optimization using efficient placement heuristics

Cheng Ye[1], Bryan Thornlow[2], Jakob McBroome[2], Angie Hinrichs[2], Nicola De Maio[3], Nick Goldman[3], Robert Lanfear[4], David Haussler[2], Russell Corbett-Detig[2], and Yatish Turakhia[1]

[1]University of California, San Diego, US. [2]University of California, Santa Cruz, US. [3]EBI-EMBL, UK. [4]Australian National University, Australia.

Motivation

Now ~2M sequences in GISAID

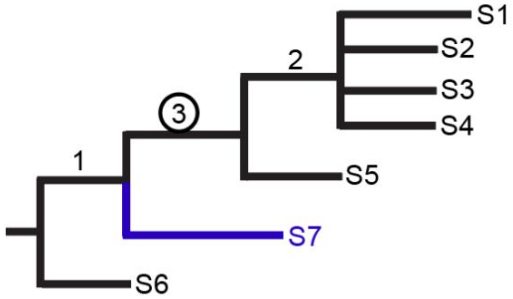


“One million coronavirus sequences: popular genome site hits mega milestone”, Nature 2021

UShER--Ultrafast Sample Placement on Existing Trees

New sample **Variants (VCF)**

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]

New mutation-annotated tree object

UShER wiki at https://usher-wiki.readthedocs.io/en/latest/UShER.html

Program	Average time to place one sample	Average peak memory used(GB)
IQ-TREE multicore v.2.1.1	46 m 31 s	12.85
EPA-ng v.0.3.8	27 m 38 s	791.82
PAGAN2 v.1.54	120 m 32 s	470.74
UShER (without preprocessed mutation-annotated tree)	1 m 43 s	1.02

(Turakhia et al. Nature Genetics 2021)

Optimization with SPR

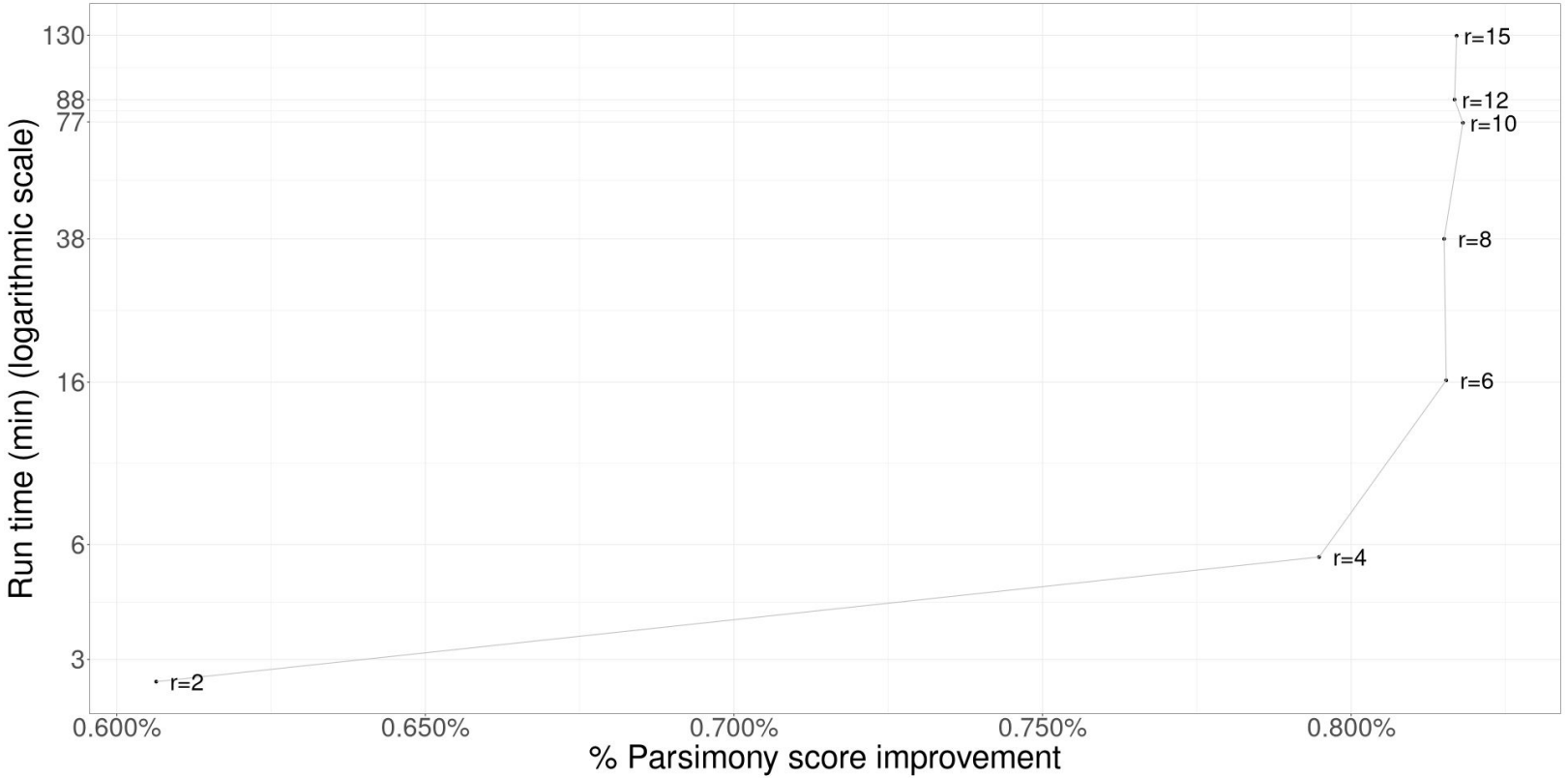
Program	Wall Time (min)	Parsimony Improvement
matOptimize R=4	6	0.81%
matOptimize R=6	16	0.83%
TNT[1] xmult	7546 (236 equiv.)	0.83%
IQ-TREE[2] May 24 dev branch R=20	25	0.51%

Using 32 threads of Xeon E7-4870 (2.4 GHz)
2021-03-18 tree with 364,428 sequences

[1]Build tree from scratch, single threaded
[2]Support for parsimony is only present in development branch

Source:https://github.com/bpt26/parsimony/tree/main/2_optimise_starting_tree

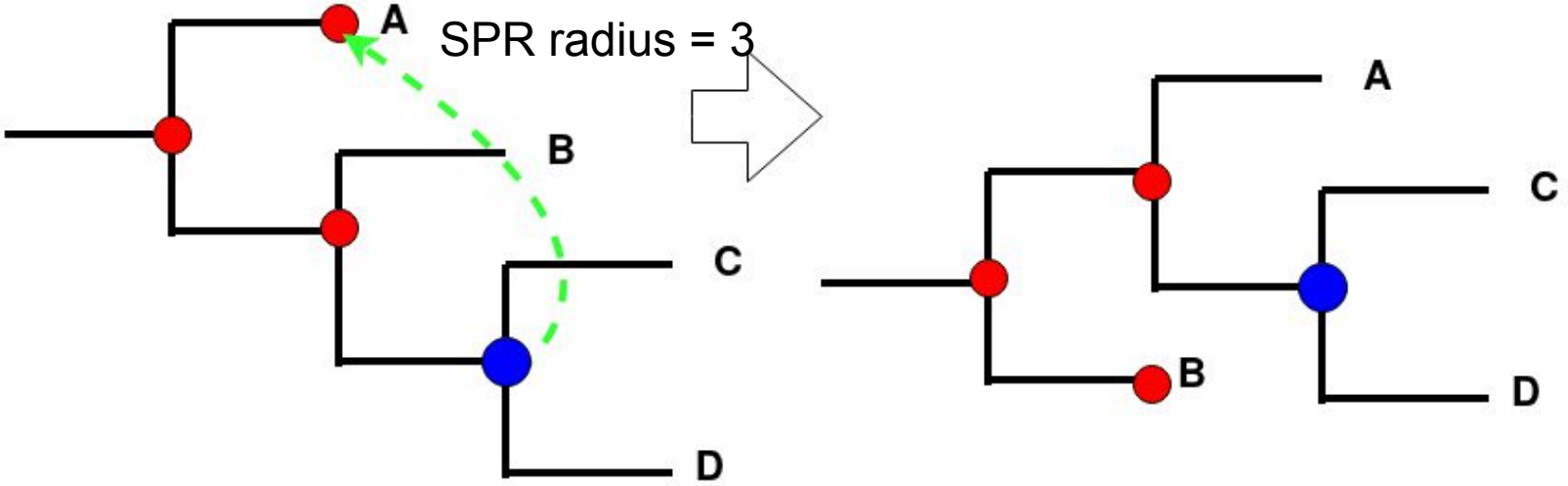
Most Parsimony Improvement Found at Small Radius With small Run Time

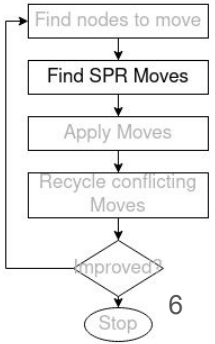


Peak Memory Usage:10 GB

Find Profitable Subtree pruning and grafting (SPR) moves

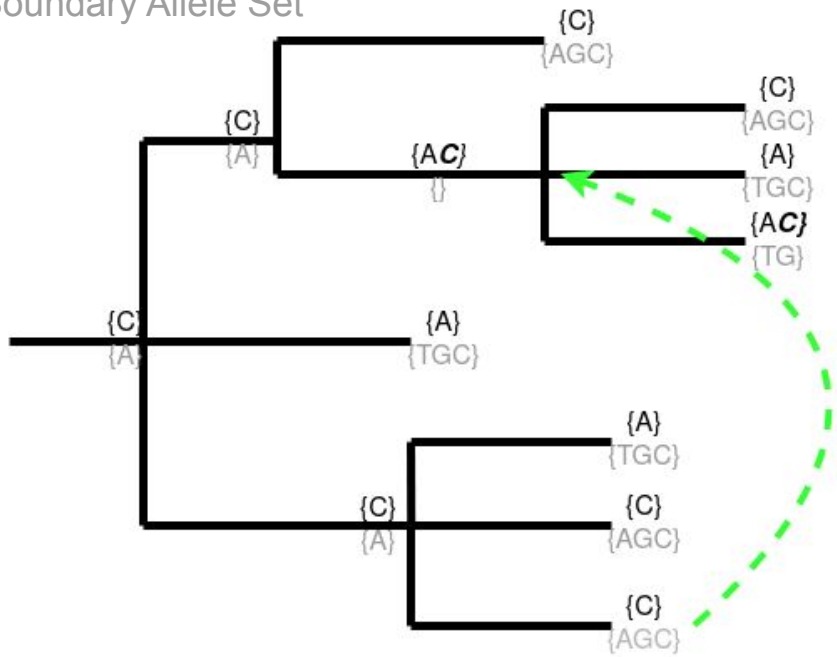
- All source nodes are searched in parallel.
- Assuming all moves are independent.
- Used the incremental update method in Gladstein, D., 1997 to calculate parsimony score change.

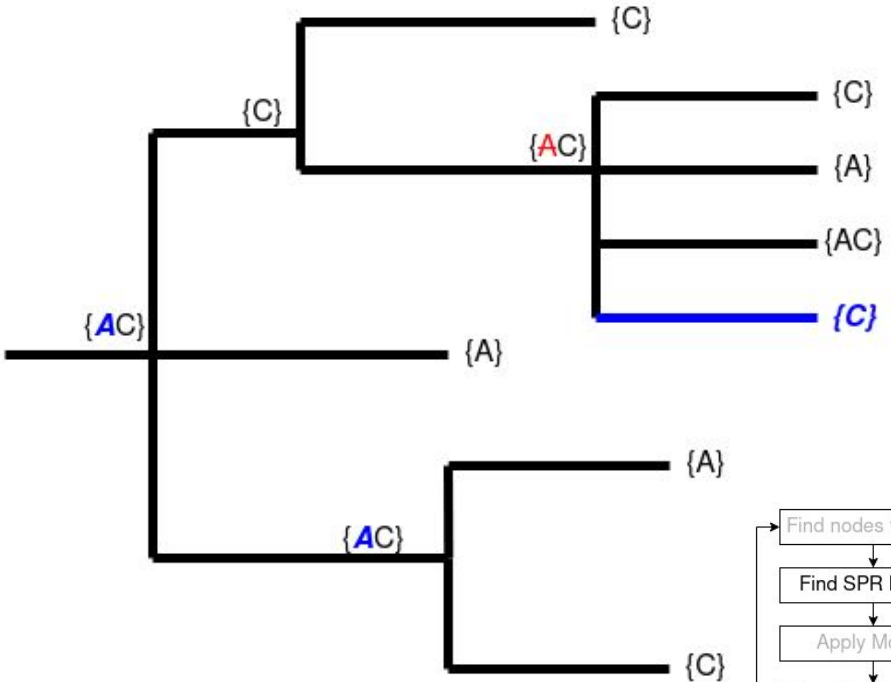


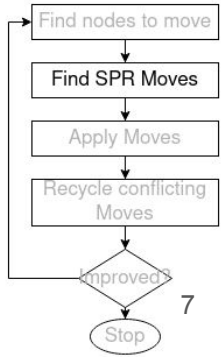


Estimating Parsimony change --Example

Legend:
Major Allele set
Boundary Allele Set

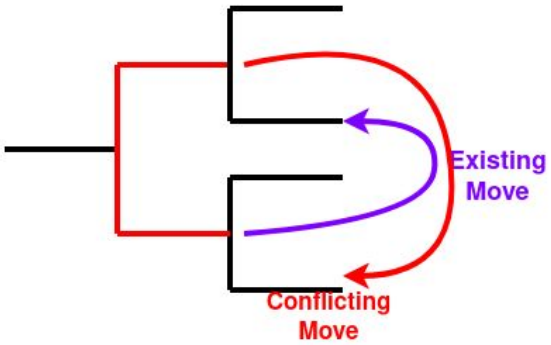




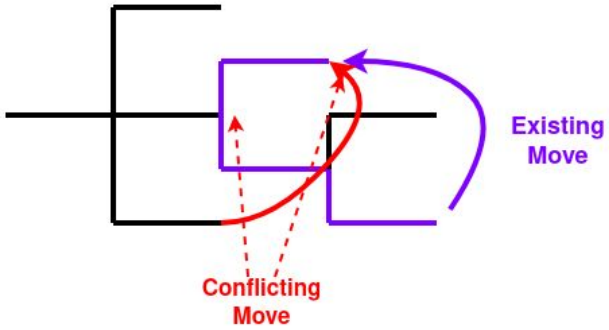


Conflicting Moves: Moves whose path intersects

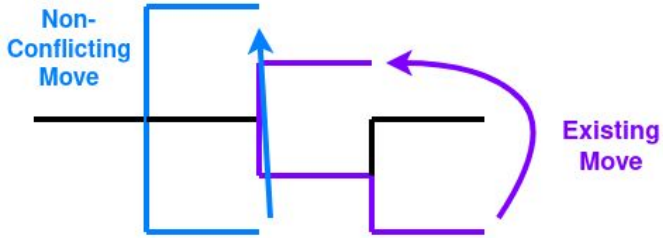
Form loop:



Can affect each other's parsimony score

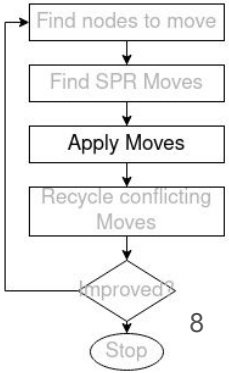


May interact (but rare), but SPR reversible



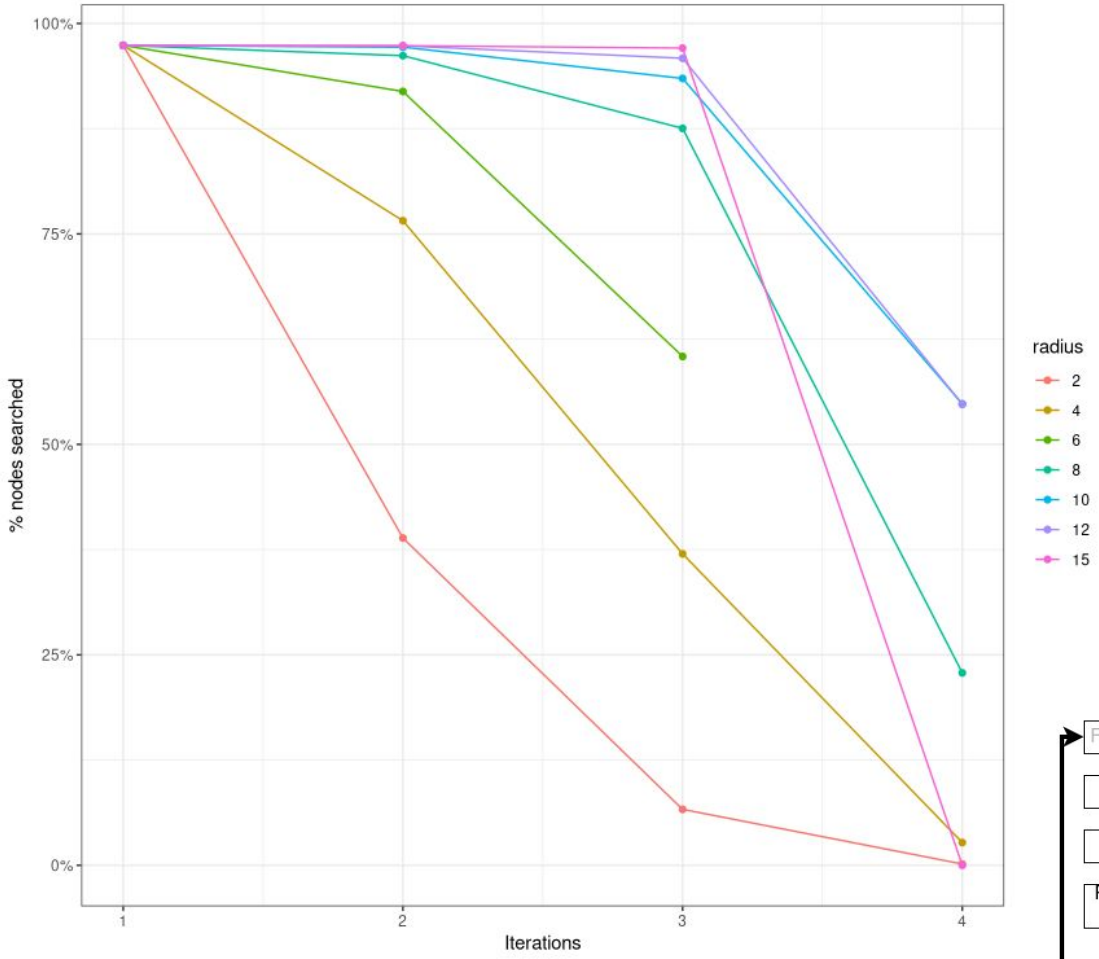
Prioritize:

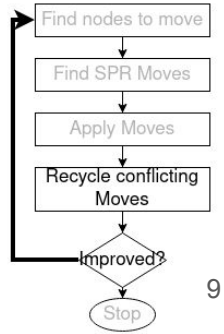
1. Moves giving the largest improvement,then
2. Moves with shorter path



Next Iteration

1. Retry Previously Conflicting moves
2. Moved Nodes & nodes within search radius





Conclusion

High efficiency can be attributed to incremental update and adaptations to SARS-COV-2 phylogeny

- Mutation Annotated Tree -> Compact representation
- Native polytomy support -> Reduce SPR move radius

Try our program matOptimize at https://github.com/yatisht/usher

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