TreeFix-VP: Phylogenetic Error-Correction for Viral Transmission Network Inference

Sledzieski, Zhang, Mandoiu, Bansal

Department of Computer Science and Engineering,
University of Connecticut,
USA

October 16, 2018

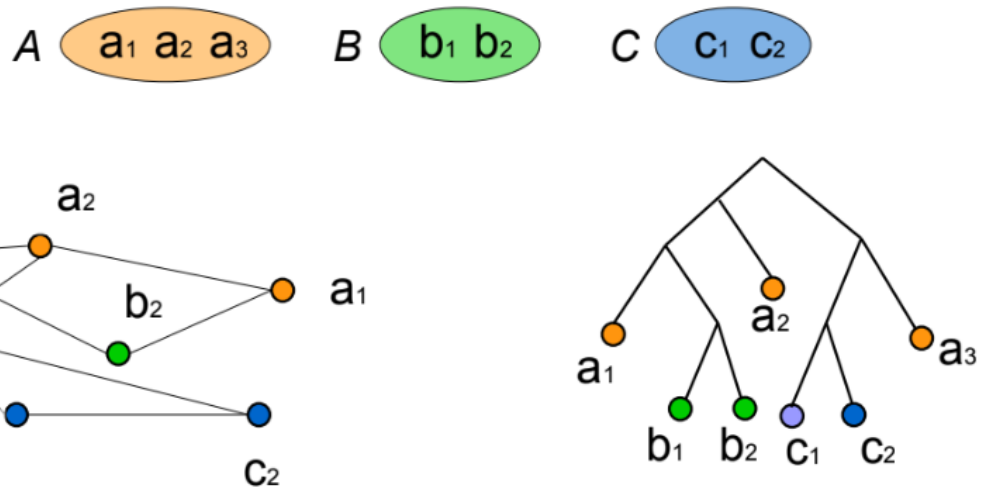
Viral Transmission Inference

Problem

Reconstruct transmission of disease

Given: *Viral sequences from infected hosts*

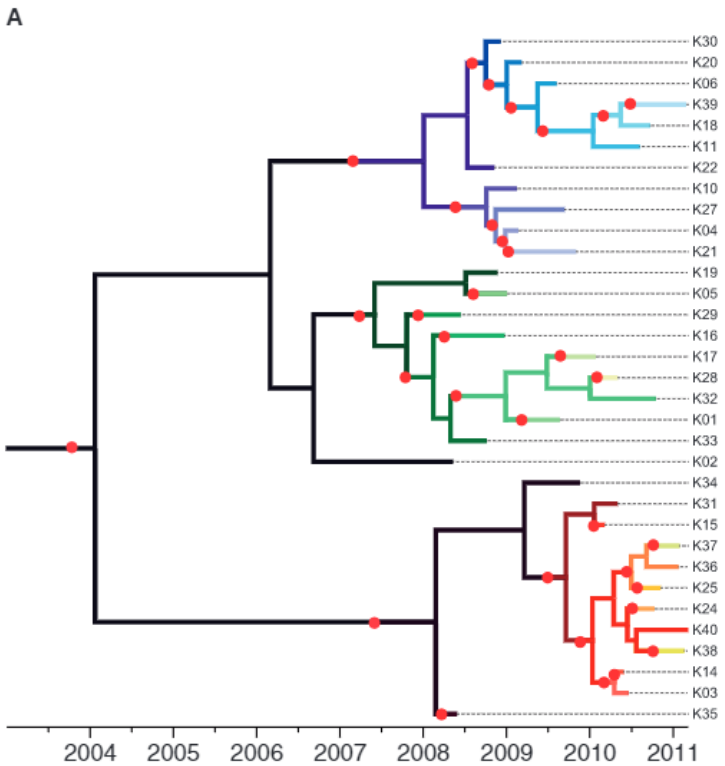
Goal: *Network $G(V, E)$ where V is the set of infected hosts, and each edge in E represents a transmission of the disease*



Bansal 2017

Phylogeny-Based Transmission Network Inference

- Label internal nodes of viral sequence phylogeny with hosts
- Didelot et. al 2014, Hall et. al 2015, Klinkengerg et al. 2017



Didelot et. al 2014

Improved Phylogenetic Inference

Goals:

- Improve downstream transmission inference
- Improve scalability by reducing the need for MCMC or coestimation of phylogeny

Approach: Error correction for reconstruction of highly accurate viral phylogenies

Improved Phylogenetic Inference

Problem

Reconstruct viral phylogeny

Given: *Viral sequences from infected hosts*

Goal: *Tree T representing evolutionary history of the virus, where leaves are labeled with infected hosts*

TreeFix-DTL

Improved gene tree error correction in the presence of horizontal gene transfer

Mukul S. Bansal[1,2,*,†], Yi-Chieh Wu[1,†], Eric J. Alm[3,4], and Manolis Kellis[1,4,*]

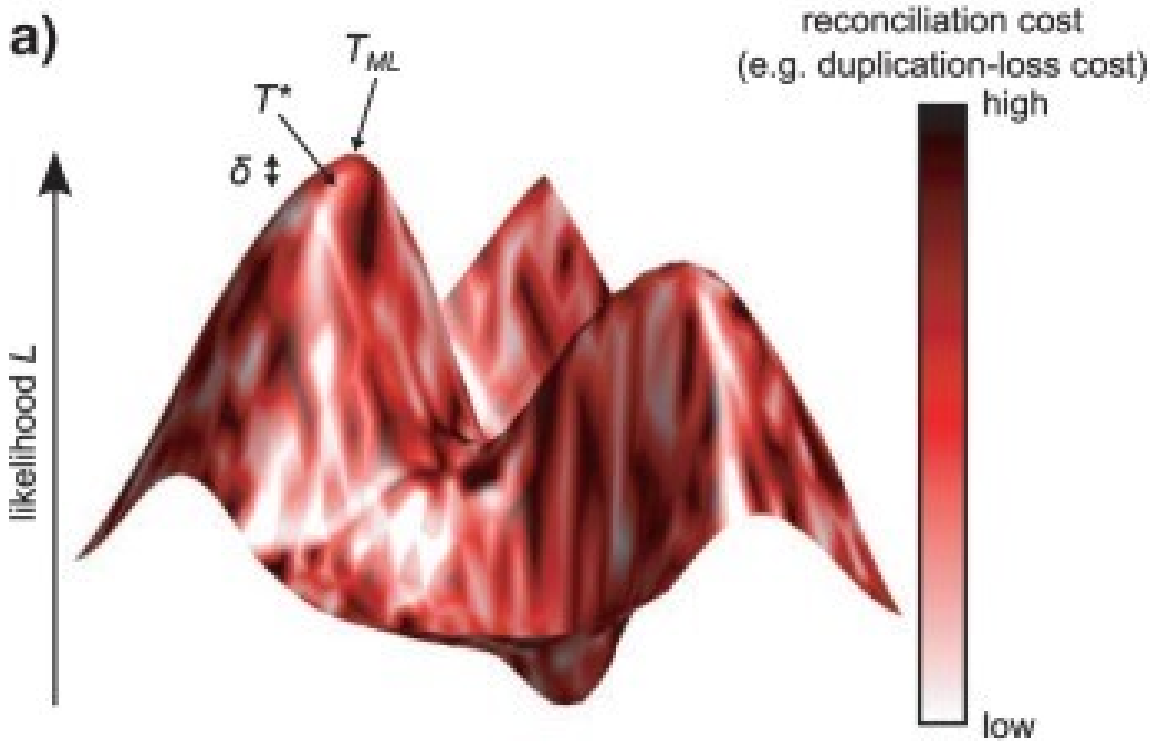
[1]Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA, USA, [2]Department of Computer Science and Engineering, University of Connecticut, Storrs, CT, USA and [3]Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge and [4]Broad Institute, Cambridge, MA, USA

[*]To whom correspondence should be addressed.

[†]The authors wish it to be known that, in their opinion, the first 2 authors should be regarded as Joint First Authors.

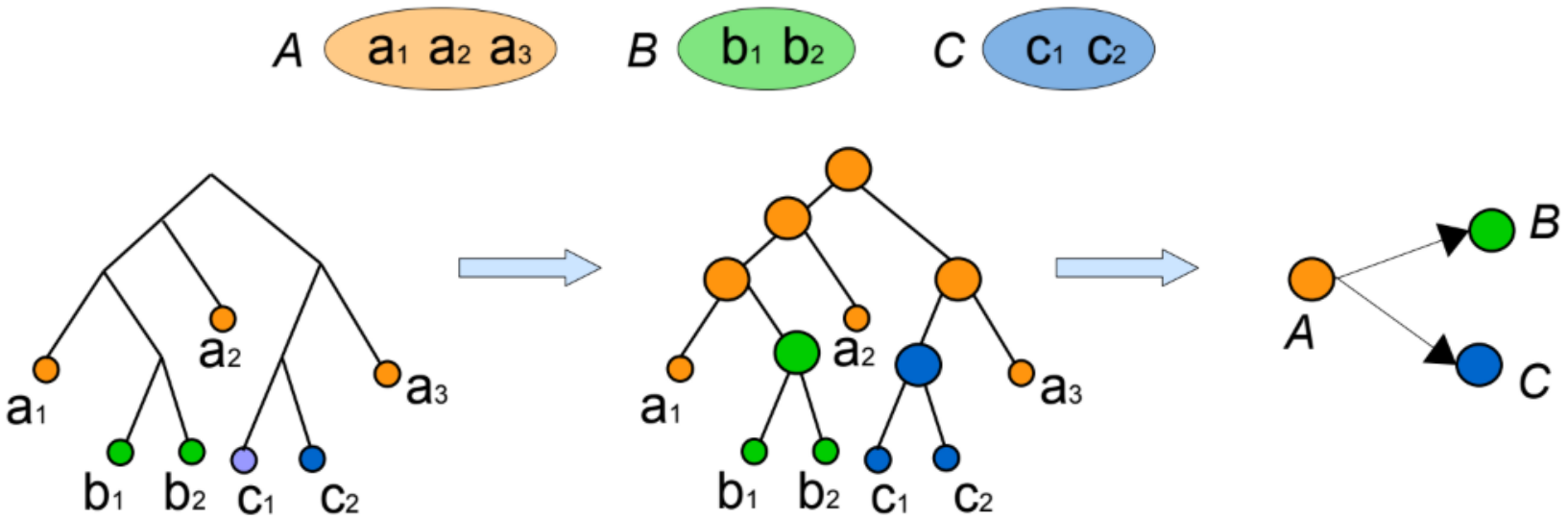
Associate Editor: David Posada

Search in Maximum Likelihood Neighborhood



Bansal and Wu et. al 2014

Multiple Sequences per Host



Bansal 2017

TreeFix-VP

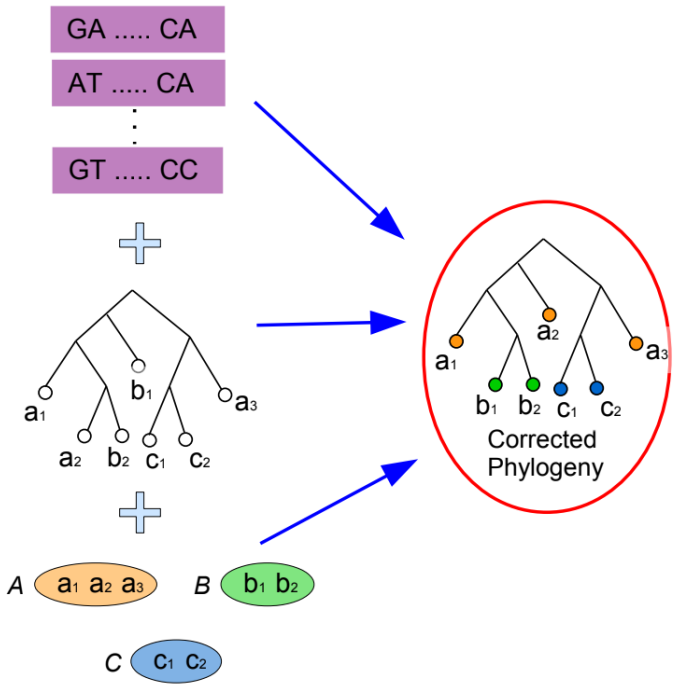
Computational method for error correction of viral sequence phylogenies

- Accurate reconstruction of phylogenies
- Increased accuracy of outbreak and transmission inference
- Scalable analysis

TreeFix-VP

Input: Maximum likelihood phylogeny, multiple sequence alignment, sequence-host mapping

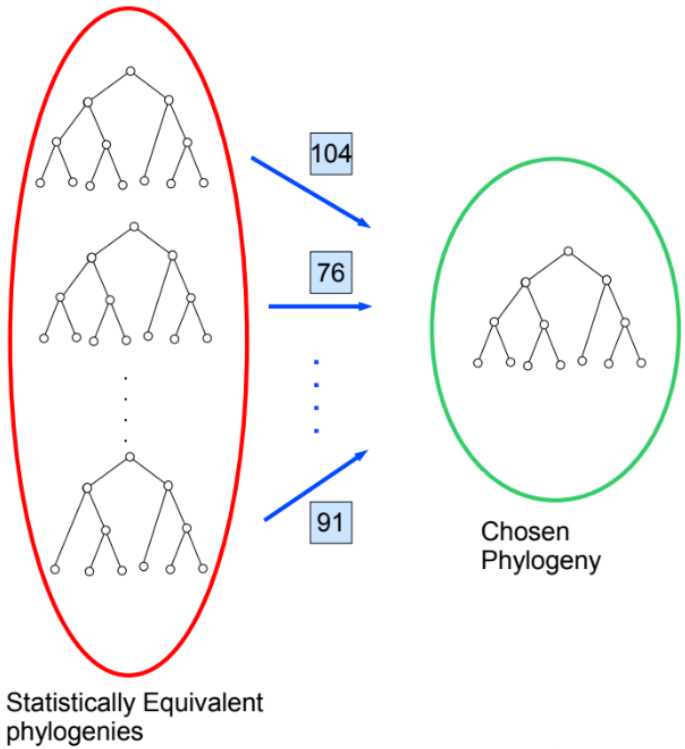
Output: Error-corrected viral phylogeny



Bansal 2017

TreeFix-VP

Approach: Use host information to select the best tree that is still well supported by sequence data.

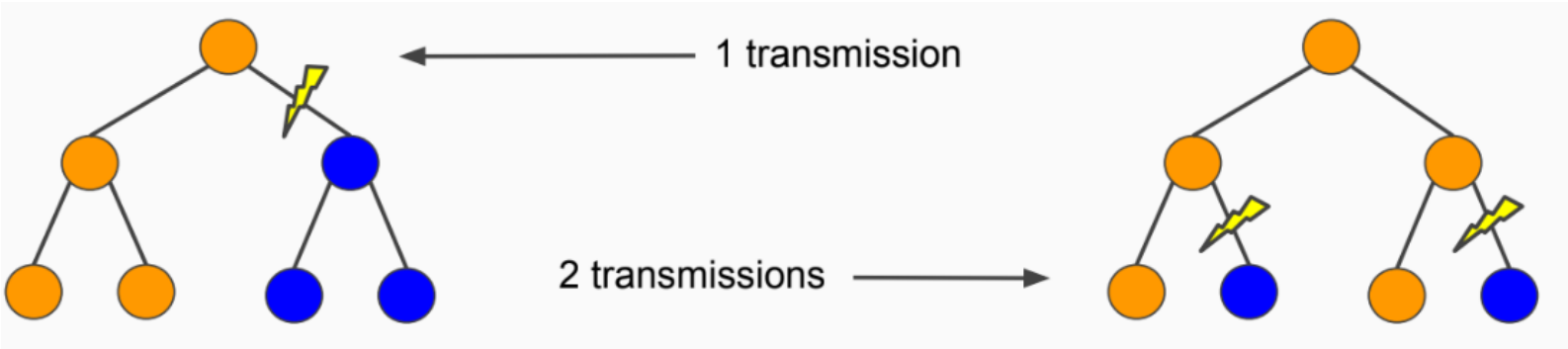


Bansal 2017

Tree Score

Question: How to determine the "best" tree?

1. Label leaves with associated hosts.
2. Use Fitch's algorithm for the small parsimony problem to calculate the minimum number of required transmissions.
   - For a tree on n leaves and k hosts, complexity $O(nk)$
   - Biologically meaningful: edges with different hosts represent transmission (direct or indirect)

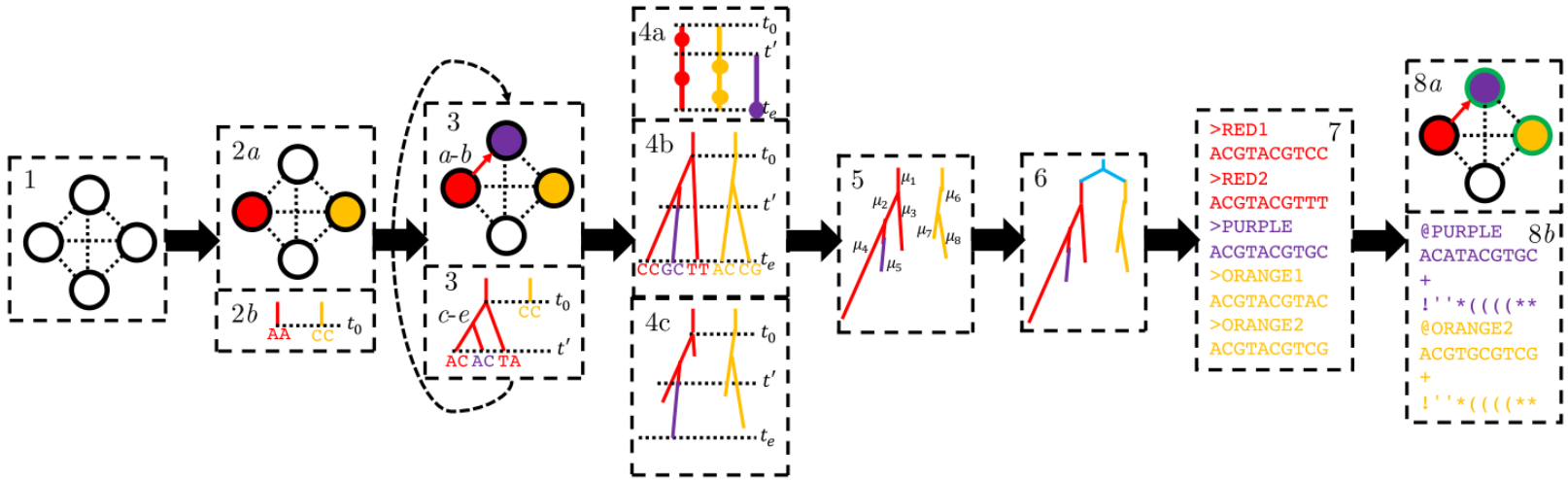


Outbreak Data

Dataset:

- 142 intra-host HCV populations from 33 outbreaks (provided by CDC)
- Outbreaks contain from 2 to 19 samples, and
- A few dozen to a few hundred sequences
- True transmission history known for 14 of the outbreaks

FAVITES (FrAmework for VIral Transmission and Evolution Simulation



Moshiri et al. 2018

Simulation Model

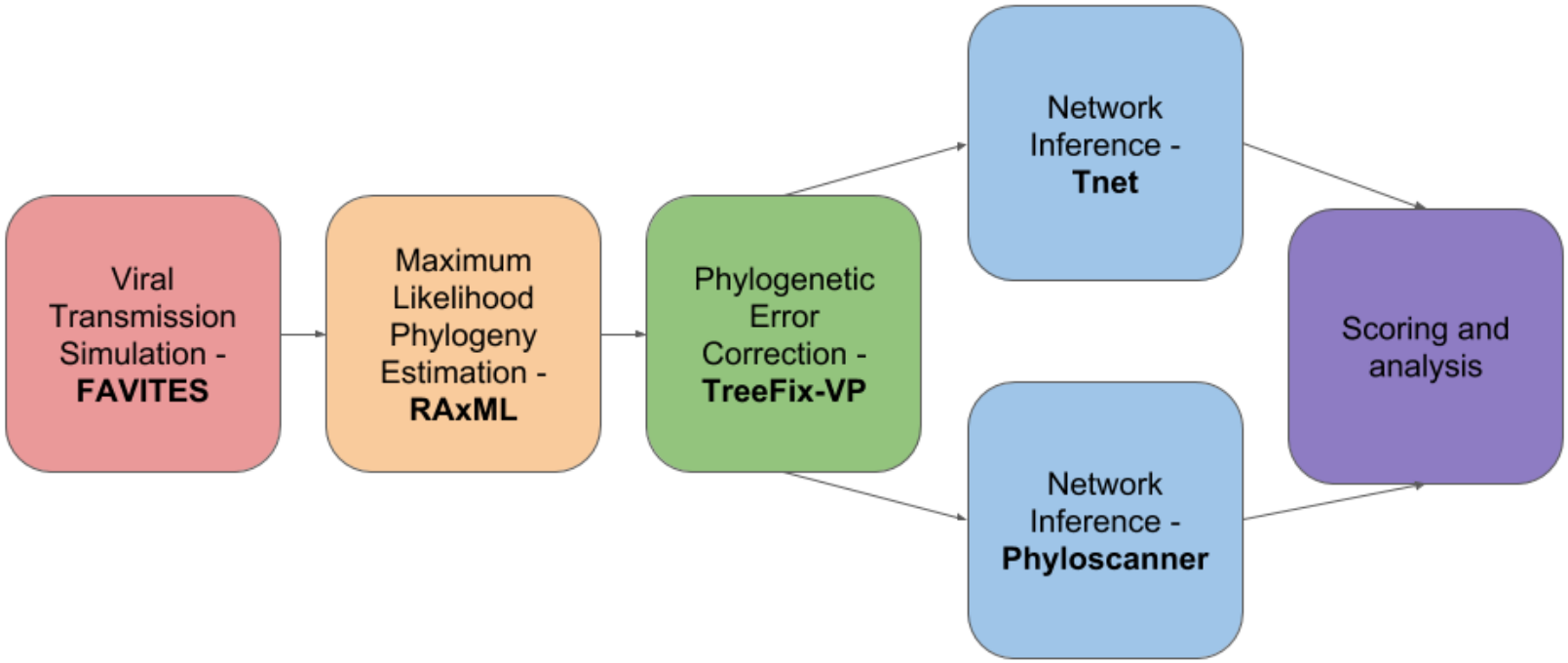
- Barabasi-Albert model for contact network generation
  - 1000 nodes
- Models of transmission
  - Used SEIR and SIR models
  - Transmission parameters chosen to evenly space transmissions
- Coalescent model with logistic growth rate for phylogeny generation
  - Coalescent parameters chosen to give even branch lengths
- GTR+Γ model of sequence evolution
  - Nucleotide frequencies and transmission rates estimated from real outbreak data

Simulation Model

Varied Parameters:

- Sequence Length
- Viruses per Host
- Mutation Rate

Analysis Pipeline

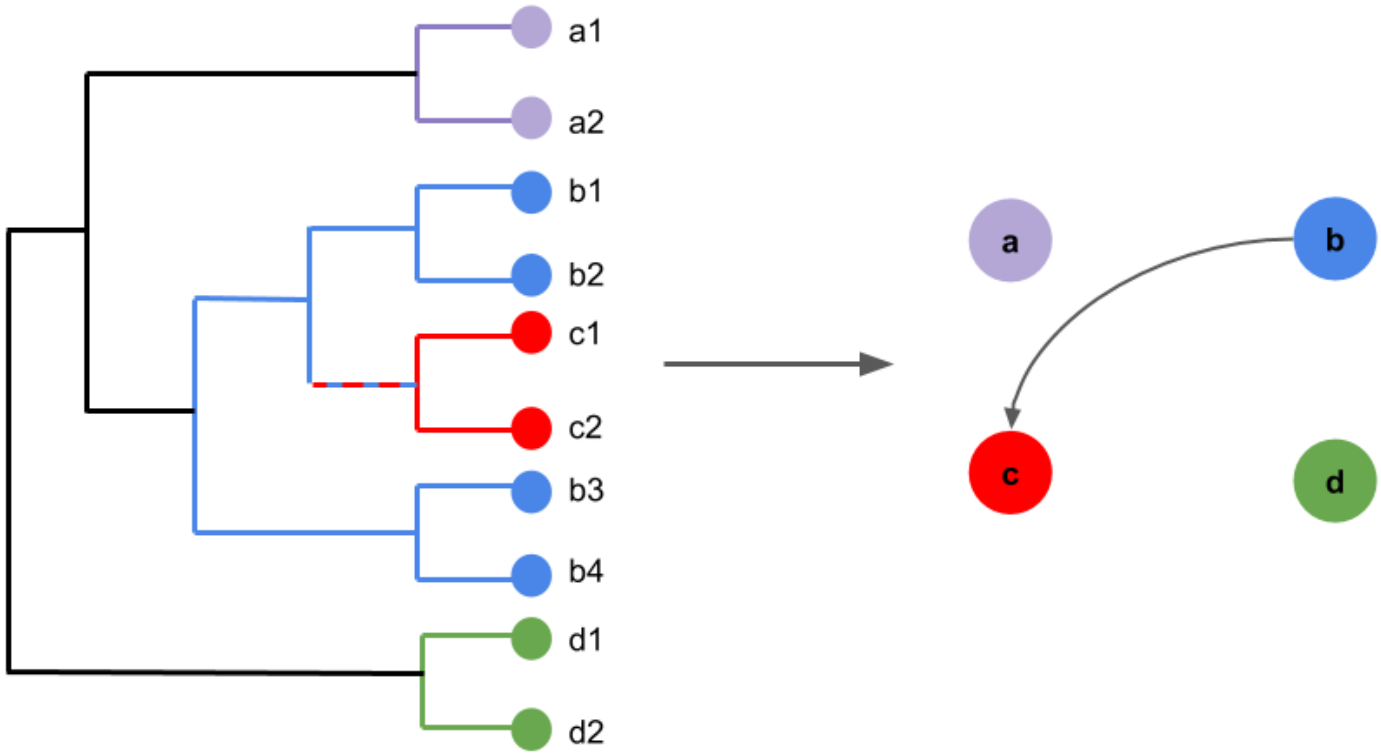


Analysis Pipeline

- **RAxML:** 25 bootstraps, GTRGAMMA model, rooted phylogeny
- **TreeFix-VP:** Run for 5000 iterations
- **Tnet:** Uses Sankoff's algorithm to label internal nodes of phylogeny and infer transmission edges
- **Phyloscanner:** Wymant and Hall et al. 2017
  - Also uses parsimony
  - Leaves some internal nodes unlabeled

Analysis Pipeline

Phyloscanner - Conservative estimation of transmissions



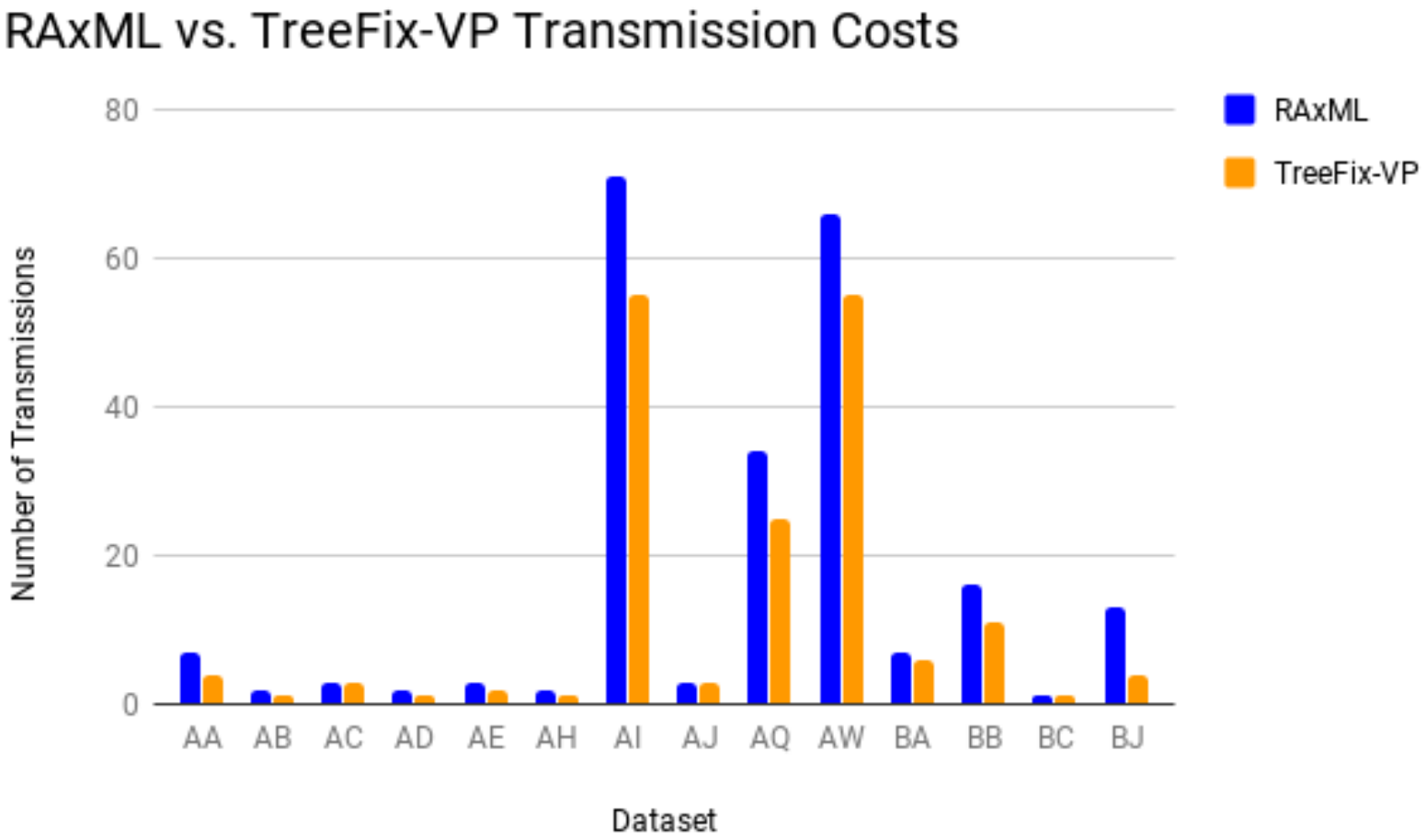


Figure: Outbreak Transmission Cost

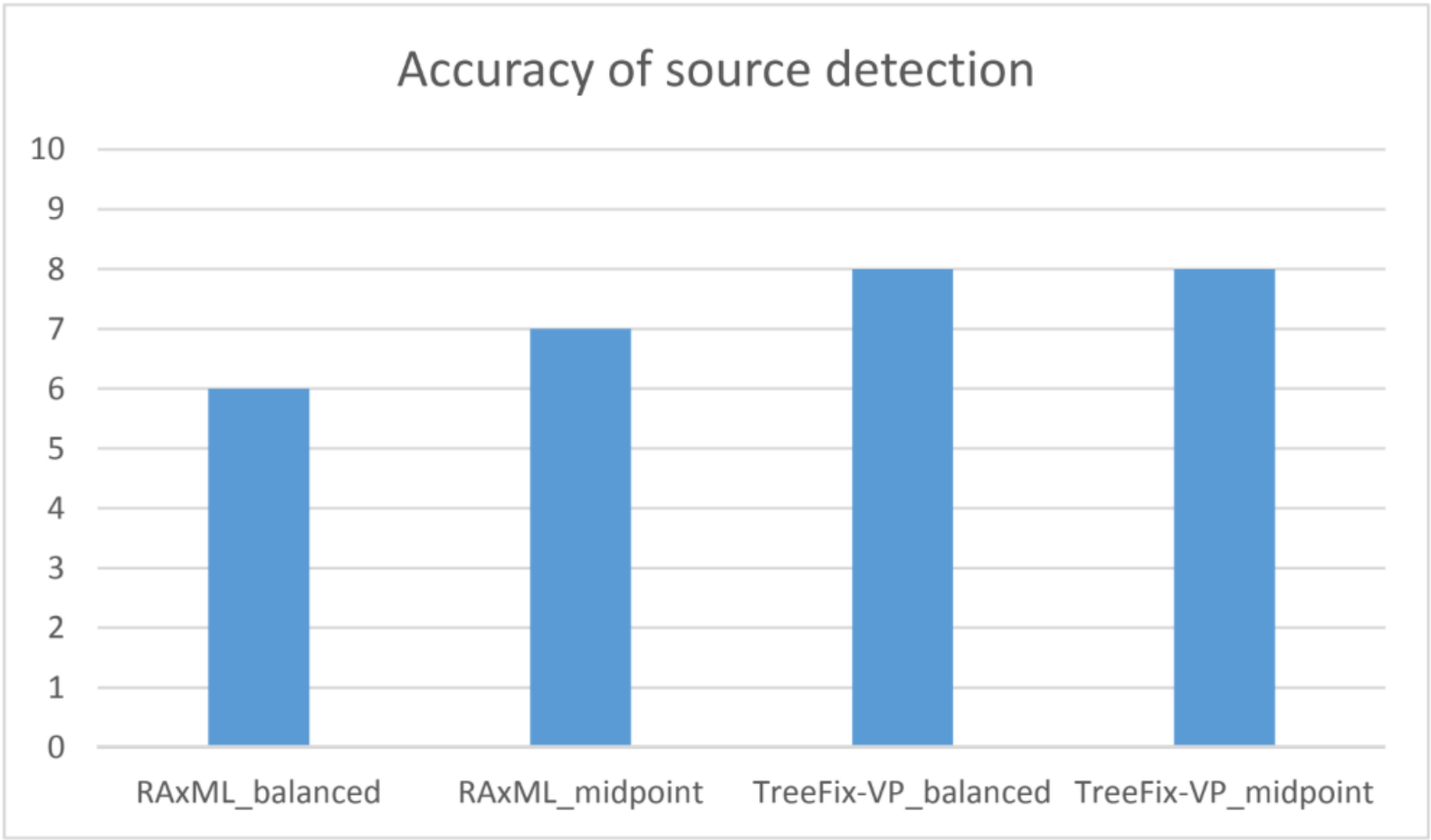


Figure: Outbreak Source Detection

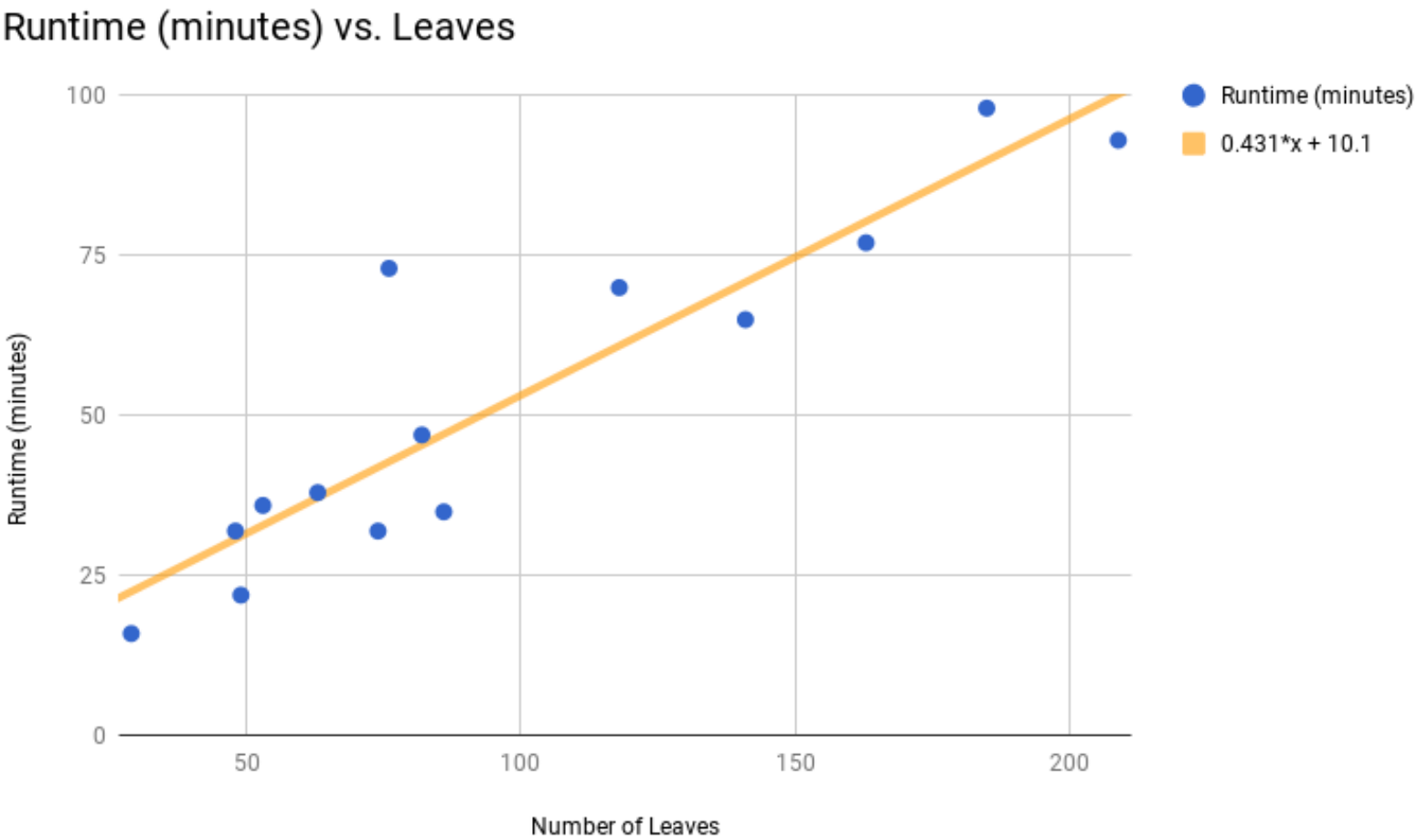


Figure: Outbreak Runtime

Simulation Parameters

Baseline

- SEIR model of transmission
- **Sequence Length:** 1000
- **Viruses per Host:** 10
- **Mutation Rate:** 0.25

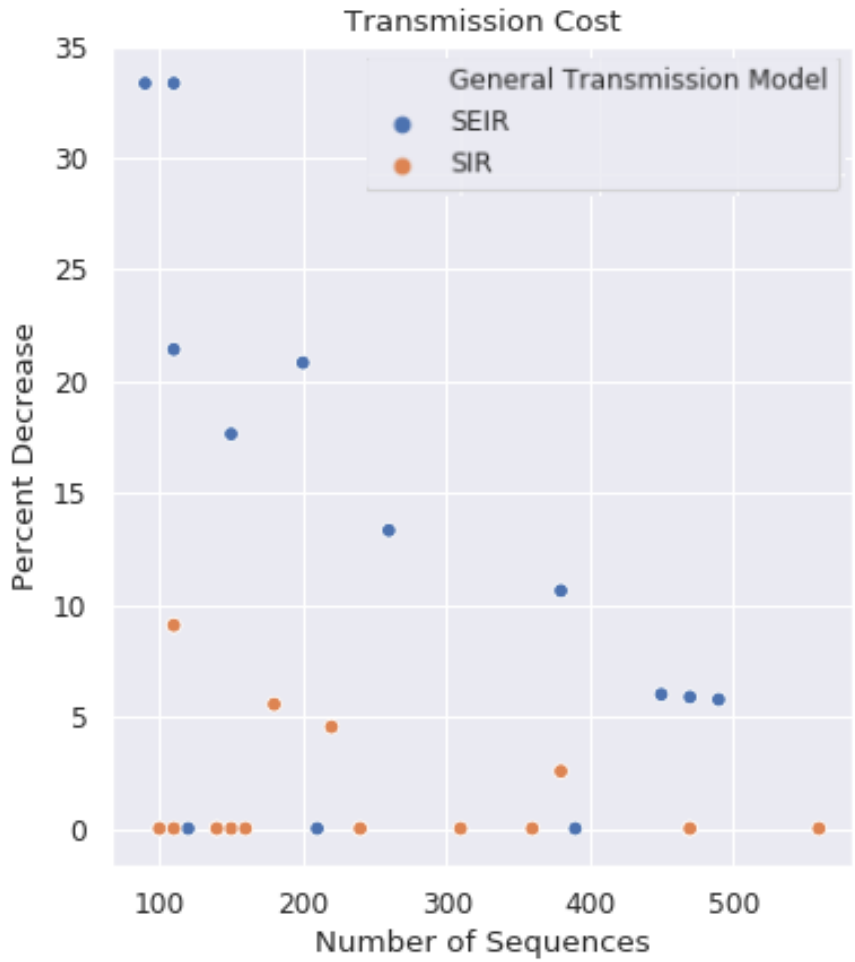


Figure: Error Corrected Transmission Cost

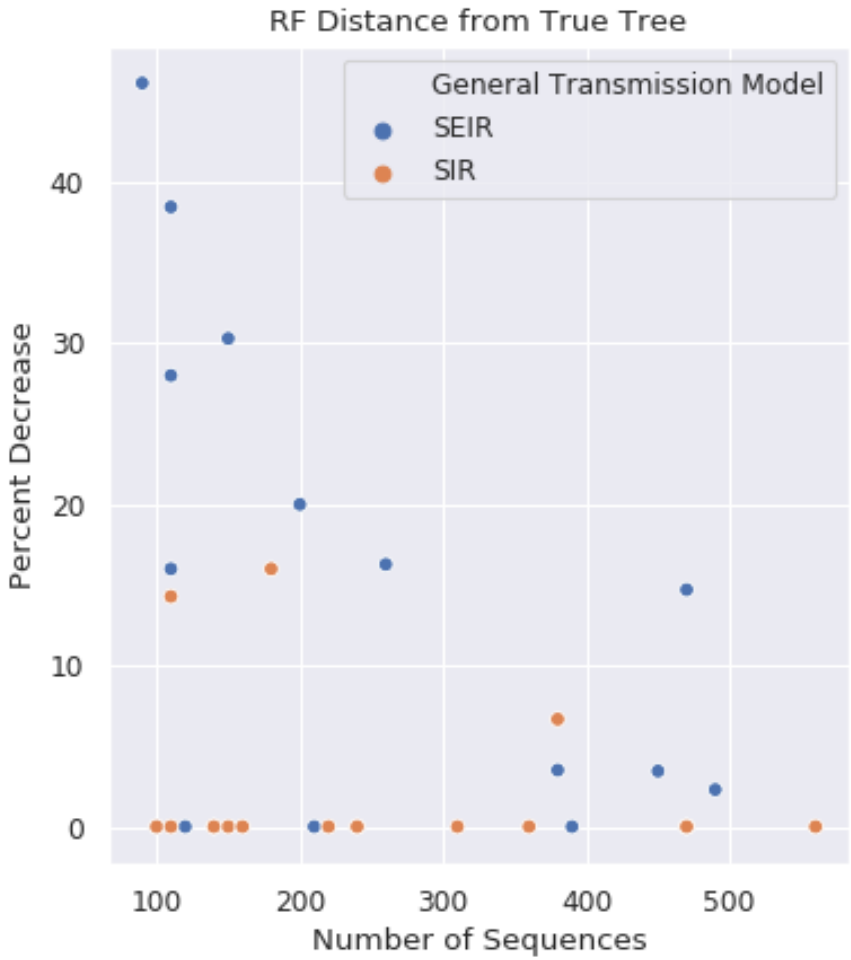


Figure: Error Corrected Robinson-Foulds Distance

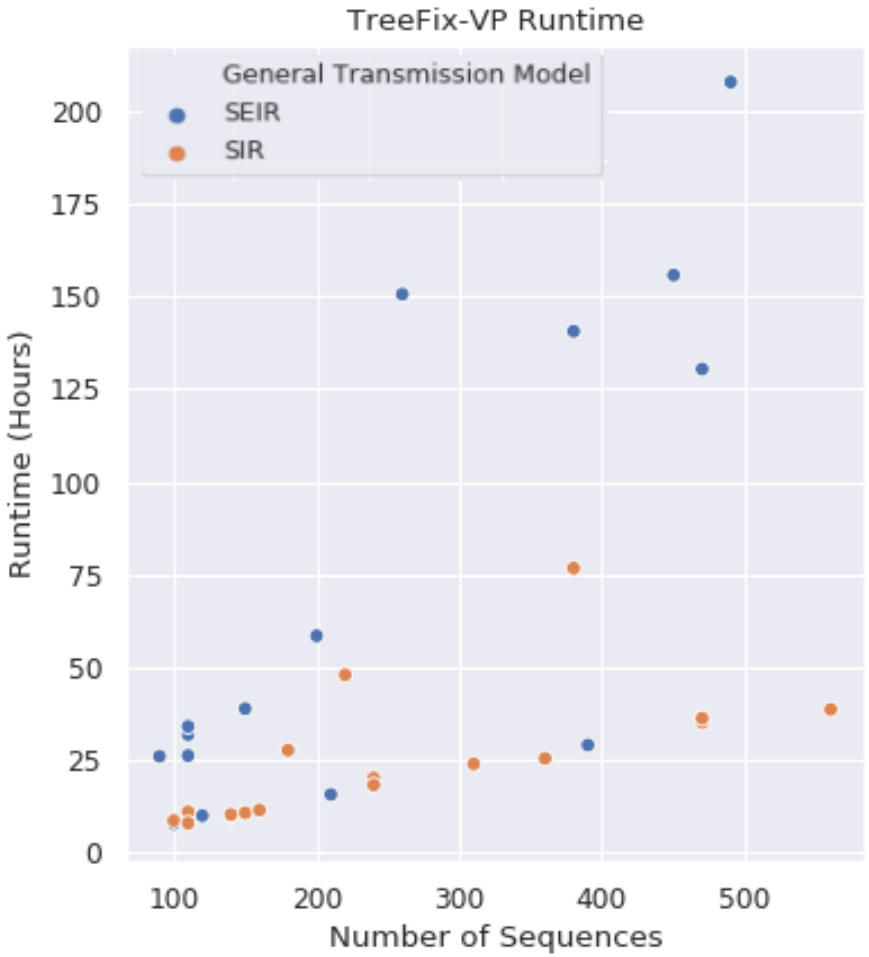


Figure: TreeFix-VP Runtime

Figure: All Runs

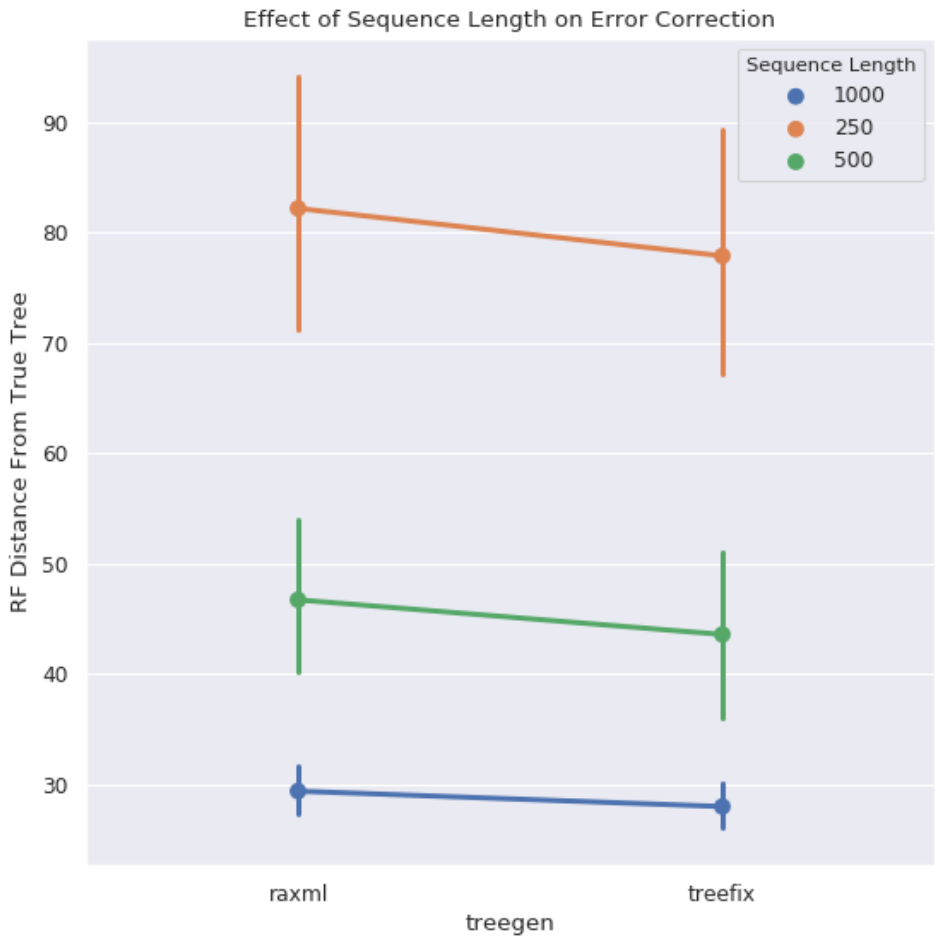


Figure: Varied Sequence Length

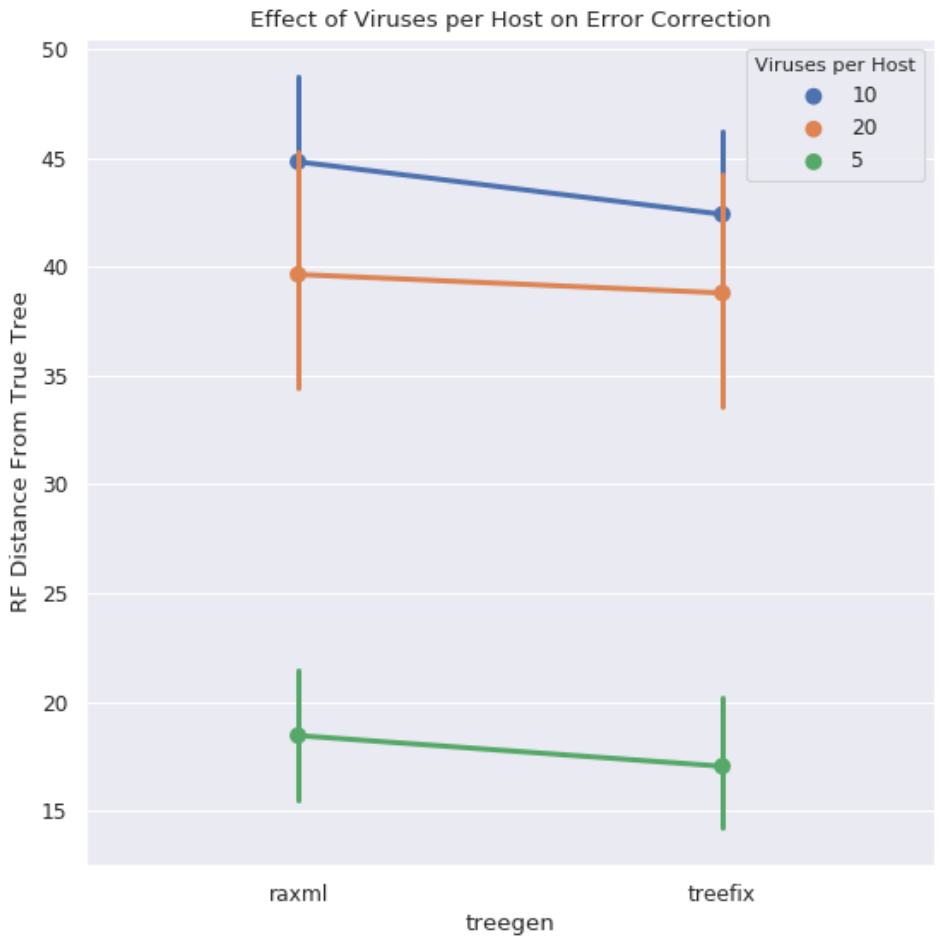


Figure: Varied Number of Viruses per Host

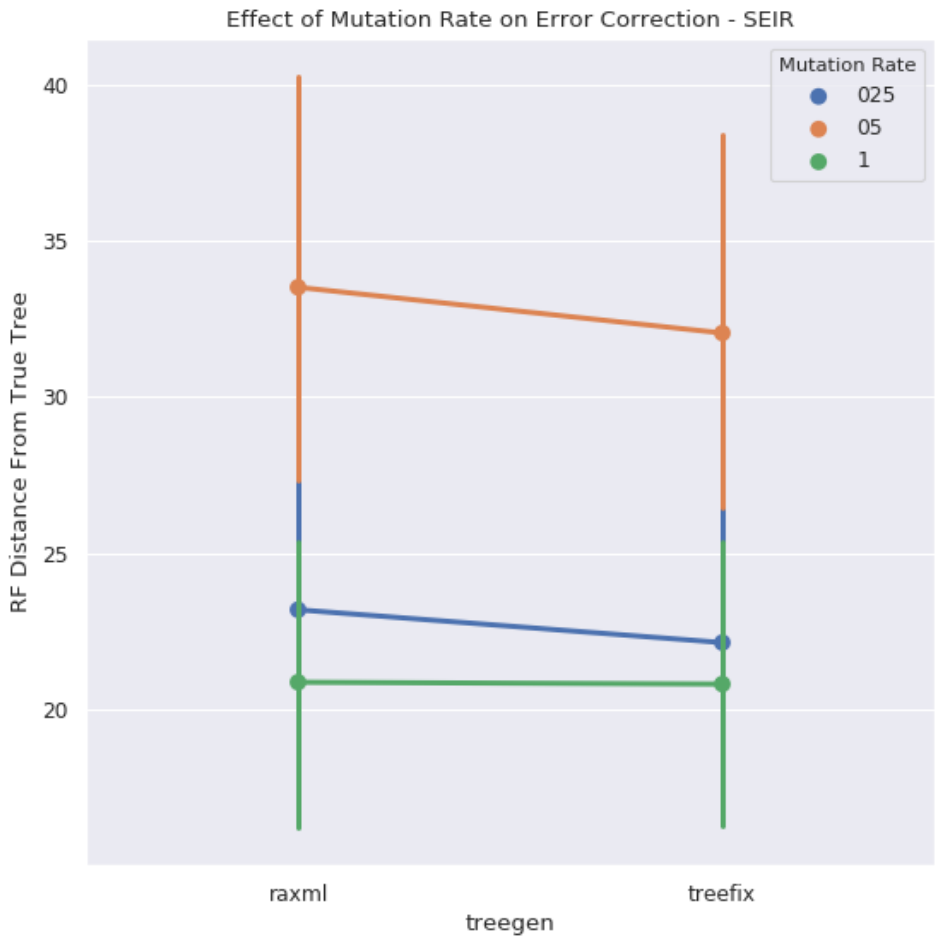


Figure: Varied Mutation Rate - SEIR

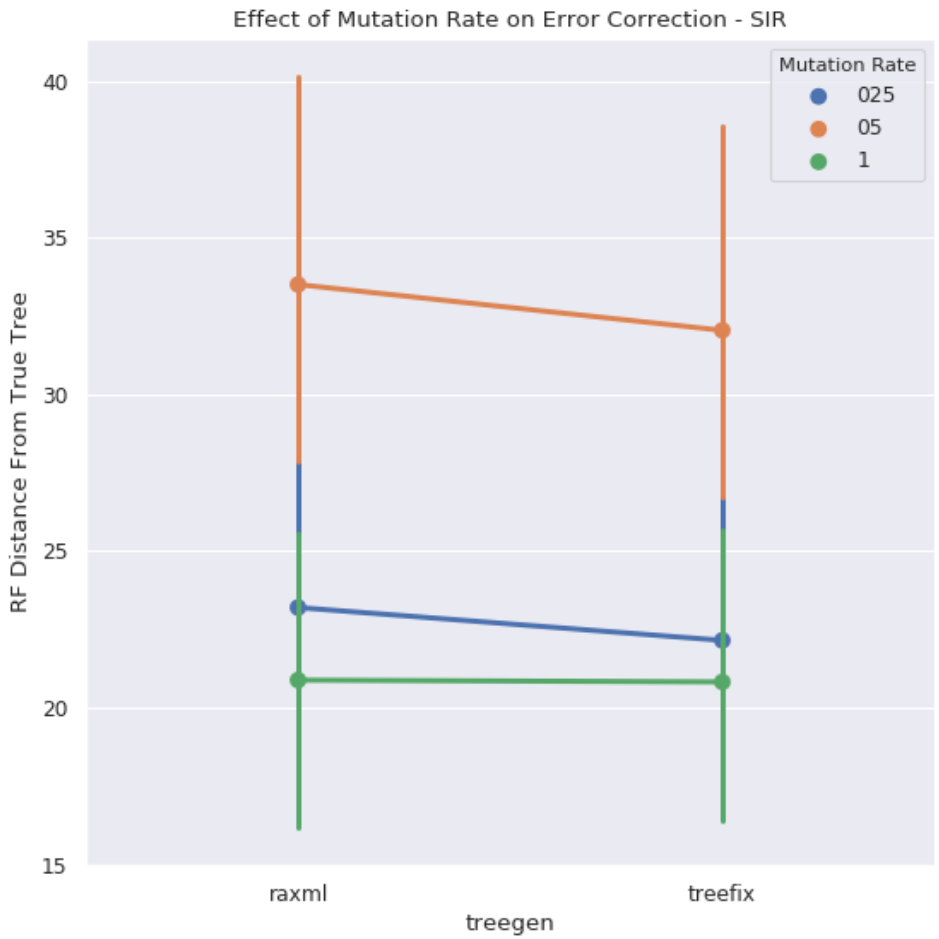


Figure: Varied Mutation Rate - SIR

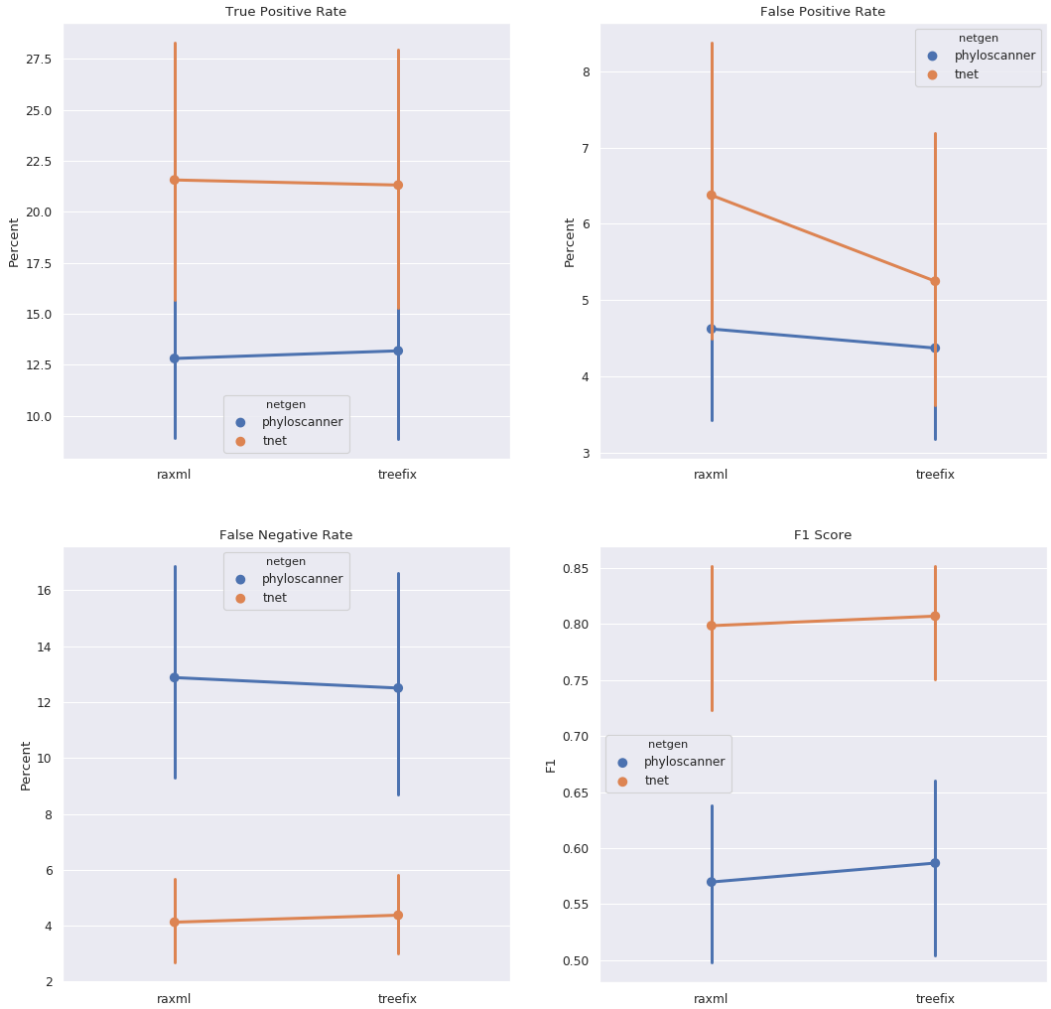


Figure: SEIR Transmission Model

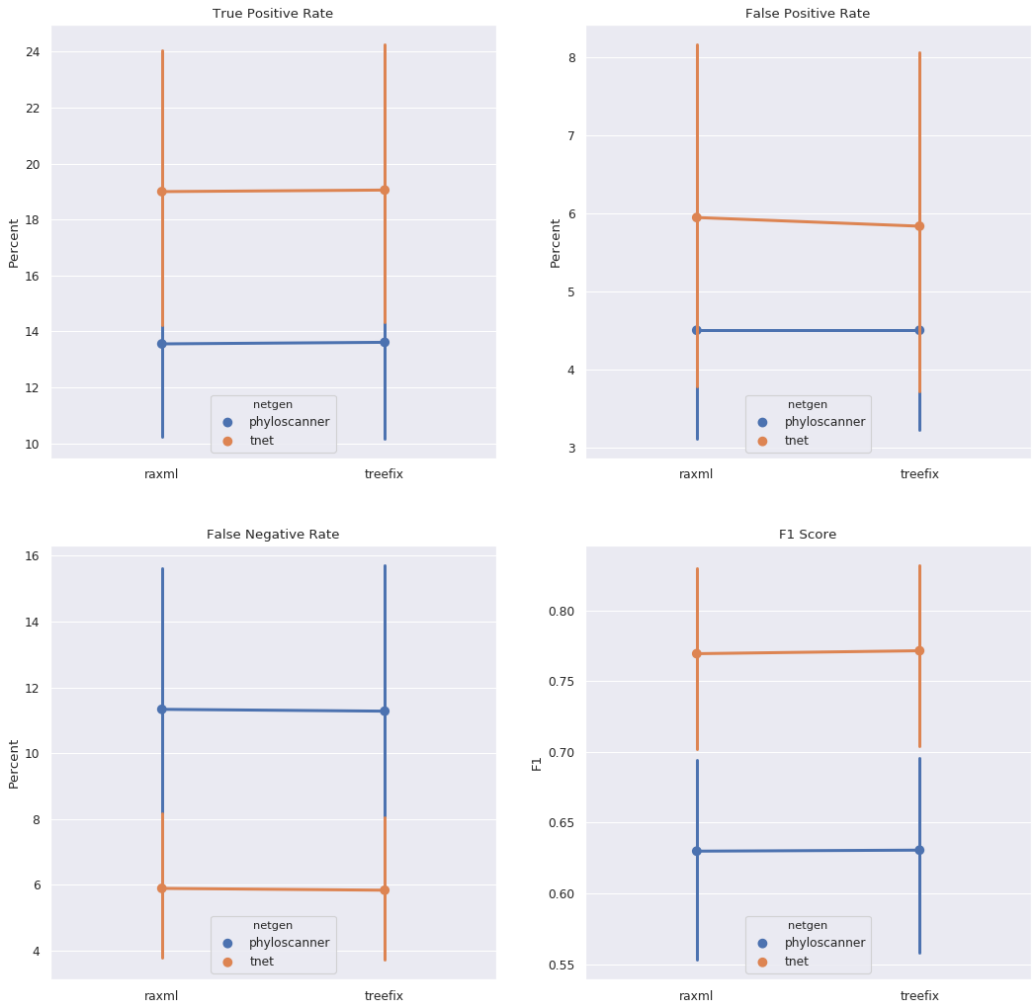


Figure: SIR Transmission Model

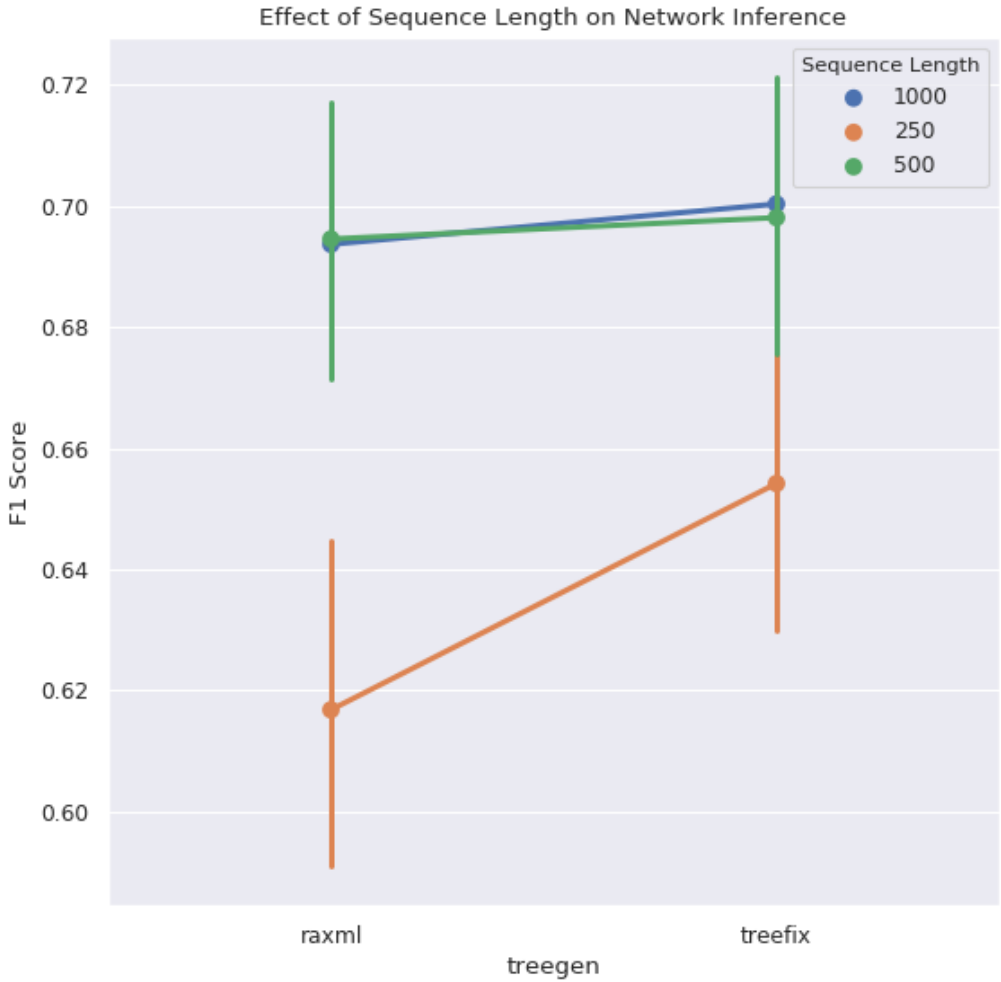


Figure: Varied Sequence Length

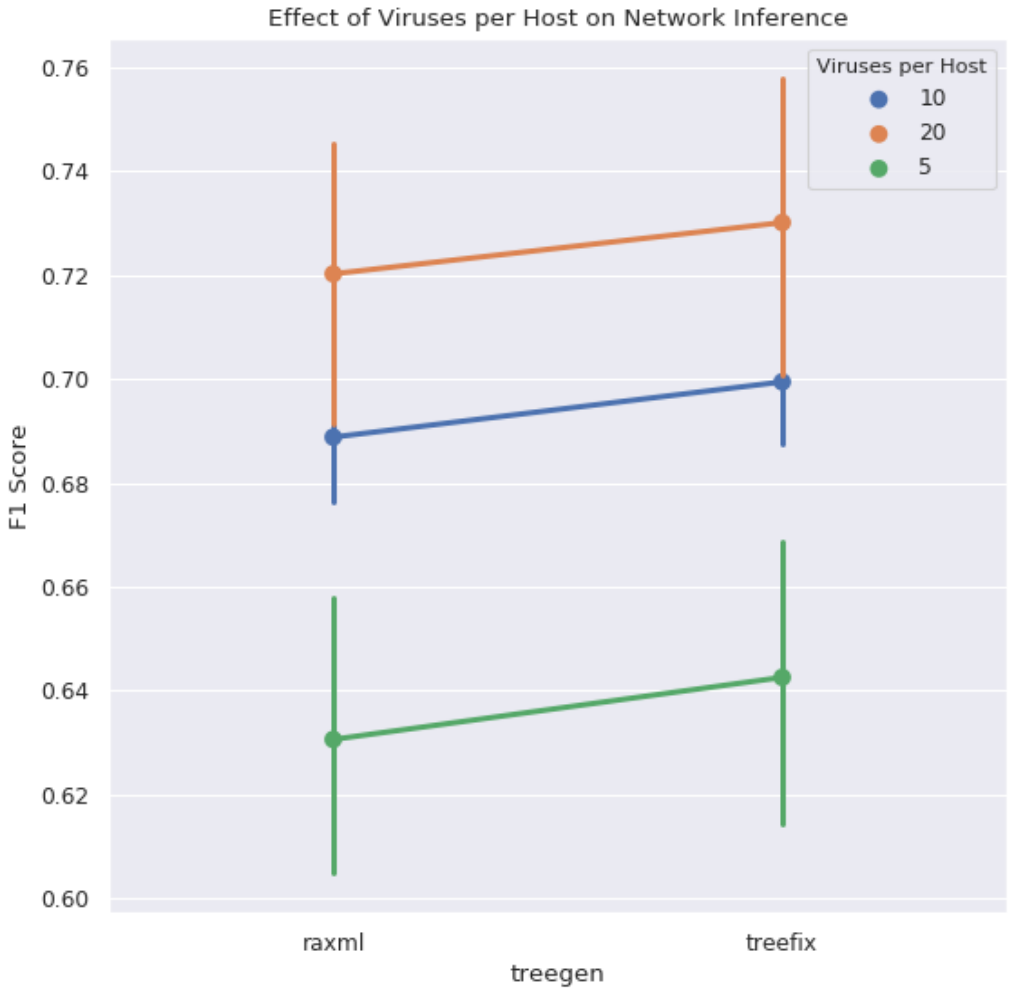


Figure: Varied Number of Viruses per Host

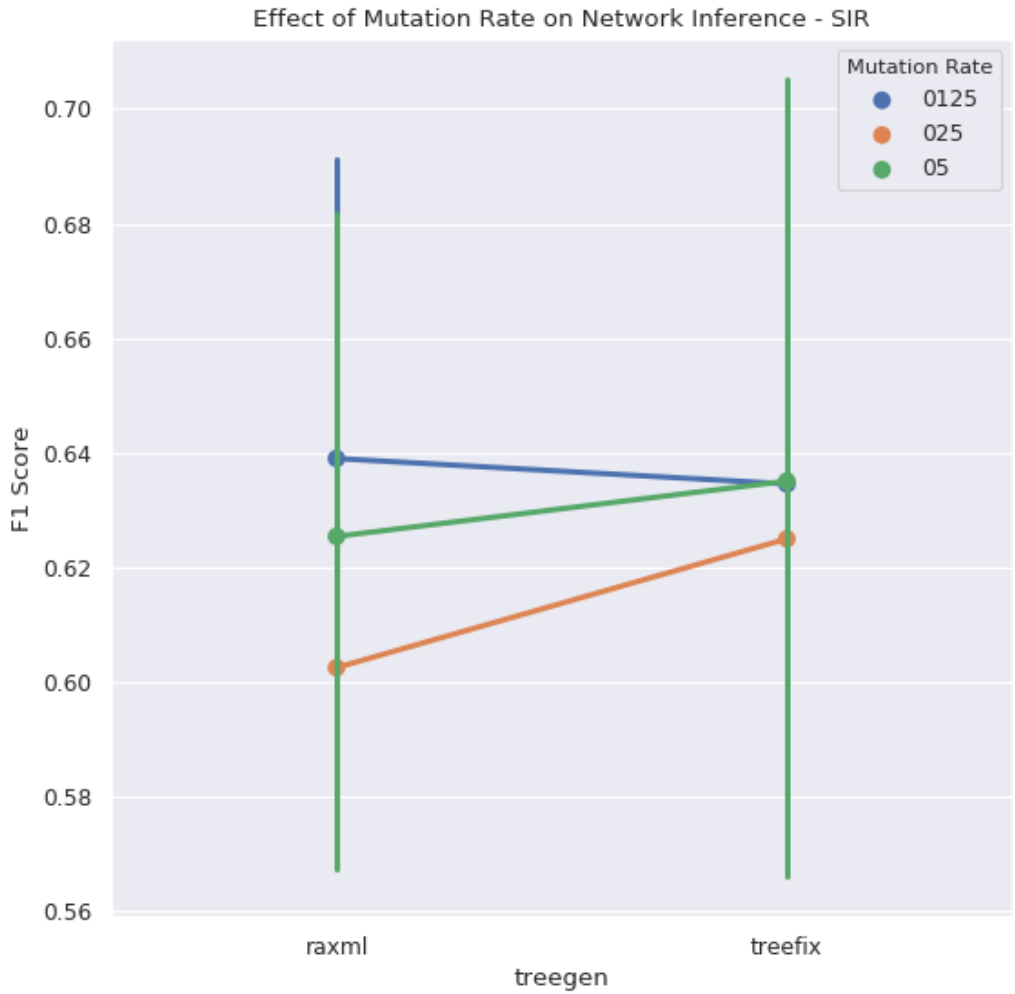


Figure: Varied Mutation Rate - SEIR

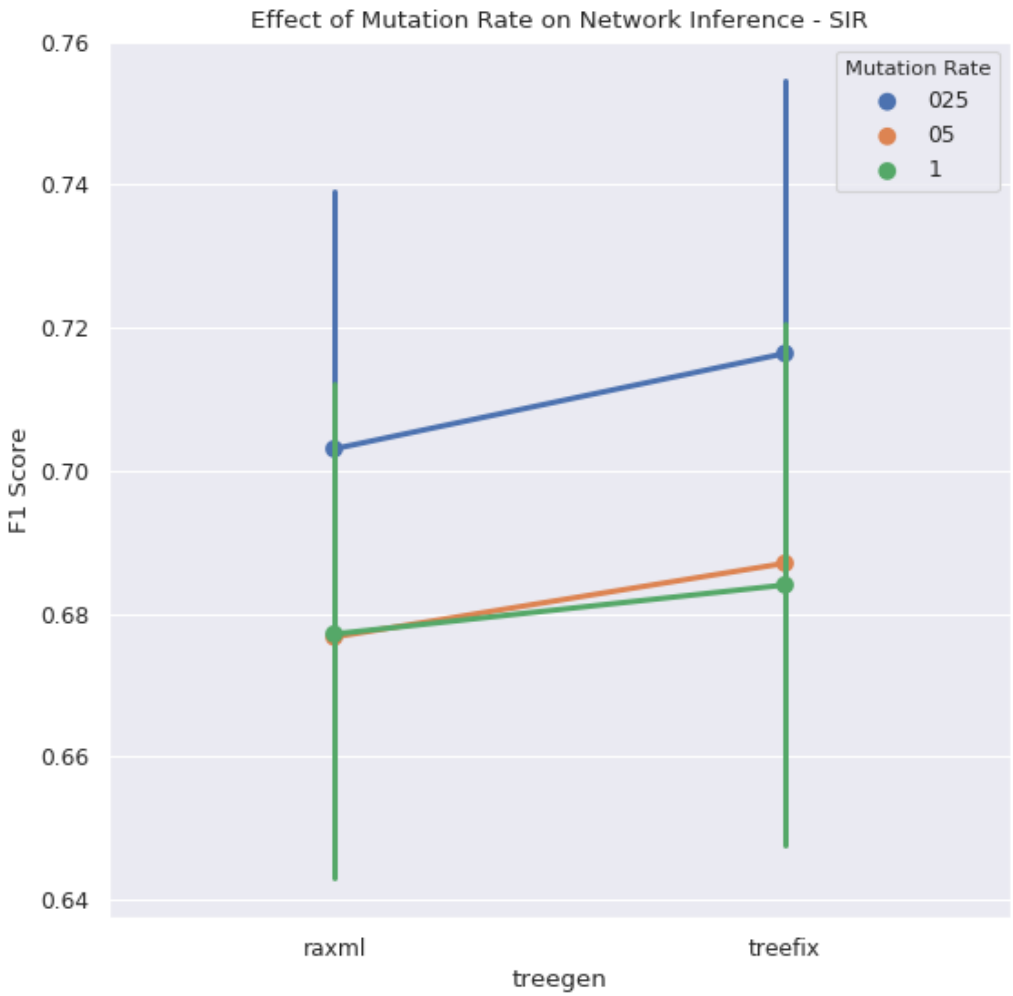


Figure: Varied Mutation Rate - SIR

Future Work

1. Evaluate performance of TreeFix-VP compared to MCMC methods using a single sequence
2. Evaluate effect of using multiple sequences per host on network inference
3. Evaluate performance of TreeFix-VP when some hosts are unsampled

References

1. Mukul S. Bansal, "Phylogenetic Error-Correction for Viral Transmission Network Inference. CAME 2017.
2. Mukul S. Bansal*, Yi-Chieh Wu*, Eric J. Alm, and Manolis Kellis. "Improved Gene Tree Error Correction in the Presence of Horizontal Gene Transfer." Bioinformatics. 2015. doi: 10.1093/bioinformatics/btu806
3. Didelot, Xavier, et al. "Genomic Infectious Disease Epidemiology in Partially Sampled and Ongoing Outbreaks." Molecular Biology and Evolution, 2017, doi:10.1093/molbev/msw275.
4. Hall, Matthew, et al. "Epidemic Reconstruction in a Phylogenetics Framework: Transmission Trees as Partitions of the Node Set." PLOS Computational Biology, vol. 11, no. 12, 2015, doi:10.1371/journal.pcbi.1004613.
5. Niema Moshiri, Manon Ragonnet-Cronin, Joel O. Wertheim, Siavash Mirarab, "FAVITES: simultaneous simulation of transmission networks, phylogenetic trees, and sequences." bioRxiv 297267; doi:10.1101/297267

References

6. Alexandros Stamatakis. RAxML-VI-HPC: Maximum Likelihood-based Phylogenetic Analyses with Thousands of Taxa and Mixed Models. Bioinformatics 22(21):2688-2690, 2006
7. Wymant C, Hall M, Ratmann O, Bonsall D, Golubchik T, de Cesare M, Gall A, Cornelissen M, Fraser C; STOP-HCV Consortium, The Maela Pneumococcal Collaboration, and The BEEHIVE Collaboration. "PHYLOSCANNER: Inferring Transmission from Within- and Between-Host Pathogen Genetic Diversity." Mol Biol Evol. 2017 Nov 23. doi: 10.1093/molbev/msx304

Acknowledgements

Collaborators: Chengchen Zhang, Mukul Bansal, Ion Mandoiu, Alex Zelikovsky, Pavel Skums, Yury Khudyakov

Funding: NSF award CCF 1618347

Questions?

Supplementary Figures - Branch Length Distribution

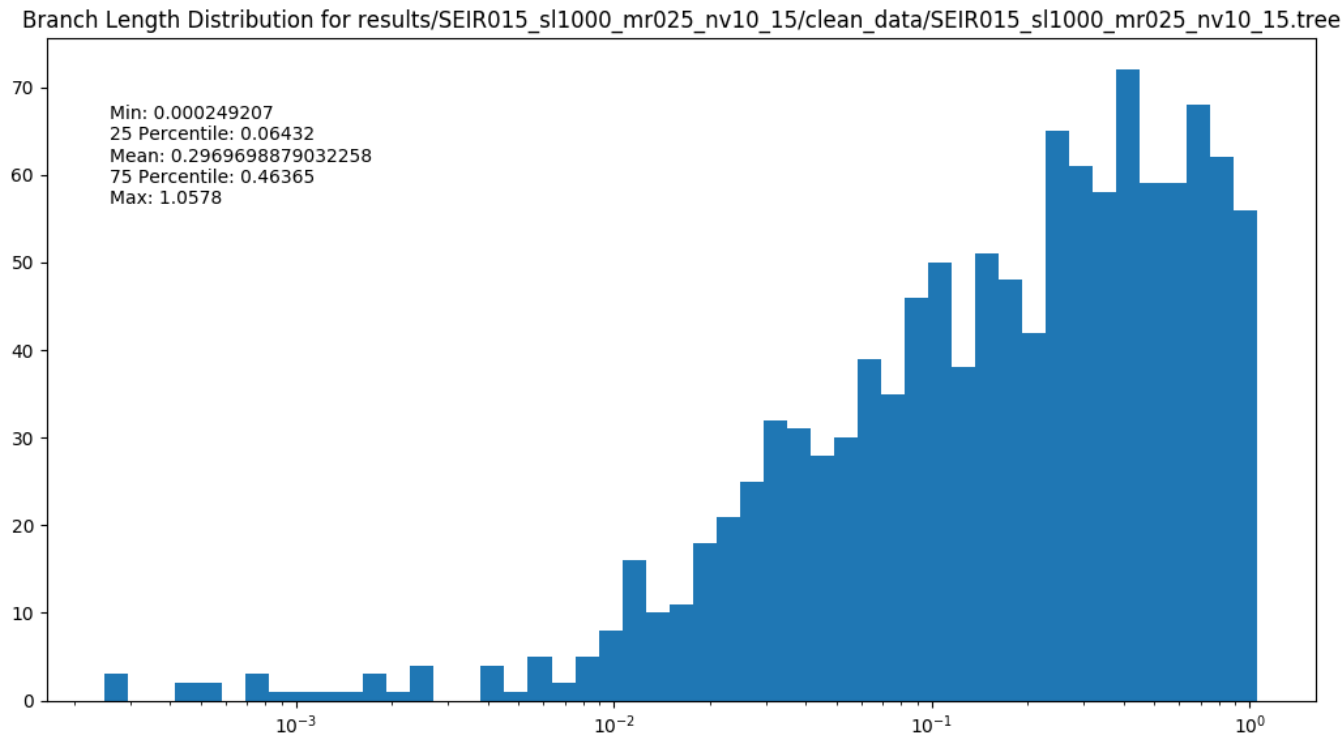


Figure: Sample Distribution of Branch Lengths