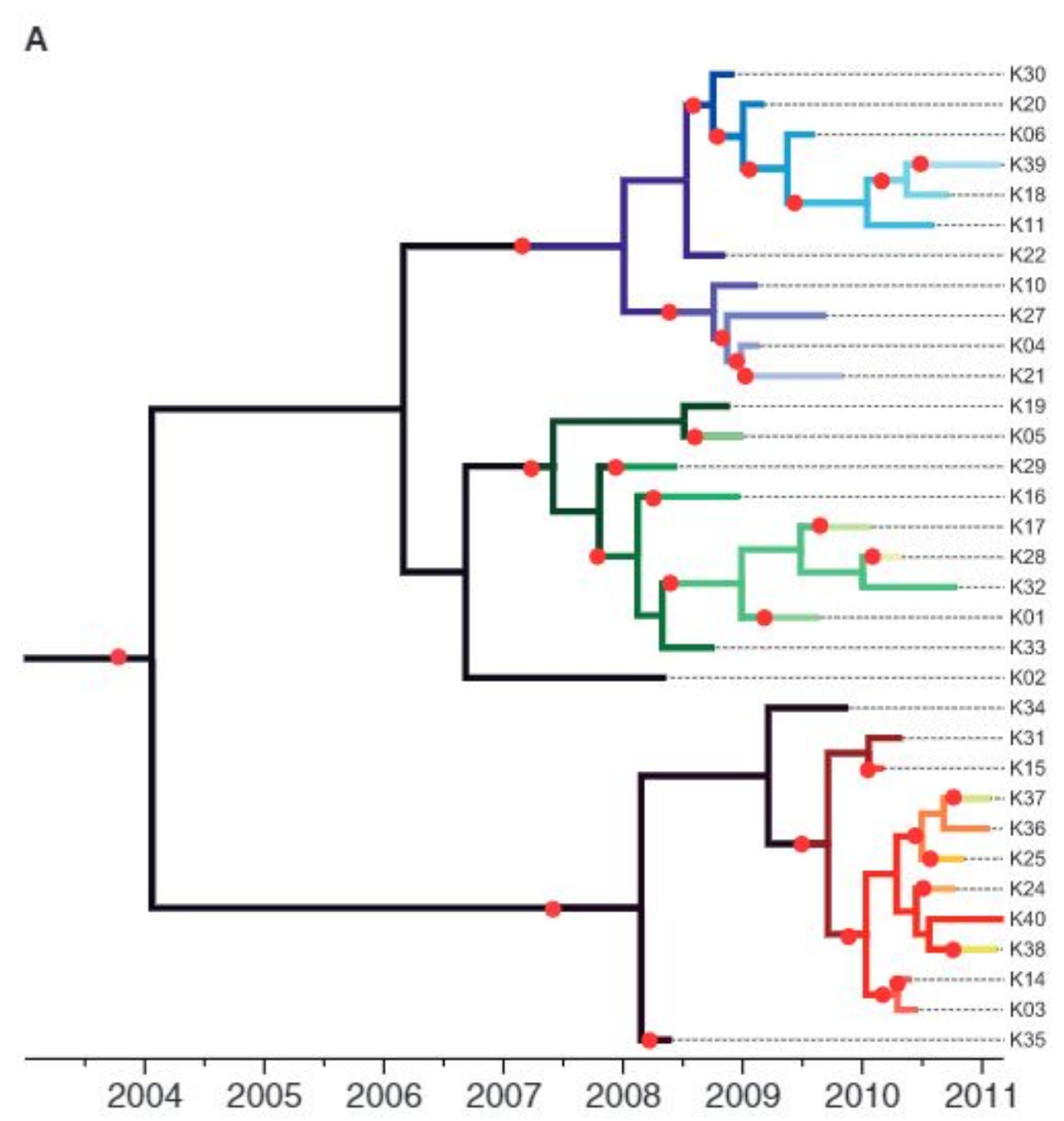
TreeFix-TP: Phylogenetic Error Correction for Infectious Disease Transmission Network Inference

Sledzieski, Zhang, Mandoiu, Bansal

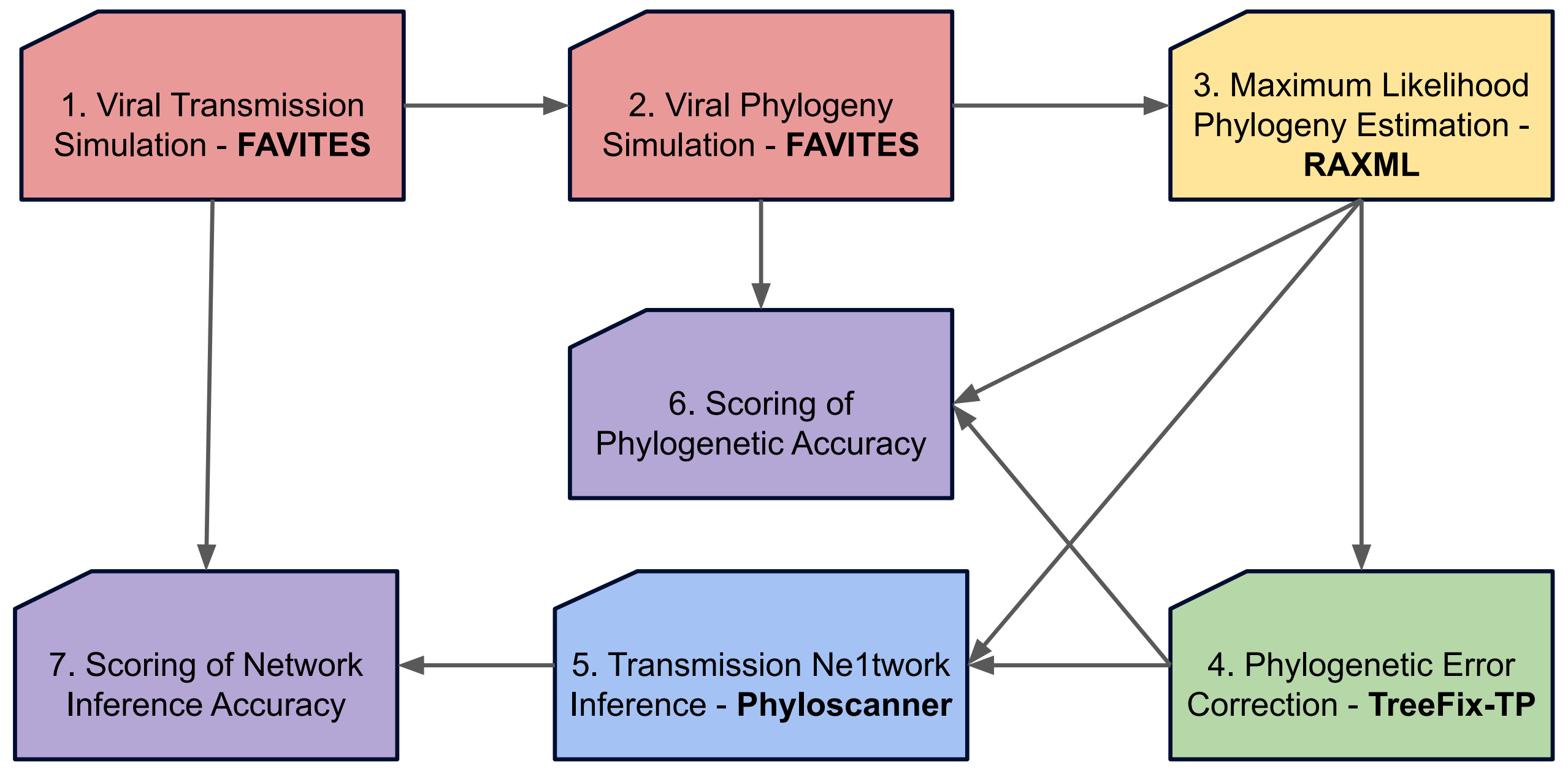
Problem

The problem of viral transmission network inference has been tackled from a variety of angles, many of which make use of sequence phylogenies [1]. The image at the right, from [4], shows how labeling internal nodes of viral phylogenies allows us to infer transmission. The goal of TreeFix-VP is to use **multiple sequences** per host to reconstruct **highly accurate phylogenies**, with the downstream goal of improving **transmission network inference**. In addition, we wish to reduce the need for MCMC or phylogenetic co-estimation and reconstruct the phylogeny in a highly scalable manner.



Testing Pipeline

1. Simulate viral transmission network using FAVITES [6]
2. Simulate viral phylogenies using FAVITES [6]
3. Reconstruct maximum likelihood phylogeny using **RAxML** [7]
4. Reconstruct error-corrected phylogeny using **TreeFix-TP**
5. Infer transmission with RAxML and TreeFix-TP trees using Phyloscanner [8], a parsimony-based program
6. Compare accuracy of RAxML tree and TreeFix-TP tree with the simulated tree, using **Robinson-Foulds distance**
7. Compare accuracy of Phyloscanner networks with the simulated network, by calculating the **F1 score**



1. Viral Transmission Simulation - **FAVITES**

2. Viral Phylogeny Simulation - **FAVITES**

3. Maximum Likelihood Phylogeny Estimation - **RAXML**

6. Scoring of Phylogenetic Accuracy

7. Scoring of Network Inference Accuracy

5. Transmission Ne1twork Inference - **Phyloscanner**

4. Phylogenetic Error Correction - **TreeFix-TP**

References

1. Andre, Mckenzie, et al. "Transmission Network Analysis to Complement Routine Tuberculosis Contact Investigations." American Journal of Public Health, vol. 97, no. 3, 2007, pp. 470–477., doi:10.2105/ajph.2005.071936.
2. Mukul S. Bansal, ``Phylogenetic Error-Correction for Viral Transmission Network Inference. CAME 2017.
3. 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
4. Didelot, Xavier, et al. ``Genomic Infectious Disease Epidemiology in Partially Sampled and Ongoing Outbreaks." Molecular Biology and Evolution, 2017, doi:10.1093/molbev/msw275.
5. 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.
6. 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
7. Alexandros Stamatakis. RAxML-VI-HPC: Maximum Likelihood-based Phylogenetic Analyses with Thousands of Taxa and Mixed Models. Bioinformatics 22(21):2688-2690, 2006
8. 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

Algorithm & Design

Inputs: Multiple sequence alignment, maximum likelihood phylogeny, host-sequence mapping
Output: Error corrected viral phylogeny

TreeFix-VP balances sequence information with host information by searching among candidate phylogenies to find the lowest cost phylogeny that is statistically equivalent to the maximum likelihood phylogeny.

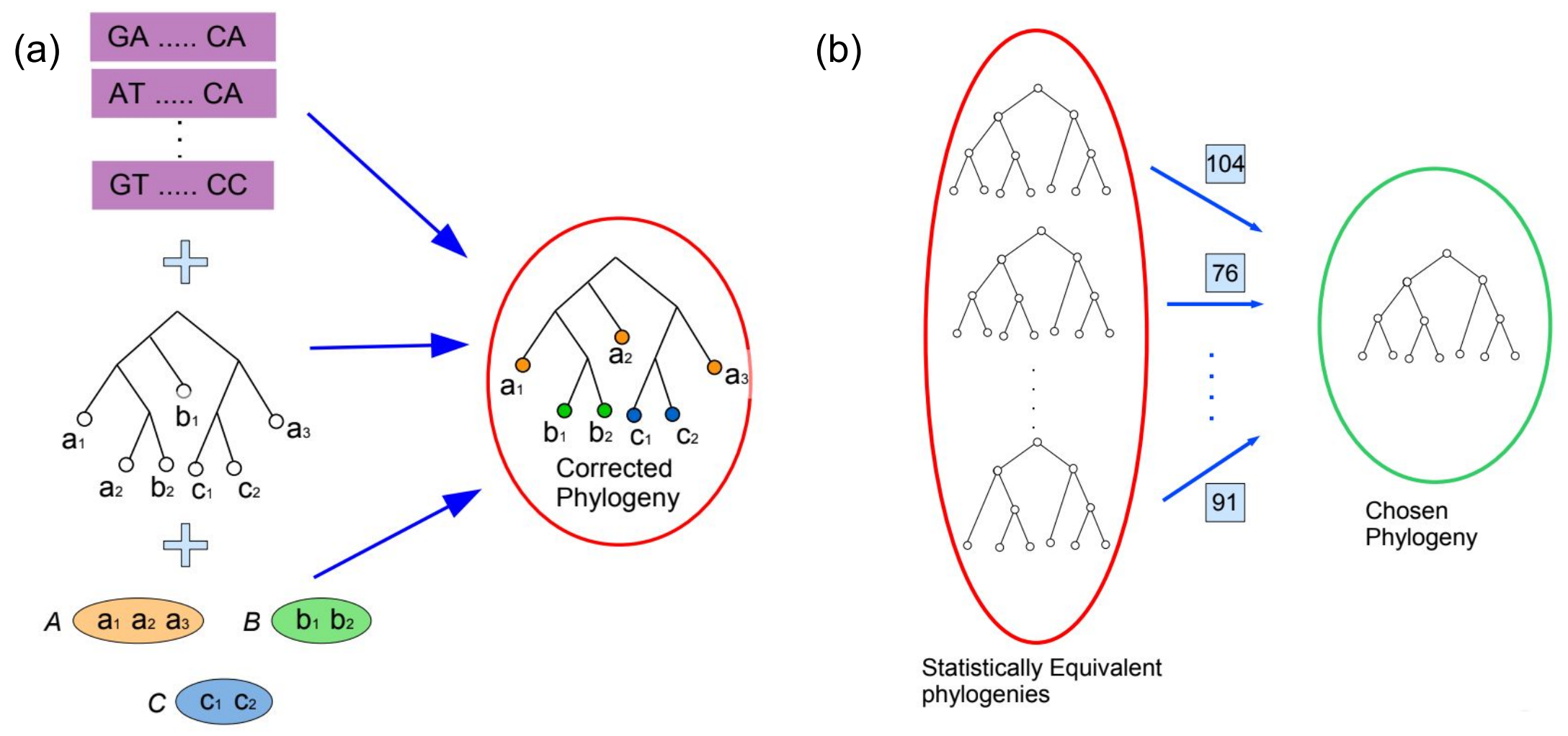


Fig 1: (a) Inputs and output of TreeFix-VP (b) TreeFix-VP selects the lowest cost candidate from statistically equivalent phylogenies. [2]

We build off of the approach used in TreeFix-DTL [3]. NNI and SPR moves are used to propose a new candidate tree, which is evaluated using the **Shimodaira-Hasegawa (SH)** statistical likelihood test. If the candidate tree is statistically equivalent to the maximum likelihood tree it is evaluated using the Fitch cost module. The parsimony score for the candidate tree is evaluated using **Fitch's algorithm** for the small parsimony problem (see below). If the calculated cost is less than the current minimum, the candidate tree is set as the current best tree. Once a preset number of candidate trees have been evaluated, the current best tree is used as the starting point for the next iteration. In our tests, we ran TreeFix-VP for 5000 iterations.

Fitch's algorithm works for a tree T by creating a set associated with each node, where the Set(i) = {Host(i)} if i ϵ L(T), or {} otherwise. **Equation 1** is applied in post-order to all nodes of T. Every time a set union is taken, we increment cost by 1. This metric works to evaluate a candidate tree because it is **biologically meaningful** and **computationally scalable**. Set union is taken when there is a change in state along an edge, which corresponds to a transmission (either direct or indirect). Therefore, the minimum cost phylogeny is the one which requires the fewest number of transmissions, and is consistent with the principles of parsimony.

Equation 1:

$$Set_N = Set_{N_l} \cap Set_{N_r} \neq \emptyset \ ? \\ Set_{N_l} \cap Set_{N_r} : Set_{N_l} \cup Set_{N_r}$$

For n sequences and k hosts, Fitch's algorithm runs in **O(nk)** time. However, for our implementation of Fitch's algorithm, we represented the set of each node as a bitstring where each bit represented a host. This allowed us to take advantage of system word operations to calculate set intersection and union in k/64 time. In practice, this means we only need 1-2 operations per internal node, so the cost calculation is reduced to **O(n)**. This optimization allows TreeFix-VP to increase the rate at which it considers candidate phylogenies.

$$\begin{aligned} A &= 01000100 = \{3,7\} \\ B &= 10010100 = \{3,5,8\} \\ A \cup B &= 11010100 = \{3,5,7,8\} \\ A \cap B &= 00000100 = \{3\} \end{aligned}$$

Fig 2: Taking advantage of bit operations allows us to represent each node set in an extremely compact way, and allows us to compute set operations in a greatly reduced number of system operations. This figure shows an example of how sets are represented, and how intersection and union are taken.

Results & Discussion

Phylogenetic Accuracy: Phylogenetic reconstruction results are highly dependent on transmission model - using the SEIR model of transmission (which mimics single-source outbreaks), we see an average improvement in RF distance of 6.779%, and as high as 46.154%. 48.6% of trials showed a decrease in RF distance, while 42.9% showed no change. 8.6% had an increase in RF distance. Using the SIR model, the average improvement in RF distance was only 3.66%. Figure 3 shows that TreeFix-VP is robust to variance in sequence length, number of viruses per host, and scale factor.

Network Inference: While we saw an overall improvement in transmission network accuracy, the results are much less consistent and the improvement in phylogenetic accuracy doesn't seem to carry over as strongly. Phyloscanner had an increase in average F1 score using the TreeFix-VP tree as compared to the RAxML tree, but it was only marginal (<0.01). In addition, network inference is not robust to varied parameters, as shown in Figure 4. We modified transmission model, sequence length, number of viruses per host, and scale factor of the tree. While TreeFix-VP was on average closer to the true tree across all parameters, the inferred networks were no more accurate and often less accurate than those inferred using the the RAxML tree.

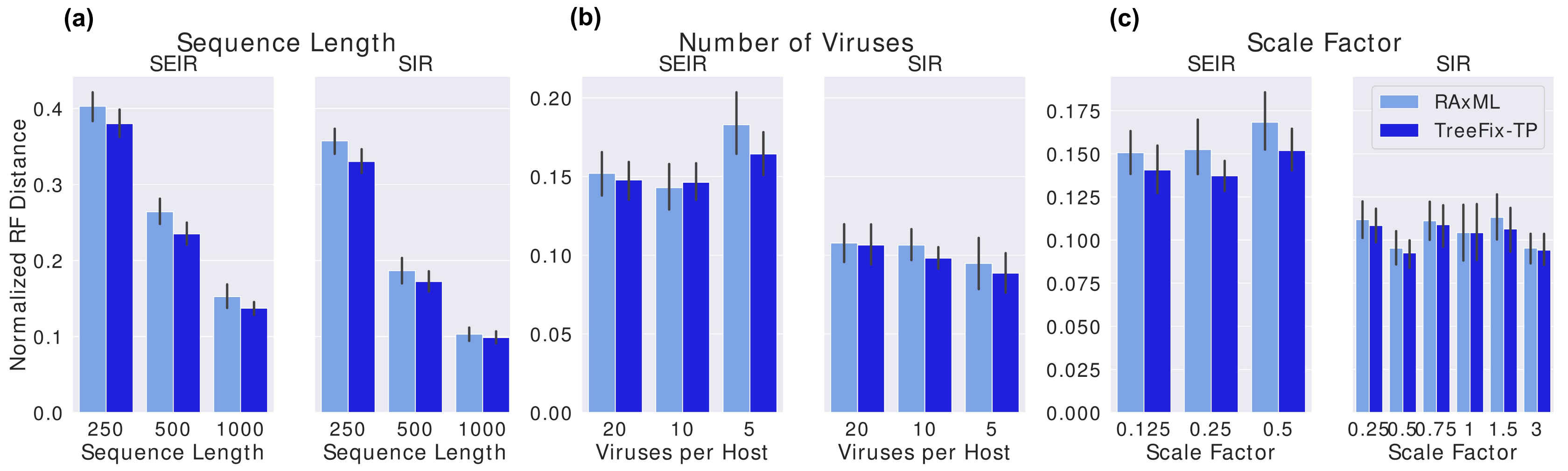


Fig. 3: Improvement in Phylogenetic Reconstruction Accuracy. Normalized Robinson-Foulds (RF) distance from the simulated phylogeny for reconstructions with both RAxML and TreeFix-TP under a variety of settings. TreeFix-TP reconstructs the most accurate trees across all data sets. **(a)** RF distance for varied sequence lengths. Trees are in general more accurate with longer sequences, and TreeFix-TP improves upon RAxML to a greater extent with shorter sequences. **(b)** RF distance for varied numbers of viruses sampled from each host. TreeFix-TP has the largest improvement when fewer viruses are sampled. **(c)** RF distance across multiple different scale factors. TreeFix-TP reconstructed the most accurate phylogenies with all scale factors.

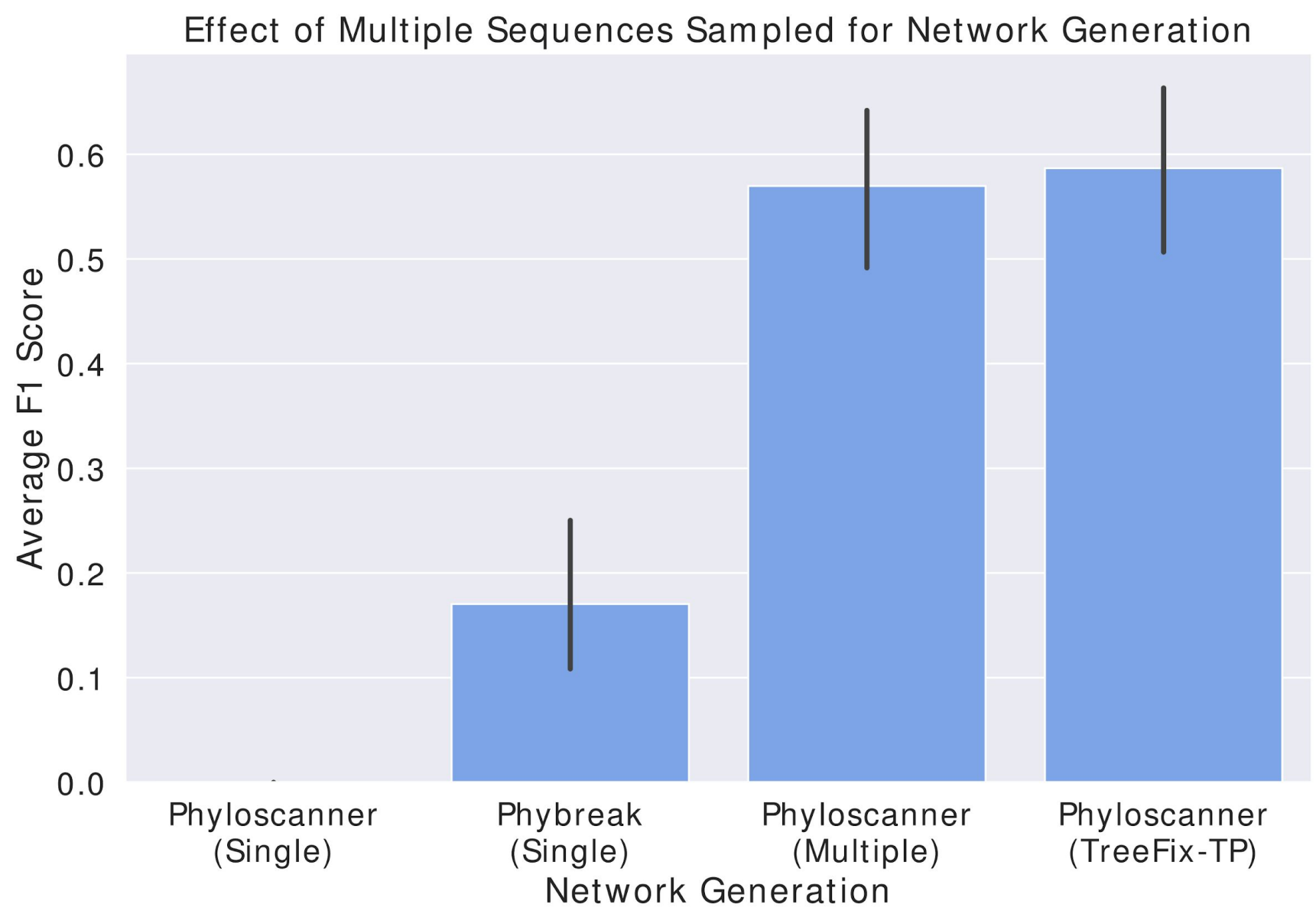


Fig. 5: Effect of Multiple Sequences on Transmission Network Inference. Average F1 score of network inference using both a single sequence per host and multiple sequences per host is shown. Networks reconstructed using Phyloscanner with multiple sequences had a significantly higher F1 score, regardless of whether RAxML or TreeFix-TP was used to reconstruct the phylogeny. When only a single sequence per host was used, network reconstruction was much less accurate. This shows that development of methods which can take advantage of multiple sequences may lead to more accurate inference of transmission networks.

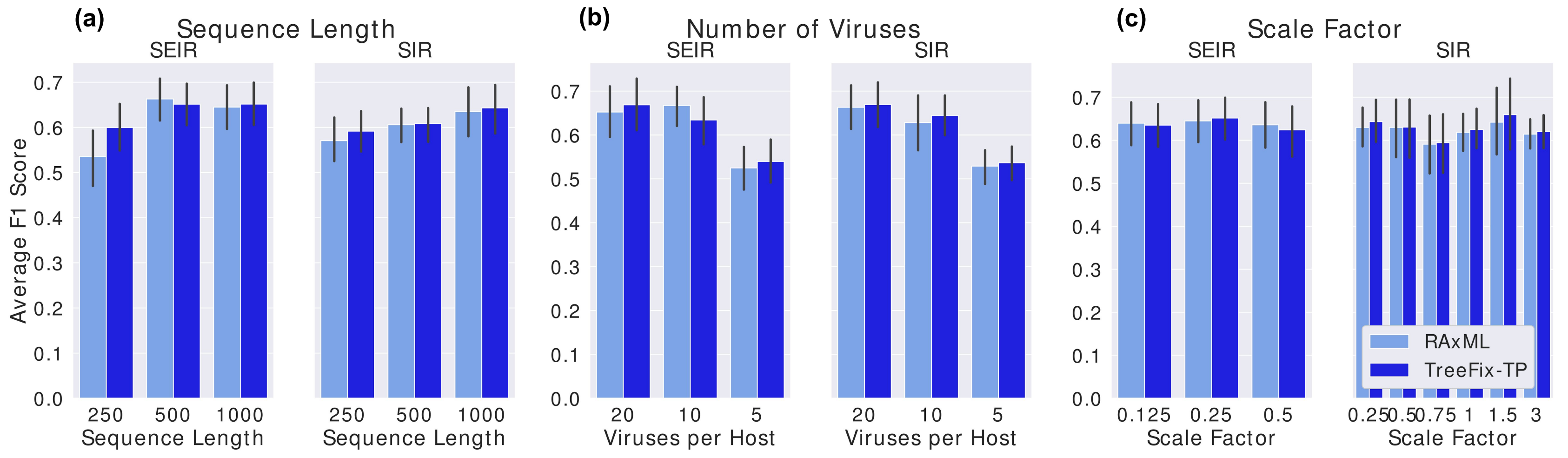


Fig. 4: Improvement in Transmission Network Inference. Average F1 score is shown for networks reconstructed from data sets with varying parameters. TreeFix-TP performs equally well or better than RAxML in most settings, and is especially useful when less sequence information is available. **(a)** Network reconstruction accuracy for multiple different sequence lengths. Reconstruction tends to be more accurate with longer sequences. **(b)** Network reconstruction accuracy for varying numbers of viral strains sampled from each infected host. Reconstruction tends to be more accurate with more sequences per host. **(c)** Network reconstruction accuracy for all different phylogeny scaling factors tested. There is no clear relationship between scale factor and reconstruction accuracy.

Discussion: Our results for phylogenetic error correction are promising. Since we are using a search heuristic, we can not expect an absolute improvement in all trials, but across most trials TreeFix-VP reconstructed a more accurate phylogeny than the maximum likelihood phylogeny. However, this improvement in accuracy didn't seem to translate to network inference. This suggests that perhaps simple parsimony is too naive to accurately recover viral transmission networks, even with improved phylogeny. Phylogeny-based inference of transmission networks is dependent on an accurate viral sequence phylogeny, and our results suggest that parsimony score, or minimum number of required transmissions, can be used as a proxy for goodness-of-fit of a given tree, however we may need to consider more sophisticated methods to take advantage of the improvement in viral phylogenetic accuracy. Our work is novel in that it takes advantage of multiple sequences per host, rather than just a single sequence. Figure 5 shows that regardless of tree reconstruction method, network reconstruction benefits from sampling multiple sequences from each infected host. Future work involves developing methods which can take advantage of multiple sequences to improve network reconstruction accuracy, such as inferring the hosts of ancestral nodes using Sankoff's algorithm.

Samuel Sledzieski
B.S. Computer Science
Bioinformatics Concentration
Molecular and Cellular Biology Minor
engr.uconn.edu/~sas14053
samuel.sledzieski@uconn.edu

github.com/samsledje/TreeFix-TP/

TreeFix-TP: Phylogenetic Error Correction for Infectious Disease Transmission Network Inference is funded by NSF award CCF 1618347