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

Samuel Sledzieski, Chengchen Zhang, Ion Mandoiu, and Mukul S. Bansal†

Department of Computer Science and Engineering, University of Connecticut
Storrs, CT 06269, USA

†*Corresponding Author: mukul.bansal@uconn.edu*

Many existing methods for estimation of infectious disease transmission networks use a phylogeny of the infecting strains as the basis for transmission network inference, and accurate network inference relies on accuracy of this underlying evolutionary history. However, phylogenetic reconstruction can be highly error prone and more sophisticated methods can fail to scale to larger outbreaks, negatively impacting downstream transmission network inference.

We introduce a new method, TreeFix-TP, for accurate and scalable reconstruction of transmission phylogenies based on an error-correction framework. Our method uses intra-host strain diversity and host information to balance a parsimonious evaluation of the implied transmission network with statistical hypothesis testing on sequence data likelihood. The reconstructed tree minimizes the number of required disease transmissions while being as well supported by sequence data as the maximum likelihood phylogeny. Using a simulation framework for viral transmission and evolution and real data from ten HCV outbreaks, we demonstrate that error-correction with TreeFix-TP improves phylogenetic accuracy and outbreak source detection. Our results show that using TreeFix-TP can lead to significant improvement in transmission phylogeny inference and that its performance is robust to variations in transmission and evolutionary parameters. TreeFix-TP is freely available open-source from `https://compbio.engr.uconn.edu/software/treefix-tp/`.

Keywords: phylogeny reconstruction, transmission network inference, infectious disease, computational epidemiology

1. Background

The study of infectious disease has benefited greatly from advances in computational molecular epidemiology. The efficacy of public health efforts to combat the spread of these pathogens has rapidly expanded as technology improves – most notably, the onset of powerful high throughput or next-generation sequencing (NGS) methods has provided molecular epidemiologists with the ability to quickly and cheaply sequence the genomes of the infecting strains (viral or bacterial)[1] which in turn has opened the door for computational analysis of these sequences and of disease transmission. By understanding disease transmission, those investigating a disease can more effectively combat its spread. Computational methods for molecular

epidemiology have had a positive impact on public health in a number of cases,[2,3] and continue to be widely used for the study of infectious disease transmission,[4] including for the ongoing COVID-19 pandemic (e.g., through the popular `https://nextstrain.org/ncov/global` web portal).

Transmission network inference is a challenging computational problem, which has been reflected in the number of new methods developed for understanding disease transmission, especially that of rapidly-evolving RNA viruses.[5–10] A key challenge with studying the transmission of rapidly evolving RNA and retroviruses[11] is that they exist in the host as "clouds" of closely related sequences. These strain variants are referred to as *quasispecies* by virologists,[12–16] and the resulting genetic diversity of the strains circulating within a host has important implications for efficiency of virus transmission, virulence, disease progression, drug/vaccine resistance, etc..[17–21] The advent of next-generation sequencing technologies, has revolutionized the study of quasispecies, but most existing transmission network inference methods are unable to make use of the ability to sequence multiple distinct strain sequences per host. However, methods that explicitly consider multiple strain sequence per host have recently started to be developed; such methods include Phyloscanner,[7] SharpTNI,[22] and TNet.[23]

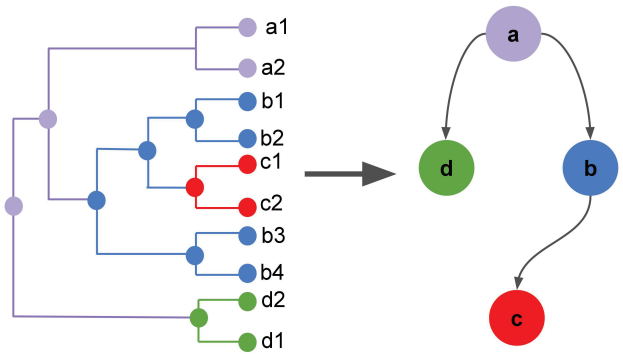

Fig. 1. **Phylogeny-based transmission network inference:** In this figure, internal nodes of the phylogenetic tree on the left are labeled by one of hosts a, b, c, or d, represented here by the different colors. This labeling of internal nodes causes some of the edges in the tree to have different labels at their two end points, and such edges represent transmission edges in the final transmission network. In the figure we see transitions from a to b, a to d, and b to c, yielding the transmission network shown on the right.

Some of the most powerful and widely used techniques for transmission network inference, including Phyloscanner,[7] SharpTNI,[22] and TNet,[23] are based on computing and using phylogenies of the infecting strains.[5–8,24] We refer to these strain phylogenies as *transmission phylogenies*. These phylogeny-based methods infer transmission networks through a host assignment for each node of the transmission phylogeny, where this phylogeny is either first constructed independently or is co-estimated along with the host assignment. Leaves of the transmission phylogeny are labeled corresponding to the host from which they are sampled, and an ancestral host assignment is then inferred for each node/edge of the phylogeny. This ancestral host assignment defines a transmission network, where transmission is inferred along any edge connecting two nodes labeled with different hosts. In the case of a rooted phylogeny, this coloring also confers direction of transmission, where the host for the ancestral sequence along a transmission edge is considered to be the source of the transmission, and the host of the child

sequence is considered to be the recipient. This is illustrated in Figure 1.

Two of the most widely-used methods for inference of transmission phylogenies are BEAST[25] and RAxML.[26] For instance, among existing transmission inference methods, TransPhylo[6] uses BEAST to infer a transmission phylogeny, while Phyloscanner[7] uses RAxML. BEAST uses Markov Chain Monte Carlo (MCMC) to estimate phylogenies and evolutionary parameters for several sophisticated models of evolution. Because the models implemented are highly complex, BEAST is prohibitively slow for use on anything other than small data sets. As a result, more scalable, but slightly less accurate, maximum likelihood based methods, such as the state-of-the-art RAxML method,[26] are often used in practice for inferring transmission phylogenies. There are also several methods which address transmission phylogeny reconstruction specifically from a transmission perspective, and use transmission information to inform phylogenetic inference. These methods often perform co-estimation of both the transmission phylogeny and network, and often model within-host evolution. BEASTlier[5] and Phybreak[8] both use Bayesian inference for co-estimation of transmission phylogeny and network, and so run into the same scalability issues as BEAST. Thus, even though accurate reconstruction of the transmission phylogeny has a direct impact on transmission network inference, all existing phylogenetic inference methods for transmission phylogenies are either prohibitively slow and unscalable or suffer from poor inference accuracy. Furthermore, none of these existing phylogenetic inference methods can take advantage of the information provided by multiple sequences from each infected host.

In this work, we introduce *TreeFix-TP*, a new method for reconstructing transmission phylogenies that is as scalable as RAxML but significantly more accurate. TreeFix-TP improves the accuracy of infectious disease transmission phylogenies using an error-correction approach. Specifically, TreeFix-TP leverages both sequence and host information to reconstruct more accurate phylogenies than maximum likelihood on its own by minimizing the number of inter-host transmissions while maintaining statistical support. Similar error correction approaches have been successfully used for reconstruction of gene trees;[27,28] however, these previous methods are based on leveraging a known species phylogeny to error-correct and improve gene trees, and they are therefore inapplicable to the current setting where the goal is to reconstruct the strain tree itself (analogous to the species tree). We address this problem by leveraging intra-host strain diversity and defining a fitness function based on minimizing the number of inter-host transmissions implied by the underlying phylogeny.

In this study, we compare the phylogenetic reconstruction accuracy of Treefix-TP to RAxML.[26] We show that TreeFix-TP reconstructs significantly more accurate transmission phylogenies than RAxML, and is robust to variations in transmission model, sequence length, rate of evolution, and number of viruses. Furthermore, we demonstrate the use of TreeFix-TP for improving source detection in 10 real-world HCV outbreaks.

2. Methods

2.1. *Minimizing inter-host transmissions*

The availability of multiple strain sequences from each host provides valuable additional information that can be used to improve the inference of transmission phylogenies. Consider an

ideal evolutionary scenario with a complete transmission bottleneck and no re-infection. In such a scenario, all sequences sampled from the same host should form a single monophyletic clade. For N hosts, this ideal case would result in a coloring with N single-color sub-graphs and would imply $N-1$ transmissions. Deviations from this ideal would be reflected in the transmission phylogeny and imply a few additional transmissions. Thus, when multiple strain sequences are available from each host, a biologically meaningful criterion for estimating the "correctness" of a transmission phylogeny is to minimize the number of implied inter-host transmissions. Note that the problem of computing the minimum number of implied inter-host transmissions on a given transmission phylogeny is equivalent to the well-known small parsimony problem in phylogenetics and can be solved very efficiently.[29] By minimizing the number of inter-host transmissions implied by a candidate phylogeny, and carefully avoiding over-fitting, we can improve the accuracy of a given phylogeny.

2.2. *Description of TreeFix-TP*

The input for TreeFix-TP is a multiple sequence alignment of infectious disease sequences, a maximum likelihood phylogeny constructed on the infectious disease sequences, and a mapping from all sequences to known hosts. TreeFix-TP aims to find the transmission phylogeny which is well supported by sequence data and has the minimum transmission cost. Using the maximum likelihood phylogeny as a starting point, we perform iterative local searches and evaluate each candidate tree using a statistical likelihood test and an evaluation of the transmission cost. Candidate phylogenies which are statistically equivalent to the maximum likelihood phylogeny, and with a lower transmission cost, are accepted and set as the starting point for the next local search iteration.

TreeFix-TP uses the Shimodaira-Hasegawa (SH) statistical likelihood test[30] to determine sequence support for a given phylogeny. This test considers two trees, in our case the maximum likelihood phylogeny and a candidate phylogeny, with the null hypothesis that the two trees are equally supported by the sequence data. The null hypothesis is rejected at a significance level α which can be defined by the user. If the null hypothesis fails to be rejected, the two trees are considered to be statistically equivalent

The transmission cost for a candidate phylogeny is calculated by solving an instance of the small parsimony problem using Fitch's algorithm.[29] The states at the leaves of a candidate phylogeny are the hosts from which each sequence is known to be sampled. Fitch's algorithm, then, calculates the minimum number of state changes required to generate the given phylogeny, which corresponds to minimizing the number of inter-host transmissions. In this case, we are concerned only with the cost of a candidate and not the internal assignments of hosts, so only the upward pass of Fitch's algorithm is performed. Full details of the algorithm and efficient implementation can be found in Section S1 in the Supplementary Material.

2.3. *Evaluation using simulated data sets*

2.3.1. *Data set generation*

To evaluate the performance of TreeFix-TP, we generated a number of simulated data sets across a variety of parameters and developed a testing pipeline to compare TreeFix-TP with RAxML (see Figure 2). Our simulated viral data sets were generated using FAVITES,[31] a recently developed framework for simultaneous simulation of viral transmission networks, phylogenetic trees, and sequences.

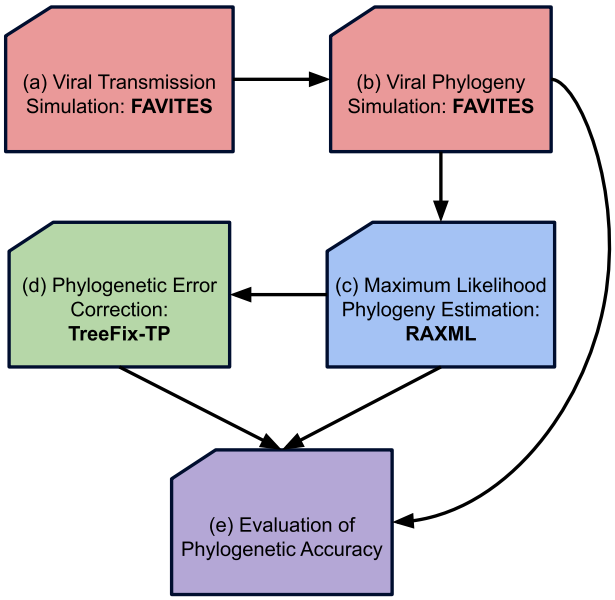

Fig. 2. **TreeFix-TP Testing Pipeline:** To evaluate TreeFix-TP, we first used FAVITES to generate a transmission network (a) and ground truth viral phylogeny (b). Maximum-likelihood phylogenies were then reconstructed from sequences using RAxML (c), and were error corrected with TreeFix-TP (d). The RAxML and TreeFix-TP phylogenies were compared using RF distance, as described in Section 2.3.2 (e).

A contact network was generated using the Barabasi-Albert model[32] with 1000 individuals each with 100 outgoing edges preferentially attached to high-degree nodes. One host was randomly selected to be the source of the infection. Transmission was simulated for a predefined amount of time, or until all hosts were recovered under one of two different compartmental models, either Susceptible-Exposed-Infected-Recovered (SEIR) or Susceptible-Infected-Recovered (SIR).[33] These models are parameterized by transition rates β, λ, and δ, where β is the rate of transition from susceptible to exposed in the SEIR model or susceptible to infected in the SIR model, λ is the rate of transition from exposed to infected (only in the SEIR model), and δ is the rate of recovery for infected individuals. In our simulation, we had four categories of data sets with variations on infection rate β to explore the effect of transmission model on reconstruction accuracy. λ and δ were set according to the infection rate. These parameters can be found in Supplementary Table S2.

Due to the simulation of latent periods, data sets generated under the SEIR model tend to exhibit an outbreak structure, where one high-degree individual infects several of its neighbors, followed by a period of low infection. When one of the newly-infected neighbors becomes infectious, another outbreak occurs. This is contrary to the SIR model, which tends to have a more periodic pattern of disease transmission. In addition to varying the transmission model, we simulated data sets with different rates of infection and recovery. This resulted in four categories of simulation with infection rates of 0.015, 0.003, 0.01, and 0.01 respectively. We group SEIR (0.015) and SEIR (0.01) together, and group SIR (0.003) and SIR (0.01) together since there was no significant difference between the transmission model parameter settings.

These transmission network parameter settings were generally based on the defaults suggested by FAVITES, with some adjustments as necessary for preventing the occurrence of long edges separating sequences from different hosts.

Internal evolution of the virus in infected hosts was simulated under a logistic-growth coalescent model. Each internal phylogeny was connected according to transmission to form a full transmission phylogeny. The branch lengths of this phylogeny were scaled to simulate different rates of sequence evolution. Sequences were simulated using the GTR + Γ model starting with a real HCV viral sequence from HCV outbreak data at the root (discussed in more detail in Section 3.2). The GTR rate matrix and gamma parameter were determined by applying RAxML to estimate parameters and construct a phylogeny for real sequences from an HCV outbreak. Thus, the simulated sequences are designed to reflect real rapidly-evolving RNA viral sequences.

By default, we simulated sequences of length 1000 nucleotides and sampled 10 sequences per infected host. We scaled the branch lengths of the phylogeny by 0.25 on data sets where the SEIR and SIR (0.003) models were used, and by 1.5 on data sets where the SIR (0.01) model was used. These scale factors were chosen so that the height of the tree would be approximately ten expected mutations per-site. We varied the sequence length, number of sequences per host, and mutation rate to quantify the robustness of TreeFix-TP to variance in sequence evolution. The list of all transmission and evolution simulation parameters can be found in Supplementary Table S2. For each of the four categories, we tested the effects of varying sequence length, number of samples per host, and scale factor, varying one of these parameters at a time from the default setting. Specifically, we simulated sequences of length 250, 500, and 1000, sampled 5, 10, and 20 sequences, and scaled the tree by double or half the default. Including the default setting, this resulted in 7 distinct parameterizations per category, or 28 total. Full specifications of parameters for each variation can also be found in Supplementary Table S2.

For each set of simulation parameters, we simulated 20 different data sets for a total of 560 simulated data sets. RAxML and TreeFix-TP were limited to 8GB of memory and 10 days, and due to these limitations we were able to reconstruct phylogenies using TreeFix-TP for 486 of these data sets. Of the 74 runs which did not complete, the simulated trees had an average of 733.43 leaves. For the 486 simulated data sets on which we obtained results, we had between 35 and 630 sequences, with an average of 223.41 leaves. The average number of transmissions was 22, and 95% of data sets had between 7 and 49 transmissions. Of the data sets for which we obtained results, only 6 had more than 60 transmissions.

2.3.2. *Evaluating reconstruction accuracy*

The accuracy of the reconstructed phylogeny was evaluated by calculating the Robinson-Foulds distance[34] between the true evolutionary history from the simulated data and both the maximum likelihood tree reconstructed by RAxML and the error-corrected tree reconstructed by TreeFix-TP. We calculated the average RF distances, normalized by the maximum possible RF distance (number of internal edges). We calculated the *RF percent decrease* as follows: Given simulated tree S, maximum likelihood tree R, and TreeFix-TP tree T, RF percent

decrease is given by $100 \times (RF(S,R) - RF(S,T))/RF(S,R)$. We calculated p-values using the one-tailed Wilcoxon Signed-Rank test implemented in Scipy 1.3.1. Additionally, we looked at the minimum transmission cost implied by the RAxML and TreeFix-TP trees. The cost of the TreeFix-TP tree is guaranteed to be no greater than that of the RAxML tree, but it is valuable to see by how much the transmission cost is decreased and the relationship between transmission cost and Robinson-Foulds distance. Note that we did not compare reconstruction accuracy against BEAST[25] since it is not scalable to data set sizes used in this study.

3. Results

3.1. *Phylogenetic error correction results*

For baseline evaluation, we compared the phylogenies reconstructed by TreeFix-TP and RAxML on 35 data sets corresponding to the SEIR transmission model, sequence length 1000, 10 sequences per host, and a mutation rate of 0.25. Among these trials, 48.6% of the data sets showed a decrease in RF distance with TreeFix-TP, while 42.86% saw no improvement and 8.57% saw an increase. The average RF percent decrease for trees which improved was 14.6%, and as high as 46.154%, while the average RF percent increase for those trees that got worse was only 3.644%. In every run where there was no improvement, the maximum likelihood tree generated with RAxML implied exactly as many or only one more transmission than the true number of transmissions, so the ability for TreeFix-TP to correct errors by minimizing transmission was limited. Across all 35 data sets, the average normalized RF distance of trees reconstructed with RAxML was 0.152, while trees reconstructed with TreeFix-TP had an average normalized RF distance of 0.137 ($p = 0.0003$, Wilcoxon Signed-Rank). The overall average RF percent decrease was 9.99%.

We also evaluated 32 data sets corresponding to the SIR transmission model, sequence length 1000, 10 sequences per host, and a mutation rate of either 1.5 or 0.25 (aggregated over both transmission rate categories). The average normalized RF distance of trees constructed with RAxML was 0.103, while trees reconstructed with TreeFix-TP had an average normalized RF distance 0.098 ($p = 0.006$, Wilcoxon Signed-Rank). The magnitude of improvement is impacted by the large number of no-change error corrections. Specifically, under the SIR model of transmission, 68.75% of runs had no-change, while 28.13% showed a decrease in RF distance, and the remaining 3.13% showed an increase. The overall average RF percent decrease was 4.36%, but those which improved had an average RF percent decrease of 14.116%, and as high as 28.57%. For those which got worse, the average RF percent increase was 9.8%. A comparison of these results across the SEIR and SIR transmission models suggests that error correction might be more effective under a model of transmission that includes a latent period, which results in transmissions patterns which more closely reflect outbreaks.

Impact of varying sequence length To evaluate the robustness of TreeFix-TP to the amount of sequence information available, we varied sequence length from the base 1000 nucleotides to 250 and 500 nucleotides (Figure 3a). Under the SEIR model, we found that TreeFix-TP continued to improve the accuracy of phylogenetic reconstruction with shorter sequence lengths, and that sequence length didn't seem to have a large effect on the ability

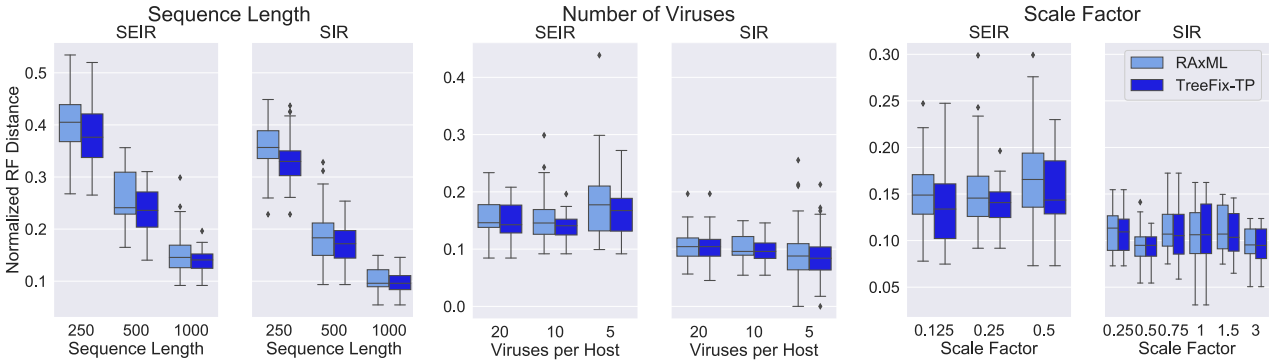

Fig. 3. **Robustness of phylogeny reconstruction to different parameters:** Normalized Robinson-Foulds (RF) distance from the simulated phylogeny for reconstructions with both RAxML and TreeFix-TP under a variety of settings. TreeFix-TP reconstructs the most accurate trees across all data sets. **(a)** 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.

of error correction to improved the accuracy of the phylogeny. At sequence length 1000, the average normalized RF distance decreased by 9.99% from 0.152 to 0.137 after error correction ($p = 0.0003$, Wilcoxon Signed-Rank). At length 500, this was a decrease of 11.03% from 0.264 to 0.235 ($p = 1e-5$, Wilcoxon Signed-Rank). At sequence length 250, the average RF distance decreased by an average of 5.68% from 0.403 to 0.380 ($p = 6e-5$, Wilcoxon Signed-Rank). As expected, the absolute error rate increases sharply, for both RAxML and TreeFix-TP, as sequence length decreases.

Under the SIR model, we found the error correction continued to have an impact at all sequence lengths, and that error correction was more effective at shorter sequence lengths. With sequence length of 1000, the average RF distance decreased by 4.36% (0.103 to 0.099 normalized RF, $p = 0.006$, Wilcoxon Signed-Rank). At length 500, there was a 7.65% decrease (0.187 to 0.172 normalized RF, $p = 0.0001$, Wilcoxon Signed-Rank), and at length 250 there was a 7.59% decrease (0.357 to 0.330 normalized RF, ($p = 9e-5$, Wilcoxon Signed-Rank). Under this model, error correction seems to be more effective with shorter sequences, likely because longer sequences contain more information which allows maximum likelihood methods to reconstruct a relatively accurate tree before any error correction occurs.

Impact of varying number of viruses We observed the effect of sampling different numbers of viruses from each infected host, from the default of 10 to 5 and 20 viral sequence samples (Figure 3b). TreeFix-TP reconstructed more accurate phylogenies in each case, with the largest overall improvement occurring for trees with 5 sequences from each host. Under the SEIR model, with 20 viruses, there was an average RF distance decrease of 2.78% (0.152 to 0.148 normalized RF, $p = 0.018$, Wilcoxon Signed-Rank). With 10 and 5 viruses, there were larger decreases of 9.99% and 10.18% respectively (0.152 to 0.137 and 0.183 to 0.164 normalized RF, $p = 0.0003$, $p = 8e-5$ Wilcoxon Signed-Rank). Under the SIR model, with 20 viruses,

there was an decrease in average RF distance of only 1.56% (0.108 to 0.106 normalized RF, $p = 0.072$, Wilcoxon Signed-Rank). This decrease was 4.36% with 10 viruses and 6.52% with 5 viruses (0.103 to 0.099 and 0.096 to 0.089 normalized RF, $p = 0.006$, $p = 0.013$ Wilcoxon Signed-Rank). Error correction seems to be more effective with fewer viruses, which matches the intuition about sequence length - that more sequence data leads to originally accurate phylogenies, and less potential for error correction.

Impact of varying scale factor We found that TreeFix-TP is also robust to various rates of sequence evolution (Figure 3c). Under the SEIR model of evolution, scale factors of 0.125, 0.25, and 0.5 resulted in a decrease in average RF distance by 6.64%, 9.99%, and 9.76% respectively (0.151 to 0.141, 0.152 to 0.137, 0.168 to 0.152 normalized RF, $p = 0.004, 0.0003, 0.0006$ Wilcoxon Signed-Rank). Under the SIR model, we used two different sets of scale factors dependent on the disease transmission parameters. Aggregated across SIR (0.003) and SIR (0.01), we tested scale factors of 0.25, 0.5, 0.75, 1, 1.5, and 3. These scale factors had average RF percent decreases of 3.05%, 2.87%, 2.02%, 0.07%, 5.97%, and 1.20% (0.111 to 0.108, 0.095 to 0.093, 0.111 to 0.109, 0.1043 to 0.1042, 0.113 to 0.106, and 0.095 to 0.094 normalized RF, $p = 0.045, 0.054, 0.034, 0.327, 0.021, 0.250$). As expected, the overall RF distances tended to be larger for very small and very large scale factors, which indicates that a reasonable rate of evolution is important to overall phylogenetic reconstruction accuracy, but plays less of an impact on error correction.

3.2. *Source recovery in HCV outbreaks*

We also evaluated the impact of using TreeFix-TP on real data sets of HCV outbreaks made available by the CDC.[9] In total, there are 10 different data sets, each representing a separate HCV outbreak. Each of these outbreak data sets contains between 2 and 19 infected hosts and a few dozen to a few hundred strain sequences. For each of these 10 outbreaks, the source host of the outbreak is known (through the CDC's epidemiological efforts). We used a simple phylogenetic pipeline to infer a source for each of these 10 data sets as follows: We constructed phylogenetic trees using RAxML and TreeFix-TP and rooted them using two of the most widely used rooting methods, balanced rooting (implemented in RAxML[26]) and midpoint rooting.[35,36] We then used Sankoff's algorithm for the small parsimony problem[37] to label the internal nodes of these phylogenies with hosts and report the host assignment at the root as the inferred source of that outbreak. (Note that Phyloscanner also uses Sankoff's algorithm to label internal nodes of the phylogeny, but we chose not to use Phyloscanner directly because it is very conservative in its host assignments and often leaves nodes unlabeled.) Using the RAxML trees, the source was correctly identified in 6 (balanced rooting) and 7 (midpoint rooting) of the 10 outbreaks. In contrast, the trees reconstructed by TreeFix-TP correctly identified the source in 8 out of the 10 outbreaks with both rooting strategies. Furthermore, the outbreaks correctly identified by RAxML were a strict subset of those identified by TreeFix-TP.

3.3. *Running time and scalability*

Using its default number of iterations (5000) TreeFix-TP required an average of approximately 37 hours for each run, but this running time varied depending on the number of tips and length of sequence. TreeFix-TP took less than an hour and a half for trees of 50-60 tips, but upwards of 200 hours for trees with more than 500 tips and 1000 nucleotide-length sequences. On average, runs took fewer than 9 minutes per tip, and scaled linearly in tree size, number of hosts, and sequence length.

4. Discussion and Conclusions

In this paper, we have introduced a new method, TreeFix-TP, for more accurate and scalable reconstruction of infectious disease transmission phylogenies when multiple strain sequences are sampled from each infected host, and demonstrated its impact on phylogenetic inference and outbreak source detection. TreeFix-TP uses an error-correction approach where it seeks to improve a given maximum-likelihood phylogeny of the infecting strains by using additional information about which host each strain was sampled from and balancing it with sequence-only likelihood using a statistical hypothesis testing framework. As our experimental results show, TreeFix-TP consistently reconstructs more accurate phylogenies than the state-of-the-art maximum-likelihood phylogeny inference method RAxML. We also demonstrate how TreeFix-TP can be used to augment existing phylogeny-based pipelines for transmission network inference by error correcting the phylogenies before they are used for network inference or outbreak source detection.

Going forward, it would be worthwhile to develop even more advanced, yet scalable, methods for construction of transmission phylogenies. As our experimental results show, even though the absolute error rate of TreeFix-TP phylogenies is often significantly lower than that of RAxML trees, this absolute error rate still remains quite high overall even after error correction. This is partly because the ability of TreeFix-TP to error-correct depends on the number of different hosts represented in the phylogeny, rather than on the size of the tree itself. In the future, it may be possible to use additional information about within-host strain evolution to further improve transmission phylogeny inference.

Acknowledgments

The authors wish to thank Dr. Pavel Skums (Georgia State University) and the Centers for Disease Control for sharing their HCV outbreak data. This work was supported in part by NSF award CCF 1618347 to IM and MSB.

Authors' Contributions

SS contributed to the theoretical results, implemented the software, performed the experimental study, analyzed the results, and contributed to the writing of the manuscript. CZ contributed to initial project development and conducted the experimental analysis on real data. IM helped supervise the research and contributed to writing the manuscript. MSB conceived the research project, supervised the research, and contributed to the writing of the manuscript.

Supplementary Material

Supplementary material can be found at:
`https://compbio.engr.uconn.edu/treefix-tp_supplement/`

References

1. S. C. Shuster, Next-generatino sequencing transforms today's biology, *Nature* **5** (dec 2007).
2. A. Grulich, A. Pinto, A. Kelleher, D. Cooper, P. Keen, F. Di Giallonardo, C. Cooper and B. Telfer, A10 Using the molecular epidemiology of HIV transmission in New South Wales to inform public health response: Assessing the representativeness of linked phylogenetic data, *Virus Evolution* **4** (04 2018).
3. D. Clutter, R. W. Shafer, S.-Y. Rhee, W. J. Fessel, D. Klein, S. Slome, B. A. Pinsky, J. L. Marcus, L. Hurley, M. J. Silverberg and S. L. Kosakovsky Pond, Trends in the Molecular Epidemiology and Genetic Mechanisms of Transmitted Human Immunodeficiency Virus Type 1 Drug Resistance in a Large US Clinic Population, *Clinical Infectious Diseases* **68**, 213 (05 2018).
4. E. O. Romero-Severson, I. Bulla and T. Leitner, Phylogenetically resolving epidemiologic linkage, *Proceedings of the National Academy of Sciences* **113**, 2690 (mar 2016).
5. M. Hall, M. Woolhouse and A. Rambaut, Epidemic Reconstruction in a Phylogenetics Framework: Transmission Trees as Partitions of the Node Set, *PLoS Computational Biology* **11**, p. e1004613 (dec 2015).
6. X. Didelot, C. Fraser, J. Gardy, C. Colijn and H. Malik, Genomic infectious disease epidemiology in partially sampled and ongoing outbreaks, *Molecular Biology and Evolution* **34**, 997 (jan 2017).
7. C. Wymant, M. Hall, O. Ratmann, D. Bonsall, T. Golubchik, M. de Cesare, A. Gall, M. Cornelissen and C. Fraser, PHYLOSCANNER: Inferring Transmission from Within- and Between-Host Pathogen Genetic Diversity, *Molecular Biology And Evolution* **35**, 719 (mar 2017).
8. D. Klinkenberg, J. A. Backer, X. Didelot, C. Colijn and J. Wallinga, *Simultaneous inference of phylogenetic and transmission trees in infectious disease outbreaks* (PLoS, 2017).
9. P. Skums, A. Zelikovsky, R. Singh, W. Gussler, Z. Dimitrova, S. Knyazev, I. Mandric, S. Ramachandran, D. Campo, D. Jha, L. Bunimovich, E. Costenbader, C. Sexton, S. O'Connor, G. L. Xia and Y. Khudyakov, QUENTIN: Reconstruction of disease transmissions from viral quasispecies genomic data, *Bioinformatics* **34**, 163 (jun 2018).
10. S. L. Kosakovsky Pond, S. Weaver, A. J. Leigh Brown and J. O. Wertheim, HIV-TRACE (TRAnsmission Cluster Engine): a Tool for Large Scale Molecular Epidemiology of HIV-1 and Other Rapidly Evolving Pathogens, *Molecular Biology and Evolution* **35**, 1812 (01 2018).
11. J. W. Drake and J. J. Holland, Mutation rates among RNA viruses, *Proceedings of the National Academy of Sciences of the United States of America* **96**, 13910 (1999).
12. E. Domingo and J. Holland, RNA virus mutations and fitness for survival, *Annu Rev Microbiol* **51**, 151 (1997).
13. E. Domingo, M.-S. E., F. Sobrino, J. de la Torre, A. Portela, J. Ortin, C. Lopez-Galindez, P. Perez-Brena, N. Villanueva and R. Najera, The quasispecies (extremely heterogeneous) nature of viral rna genome populations: biological relevance – review, *Gene* **40**, 1 (1985).
14. M. E. M, J. McCaskill and P. Schuster, The molecular quasi-species, *Adv Chem Phys* **75**, 149 (1989).
15. M. Martell, J. Esteban, J. Quer, J. Genesca, A. Weiner, R. Esteban, J. Guardia and J. Gomez, Hepatitis c virus (hcv) circulates as a population of different but closely related genomes: quasispecies nature of hcv genome distribution, *Journal of Virology* **66**, 3225 (1992).
16. D. Steinhauer and J. Holland, Rapid evolution of rna viruses, *Annual Review of Microbiology* **41**, 409 (1987).
17. N. Beerenwinkel, T. Sing, T. Lengauer, J. Rahnenfuehrer and K. R. et al., Computational meth-

ods for the design of effective therapies against drug resistant HIV strains, *Bioinformatics* **21**, 3943 (2005).
18. N. G. Douek DC, Kwong PD, The rational design of an AIDS vaccine., *Cell* **124**, 677 (2006).
19. B. Gaschen, J. Taylor, K. Yusim, B. Foley and F. G. et al., Diversity considerations in HIV-1 vaccine selection, *Science* **296**, 2354 (2002).
20. J. Holland, J. de la Torre and D. Steinhauer, Rna virus populations as quasispecies, *Current Topics in Microbiology and Immunology* **176**, 1 (1992).
21. S.-Y. Rhee, T. Liu, S. Holmes and R. Shafer, HIV-1 subtype B protease and reverse transcriptase amino acid covariation, *PLoS Comput Biol* **3**, p. e87 (2007).
22. P. Sashittal and M. El-Kebir, SharpTNI: Counting and Sampling Parsimonious Transmission Networks under a Weak Bottleneck, *bioRxiv* (2019).
23. S. Dhar, C. Zhang, I. Mandoiu and M. S. Bansal, Tnet: Phylogeny-based inference of disease transmission networks using within-host strain diversity, in *Bioinformatics Research and Applications*, eds. Z. Cai, I. Mandoiu, G. Narasimhan and P. Skums (Springer Nature, 2020)
24. E. M. Volz, K. Koelle and T. Bedford, Viral Phylodynamics, *PLoS Computational Biology* **9**, p. e1002947 (mar 2013).
25. A. J. Drummond and A. Rambaut, BEAST: Bayesian evolutionary analysis by sampling trees, *BMC Evolutionary Biology* **7**, p. 214 (nov 2007).
26. A. Stamatakis, RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies, *Bioinformatics* **30**, 1312 (may 2014).
27. Y.-C. Wu, M. D. Rasmussen, M. S. Bansal and M. Kellis, TreeFix: Statistically Informed Gene Tree Error Correction Using Species Trees, *Systematic Biology* **62**, 110 (jan 2013).
28. M. S. Bansal, Y. C. Wu, E. J. Alm and M. Kellis, Improved gene tree error correction in the presence of horizontal gene transfer, *Bioinformatics* **31**, 1211 (apr 2015).
29. W. Fitch, Towards defining the course of evolution: minimum change for a specified tree topology, *Syst. Zool.* **20**, 406 (1971).
30. H. Shimodaira and M. Hasegawa, Multiple comparisons of log-likelihoods with applications to phylogenetic inference, *Molecular Biology and Evolution* **16**, p. 1114 (1999).
31. N. Moshiri, J. O. Wertheim, M. Ragonnet-Cronin and S. Mirarab, FAVITES: simultaneous simulation of transmission networks, phylogenetic trees and sequences, *Bioinformatics* **35** (11 2018).
32. R. Albert and A.-L. Barabási, Statistical mechanics of complex networks, *Rev. Mod. Phys.* **74**, 47 (Jan 2002).
33. W. O. Kermack, A. G. McKendrick and G. T. Walker, A contribution to the mathematical theory of epidemics, *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character* **115**, 700 (1927).
34. D. F. Robinson and L. R. Foulds, Comparison of phylogenetic trees, *Math. Biosci.* **53**, 131 (feb 1981).
35. J. S. Farris, Estimating phylogenetic trees from distance matrices, *The American Naturalist* **106**, 645 (1972).
36. D. Swofford, G. Olsen, P. Waddell and D. Hillis, Phylogenetic inference, in *Molecular systematics*, eds. D. Hillis, C. Moritz and e. B. Mabl (Sinauer Associates, 1996) pp. 407–514.
37. D. Sankoff, Minimal mutation trees of sequences, *SIAM Journal on Applied Mathematics* **28**, 35 (1975).

Supplementary Section S1 - Algorithmic Details of TreeFix-TP

The input for TreeFix-TP is a multiple sequence alignment A, maximum likelihood phylogeny T_{ML}, and a mapping $M : L(T_{ML}) \rightarrow H$, where H is the set of infected hosts and $L(T_{ML})$ denotes the set of leaves of T_{ML}. The output is a transmission phylogeny T^*. We perform a local search in a manner similar to TreeFix and TreeFix-DTL.[27,28] Using $T^{'} \leftarrow T_{ML}$ as a starting point, we set $C^{'}$ to be the transmission cost of $T^{'}$, and a candidate tree T is proposed by performing random NNI and SPR operations on $T^{'}$. Using the SH test, we find a p-value for the null hypothesis that T and $T^{'}$ are equivalent.

Then, we calculate the transmission cost $C(T)$ of T. $C(T)$ is calculated using Fitch's algorithm for the small parsimony problem, where the states are the possible hosts. Set intersection and unions are done efficiently by implementing each nodes set of potential hosts as a bit set. This allows us to take advantage of bit-level parallelism and use bitwise AND and OR word operations to efficiently calculate the parsimony cost of a given phylogeny. T is accepted if T is statistically equivalent to $T^{'}$ $(p < \alpha)$, and has a lower transmission cost than $T^{'}$ $(C(T) < C^{'})$. Otherwise, T is accepted with some predefined probability. If T is accepted, we set $T^{'} \leftarrow T, C^{'} \leftarrow C(T)$, and perform another iteration of local search. By default, TreeFix-TP runs for 5000 iterations, which is often sufficient for trees with a few hundred leaves. For larger trees, more iterations may be necessary to fully explore the local space. At the end of the local search, TreeFix-TP outputs the tree T^* which is statistically equivalent to T_{ML} and has the lowest transmission cost found during the search. If multiple trees were found with the same minimum transmission cost, TreeFix-TP outputs the tree with the highest likelihood.

The search can be modified in a number of ways. The model under which likelihood is calculated, likelihood test used, and significance level α can all be changed. By default, we use the $GTR + \Gamma$ model, and the SH-test implemented in RAxML[26] with a significance level of 0.05. A custom transmission cost module can also be specified. One possible application of this is to use epidemiological data to weight certain transmissions higher, which would cause TreeFix-TP to prefer phylogenies where that transmission does not occur. Finally, the user can specify the number of iterations to be performed.

Supplementary Table S2 This table shows the set of parameters used for generating the 28 different types of data sets used in our simulation study. For each of these 28 data set types, 20 unique data sets were simulated. These parameters were chosen to mimic realistic viral evolution and to evaluate TreeFix-TP under a wide range of conditions.

Transmission Model	Transmission Parameters	Sequence Length	Viruses per Host	Mutation Rate
SEIR	$\beta = 0.015,\ \lambda = 0.01,\ \delta = 0.45$	1000 500 2000	10	0.25
		1000	5 20	
			10	0.125 0.5
	$\beta = 0.01,\ \lambda = 0.06,\ \delta = 0.93$	1000 500 2000	10	0.25
		1000	5 20	
			10	0.125 0.5
SIR	$\beta = 0.003,\ \delta = 0.95$	1000 500 2000	10	0.5
		1000	5 20	
			10	0.25 1
	$\beta = 0.01,\ \delta = 0.09$	1000 500 2000	10	1.5
		1000	5 20	
			10	0.75 3