TNet: Phylogeny-Based Inference of Disease Transmission Networks Using Within-Host Strain Diversity

Saurav Dhar[1], Chengchen Zhang[1,3], Ion Mandoiu[1,2], and Mukul S. Bansal[1,2]

[1] Department of Computer Science and Engineering, University of Connecticut, Storrs, USA
[2] Institute for Systems Genomics, University of Connecticut, Storrs, USA
[3] Computer Science and Engineering, University of California, San Diego, USA
ion.mandoiu@uconn.edu, mukul.bansal@uconn.edu

Abstract. The inference of disease transmission networks from genetic sequence data is an important problem in epidemiology. One popular approach for building transmission networks is to reconstruct a phylogenetic tree using sequences from disease strains sampled from (a subset of) infected hosts and infer transmissions based on this tree. However, most existing phylogenetic approaches for transmission network inference cannot take within-host strain diversity into account, which affects their accuracy, and, moreover, are highly computationally intensive and unscalable.

In this work, we introduce a new phylogenetic approach, TNet, for inferring transmission networks that addresses these limitations. TNet uses multiple strain sequences from each sampled host to infer transmissions and is simpler and more accurate than existing approaches. Furthermore, TNet is highly scalable and able to distinguish between ambiguous and unambiguous transmission inferences. We evaluated TNet on a large collection of 560 simulated transmission networks of various sizes and diverse host, sequence, and transmission characteristics, as well as on 10 real transmission datasets with known transmission histories. Our results show that TNet outperforms two other recently developed methods, phyloscanner and SharpTNI, that also consider within-host strain diversity using a similar computational framework. TNet is freely available open-source from https://compbio.engr.uconn.edu/software/TNet/.

1 Introduction

The accurate inference of disease transmission networks is fundamental to understanding and containing the spread of infectious diseases [2, 10, 16]. A key challenge with inferring transmission networks, particularly those of rapidly evolving RNA and retroviruses [7], is that they exist in the host as "clouds" of closely related sequences. These variants are referred to as *quasispecies* [6,22], and the resulting genetic diversity of the strains circulating within a host has important implications for efficiency of transmission, disease progression, drug/vaccine resistance, etc. The availability of quasispecies, or sequences from multiple strains per

infected host, also has direct relevance for inferring transmission networks and has the potential to make such inference easier and far more accurate [18,20,23]. Yet, while the advent of next-generation sequencing technologies has revolutionized the study of quasispecies, most existing transmission network inference methods cannot use multiple distinct strain sequences per host.

Existing methods for inferring transmission networks can be classified into two categories: Those based on constructing and analyzing sequence similarity or relatedness graphs, and those based on constructing and analyzing phylogenetic trees for the infecting strains. Many methods based on sequence similarity or relatedness graph analysis exist and several recently developed methods in this category are also able to take into account multiple distinct strain sequences per host [9,14,19]. However, similarity/relatedness based methods can suffer from a lack of resolution and are often unable to infer transmission directions or complete transmission histories. Phylogeny-based methods [5,11,13,16,23] attempt to overcome these limitations by constructing and analyzing phylogenies of the infecting strains. We refer to these strain phylogenies as *transmission phylogenies*. These phylogeny-based methods infer transmission networks by computing 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 labelled by 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 the transmission network, where transmission is inferred along any edge connecting two nodes labeled with different hosts.

Several sophisticated phylogeny-based methods have been developed over the last few years. These include BEASTlier [11], SCOTTI [4], phybreak [13], TransPhylo [5], and phyloscanner [23], BadTrIP [3]. Among these, only SCOTTI [4], BadTrIP [3], and phyloscanner [23] can explicitly consider multiple strain sequences per host. BEASTlier also allows for the presence of multiple sequences per host, but requires that all sequences from the same host be clustered together on the phylogeny, a precondition that is often violated in practice. Among the methods that explicitly consider multiple strain sequences per host, SCOTTI, BadTrIP, and BEASTlier are model-based and highly computationally intensive, relying on the use of Markov Chain Monte Carlo (MCMC) algorithms for inference. These methods also require several difficult-to-estimate epidemiological parameters, such as infection times, and make several strong assumptions about pathogen evolution and the underlying transmission network. Thus, phyloscanner [23] is the only previous method that takes advantage of multiple sequences per host and that is also computationally efficient, easy to use, and scalable to large datasets.

In this work, we introduce a new phylogenetic approach, TNet, for inferring transmission networks. TNet uses multiple strain sequences from each sampled host to infer transmissions and is simpler and more accurate than existing approaches. TNet uses an extended version of the classical Sankoff algorithm [17] from the phylogenetics literature for ancestral host assignment, where the ex-

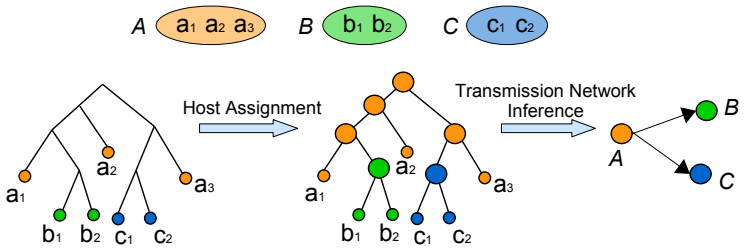

Fig. 1. Phylogeny-based transmission network inference. The figure shows a simple example with three infected individuals A, B, and C, represented here by the three different colors, where A has three viral variants while B and C have two each. The tree on the left depicts the transmission phylogeny for the seven sampled strains, with each of these strains colored by the host from which it was sampled. The tree in the middle shows a hypothetical assignment of hosts to ancestral nodes of the transmission phylogeny. This ancestral host assignment can then be used to infer the transmission network shown on the right.

tension makes it possible to efficiently compute support values for individual transmission edges based on a sampling of optimal host assignments where the number of back-transmissions (or reinfections by descendant disease strains) is minimized. TNet is parameter-free and highly scalable and can be easily applied within seconds to datasets with hundreds of strain sequences and hosts. In recent independent work, Sashittal et al. [18] developed a new method called SharpTNI that is based on similar ideas to TNet. SharpTNI is based on an NP-hard problem formulation that seeks to find parsimonious ancestral host assignments minimizing the number of co-transmissions [18].The authors provide an efficient heuristic for this problem that is based on uniform sampling of parsimonious ancestral host assignments (not necessarily minimizing co-transmissions) and subsequently filtering them to only keep those assignments among the samples that minimize co-transmissions [18]. Thus, both TNet and SharpTNI are based on the idea of parsimonious ancestral host assignments and on aggregating across the diversity of possible solutions obtained through some kind of sampling of optimal solutions. The primary distinction between the two methods is the strategy employed for sampling of the optimal solutions, with SharpTNI minimizing co-transmissions and TNet minimizing back-transmissions.

We evaluated TNet, SharpTNI, and phyloscanner on a large collection of 560 simulated transmission networks of various sizes and representing a wide range of host, sequence, and transmission characteristics, as well as on 10 real transmission datasets with known transmission histories. We found that both TNet and SharpTNI significantly outperformed phyloscanner under all tested conditions and all datasets, yielding more accurate transmission networks for both simulated and real datasets. Between TNet and SharpTNI, we found that both methods performed similarly on the real datasets but that TNet showed

better accuracy on the simulated datasets. TNet is freely available open-source from: `https://compbio.engr.uconn.edu/software/TNet/`

2 Basic Definitions and Preliminaries

Given a rooted tree T, we denote its node set, edge set, and leaf set by $V(T)$, $E(T)$, and $Le(T)$ respectively. The root node of T is denoted by $rt(T)$, the parent of a node $v \in V(T)$ by $pa_T(v)$, its set of children by $Ch_T(v)$, and the (maximal) subtree of T rooted at v by $T(v)$. The set of *internal nodes* of T, denoted $I(T)$, is defined to be $V(T) \setminus Le(T)$. A rooted tree is *binary* if all of its internal nodes have exactly two children. In this work, the term *tree* refers to a rooted binary tree.

2.1 Problem formulation

Let T denote the transmission phylogeny constructed from the genetic sequences of the infecting strains (i.e., pathogens) sampled from the infected hosts under consideration. Note that such trees can be easily constructed using standard phylogenetic methods such as RAxML [21]. These trees can also be rooted relatively accurately using either standard phylogenetic rooting techniques or by using a related sequence from a previous outbreak of the same disease as an outgroup. Let $H = \{h_1, h_2, \ldots, h_n\}$ denote the set of n hosts under consideration. We assume that each leaf of T is labeled with the host from H from which the corresponding strain sequence was obtained. Figure 1 shows an example of such a tree and its leaf labeling, where the labeling is depicted using the different colors.

Observe that each internal node of T represents an ancestral strain sequence that existed in some infected host. Moreover, each internal node (or bifurcation) represents either intra-host diversification and evolution of that ancestral strain or a transmission event where that ancestral strain is transmitted from one host to another along one of the child edges. Thus, each node of T is associated with an infected host. Given $t \in V(T)$, we denote the host associated with node t by $h(t)$. Note that internal nodes may represent strains from hosts that do not appear in H, i.e., strains from unsampled hosts, and so there may be $t \in I(T)$ for which $h(t) \notin H$. Given an ancestral host assignment for T, i.e., given $h(t)$ for each $t \in I(T)$, the implied transmission network can be easily inferred as follows: A transmission edge is inferred from host x to host y if there is an edge $(pa(t), t) \in E(T)$, where $h(pa(t)) = x$ and $h(t) = y$. Note that each transmission edge in the reconstructed transmission network may represent either direct transmission or indirect transmission through one or more unsampled hosts. Thus, to reconstruct transmission networks it suffices to compute $h(t)$ for each $t \in I(T)$.

TNet (along with SharpTNI) is based on finding ancestral host assignments that minimize the number of inter-host transmission events on T. The utility of such parsimonious ancestral host assignment for transmission network inference when multiple strain sequences per host are available was first systematically demonstrated by Romero-Severson et al. [16] and later developed further

by Wymant et al. [23] in their phyloscanner method. The basic computational problem under this formulation can be stated as follows:

Problem 1 (Optimal ancestral host assignment) *Given a transmission phylogeny T on strain sequences sampled from a set $H = \{h_1, h_2, \ldots, h_n\}$ of n infected hosts, compute $h(t)$ for each $t \in I(T)$ such that the number of edges $(t', t'') \in E$ for which $h(t') \neq h(t'')$ is minimized.*

Problem 1 is equivalent to the well-known small parsimony problem in phylogenetics and can be solved efficiently using the classical Fitch [8] and Sankoff [17] algorithms. In TNet, we solve a modified version of the problem above that considers all possible optimal ancestral host assignments and samples greedily among them to minimize the number of back-transmissions (or reinfection by a descendant disease strain). To accomplish this goal efficiently, TNet uses an extended version of Sankoff's algorithm.

3 Algorithmic Details

A primary methodological and algorithmic innovation responsible for the improved accuracy of TNet (and also of SharpTNI) is the explicit and principled consideration of variability in optimal ancestral host assignments. More precisely, TNet recognizes that there are often a very large number of distinct optimal ancestral host assignments and it samples the space of all optimal ancestral host assignments in a manner that preferentially preserves optimal ancestral host assignments (described in detail below). TNet then aggregates across these samples to compute a support value for each edge in the final transmission network. This approach is illustrated in Figure 2. Thus, the core computational problem solved by TNet can be formulated as follows:

Definition 1 (Back-Transmission). *Given a transmission network N on n infected hosts $H = \{h_1, h_2, \ldots, h_n\}$, we say that there exists a back-transmission for host h_i if there exists a directed cycle containing h_i in N. The total number of back-transmissions implied by N equals the number of hosts with back-transmissions.*

Problem 2 (Minimum back-transmission sampling) *Given a transmission phylogeny T on strain sequences sampled from a set $H = \{h_1, h_2, \ldots, h_n\}$ of n infected hosts, let $\mathcal{O}$ denote the set containing all distinct ancestral host assignments for T. Further, let $\mathcal{O}'$ denote the subset of $\mathcal{O}$ that implies the fewest back-transmissions in the resulting transmission network. Compute an optimal ancestral host assignment from $\mathcal{O}'$ such that each element of $\mathcal{O}'$ has an equal probability of being computed.*

Observe that the actual number of optimal ancestral host assignments (both $\mathcal{O}$ and $\mathcal{O}'$) can grow exponentially in the number of hosts n. By addressing the

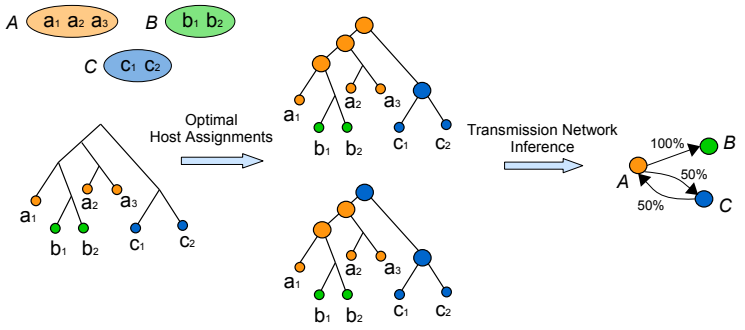

Fig. 2. Accounting for multiple optima in transmission network inference. The tree on the left depicts the transmission phylogeny for the seven strains sampled from three infected individuals A, B, and C, represented here by the three different colors. This tree admits two distinct optimal ancestral host assignments as shown in the figure. These two optimal ancestral host assignments can then be together used to infer a transmission network, as shown on the right, in which each edge has a support value. The support value of a transmission edge is define to be the percentage of optimal ancestral host assignments that imply that transmission edge.

sampling problem above instead, TNet seeks to efficiently account for the diversity within optimal ancestral host assignments with minimum back-transmissions, without explicitly having to enumerate them all.

Note that SharpTNI performs a similar sampling among all optimal ancestral host assignments, but employs a different optimality objective. Specifically, SharpTNI seeks to sample optimal ancestral host assignments that minimize the number of *co-transmissions*, i.e., minimize the number of inter-host edges in the transmission network.

3.1 Minimum back-transmission sampling of optimal host assignments

TNet approximates minimum back-transmission sampling by combining uniform sampling of ancestral host assignments with a greedy procedure to assign specific hosts to internal nodes. This is accomplished by suitably extending and modifying Sankoff's algorithm. This extended Sankoff algorithm computes, for each $t \in V(T)$ and $h_i \in H$, the number of distinct optimal host assignments for the subtree $T(t)$ under the constraint that $h(t) = h_i$, denoted by $N(t, h_i)$. After all $N(\cdot, \cdot)$ numbers have been computed, we perform our greedy sampling procedure using probabilistic backtracking. The basic idea is to perform a pre-order traversal of T and make final host assignment at the current node based on the number of optimal ancestral host assignments available for each optimal choice at that node, while preferentially preserving the parent host assignment. This is described in detail in Procedure *GreedyProbabilisticBacktracking* below.

Procedure *GreedyProbabilisticBacktracking*

Let $\alpha = \min_i\{C(rt(T), h_i)\}$.
for each $t \in I(T)$ in a pre-order traversal of T **do**
  if $t = rt(T)$ **then**
    Let $X = \{h_i \in H \mid C(rt(T), h_i) = \alpha\}$.
    For each $h_i \in X$, assign $h(t) = h_i$ with probability $\frac{N(t,h_i)}{\sum_{h_j \in X} N(t,h_j)}$.
  if $t \neq rt(T)$ **then**
    Let $X = \{h_i \in H \mid C(t, h_i) + p(h(pa(t)), h_i)$ is minimized$\}$.
    if $h(pa(t)) \in X$ **then**
      Assign $h(t) = h(pa(t))$.
    if $h(pa(t)) \notin X$ **then**
      For each $h_i \in X$, assign $h(t) = h_i$ with probability $\frac{N(t,h_i)}{\sum_{h_j \in X} N(t,h_j)}$.

The procedure above preferentially assigns each internal node the same host assignment as that node's parent, if such an assignment is optimal. This strategy is based on the following straightforward observation: If the host assignment of an internal node t *could* be the same as that of its parent (while remaining optimal), i.e., $h(t) = h(pa(t))$ is optimal, then assigning a different optimal mapping $h(t) \neq h(pa(t))$ can result in a transmission edge back to $h(pa(t))$, effectively implying a reinfection of host $h(pa(t))$ by a descendant disease strain. Thus, the goal of TNet's sampling strategy is to strike a balance between sampling the diversity of optimal ancestral host assignments but avoiding sampling solutions with unnecessary back-transmissions.

3.2 Additional methodological details

Aggregation across multiple optimal ancestral host assignments. As illustrated in Figure 2, aggregating across the sampled optimal ancestral host assignments can be used to improve transmission network inference by distinguishing between high-support and low-support transmission edges. Specifically, each directed edge in the transmission network can be assigned a support value based on the percentage of sampled optimal ancestral host assignments that imply that transmission edge. By executing TNet multiple times on the same transmission phylogeny (100 times per tree in our experimental study), these support values for edges can be estimated very accurately.

Accounting for phylogenetic inference error. In addition to capturing the uncertainty of minimum back-transmission ancestral host assignments, which we show how to handle above, a second key source of inference uncertainty is phylogenetic error, i.e., errors in the inferred transmission phylogeny. Phyloscanner [23] accounts for such phylogenetic error by aggregating results across multiple transmission phylogenies (e.g., derived from different genomic regions of the samples strains, bootstrap replicates, etc.). We employ the same approach with TNet, aggregating the transmission network across multiple transmission phylogenies, in addition to the aggregation across multiple optimal ancestral host assignments per transmission phylogeny.

4 Datasets and Evaluation Methodology

Simulated datasets. To evaluate the performance of TNet, SharpTNI, and phyloscanner, we generated a number of simulated viral transmission data sets across a variety of parameters. These datasets were generated using FAVITES [15], which can simultaneous simulate transmission networks, phylogenetic trees, and sequences. The simulated contact networks consisted of 1000 individuals, with each individual connected to other individuals through 100 outgoing edges preferentially attached to high-degree nodes using the Barabasi-Albert model [1]. On these contact networks, we simulated datasets with (i) four types of transmission networks using both Susceptible-Exposed-Infected-Recovered (SEIR) and Susceptible-Infected-Recovered (SIR) [12] models with two different infection rates for each, (ii) number of viruses sampled per host (5, 10, and 20), (iii) three different nucleotide sequence lengths (1000nt, 500nt, and 250nt), and (iv) three different rates of with-in host sequence evolution (normal, half, and double). This resulted in 560 different transmission network datasets representing 28 different parameter combinations. Further details on the construction and specific parameters used for these simulated datasets appear in [20]. These 560 simulated datasets had between 35 and 1400 sequences (i.e., leaves in the corresponding transmission phylogeny), with an average of 287.44 leaves. The maximum number of hosts per tree was 75, with an average of 26.72.

Data from real HCV outbreaks. We also evaluated the accuracies of TNet, SharpTNI, and phyloscanner on real datasets of HCV outbreaks made available by the CDC [19]. This collection consists 10 different datasets, 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. The approximate transmission network is known for each of these datasets through CDC's monitoring and epidemiological efforts. In each of the 10 cases, this estimated transmission network consists of a single known host infecting all the other hosts in that network.

Evaluating transmission network inference accuracy. For all simulated and real datasets, we constructed transmission phylogenies using RAxML and used RAxML's own balanced rooting procedure to root them [21]. Note that TNet, SharpTNI, and phyloscanner all require rooted transmission phylogenies. To account for phylogenetic uncertainty and error, we computed 100 bootstrap replicates for each simulated and real dataset. For SharpTNI we used the efficient heuristic implementation for evaluation (not the exponential-time exact solution). All TNet results were based on aggregating across 100 sampled optimal host assignments per transmission phylogeny, and all SharpTNI results were aggregated across that subset of 100 samples that had minimum co-transmission cost, per transmission phylogeny. Results for all methods were aggregated across the different bootstrap replicates to account for phylogenetic uncertainty and yield edge-weighted transmission networks. To convert such edge-weighted transmission networks into unweighted transmission networks, we used the same 0.5 (or 50%) tree-support threshold used by phyloscanner in [23]. Thus, all directed

edges with an edge-weight of at least 0.5 (or 50%) tree-support were retained in the final inferred transmission network and other edges were deleted. For a fair evaluation, none of the methods were provided with any epidemiological information such as sampling times or infection times. Finally, since both TNet and SharpTNI build upon uniform sampling procedures for optimal ancestral host assignments (minimizing the total number of inter-host transmissions), we also report results for uniform random sampling of optimal ancestral host assignments as a baseline.

To evaluate the accuracies of these final inferred transmission networks, we computed *precision* (i.e., the fraction of inferred edges in the transmission network that are also in the true network), *recall* (i.e., the fraction of true transmission network edges that are also in the inferred network), and *F1 scores* (i.e., harmonic mean of precision and recall).

5 Experimental Results

5.1 Simulated data results

Accuracy of single samples. We first considered the impact of inferring the transmission network using only a single optimal solution, i.e., without any aggregation across samples or bootstrap replicates. Figure 3 shows the results of this analysis. As the figure shows, TNet has by far the best overall accuracy, with precision, recall, and F1 scores of 0.72, 0.75, and 0.73, respectively. Phyloscanner showed the greatest precision at 0.828 but had significantly lower recall and F1 at 0.522 and 0.626, respectively. SharpTNI performed slightly better than a random optimal solution (uniform sampling), with precision, recall, and F1 scores of 0.68, 0.71, and 0.694, respectively, compared to 0.67, 0.71, and 0.687, respectively, for a randomly sampled optimal solution.

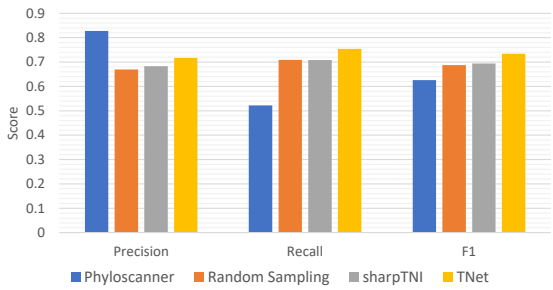

Fig. 3. Accuracy of methods using single samples. This figure plots precision, recall, and F1 scores for the different methods without any aggregation of results across multiple samples or bootstrap replicates. Results are averaged across the 560 simulated datasets.

Impact of sampling multiple optimal solutions on TNet and SharpTNI. For improved accuracy, both TNet and SharpTNI rely on aggregation across multiple samples per transmission phylogeny. Note that, when aggregating across multiple optimal ancestral host assignments, the final transmission network is obtained by applying a cutoff for the edge support values. For example, in Figure 2, at a cutoff threshold of 100%, only a single transmission from $(A \rightarrow B)$ would be inferred, while with a cutoff threshold of 50%, all three transmission edges shown in the figure would be inferred. We studied the impact of multiple sample aggregation by considering two natural sampling cutoff thresholds: 50% and 100%. As Figure 4 shows, results improve as multiple optimal are considered. Specifically, for the 50% sampling cutoff threshold, we found that the overall accuracy of all methods improves as multiple samples are considered. For TNet, precision, recall, and F1 score all increase to 0.73, 0.75, and 0.74, respectively. For SharpTNI, precision and F1 score increase significantly to 0.76 and 0.72, respectively, while recall decreases slightly to 0.706. Surprisingly, we found that uniform random sampling outperformed SharpTNI, with precision, recall, and F1 score of 0.77, 0.70, and 0.73, respectively.

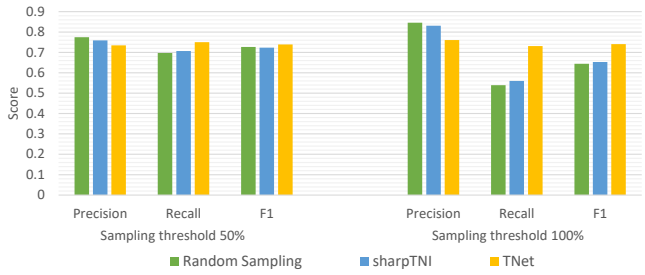

Fig. 4. Accuracy of methods using multiple samples on a single transmission phylogeny. This figure plots average precision, recall, and F1 scores for random sampling, sharpTNI, and TNet when 100 samples are used on a single transmission phylogeny. Values reported are averaged across all 560 simulated datasets, and results are shown for both 50% and 100% sampling cutoff thresholds.

We also see a clear tradeoff between precision and recall as the sampling cutoff threshold is increased. Specifically, for the 100% sampling cutoff threshold, the precision of all methods increases significantly, but overall F1 score falls to 0.65 and 0.64 for SharpTNI and random sampling, respectively. Surprisingly, recall only decreases slightly for TNet, and its overall F1 score remains 0.74 even for the 100% sampling cutoff threshold.

Accuracy on multiple bootstrapped transmission phylogenies. To further improve inference accuracy, results can be aggregated across the different bootstrap replicates to account for phylogenetic uncertainty. We therefore ran phyloscanner, TNet, and SharpTNI with 100 transmission phylogeny estimates (bootstrap replicates) per dataset. (We tested for the impact of using varying

numbers of bootstrap replicates, trying 25, 50, and 100, but found that results were roughly identical in each case. We therefore report results for only the 100 bootstrap analyses.) As figure 5 shows, for the 50% sampling cutoff threshold, the accuracies of all methods improve over the corresponding single-tree results, with particularly notable improvements in precision. For the 100% sampling cutoff threshold, the precision of all methods improves further, but for phyloscanner and SharpTNI this comes at the expense of large reductions in recall. TNet continues to be best performing method overall for both samplng cutoff thresholds, with precision, recall, and F1 score of 0.79, 0.73, and 0.76, respectively, at the 50% sampling cutoff threshold, and 0.82, 0.71, and 0.754, respectively at the 100% sampling cutoff threshold.

Precision-recall characteristics of SharpTNI and TNet. The results above shed light on the differences between the sampling strategies (i.e, objective functions) used by SharpTNI and TNet, revealing that SharpTNI tends to have higher precision but much lower recall. Thus, depending on use case, either SharpTNI or TNet may be the method of choice. We also note that random sampling shows similar accuracy and precision-recall characteristics as SharpTNI, suggesting that SharpTNI may not offer much improvement over the much simpler random sampling strategy.

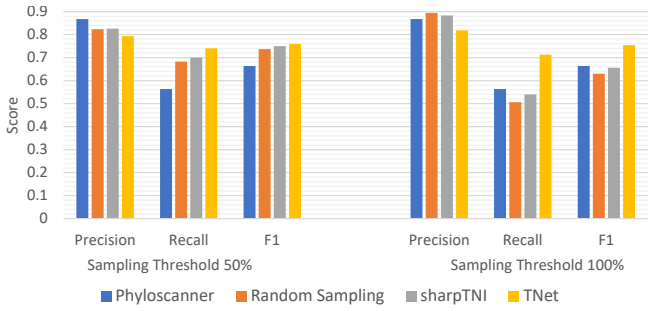

Fig. 5. Transmission network inference accuracy when multiple transmission phylogenies are used. This figure plots average precision, recall, and F1 scores for phyloscanner, random sampling, sharpTNI, and TNet when 100 bootstrap replicate transmission phylogenies are used for transmission network inference. Values reported are averaged across all 560 simulated datasets, and results are shown for both 50% and 100% sampling cutoff thresholds.

5.2 Real data results

We applied TNet, SharpTNI, and phyloscanner to the 10 real HCV datasets using 100 bootstrap replicates per dataset. We found that both TNet and SharpTNI performed almost identically on these datasets, and that both dramatically outperformed phyloscanner on the real datasets in terms of both precision and recall

(and, consequently, F1 scores). Figure 6 shows these results averaged across the 10 real datasets. As the figure shows, both TNet and SharpTNI have identical F1 scores for the 50% and 100% sampling cutoff thresholds, with both methods showing F1 scores of 0.57 and 0.56, respectively. In contrast, phyloscanner shows much lower precision and recall, with an F1 score of only 0.22. Random sampling had slightly worse performance than TNet and SharpTNI at both the 50% and 100% sampling cutoff thresholds. At the 100% sampling cutoff threshold, we observe the same precision-recall characteristics seen in the simulated datasets, with SharpTNI showing higher precision but lower recall.

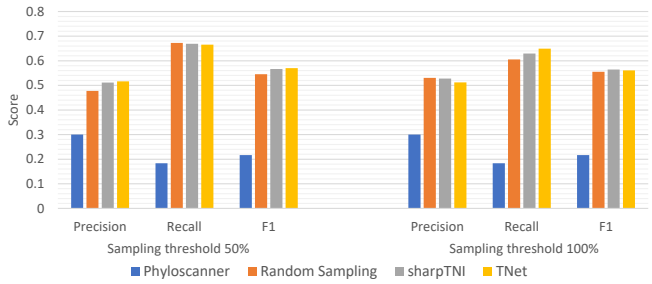

Fig. 6. Transmission network inference accuracy across the 10 real HCV datasets. This figure plots average precision, recall, and F1 scores for phyloscanner, random sampling, sharpTNI, and TNet on the 10 real HCV datasets with known transmission histories. Results are shown for both 50% and 100% sampling cutoff thresholds.

6 Discussion

In this paper, we introduced TNet, a new method for transmission network inference when multiple strain sequences are sampled from the infected hosts. TNet has two distinguishing features: First, it systematically accounts for variability among different optimal solutions to efficiently compute support values for individual transmission edges and improve transmission inference accuracy, and second, its objective function seeks to find those optimal host assignments that minimize the number of back-transmissions. TNet is based on a relatively simple parsimony-based formulation and is parameter-free and highly scalable. It can be easily applied within seconds to datasets with many hundreds of strain sequences and hosts. As our experimental results on both simulated and real datasets show, TNet is highly accurate and significantly outperforms phyloscanner. We find that TNet also outperforms SharpTNI, a distinct but very similar method developed independently and published recently.

Going forward, several aspects of TNet can be tested and improved further. The simulated datasets used in our experimental study assume that all infected hosts have been sampled. It would be useful to test how accuracy decreases

as fewer and fewer infected hosts are sampled. Phyloscanner employs a simple technique to estimate if an ancestral host assignment may be to an unsampled host, and a similar technique could be used in TNet. Currently, TNet does not use branch lengths or overall strain diversity within hosts, and these could be used to further improve the accuracy of ancestral host assignment and transmission network inference. Finally, our results suggest that, despite their conceptual similarities, SharpTNI and TNet, show different precision-recall characteristics. It may be possible to meaningfully combine the objective functions of SharpTNI and TNet to create a more accurate hybrid method.

Acknowledgements The authors wish to thank Dr. Pavel Skums (Georgia State University) and the Centers for Disease Control for sharing their HCV outbreak data. We also thank Samuel Sledzieski for creating and sharing the simulated transmission network datasets used in this work.
Funding This work was supported in part by NSF award CCF 1618347 to IM and MSB.

References

1. R. Albert and A.-L. Barabási. Statistical mechanics of complex networks. *Rev. Mod. Phys.*, 74:47–97, Jan 2002.
2. 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(2):213–221, 05 2018.
3. N. De Maio, C. J. Worby, D. J. Wilson, and N. Stoesser. Bayesian reconstruction of transmission within outbreaks using genomic variants. *PLOS Comp. Biol.*, 14(4):1–23, 04 2018.
4. N. De Maio, C.-H. Wu, and D. J. Wilson. Scotti: Efficient reconstruction of transmission within outbreaks with the structured coalescent. *PLOS Comp. Biol.*, 12(9):1–23, 09 2016.
5. 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(4):997–1007, jan 2017.
6. E. Domingo and J. Holland. RNA virus mutations and fitness for survival. *Annu Rev Microbiol*, 51:151–178, 1997.
7. 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(24):13910–13913, 1999.
8. W. Fitch. Towards defining the course of evolution: minimum change for a specified tree topology. *Syst. Zool.*, 20:406–416, 1971.
9. O. Glebova, S. Knyazev, A. Melnyk, A. Artyomenko, Y. Khudyakov, A. Zelikovsky, and P. Skums. Inference of genetic relatedness between viral quasispecies from sequencing data. *BMC Genomics*, 18(Suppl 10):918, 2017.
10. 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(suppl_1), 04 2018.

11. M. Hall, M. Woolhouse, and A. Rambaut. Epidemic Reconstruction in a Phylogenetics Framework: Transmission Trees as Partitions of the Node Set. *PLoS Comp. Biol.*, 11(12):e1004613, dec 2015.
12. 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(772):700–721, 1927.
13. D. Klinkenberg, J. A. Backer, X. Didelot, C. Colijn, and J. Wallinga. Simultaneous inference of phylogenetic and transmission trees in infectious disease outbreaks. *PLoS Comp. Biol.*, 13:1–32, 2017.
14. 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. *Mol. Biol. Evol.*, 35(7):1812–1819, 01 2018.
15. N. Moshiri, J. O. Wertheim, M. Ragonnet-Cronin, and S. Mirarab. FAVITES: simultaneous simulation of transmission networks, phylogenetic trees and sequences. *Bioinformatics*, 11 2018.
16. E. O. Romero-Severson, I. Bulla, and T. Leitner. Phylogenetically resolving epidemiologic linkage. *Proceedings of the National Academy of Sciences*, 113(10):2690–2695, mar 2016.
17. D. Sankoff. Minimal mutation trees of sequences. *SIAM Journal on Applied Mathematics*, 28(1):35–42, 1975.
18. P. Sashittal and M. El-Kebir. SharpTNI: Counting and sampling parsimonious transmission networks under a weak bottleneck. *bioRxiv*, 2019.
19. 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(1):163–170, jun 2018.
20. S. Sledzieski, C. Zhang, I. Mandoiu, and M. S. Bansal. TreeFix-TP: Phylogenetic Error-Correction for Infectious Disease Transmission Network Inference. *bioRxiv*, 2019.
21. A. Stamatakis. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30(9):1312–1313, may 2014.
22. D. Steinhauer and J. Holland. Rapid evolution of rna viruses. *Annual Review of Microbiology, 41*, pages 409–433, 1987.
23. 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(3):719–733, mar 2017.