Bioinformatics, 34(1), 2018, 163–170
doi: 10.1093/bioinformatics/btx402
Advance Access Publication Date: 19 June 2017
Original Paper

OXFORD

Sequence analysis

QUENTIN: reconstruction of disease transmissions from viral quasispecies genomic data

Pavel Skums[1,2,*], Alex Zelikovsky[1,*], Rahul Singh[3,*], Walker Gussler[2], Zoya Dimitrova[2], Sergey Knyazev[1], Igor Mandric[1], Sumathi Ramachandran[2], David Campo[2], Deeptanshu Jha[3], Leonid Bunimovich[4], Elizabeth Costenbader[5], Connie Sexton[2,6], Siobhan O'Connor[2,7], Guo-Liang Xia[2] and Yury Khudyakov[2,*]

[1]Department of Computer Science, Georgia State University, [2]Centers for Disease Control and Prevention, Division of Viral Hepatitis, Atlanta, GA 30303, USA, [3]Department of Computer Science, San Francisco State University, San Francisco, CA 94132, USA, [4]School of Mathematics, Georgia Institute of Technology, Atlanta, GA 30313, USA, [5]FHI 360, Durham, NC 27701, USA, [6]Division of Global HIV and TB, Centers for Disease Control and Prevention, Atlanta, GA 30333, USA and [7]Division of HIV/AIDS Prevention, Centers for Disease Control and Prevention, Atlanta, GA 30333, USA

*To whom correspondence should be addressed.

Associate Editor: Cenk Sahinalp

Received and revised on April 7, 2017; editorial decision on June 14, 2017; accepted on June 15, 2017

Abstract

Motivation: Genomic analysis has become one of the major tools for disease outbreak investigations. However, existing computational frameworks for inference of transmission history from viral genomic data often do not consider intra-host diversity of pathogens and heavily rely on additional epidemiological data, such as sampling times and exposure intervals. This impedes genomic analysis of outbreaks of highly mutable viruses associated with chronic infections, such as human immunodeficiency virus and hepatitis C virus, whose transmissions are often carried out through minor intra-host variants, while the additional epidemiological information often is either unavailable or has a limited use.

Results: The proposed framework QUasispecies Evolution, Network-based Transmission INference (QUENTIN) addresses the above challenges by evolutionary analysis of intra-host viral populations sampled by deep sequencing and Bayesian inference using general properties of social networks relevant to infection dissemination. This method allows inference of transmission direction even without the supporting case-specific epidemiological information, identify transmission clusters and reconstruct transmission history. QUENTIN was validated on experimental and simulated data, and applied to investigate HCV transmission within a community of hosts with high-risk behavior. It is available at https://github.com/skumsp/QUENTIN.

Contact: pskums@gsu.edu or alexz@cs.gsu.edu or rahul@sfsu.edu or yek0@cdc.gov

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Understanding the transmission dynamics of pathogens is critically important for development of effective public health interventions and disease eradication strategies. In the last decade, epidemiology has been greatly transformed by sequencing technologies. Genomic analysis has become a major tool for microbial transmission detection and outbreak investigations (Collier *et al.*, 2014; Grabowski and Redd, 2014; Holodniy *et al.*, 2012). From the computational point of view, disease outbreak investigation requires solving the following two problems:

1. Detection of transmission clusters consisting of hosts involved in outbreaks;
2. Inference of transmission history of each outbreak.

This article focuses on RNA viruses such as human immunodeficiency virus (HIV) and hepatitis C virus (HCV). The hallmark of these viruses is their extreme genetic heterogeneity that allows them to exist in infected hosts as populations of genetically related variants or *quasispecies* (Domingo *et al.*, 2012). Extreme diversity of intra-host viral populations plays crucial role in disease progression and epidemic spread (Beerenwinkel *et al.*, 2005). Moreover, diseases caused by HIV and HCV are initially asymptomatic, which impedes their early detection. High-throughput sequencing allows for extensive sampling of intra-host viral populations (Beerenwinkel and Zagordi, 2011), providing rich information, which could be used for outbreak investigations. However, the nature of RNA viruses poses the following challenges:

1. **Intra-host population diversity.** Minor viral variants are frequently responsible for transmission (Apostolou *et al.*, 2015; Fischer *et al.*, 2010). Thus, traditional approaches that rely on consensus sequences (Bartlett *et al.*, 2016; Wertheim *et al.*, 2014) could be inaccurate without consideration of intra-host viral diversity. Figure 1 provides an example. It shows a phylogenetic tree of intra-host HCV populations obtained from seven cases, each identified by different color, involved in a single outbreak (Fischer *et al.*, 2010). Here, a host shown in black has infected all other cases. This finding is supported by observation of intermixing among intra-host variants from the six cases with minority variants from the source in the tree. A consensus sequence from the source belongs to the dominant (right) clade in the tree while consensus sequences from the six cases are located in the left clade, indicating that the use of consensus sequences does not allow for detecting intermixing of HCV sequences in a single clade, which is most suggestive of linkage by transmission. Moreover, a greater heterogeneity of HCV population from the source can be used to detect the direction of transmission, indicating a more efficient utilization of viral diversity to the investigation of transmission as compared to consensus sequences.

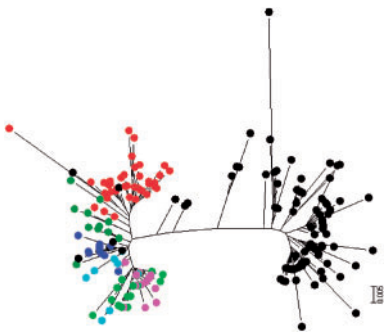

Fig. 1. Phylogenetic tree of HCV quasispecies of infected hosts from an outbreak (Fischer *et al.*, 2010). Colors represent variants from different hosts (Color version of this figure is available at *Bioinformatics* online.)

2. **Limited use of temporal epidemiological data.** For transmission history inference, existing methods rely on epidemiological data indicating times of sample collection and/or exposure intervals (De Maio *et al.*, 2016). However, HIV and HCV cause mainly chronic infections, whereas, during outbreak investigations, samples are often either collected simultaneously or the differences between collection times are negligible in comparison with times since exposure. Owing to the initial asymptomatic stage of infection, the time of sample collection may not accurately reflect the duration of infection in the host, which is in contrast to underlying assumptions of many existing methods, thus, confounding estimates of the direction of transmission among infected cases. For outbreaks associated with a high transmission rate (as among injection drug users), the time of persistence of infection in the population exceeds the time intervals between transmission events, preventing the use of exposure intervals for investigation.

3. **Viral dynamics complexity.** Viral evolution is a non-linear process, which is governed by transmission bottlenecks, impact of host's immune system and changes in selection pressure during the course of infection. However, intra-host viral dynamics is often modeled as a constant-population or linear growth process (Romero-Severson *et al.*, 2016).

Traditionally, transmission clusters were inferred by identifying tight clades in phylogenetic trees (Harris *et al.*, 2010; Holodniy *et al.*, 2012). While such an approach can be effective, it also has a number of drawbacks, e.g. it frequently fails to distinguish between recent transmission clusters and genetically close but distinct viral populations (De Maio *et al.*, 2016; Wertheim *et al.*, 2014). Another family of methods identifies potential transmissions by linking host, if the genetic distance between their viral isolates does not exceed a predefined *relatedness threshold* (Campo *et al.* 2016; Walker *et al.*, 2014; Wertheim *et al.*, 2014). Such an approach could be efficient in detection of transmission clusters, but it does not allow to infer transmission directions, cannot distinguish between direct and indirect transmissions and, as a result, cannot be used to reconstruct the transmission history.

In the recent years, several computational tools for reconstruction of viral transmission history have been published (Aldrin *et al.*, 2011; Didelot *et al.*, 2014; Cottam *et al.*, 2008; De Maio *et al.*, 2016; Jombart *et al.*, 2011, 2014; Mollentze *et al.*, 2014; Morelli *et al.*, 2012; Ypma *et al.*, 2013). Many of these tools utilize a combination of genetic and epidemiological evidences, and have been shown to be efficient in many cases, especially for viruses causing acute disease with clearly manifested symptoms, such as Influenza. However, the aforementioned challenges impede application of these tools to RNA viruses. Although some of these methods take into account one of the above challenges, to the best of our knowledge, none of them addresses any two of them (De Maio *et al.*, 2016).

Contributions. In this article, we address these problems with a unified computational approach for the inference of transmission clusters, direction of transmission, source of outbreak and transmission history. The proposed framework QUasispecies Evolution, Network-based Transmission Inference (QUENTIN) is based on modeling intra-host viral evolution together with utilization of general properties of inter-host social networks. QUENTIN uses a network-based approach utilizing plethora of powerful methods of graph and network theories. In particular, a quasispecies logistical model for intra-host viral dynamics allows incorporation into analysis the observed structure of quasispecies populations and non-linearity of viral evolution. QUENTIN was validated on simulated data and

experimental data from HCV outbreaks investigated by Centers for Disease Control and Prevention (CDC) in recent years. We also report results of application of our approach to HCV sequencing data from a high-risk community.

2 Materials and methods

QUENTIN is a graph-based (or network-based) approach, which models viral evolution and epidemic spread using the following graphs: (Fig. 2):

- *Genetic network* $\mathcal{G}_g$, which is an undirected graph with vertices corresponding to viral genomes, and edges connecting genomes, which differ by a single mutation
- *Host network* $\mathcal{G}_h$, which is a tournament with infected hosts being its vertices and arcs representing possible transmission directions. Arc weights $\mathcal{W} = (W_e)_{e \in E(\mathcal{G}_h)}$ are equal to genetic distances between corresponding viral populations.
- *Transmission tree* $\mathcal{T}$, which represents a transmission history of an outbreak. It is a rooted binary labeled tree with leafs representing infected hosts and interior nodes representing transmission events: an interior node with a label x and its children with labels x and y represent an infection of a host y by a host x. For a given transmission tree, *transmission network* $\mathcal{G}_\mathcal{T}$ indicates who infected whom: the vertices of $\mathcal{G}_\mathcal{T}$ are infected hosts, and arcs connect hosts, that are linked by transmission.

QUENTIN is organized as a pipeline that consists of two stages: (i) construction of host network, detection of outbreaks and their sources and (ii) inference of transmission trees and transmission networks. Further we describe these stages.

2.1 Estimation of genetic distances between intra-host viral populations, construction of host network and inference of transmission clusters

Consider the collection $\mathcal{P} = \{P_1, \ldots, P_n\}$ of n sets of aligned sequences representing intra-host viral populations sampled from infected individuals. We first calculate a matrix $\mathcal{D} = (D_{i,j})_{i,j=1}^n$ of pairwise distances between the populations. Let P_1 and P_2 be two fixed populations. We consider a genetic network $\mathcal{G}_g$ that contains both populations P_1 and P_2, as well as additional set of vertices $U_{1,2}$ corresponding to unsampled viral variants. The set $U_{1,2}$ represent variants, that existed between the moment of transmission and the moment of sampling, but have not been detected due to the continuous evolution or sampling bias. We estimate the set $U_{1,2}$ using Median Joining network (Bandelt *et al.*, 1999) implemented in SplitsTree (Huson and Bryant, 2006).

Viral evolution can be considered as a random process on a genetic network, and the distance between populations P_1 and P_2 can be measured using an analogue of cover time for graph random walks. Namely, for the fixed model of evolution, the distance $D_{1,2}$ between P_1 and P_2 is the expected time of an evolutionary process starting at vertices of P_1 in the graph $\mathcal{G}_g$ to reach each vertex of P_2. We use the following model:

$$\mathbf{x}^t = \left(1 - \sum_{i=1}^{n} x_i^{t-1}/M\right)((1+r)E + qA)\mathbf{x}^{t-1}. \quad (1)$$

Here $\mathbf{x}^t = (x_1^t, \ldots, x_n^t)^T$ is a vector representing the state of a viral population at time t, where x_i^t is the expected number of virions with the ith genome; M is the maximal population size; E is an identity matrix and A is an adjacency matrix of the genetic network $\mathcal{G}_g$. Given the mutation rate ϵ and the genome length L, r and q are the probabilities of replication without mutations and with a single mutation, which are calculated as $r = (1-\epsilon)^L$, $q = (\epsilon/3)(1-\epsilon)^{L-1}$. The logistic term $1 - \sum_{i=1}^n x_i^{t-1}/M$ is used to take into account nonlinearity of intra-host viral evolution, which is manifested by an exponential growth in pathogen population size during an acute stage followed by a slower growth or saturation caused by host's immune response activation or exhaustion of available resources.

We simulate viral evolution using the model (1) with the initial conditions $x_i^0 = \delta_0$, if the genome i belongs to the population P_1 and $x_i^0 = 0$, otherwise. The distance $D_{1,2}$ between populations P_1 and P_2 is defined as follows:

$$D_{1,2} = \min\{t : x_i^t \geq \delta_0, \text{ for all } i \in P_2\}. \quad (2)$$

Note that in general the distance matrix $\mathcal{D}$ is non-symmetric, i.e. it is possible that $D_{1,2} \neq D_{2,1}$. This fact combined with a minimal evolution principle (Rzhetsky and Nei, 1993) can be used to identify possible transmission direction and the direction of the corresponding arc of a host network: $(1,2) \in E(\mathcal{G}_h)$, whenever $D_{1,2} < D_{2,1}$; the weight of that arc is $W_{1,2} = D_{1,2}$.

Transmission clusters (outbreaks) are identified as weakly connected components of the graph which is obtained from the host network by removal of arcs with weights exceeding the specified threshold T^*. In addition, source of each outbreak is inferred either as a vertex with indegree 0 or as a vertex with the highest eigenvector centrality.

2.2 Inference of transmission trees and networks

The algorithm described below infers transmission tree for every weakly connected component of a host network calculated at the previous stage. Let $\mathcal{G}_h$ and $\mathcal{W}$ be a host network and arc weight function corresponding to a such component. The objective of this stage is to find a transmission tree $\mathcal{T}$, that maximizes the probability $p(\mathcal{T}|$

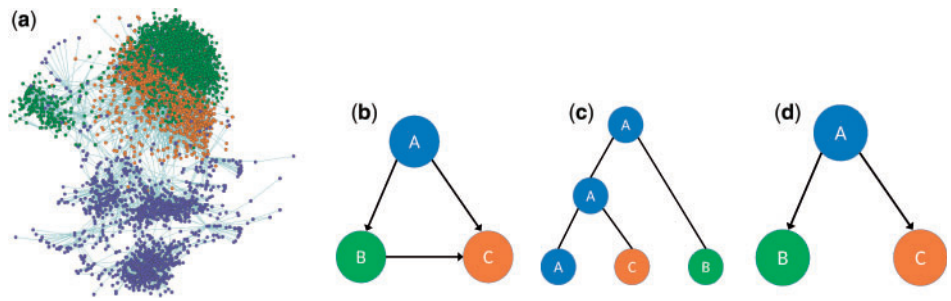

Fig. 2. Graphs used by QUENTIN. In genetic network, viral quasispecices of hosts A, B and C are shown in different colors. Networks are constructed using NGS data from HCV outbreak (Warner *et al.*, 2015). (**a**) Genetic network $\mathcal{G}_g$. (**b**) Host network $\mathcal{G}_h$. (**c**) Transmission tree $\mathcal{T}$. (**d**) Transmission network $\mathcal{G}_\mathcal{T}$ (Color version of this figure is available at *Bioinformatics* online.)

$\mathcal{G}_h, \mathcal{W})$ of observing $\mathcal{T}$ given $\mathcal{G}_h$ and $\mathcal{W}$. This probability is estimated in a Bayesian fashion as follows:

$$p(\mathcal{T}|\mathcal{G}_h, \mathcal{W}) \propto p(\mathcal{W}|\mathcal{G}_h, \mathcal{T})p(\mathcal{T}|\mathcal{G}_h), \tag{3}$$

where $p(\mathcal{W}|\mathcal{G}_h, \mathcal{T})$ is a likelihood of distances $\mathcal{W}$ given a tree $\mathcal{T}$ and possible transmission directions represented by $\mathcal{G}_h$, and $p(\mathcal{T}|\mathcal{G}_h)$ is a prior probability of $\mathcal{T}$ given $\mathcal{G}_h$.

Likelihood of estimated genetic distances. We estimate the likelihood $p(\mathcal{W}|\mathcal{G}_h, \mathcal{T})$ by assessing, how distances $\mathcal{W}$ correlate with a topology of the tree $\mathcal{T}$. We use a least square approach (Fitch *et al.*, 1967) under the assumption, that the differences between collection times of viral samples are small in comparison with times since transmission events. We first solve the following constrained linear least-squares problem:

$$\sum_{ij \in E(\mathcal{G}_h)} (X_{i,j} - W_{i,j})^2 / W_{i,j}^2 \to \min, \tag{4}$$

$$\sum_{e \in C_i} \alpha_e x_e = \sum_{e \in C_{i+1}} \alpha_e x_e, \quad i = 1, \ldots, n-1, \tag{5}$$

$$x_e \geq 0, \quad e \in E(\mathcal{T}). \tag{6}$$

Here variables x_e are lengths of edges of $\mathcal{T}$ measured in viral generations, $X_{i,j} = \sum_{e \in C_{i,j}} x_e$ is the length of the path $C_{i,j}$ between leafs corresponding to the hosts i and j; C_i is a path between a leaf corresponding to a host i and the root. The terms $\sum_{e \in C_i} \alpha_e x_e$ represent physical times of sample collection, where coefficients α_e are used for adjustment of evolutionary time (in generations) and physical time. In our model, these coefficients reflect the difference in viral evolution speed between donor and recipient intra-host populations following the transmission event represented by the parent vertex of e. Speed of evolution for recipient is generally higher due to the bottleneck effect and initial immune response absence (De Maio *et al.*, 2016); therefore assuming that e connects nodes corresponding to hosts i and j, we put $\alpha_e = 1$, if $i = j$ and $\alpha_e = \alpha \in (0, 1]$, otherwise (here α is a constant).

Finally, the likelihood $p(\mathcal{W}|\mathcal{G}_h, \mathcal{T})$ is estimated as the value $\max\{r(\mathcal{W}, \mathcal{X}), 0\}$, where $r(\mathcal{W}, \mathcal{X})$ is a Pearson correlation between vectors $(D_{i,j})_{i,j \in E(\mathcal{G}_h)}$ and $(X_{i,j})_{i,j \in E(\mathcal{G}_h)}$ (for interpretation of positive Pearson correlation as probability see e.g. Falk and Well, 1997).

Prior probability of transmission tree. Let $\mathcal{T}$ be a transmission tree and $\mathcal{G}_\mathcal{T}$ be the corresponding transmission network, which could be straightforwardly constructed by adding an arc $(x, y) \in E(\mathcal{G}_\mathcal{T})$ for every internal node of $\mathcal{T}$ with the label x and two children with labels x and y. Tree $\mathcal{T}$ *agrees* with host network $\mathcal{G}_h$, if $\mathcal{G}_\mathcal{T}$ is a subgraph of $\mathcal{G}_h$.

If $\mathcal{T}$ does not agree with $\mathcal{G}_h$, then $p(\mathcal{T}|\mathcal{G}_h) = 0$. Otherwise, for estimation of the prior probability $p(\mathcal{T}|\mathcal{G}_h)$ we use the fact that generally RNA virus transmission networks are social networks, which are usually scale-free (Brown *et al.*, 2011; Wertheim *et al.*, 2014). We assume, that transmission trees that agree with $\mathcal{G}_h$ are distributed in such a way, that trees corresponding to scale-free transmission networks have higher probabilities to be observed.

To measure "scale-freeness" of the transmission network $\mathcal{G}_\mathcal{T}$, we utilize an idea proposed in Li *et al.* (2005), which is based on a graph parameter called *s-metric*:

$$s(\mathcal{G}_\mathcal{T}) = \sum_{(i,j) \in E(\mathcal{G}_\mathcal{T})} d_i d_j, \tag{7}$$

where d_i is a (undirected) degree of a vertex i. In the statistical ensemble $\mathbb{G}(d)$ of random graphs with the same expected degree sequence d, high s-metric indicates presence of most of the typical properties of scale-free networks (Li *et al.*, 2005). Moreover, if s^* is the maximal s-metric for graphs from $\mathbb{G}(d)$, then the value $s(\mathcal{G}_\mathcal{T})/s^*$ is proportional to a relative log-likelihood of the graph $\mathcal{G}_\mathcal{T}$ under the Generalized Random Graph model (Li *et al.*, 2005).

Since we have no prior knowledge about the degree sequence of a real transmission network, we use a wider statistical ensemble $\mathbb{H}_k$ consisting of graphs with the same expected number of hubs (high-degree vertices) k. Inside $\mathbb{H}_k$, the maximal s-metric $s^*(k) = (\lfloor n/k \rfloor + k - 2)(n-1) + (k-1)(\lfloor n/k \rfloor) - 1)n/k$ is achieved on graph H_k obtained by taking disjoint union of k stars $K_{1,\lfloor \frac{n}{k} \rfloor - 1}$ and connecting the center of one of them to centers of all others.

Let κ and ρ be constants. Based on considerations above, the prior probability $p(\mathcal{T}|\mathcal{G}_h)$ is estimated as

$$p(\mathcal{T}|\mathcal{G}_h) = \begin{cases} \kappa e^{-\rho \left| 1 - \frac{s(\mathcal{G}_\mathcal{T})}{s^*(k)} \right|}, & \text{if } \mathcal{T} \text{ agrees with } \mathcal{G}_h; \\ 0, & \text{otherwise}. \end{cases} \tag{8}$$

Most probable transmission tree inference. Transmission tree maximizing the probability (3) is estimated using Markov Chain Monte Carlo (MCMC) algorithm with nearest neighbor interchange modification operation (Huson *et al.*, 2010). Several MCMC instances are run in parallel, and the most probable tree $\mathcal{T}$ generated during these runs is chosen. For each MCMC instance, the constant α for (4)–(6) is chosen randomly from the interval $[0.5, 1]$. For every tree considered by MCMC, labels of internal nodes are recalculated in order to make them agreed with the host network $\mathcal{G}_h$. It is done recursively using the following rule: an internal node with children having labels x and y receives the label x if $(x, y) \in A(\mathcal{G}_h)$. Due to (8), it allows to eliminate the search over all possible tree labelings.

In order to use (8), the expected number of hubs k of a true transmission network should be estimated. We use the following heuristic approach. First, a neighbor-joining tree $\mathcal{T}_0$ based on distances $\mathcal{W}$ and the corresponding transmission network $\mathcal{G}_{\mathcal{T}_0}$ are constructed. The degree sequence of $\mathcal{G}_{\mathcal{T}_0}$ is partitioned into two clusters using hierarchical clustering, and the cardinality of a cluster containing highest degrees is used as an estimation for k.

3 Results

We evaluated QUENTIN's ability to identify transmission directions, transmission clusters, sources of outbreaks and infer transmission history using experimental and simulated data. The following evaluation metrics were used. For transmission directions and sources of outbreaks, we calculated the percentages of correct answers among all cases, when such answers were known. To measure the similarity between true and estimated partitions into transmission clusters, *Fowlkes-Mallows (FM) index* (Amigó *et al.*, 2009) was used. For transmission network inference, we measure algorithm accuracy by proportion of correctly inferred transmission links (i.e. direct transmissions) and transmission ancestries (i.e. pairs ancestor-descendant).

3.1 Experimental data

The data consist of 335 intra-host HCV populations, including 142 populations from 33 outbreaks reported to CDC in 2008–2013 and 193 populations from infected individuals without any known epidemiological relationship, all obtained from national collections and other research projects (Campo *et al.*, 2016). Outbreak collections contain from 2 to 19 samples, with transmission histories being known for 10 outbreaks as a result of epidemiological investigations. For all samples, HCV hypervariable region 1 (HVR1) was

sequenced. In order to eliminate bias caused by different numbers of sequences sampled from different hosts, all populations were partitioned into fixed number of clusters and each cluster was replaced by its center.

First, we tested QUENTIN ability to identify transmission directions, sources in outbreaks and transmission clusters. The obtained results were compared to results from a consensus-based method [e.g. HIV-Trace (Bartlett *et al.*, 2016; Wertheim *et al.*, 2014) (Currently available online version of HIV-Trace supports only distance thresholds that do not exceed 0.02. For HCV HVR1, higher thresholds are required, therefore we used our own version of consensus-based algorithm implemented in Matlab.)] and a quasispecies-based method from Campo *et al.* (2016) (further called Mindist), which estimates the distance between viral populations as a minimal distance between their members. These methods were chosen for comparison since they, as QUENTIN, do not require additional epidemiological information to run. Both methods identify transmissions using a relatedness threshold. We ran all three algorithms with different values of relatedness thresholds, and the best threshold T^* for each method (according to FM index) has been chosen: $T^* = 1100$ for QUENTIN, $T^* = 0.0377$ for Mindist and $T^* = 0.06$ for Consensus. The results are reported in Table 1.

QUENTIN correctly identified directions of transmissions in 87% of cases and sources of outbreaks in 90% of cases (Table 1a and b). As discussed above, both Mindist and Consensus are not able to identify transmission directions, and the centrality-based method for the source detection combined with Mindist and Consensus allowed for correct identification of the source in 40 and 20% of cases, respectively. It is interesting to note, that all errors made by QUENTIN are associated with a single outbreak, where virus was transmitted through blood transfusions, while all other outbreaks were associated with unsafe injection practices or sexual contacts. The major difference between transmission through blood transfusion and all other modes of transmission is that in the former case viruses undergo a much wider transmission bottleneck. These findings demonstrate that algorithms may benefit from utilization of models specifically targeting various modes of transmission.

Table 1. Results on experimental data with best thresholds: $T^* = 1100$ (QUENTIN), $T^* = 0.0377$ (Mindist), $T^* = 0.06$ (Consensus)

Methods	Evaluation metric					
	(a)	(b)	(c)	(d)	(e)	(f)
QUENTIN	**0.87**	**0.9**	0.996	**0.992 (0.005)**	**0.78**	**0.98**
Mindist	—	0.4	**1**	0.814 (0.263)	—	—
Consensus	—	0.2	0.959	0.570 (0.327)	—	—

Note: (a) Transmission direction estimation accuracy, (b) outbreak source inference accuracy, (c) FM index for transmission clusters detection accuracy, (d) robustness to threshold variation. Values represent mean and standard deviation (in parentheses) of FM over different thresholds from the interval $[\frac{1}{2}T^*; 2T^*]$, (e) accuracy of transmission links inference and (f) accuracy of transmission ancestries inference. The best result for each parameter is shown in bold.

QUENTIN and Mindist were almost equally accurate in transmission clusters detection, both achieving high accuracy and outperforming Consensus (Table 1c). In particular, while Mindist was perfectly accurate and QUENTIN was not able to correctly link 2 hosts to their transmission clusters, Consensus was wrong for 17 hosts.

QUENTIN was significantly more robust to the threshold variation than Mindist and Consensus (Table 1d), P values 1.3×10^{-5} and 7.8×10^{-24}, respectively, two-sample Kolmogorov–Smirnov test). Figure 3a illustrates the algorithms' results for 50 different thresholds from the interval $[\frac{1}{2}T^*; 2T^*]$. While the results of Mindist and Consensus quickly deteriorate when thresholds deviate from T^*, QUENTIN shows consistent performance for all threshold values. This finding suggests that QUENTIN could be readily used in different settings, while Mindist and Consensus require fine tuning of the threshold using extensive training sets. In general, increase in proportion of non-sampled hosts does not significantly impair QUENTIN accuracy (Fig. 3b), while the accuracy of Consensus is considerably more affected.

Furthermore, we studied QUENTIN accuracy of transmission networks inference (Fig. 4a; Table 1e and f). With all hosts being sampled, QUENTIN correctly reconstructs ~78% of transmission links and ~98% of transmission ancestries. Higher accuracy of ancestries estimation is due to the fact, that the most common error made by QUENTIN is creation of too long transmission paths instead of assignment of links to superspreaders. The accuracy is affected by missing samples, although the effect is non-linear: while the accuracy decrease, when 10 and 30% of hosts are not sampled, it increases when 50% of samples are missing. Similar effect was observed for other transmission inference algorithms (De Maio *et al.*, 2016). Prior probability of a tree is essential for the algorithm: without it [$\rho = 0$ in (8)] link detection accuracy without missing samples falls from 78 to 41%.

Given that certain viruses are prone to genetic recombination, we tested its impact on QUENTIN performance. HCV recombination is rare. Therefore, it was simulated at different rates θ, starting from $\theta = 10^{-5}$, which corresponds to the estimated effective recombination rate of HIV (Neher and Leitner, 2010) (Table 2). Although the results were affected by increase in the recombination rate, reduction in accuracy of cluster, direction and source inferences was inconsequential. This diminished effect of recombination may be explained by close genetic relatedness of intra-host viral variants usually found among recently infected cases frequently identified during outbreak investigation. Recombination between such variants does

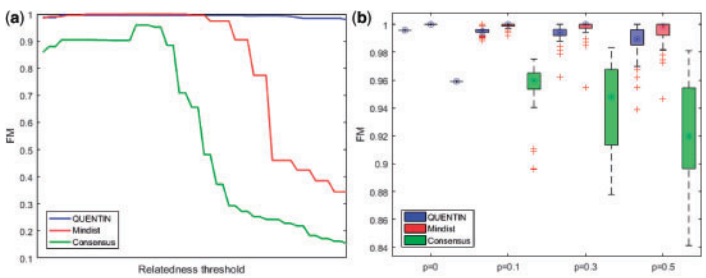

Fig. 3. Transmission clusters estimation robustness. (**a**) Threshold variation robustness. Graphs show values of FM over different thresholds from the interval $[\frac{1}{2}T^*; 2T^*]$. (**b**) Sampling robustness. Box plots show values of FM, when some hosts are not sampled. p is a proportion of unsampled hosts (Color version of this figure is available at *Bioinformatics* online.)

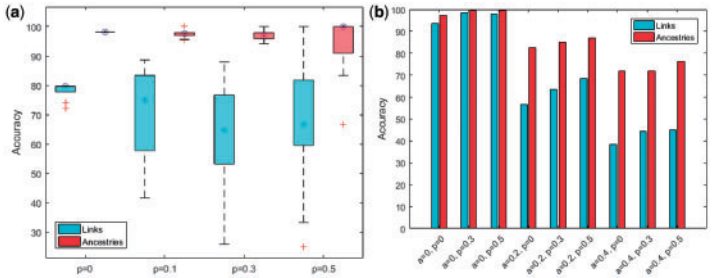

Fig. 4. Transmission history inference accuracy. (**a**) Experimental data. Here p is percentage of unsampled hosts. (**b**) Simulated data, random contact networks. p is percentage of unsampled hosts and a is distance estimation noise (Color version of this figure is available at *Bioinformatics* online.)

Table 2. Results on experimental data with simulated recombination

Methods	Evaluation metric				
	(a)	(b)	(c)	(e)	(f)
$\theta = 0$	0.87	0.9	0.996	0.78	0.98
$\theta = 10^{-5}$	0.85	0.9	0.980	0.74	0.87
$\theta = 10^{-4}$	0.84	0.8	0.979	0.66	0.77
$\theta = 10^{-2}$	0.81	0.8	0.978	0.58	0.66

Note: (a) Transmission direction estimation accuracy, (b) outbreak source inference accuracy, (c) FM index for transmission clusters detection accuracy, (d) accuracy of transmission links inference and (e) accuracy of transmission ancestries inference.

not significantly alter intra-host population structure, thus, producing a limited effect on estimates of distances between populations and direction of transmission between chronically and recently infected hosts, which form most of transmission links. However, for large θ accuracy of transmission history inference was found to be detectably reduced, suggesting its sensitivity to the local changes in the structure of the host network produced by recombination.

3.2 Simulated data

To the best of our knowledge, currently there is no tools for viral quasispecies evolution simulation, that reproduces real evolution with sufficient accuracy. Therefore we tested only QUENTIN subroutine for inference of transmission networks, when a host network $\mathcal{G}_h$ and distances $\mathcal{W}$ are given. Epidemic spread on given contact networks consisting of 10–100 individuals was simulated using SI model, $\mathcal{G}_h$ was constructed by connecting hosts with earlier dates of infection to the hosts with later dates, and the distance $W_{i,j}$ was defined as the doubled distance to the most recent common ancestor of viral strains infecting hosts i and j. As contact networks, we used either random scale-free networks or the network of individuals with potential higher risk of HIV/HCV infection inferred using social media mining (see Supplementary Material). To simulate uncertainty in distance estimation, random noise was introduced to $\mathcal{W}$ by perturbing all distances by $100a\%$. In addition, $100p\%$ of hosts were removed. The results are reported in Figure 4b and Supplementary Figure S4. As expected, algorithm accuracy is negatively affected by the noise in distances estimation. At the same time, as for the experimental data, an increase in proportion of unsampled hosts may lead to an increase of accuracy. It could be attributed to the combination of two factors: smaller number of vertices allow for more efficient traversing of the set of all possible solutions; removal of random subset of vertices with high probability does not significantly affect the transmission network topology due to the fact that most vertices of scale-free networks have low degrees.

3.3 Analysis of HCV data from a high-risk community

The data consist of HCV samples collected in 2008–09 from 34 people with high-risk behavior who reside in Ho Chi Minh City, Vietnam. Persons enrolled in the study were either commercial sex workers or injection drug users. They were selected using a chain introduction system, when a person enrolled receives a certificate for each other person who he/she introduces to the study. HCV HVR1 was sequenced using 454 GS Junior System (454 Life Sciences, Branford, CT) from each infected person.

Among tested, 17 and 15 hosts were infected with subtypes 1a and 1b, respectively. Additionally, two hosts were infected by both subtypes 1a and 1b. We separated intra-host populations from these two hosts between subtypes, and each subpopulation was treated as a separate sample. QUENTIN identified two large clusters L_b and L_a consisting of 14 and 8 hosts, respectively, and two small clusters S_a and S_b each consisting of two hosts. Hosts from clusters L_a, S_a and L_b, S_b were infected with subtypes 1a and 1b, respectively. QUENTIN was used to reconstruct transmission networks in both large clusters (Fig. 5).

The number of hosts tested negative for drug usage in the cluster L_a was significantly larger than expected (permutation test $P = 0.0127$). Additionally, sources of all arcs in the transmission tree of L_a were drug negative. These two observations suggest that sexual transmission was predominant in the cluster L_a. At the same time, although sources of all arcs in the transmission tree of the cluster L_b are drug use associated, the hypothesis that this cluster is mostly associated with unsafe injection practices was not statistically significant ($P = 0.147$).

Two samples corresponding to host VNHCV106 belong to two clusters. In the cluster L_b, this host was identified as an initial source of infection, which implies that it was infected with a subtype 1b for a long time. At the same time, QUENTIN suggests that in the cluster S_a the host VNHCV106 recently has been infected with a subtype 1a by another individual. This scenario agrees with the observation that HCV co-infection by two subtypes is unstable, with one subtype usually promptly supplanting another subtype (Webster *et al.*, 2013).

In clusters L_a and L_b, two out of four superspreaders (vertices with outdegrees > 1) were co-infected with both subtypes 1a and 1b. This result is statistically significant and in accordance with previous findings that a mixed HCV infection is an indicator of high-risk behaviors associated with associated with frequent HCV transmissions and re-exposures (Cunningham *et al.*, 2015). It is noteworthy that the three highest outdegree superspreaders were male. This finding was not statistically significant ($P = 0.078$), perhaps because there were a limited number of analyzed samples, but it does align with previous study findings demonstrating that HCV transmissions from men are more efficient than transmissions from women (Halfon *et al.*, 2001).

4 Discussion

While NGS technologies significantly facilitated pathogen research, application of genomic data to study of viral transmissions faces

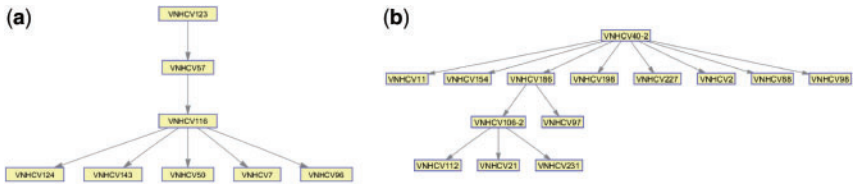

Fig. 5. Transmission history for two largest cluster from Vietnam data. (**a**) Transmission network for the cluster L_a. (**b**) Transmission network for the cluster L_b

significant computational challenges. Here, we presented QUENTIN—a framework for inference of viral transmissions, which addresses these challenges by incorporating into analysis the structures of quasispecies populations and general properties of social networks relevant to infection dissemination. QUENTIN uses a network-based approach to the intra-host population dynamics, which allows to estimate direction of transmissions and use a more complex quasispecies logistical model for intra-host viral dynamics instead of a simpler linear growth model. OUENTIN uses the fact that generally virus transmission networks are social networks with a specific properties such as power law degree distribution, small diameter and presence of hubs (Brown *et al.*, 2011; Wertheim *et al.*, 2014). We consider a transmission network to be more probable if it is close to being scale-free. Standard measures of 'scale-freeness' are based on statistical estimations, which are poorly applicable to inference of epidemic history since real transmission networks may have limited number of vertices. We utilize a combinatorial approach, which makes the proposed method universally applicable to outbreaks of various sizes and structures.

QUENTIN validation on experimental and simulated data shows its ability for the accurate inference of transmission clusters, directions of transmission, sources of outbreaks and transmission history. QUENTIN is most useful for investigation of extensively sampled outbreaks caused by RNA viruses, when additional epidemiological data are unavailable. Superior performance of the new algorithm over the consensus-based approach indicates importance of quasispecies analysis for viral molecular surveillance and outbreak investigation. Application of QUENTIN to the molecular surveillance data from a high-risk community demonstrates that its results are biologically and epidemiologically sound.

In most cases, each infected individual has a single source of infection. However, occasionally certain hosts might be infected by multiple sources. Such situation is possible among hosts with high-risk behavior, such as injection drug users. Since hosts with multiple sources are usually infected with several distinct viral subpopulations, one of the possible ways to resolve this problem is to separate these subpopulations using some clustering method and consider each subpopulation as a separate sample. We applied such approach to samples with mixed HCV subtypes from Vietnam (Section 3.3). In general, we believe that this problem requires a separate study.

Although QUENTIN is a general framework applicable to a wide variety of viruses, it has limitations and potential improvements that we will address in a future work. In particular, some RNA viruses, such as HIV, have a high recombination rate. Although recombination was shown to have a limited effect on accuracy of detection of transmission clusters and transmission direction in outbreak settings, it may affect the QUENTIN performance in more general surveillance settings. Moreover, extremely high recombination rates may alter inference of transmission history even in outbreak settings. Currently available outbreak inference tools recommend advance filtering out of recombinants (Jombart *et al.*, 2014). One of the possible ways to deal with the recombination-induced heterogeneity within our framework and without filtering is separation of each intra-host population into subpopulations, as described above. However, the best solution would be the incorporation of recombination into the mutation-based evolutionary model used by QUENTIN (see e.g. Boerlijst *et al.*, 1996). The model could be also improved by taking into account epistatic connectivity and functional differences among viral variants. The transmission tree inference could be improved by incorporation of unobserved hosts explicitly into the model, as well as by utilization of available information about the behavioral patterns of the hosts.

Further improvement of the presented framework can be achieved by integration of viral genomic data with the data extracted from actual social networks. QUENTIN utilizes most general properties of social networks, which are not specific for particular epidemiological settings. Identification of the structure and specific properties of real social networks relevant to disease dissemination is, though, a complex and labor-intensive process. However, certain important features of social network for specific types of outbreaks may be identified or accurately estimated using social media analysis (Sadilek *et al.*, 2012). Such data could be used as a more specific guide to reconstruction of transmission networks from genomic data, especially among large local groups of people with high-risk of infections such as groups of injection drug users or men who have sex with men (MSM). In this context, co-location, social ties, and behavioral patterns of the susceptible individuals are crucial in the spread of viruses like HIV and HCV. Estimation of these factors can be achieved using social media data. Our preliminary research (see Supplementary Material) on reconstruction of a network of hosts using informational indicators of HCV or HIV status, addiction, sexual identity, geographic location and connection among users of Twitter shows that the social media data can be successfully analyzed to extract contact networks of communities of potentially susceptible hosts and determine their topological parameters such as degree distribution and number of hubs. In particular, the latter parameter could be used for estimation of the expected number of hubs used by QUENTIN [see (8)] for an outbreak in a small geographic area among hosts sharing patterns of behavior relevant to disease transmission.

Acknowledgements

IM, SK and AZ were partially supported from NSF Grants 1564899 and 16119110, IM and SK were partially supported by GSU Molecular Basis of Disease Fellowship.

Conflict of Interest: none declared.

References

Aldrin,M. *et al.* (2011) Modelling the spread of infectious salmon anaemia among salmon farms based on seaway distances between farms and genetic relationships between infectious salmon anaemia virus isolates. *J. Roy. Soc. Interf.*, **8**, 1346–1356.

Amigó,E. *et al.* (2009) A comparison of extrinsic clustering evaluation metrics based on formal constraints. *Inf. Retrieval.*, **12**, 461–486.

Apostolou,A. *et al*. (2015) Transmission of hepatitis c virus associated with surgical procedures-New Jersey 2010 and Wisconsin 2011. *Morb. Mortal. Wkly. Rep*., **64**, 165–170.

Bandelt,H.-J. *et al*. (1999) Median-joining networks for inferring intraspecific phylogenies. *Mol. Biol. Evol*., **16**, 37–48.

Bartlett,S. *et al*. (2016) A molecular transmission network of recent hepatitis C infection in people with and without HIV: Implications for targeted treatment strategies. *J. Viral Hepat*., **24**, 404–411

Beerenwinkel,N. and Zagordi,O. (2011) Ultra-deep sequencing for the analysis of viral populations. *Curr. Opin. Virol*., **1**, 413–418.

Beerenwinkel,N. *et al*. (2005) Computational methods for the design of effective therapies against drug resistant HIV strains. *Bioinformatics*, **21**, 3943–3950.

Boerlijst,M.C. *et al*. (1996) Viral quasi-species and recombination. *Proc. Roy. Soc. Lond. B Biol. Sci*., **263**, 1577–1584.

Brown,A.J.L. *et al*. (2011) Transmission network parameters estimated from HIV sequences for a nationwide epidemic. *J. Infect. Dis*., jir550.

Campo,D.S. *et al*. (2016) Accurate genetic detection of hepatitis c virus transmissions in outbreak settings. *J. Infect. Dis*., **213**, 957–965.

Collier,M.G. *et al*. (2014) Outbreak of hepatitis a in the USA associated with frozen pomegranate arils imported from turkey: an epidemiological case study. *Lancet Infect. Dis*.., **14**, 976–981.

Cottam,E.M. *et al*. (2008) Integrating genetic and epidemiological data to determine transmission pathways of foot-and-mouth disease virus. *Proc. Roy. Soc. Lond. B Biol. Sci*., **275**, 887–895.

Cunningham,E.B. *et al*. (2015) Mixed HCV infection and reinfection in people who inject drugs – impact on therapy. *Nat. Rev. Gastroenterol. Hepatol*., **12**, 218–230.

De Maio,N. *et al*. (2016) Scotti: Efficient reconstruction of transmission within outbreaks with the structured coalescent. *PLoS Comput. Biol*., **12**, e1005130.

Didelot,X. *et al*. (2014) Bayesian inference of infectious disease transmission from whole-genome sequence data. *Mol. Biol. Evol*., **31**, 1869–1879.

Domingo,E. *et al*. (2012) Viral quasispecies evolution. *Microbiol. Mol. Biol. Rev*., **76**, 159–216.

Falk,R., and Well,A.D. (1997) Many faces of the correlation coefficient. *J. Stat. Educ*., **5**, 1–18.

Fischer,G.E. *et al*. (2010) Hepatitis C virus infections from unsafe injection practices at an endoscopy clinic in Las Vegas, Nevada, 2007–2008. *Clin. Infect. Dis*., **51**, 267–273.

Fitch,W.M. *et al*. (1967) Construction of phylogenetic trees. *Science*, **155**, 279–284.

Grabowski,M.K., and Redd,A.D. (2014) Molecular tools for studying hiv transmission in sexual networks. *Curr. Opin. HIV AIDS*, **9**, 126–133.

Halfon,P. *et al*. (2001) Molecular evidence of male-to-female sexual transmission of hepatitis c virus after vaginal and anal intercourse. *J. Clin. Microbiol*., **39**, 1204–1206.

Harris,S.R. *et al*. (2010) Evolution of mrsa during hospital transmission and intercontinental spread. *Science*, **327**, 469–474.

Holodniy,M. *et al*. (2012) Results from a large-scale epidemiologic look-back investigation of improperly reprocessed endoscopy equipment. *Infect. Control*., **33**, 649–656.

Huson,D.H., and Bryant,D. (2006) Application of phylogenetic networks in evolutionary studies. *Mol. Biol. Evol*., **23**, 254–267.

Huson,D.H. *et al*. (2010). *Phylogenetic Networks: concepts, Algorithms and Applications*. Cambridge University Press.

Jombart,T. *et al*. (2011) Reconstructing disease outbreaks from genetic data: a graph approach. *Heredity*, **106**, 383–390.

Jombart,T. *et al*. (2014) Bayesian reconstruction of disease outbreaks by combining epidemiologic and genomic data. *PLoS Comput. Biol*., **10**, e1003457.

Li,L. *et al*. (2005) Towards a theory of scale-free graphs: definition, properties, and implications. *Internet Math*., **2**, 431–523.

Mollentze,N. *et al*. (2014) A bayesian approach for inferring the dynamics of partially observed endemic infectious diseases from space-time-genetic data. *Proc. Roy. Soc. Lond. B Biol. Sci*., **281**, 20133251.

Morelli,M.J. *et al*. (2012) A Bayesian inference framework to reconstruct transmission trees using epidemiological and genetic data. *PLoS Comput. Biol*., **8**, e1002768.

Neher,R.A. and Leitner,T. (2010) Recombination rate and selection strength in hiv intra-patient evolution. *PLoS Comput. Biol*., **6**, e1000660.

Romero-Severson,E.O. *et al*. (2016) Phylogenetically resolving epidemiologic linkage. *Proc. Natl. Acad. Sci*., 201522930.

Rzhetsky,A. and Nei,M. (1993) Theoretical foundation of the minimum-evolution method of phylogenetic inference. *Mol. Biol. Evol*., **10**, 1073–1095.

Sadilek,A. *et al*. (2012) Modeling spread of disease from social interactions. In: *Proc. of 6th International AAAI Conference on Weblogs and Social Media*, 4 June 2012, Dublin, Ireland, pp. 322–329.

Walker,T.M. *et al*. (2014) Assessment of mycobacterium tuberculosis transmission in oxfordshire, uk, 2007–12, with whole pathogen genome sequences: an observational study. *Lancet Respir. Med*., **2**, 285–292.

Warner,A.E. *et al*. (2015) Outbreak of hepatitis c virus infection associated with narcotics diversion by an hepatitis c virus–infected surgical technician. *Am J. Infect. Control*., **43**, 53–58.

Webster,B. *et al*. (2013) Evasion of superinfection exclusion and elimination of primary viral rna by an adapted strain of Hepatitis C virus. *J. Virol*., **87**, 13354–13369.

Wertheim,J.O. *et al*. (2014) The global transmission network of hiv-1. *J. Infect. Dis*., **209**, 304–313.

Ypma,R.J. *et al*. (2013) Relating phylogenetic trees to transmission trees of infectious disease outbreaks. *Genetics*, **195**, 1055–1062.