Glebova *et al. BMC Genomics* 2017, **18**(Suppl 10):918
DOI 10.1186/s12864-017-4274-5

BMC Genomics

RESEARCH Open Access

Inference of genetic relatedness between viral quasispecies from sequencing data

Olga Glebova[1*†], Sergey Knyazev[1†], Andrew Melnyk[1], Alexander Artyomenko[1], Yury Khudyakov[2], Alex Zelikovsky[1] and Pavel Skums[1,2]

From 6th IEEE International Conference on Computational Advances in Bio and Medical Sciences (ICCABS) Atlanta, GA, USA. 13–15 October 2016

Abstract

Background: RNA viruses such as HCV and HIV mutate at extremely high rates, and as a result, they exist in infected hosts as populations of genetically related variants. Recent advances in sequencing technologies make possible to identify such populations at great depth. In particular, these technologies provide new opportunities for inference of relatedness between viral samples, identification of transmission clusters and sources of infection, which are crucial tasks for viral outbreaks investigations.

Results: We present (i) an evolutionary simulation algorithm *V*iral *O*utbreak *I*nferen*CE* (*VOICE*) inferring genetic relatedness, (ii) an algorithm MinDistB detecting possible transmission using minimal distances between intra-host viral populations and sizes of their relative borders, and (iii) a non-parametric recursive clustering algorithm *Re*latedness *D*epth (*ReD*) analyzing clusters' structure to infer possible transmissions and their directions. All proposed algorithms were validated using real sequencing data from HCV outbreaks.

Conclusions: All algorithms are applicable to the analysis of outbreaks of highly heterogeneous RNA viruses. Our experimental validation shows that they can successfully identify genetic relatedness between viral populations, as well as infer transmission clusters and outbreak sources.

Keywords: Genetic relatedness, Transmission networks, Outbreaks investigations, Simulation, Clustering

Background

Inferring transmission clusters, transmission directions, and sources of outbreaks from viral sequencing data are crucial for viral outbreaks investigation. Outbreaks of RNA viruses, such as Human Immunodeficiency Virus (HIV) and Hepatitis C virus (HCV), are particularly dangerous and pose a significant problem for public health. It is well known that genomes of RNA viruses mutate at extremely high rates [1]. As a result, RNA viruses exist in infected hosts as populations of closely related variants called *quasispecies* [2, 3]. However, only recently with the progress of sequencing technologies, it became possible to identify and sample quasispecies at great depth [4–9]. Consequently, a contribution of sequencing technologies to molecular surveillance of viral disease epidemic spread becomes more and more substantial [10, 11].

Computational methods can be used to infer transmission characteristics from sequencing data. The first question usually is whether two viral populations belong to the same outbreak. The methods typically utilize the simple observation that all samples from the same outbreak are genetically related, so they use some measure of genetic relatedness as a predictor for epidemiological relatedness [10–12].

The second question is which samples constitute isolated outbreaks. For this purposes, we define a transmission cluster as a connected set of genetically related viral populations. The third questions we address in this article is "Who is the source of infection?". This questions is the most difficult to answer, and there were

*Correspondence: glebova@cs.gsu.edu
†Equal contributors
[1]Computer Science Department, Georgia State University, 25 Park Place NE, 30303 Atlanta, GA, USA
Full list of author information is available at the end of the article

only a few attempts to do it computationally using solely genomic data [13] without invoking additional epidemiological information [14]. To the best of our knowledge, there is still no freely available computational tool for this problem.

Computational methods for detection of viral transmissions and inference of transmission clusters are often consensus-based, i.e. they analyze only a single representative sequence per intra-host population (for example, consensus sequence). Such methods assign two hosts into one transmission cluster, if the distances between corresponding sequences do not exceed a predefined threshold [10, 11]. Although consensus-based methods proved to be useful, they do not take into account intra-host viral diversity. Inclusion of whole intra-host populations into analysis is important, because minor viral variants are frequently responsible for transmission of RNA viruses [15, 16].

Recently published computational approach (further referred to as MinDist) [12] uses the minimal genetic distance between sequences of two viral populations as a measure of genetic relatedness of intra-host viral populations. Since minimal genetic distances between different pairs of populations can be achieved on various pairs of sequences, this approach takes into account intra-host diversity.

However, both consensus-based and MinDist approaches have further limitations. First of all, they do not allow to detect directions of transmissions, which is crucial for detection of outbreak sources and transmission histories. Secondly, distance thresholds utilized by both approaches could be derived from analysis of limited or incomplete experimental data and highly data- and situation-specific, with different viruses or even different genomic regions of the same virus requiring specifically established thresholds.

In this paper, we address the above limitations by proposing two novel algorithms *ReD* and *VOICE*, as well as by suggesting an improvement of the MinDist algorithm. The new algorithms allow to infer important epidemiological characteristics, including genetic relatedness, directions of transmissions and transmission clusters.

- *R*elatedness *D*epth (*ReD*) method uses clustering-based analysis of intra-host viral populations. It is a non-parametric algorithm, so it does not rely on any virus-specific threshold values to predict epidemiological characteristics.
- *V*iral *O*utbreak *I*nferen*CE* (*VOICE*) is a simulation-based method which imitates viral evolution as a Markov process in the space of observed viral haplotypes
- MinDistB method is a modification of MinDist [12], which takes into account the sizes of relative borders of each pair of viral populations.

The proposed algorithms were validated on the experimental data obtained from HCV outbreaks. Comparative results suggest that our methods are efficient in epidemiological characteristics inference.

Methods

Relatedness depth (*ReD*) algorithm

ReD is a deterministic algorithm based on deterministic hierarchical clustering. The key concept of this method is a *k-clustered intersection* of viral populations (we used similar idea previously for combinatorial pooling [17]). For two sets of viral sequences P_1 and P_2, their k-clustered intersection $P_1 \overline{\cap} P_2$ is calculated as follows:

1) Partition the union $P_1 \cup P_2$ into k clusters $C_1, ..., C_k$;
2) $P_1 \overline{\cap} P_2 = \bigcup_{i \in B} C_i$, where
$B = \{i \in \{1, ..., k\} : C_i \cap P_1 \neq \emptyset, C_i \cap P_2 \neq \emptyset\}$, i.e. $P_1 \overline{\cap} P_2$ is the union of clusters, which contain sequences from both P_1 and P_2 (see Fig. 1);

The parameter k is a *scale* of clustering. In particular, populations P_1 and P_2 are *separable*, if $P_1 \overline{\cap} P_2 = \emptyset$, while the fact that $P_1 \overline{\cap} P_2 \neq \emptyset$ indicates that they may

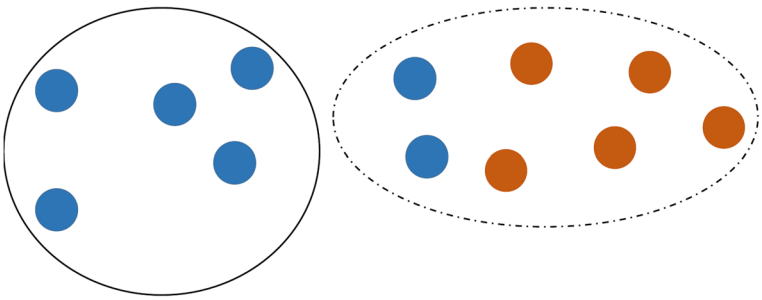

Fig. 1 *k*-clustered intersection of two viral populations (blue and red). Union of populations is partitioned into $k = 2$ clusters (dashed and solid). Dashed cluster is the *k*-clustered intersection. Direction of transmission is from the blue population to the red population

be genetically related. In the most extreme case $P_1 \overline{\cap} P_2 = P_1 \cup P_2$, i.e. populations are *completely inseparable* under the scale k.

The degree of confidence that the samples are genetically close is represented by the *relatedness depth* $d(P_1, P_2)$, which is calculated by Algorithm 1. Simply speaking, Algorithm 1 tries to recursively separate populations P_1 and P_2. At each iteration, k-clustered intersection is calculated. If two populations are separable, then the algorithm stops. Otherwise, it continues the separation of sequences from P_1 and P_2 within their k-clustered intersection. The separation depth is a depth of this recursion. It is possible that at some iterations of Algorithm 1 two populations are completely inseparable under a current clustering scale. In this case, the scale k is increased and k-clustered intersection is recalculated. The initial value of k used by Algorithm 1 is $k = 2$.

Algorithm 1 *ReD* (relatedness depth calculation)

Input Two sets of viral sequences P_1, P_2.
Output Relatedness depth $d = d(P_1, P_2)$

$d \leftarrow 0$
$k \leftarrow 2$
$I \leftarrow P_1 \overline{\cap} P_2$
while $I \neq \emptyset$ and $k \leq |P_1| + |P_2|$ **do**
  $d \leftarrow d + 1$
  if $I \neq P_1 \cup P_2$ **then**
    $P_1 \leftarrow P_1|_I$, $P_2 \leftarrow P_2|_I$ (restrictions of P_1 and P_1 on I)
    $k \leftarrow 2$
  else
    $k \leftarrow k + 2$
  end if
  $I \leftarrow P_1 \overline{\cap} P_2$
end while

k-clustered intersections depend on a clustering method. Our implementation uses a hierarchical clustering based on neighbor-joining tree (as implemented in Matlab (MathWorks, Natick, MA)). The algorithm utilizes a standard Jukes-Cantor distance which is based on the simplest substitution-based evolutionary model.

Clustered intersections also allow for estimating the direction of transmissions. It is reasonable to assume that if two hosts share a population, then a host with more heterogeneous population is more likely to be the transmission source [18]. Formally, if $I = P_1 \overline{\cap} P_2$, $P_1 \subseteq I$ and $P_2 \backslash I \neq \emptyset$, then we assume that probable transmission direction is from P_2 to P_1 (see Fig. 1). The direction is defined according to the first occurrence of such situation during execution of Algorithm 1. Note that in some cases direction may not be identified.

Given the collection of viral populations $\mathcal{P} = \{P_1, ..., P_n\}$, *ReD* produces the weighted directed genetic relatedness graph $G = (V, A, d)$ with $V = \mathcal{P}$. An arc (P_i, P_j) is in A whenever populations P_i and P_j are genetically related, i.e., have sufficiently high relatedness depth; the direction of an arc corresponds to the estimated direction of transmission and its weight to the relatedness depth. Transmission clusters are calculated as weakly connected components of the digraph G. To determine transmission clusters, the simplest depth cutoff $T = 1$ can be used. In addition, only components containing at least one arc a of weight $d(a) \geq 2$ were considered as reliable. For each reliable component, a source s of the corresponding outbreak is identified as a vertex with highest eigenvector centrality.

Viral outbreak inference (*VOICE*) simulation method

VOICE is another approach to predict epidemiological characteristics. Unlike *ReD*, it is not deterministic. Instead, it simulates the process of evolution from one viral population (source) into another (recipient) as a Markov process on a union of both populations. VOICE starts evolution from a subset of source sequences called the *border set* and estimates the number of generations required to acquire a genetic heterogeneity observed in the recipient.

Formally, given two sets of viral sequences P_1 and P_2, *VOICE* simulates viral evolution to estimate times t_{12} and t_{21} needed to cover all sequences from the recipient population under the assumptions that first and second host were sources of infection. Based on the value $\min\{t_{12}, t_{21}\}$, the algorithm decides whether the populations are related. The direction of possible transmission between the related pair is assumed to follow the direction which requires less time.

The simulation starts from the δ-*border set* B_1, which contains viral variants that are likely the closest to variants transmitted between P_1 and P_2. It is defined as the set of vertices of P_1 minimizing pairwise Hamming distance D between vertices from P_1 and P_2 up to a constant δ:

$$B_1 = \left\{ u \in P_1 : \exists v \in P_2 \;\; D(u, v) = \min_{x \in P_1, y \in P_2} D(x, y) + \delta \right\}$$

(see Fig. 2). The constant δ is a parameter, with the default value 1.

The simulated evolutionary process is carried out in the evolutionary space represented by the *variant graph* $G(B_1, P_2)$, which is constructed as follows. First, construct a union of all minimal spanning trees of the complete graph on a vertex set $B_1 \cup P_2$ with the edge weights equal to Hamming distances between variants (sometimes referred to as a pathfinder network $PFNet(n-1, \infty)$ [19, 20]). Then substitute every edge in graph with two directed edges of the same weight. Next, subdivide each

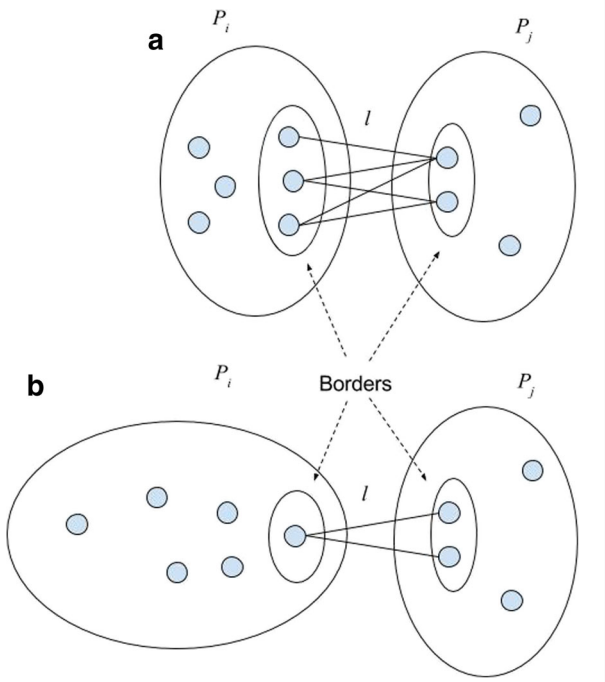

Fig. 2 δ-Crossing between two viral populations P_1 and P_2 $l \leq d(u,v) + \delta$; (**a**) $|B_\delta| = 5$; (**b**) $|B_\delta| = 2$

edge (u_1, u_2) of weight $w \geq 2$ with $w-1$ vertices $v_1, ..., v_{w-1}$ and add multiple directed edges as follows: add $w-1$ edges between vertices u_1 and v_1; $w-2$ edges between v_1 and v_2; and so forth as shown on Fig. 3. This model can be explained as follows: to mutate from vertex u_1 to u_2 during simulation, there should occur mutations at w positions that are different between u_1 and u_2. During the first step, simulation can mutate any of w positions, then any of $w-1$ positions on the second step and so forth.

The simulation starts from all border vertices B_1 and runs until all the vertices of the population P_2 are reached. At the beginning of the simulation, border vertices get count equal to 1, and the rest of the vertices get count 0. Each tact simulates variants replication by updating vertex counts according to one of the three following scenarios happening with the specified probabilities (see Fig. 4). First, if during replication there are no mutations, then the vertex v replicates itself and its count label is incremented. This happens with the probability p_1 (1). Second, the vertex can mutate into one of its neighboring vertices with probability p_2 (see Eq. (2)), in which case the count of the neighbor is incremented. Finally, with probability p_3, vertex does not produce any viable offspring, in which case vertex counts are not changed. If the count of a vertex reaches the maximum allowed variant population size C_{max}, then it is not increased. The probabilities of these scenarios are calculated as follows:

$$p_1 = (1-3\epsilon)^L \quad (1)$$

$$p_2 = p_1 \frac{\epsilon}{1-3\epsilon} \quad (2)$$

$$p_3 = 1 - p_1 - p_2 \deg^-(v) \quad (3)$$

where ϵ is the mutation rate, L is the genome length and $\deg^-(v)$ is an outdegree of a vertex v.

Algorithm 2 represents the flow of the method. The time t_{12} is computed as the average over s simulations. The same procedure is repeated for the opposite direction of the transmission with its border set B_2 and the time t_{21} is computed. The value $\min\{t_{12}, t_{21}\}$ determines which direction of transmission is more likely.

Algorithm 2 *VOICE* (Viral Outbreak InferenCE)

Input Two sets of viral variants P_1, P_2.
Output Time $t_{1,2}$ to evolve from P_1 to P_2.

find the δ-border set B_1
build the variant graph $G = G(B_1, P_2)$
$t \leftarrow 0$
Assign the number of copies $c_v^t \leftarrow 1$ to each variant $v \in B_1$ and $c_v^t \leftarrow 0$ to each variant $v \in P_2 \setminus B_1$
while there are variants $v \in P_2$ with $c_v^t = 0$ **do**
    $c_v^{t+1} \leftarrow c_v^t$ for every $v \in V(G)$
    for each variant $v \in V(G)$ **do**
        for $i = 1, ..., c_v^t$ **do**
            with a probability p_1, $c_v^{t+1} \leftarrow \min\{c_v^{t+1} + 1, C_{max}\}$
            with a probability p_2, $c_u^{t+1} \leftarrow \min\{c_u^{t+1} + 1, C_{max}\}$, where u is a randomly chosen neighbor of v
        end for
    end for
    $t \leftarrow t + 1$
end while
$t_{1,2} \leftarrow t$

Data normalization

The sizes of observed intra-host viral populations may significantly vary due to sampling and sequencing biases. Since the larger population will require more time to cover, the estimation of t_{12} and t_{21} could be biased. VOICE avoids such biases by normalizing the intra-host population sizes. The deterministic normalization partitions each viral population into q clusters using hierarchical clustering and each cluster is replaced with the consensus of its members. The subsampling normalization randomly chooses q sequences from each population. The procedure is repeated r times, and the final result is an average over all subsamplings.

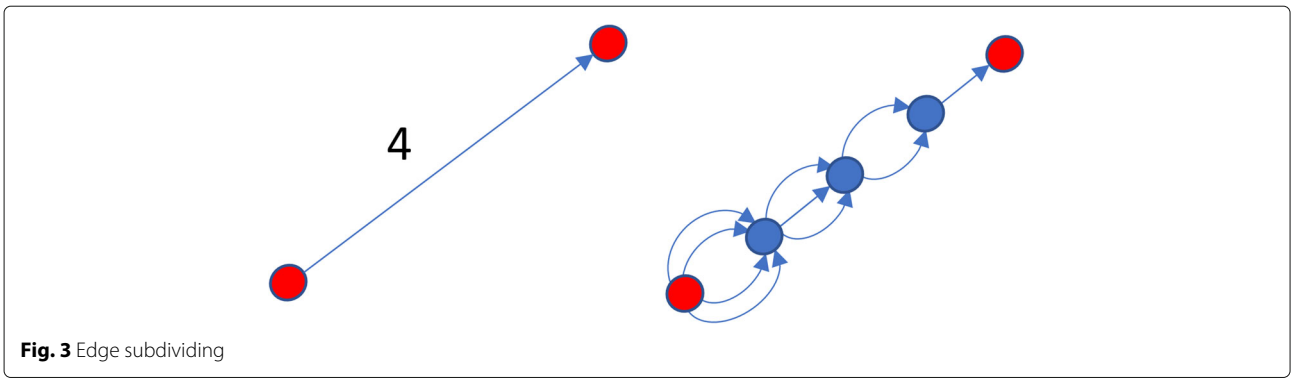

Fig. 3 Edge subdividing

Identification of genetic relatedness, transmission directions, clusters and sources of outbreaks

Analogously to *ReD*, *VOICE* produces a weighted directed genetic relatedness graph $G = (V, A, w)$ with $V = \mathcal{P}$. An arc P_iP_j is in A whenever populations P_i and P_j are genetically related, i.e., value $\min\{t_{ij}, t_{ji}\}$ is less than a threshold. Weakly connected components of G represent transmission clusters or outbreaks. To determine the source of each outbreak, we build a Shortest Paths Tree (SPT) for every vertex in the corresponding component. The source is estimated as the vertex with an SPT of minimal weight.

MinDistB method

The method extends the MinDist approach proposed in [12], which defines the distance between viral populations as the minimum Hamming distance between their representatives. The new approach also takes into account sizes of border sets, on which the minimum distance is achieved.

Formally, given an integer δ (by default $\delta = 1$), the δ-crossing between populations P_1 and P_2 is the set of pairs of variants (u, v) from different populations, the Hamming distance $D(u, v)$ between which is within δ from the minimum Hamming distance:

$$B_\delta(P_1, P_2) = \left\{(u, v) : u \in P_1, v \in P_2, D(u, v) \leq \min_{x \in P_1, y \in P_2} D(x, y) + \delta\right\}$$

(see Fig. 2). Our empirical study shows that in case when the crossing is large (see Fig. 2a), then the populations are less likely to be related than in case when the borders are small (see Fig. 2b).

This effect can be intuitively explained. Two related populations likely diverge away from the common ancestor and from each other, and their borders are formed by few old survived variants closest to the common ancestor. Two unrelated populations diverging from two different ancestors may in time reduce minimum distance from each other randomly and closest variants are relatively young and abundant (see Fig. 5).

We define a δ-distance between populations P_1 and P_2 as follows:

$$D_\delta(P_1, P_2) = D(P_1, P_2) + c \ln(|B_\delta(P_1, P_2)|) \quad (4)$$

where $c = 3$ is an empirically chosen constant.

Identification of genetic relatedness, transmission clusters and sources of outbreaks

For MinDistB methods, genetic relatedness graph $G = (V, E, w)$ is a weighted undirected graph with the vertex set $V = \mathcal{P}$ and an edge of weight $w_{i,j}$ connecting populations P_i, P_j whenever $w_{i,j} = D_\delta(P_1, P_2)$ does not exceed a threshold. Transmission clusters are estimated as connected components of the graph G. For each transmission cluster its source could be inferred either as a vertex with maximum eigenvector centrality or as a vertex with the shortest paths tree of minimal weight.

Results and discussions

ReD, VOICE and MinDistB were validated using experimental outbreak sequencing data, and their predictions were compared with the previously published MinDist method [12].

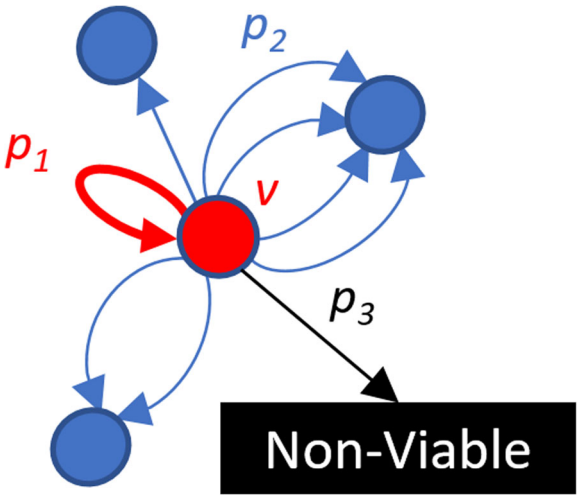

Fig. 4 All possible moves of a vertex *v*

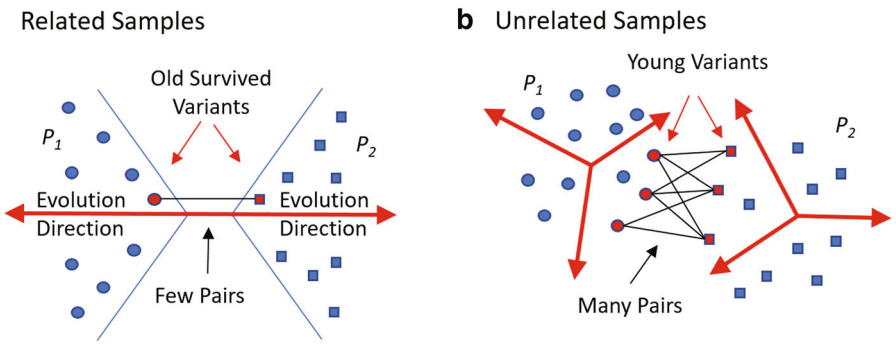

Fig. 5 Intuition behind the MinDistB method. **a** Related samples – crossing is between old survived variants. **b** Unrelated samples –crossing is between many young variants which are close to each other by chance

Data sets

We used the benchmark data presented in [12], which is a collection of HCV intra-host populations sampled from 335 infected individuals.

- Outbreak collection contains 142 HCV samples from 33 epidemiologically curated outbreaks reported to Centers for Disease Control and Prevention in 2008–2013. Outbreaks contain from 2 to 19 samples. Epidemiological histories, including sources of infection, are known for 10 outbreaks.
- Collection of 193 epidemiologically unrelated HCV samples.

All viral sequences represent a fragment of E1/E2 genomic region of length 264 bp.

Prediction of epidemiological characteristics

The proposed methods were used to infer the following epidemiological characteristics:

- genetic relatedness between populations;
- transmission clusters representing outbreaks and isolated samples;
- sources of outbreaks;
- transmission directions between pairs of samples.

Comparison results are collected in Table 1. The variants of VOICE with deterministic and subsampling normalizations are referred to as $VOICE-D$ and $VOICE-S$, and for them we used the normalization constants $q = 10$ and $q = 4$, respectively. For all VOICE runs, five independent simulations were performed, and the averages over that simulations are reported. For each simulation, VOICE-S performs 50 subsamplings, and the results of the algorithm are averaged over all subsamplings. For MinDist, sources of outbreaks were identified as vertices with highest eigenvector centralities in the corresponding genetic relatedness graphs, since for MinDist this method outperform the shortest path tree-based approach.

Genetic relatedness between populations

Viral populations from two samples are genetically related if they belong to the same outbreak and unrelated, otherwise. The genetic relatedness is validated on the union of both collections containing all outbreaks and unrelated samples. There are 55945 pairs of samples, and 479 of them are related. For all algorithms we choose the best thresholds, which produce no false positives, i.e. no unrelated populations are predicted to be related. The values of thresholds T are: $ReD : T = 2$; $MinDist : T = 11$; $MinDistB : T = 28.4$; $Voice - D : T = 1710$; $Voice - S : T = 4585$.

Table 1 Validation results

Methods	MinDist	MinDistB	*ReD*	VOICE-D	VOICE-S
Relatedness					
Sensitivity, %	90%	92.9%	55.3%	85.2%	86.8%
AUROC	0.992	0.996	N/A	0.993	0.990
Clustering					
Sensitivity, %	100%	100%	96.3%	98.2%	98.2%
Source					
Accuracy, %	50%	40%	90%	80%	90%
Directions					
Accuracy, %	N/A	N/A	87.1%	83.9%	87.1%

For each method, the sensitivity (i.e. the percentage of detected related pairs) was calculated (Table 1). The highest sensitivity is achieved by MinDistB method. Figure 6 depict ROC curve for the tested methods (*ReD* is not present, since for this method only few viable discrete thresholds are possible). *MinDistB* and $VOICE - D$ have highest areas under a curve value followed by *MinDist* and $VOICE - S$.

Detection of transmission clusters

The similarities between true and estimated partitions into transmission clusters were measured using an editing metric [21], which is defined as the minimum number of elementary operations required to transform one partition into another. An elementary operation is either merging (joining of two clusters into a single cluster) or division (partition of a cluster into two clusters) [21]. We calculate sensitivity by normalizing an editing distance E by dividing it by the number N of elementary operations required to transform trivial partition (i.e. the partition into singleton sets) into the true partition. The number N is equal to $n - k$, where n is the total number of samples and k is the number of true clusters:

$$Sensitivity = \frac{E}{n-k} \times 100\%. \quad (5)$$

Table 1 shows that MinDistB and MinDist demonstrate the highest sensitivity.

Source identification

The accuracy of the source identification is defined as the percentage of correctly predicted sources for outbreaks, where the correct sources are known. The Source section of Table 1 shows that the best results are achieved by *ReD* and $VOICE - S$ which were able to detect sources in 90% of cases. At the same time, MinDist and MinDistB, which are not able to identify transmission directions, were significantly less accurate.

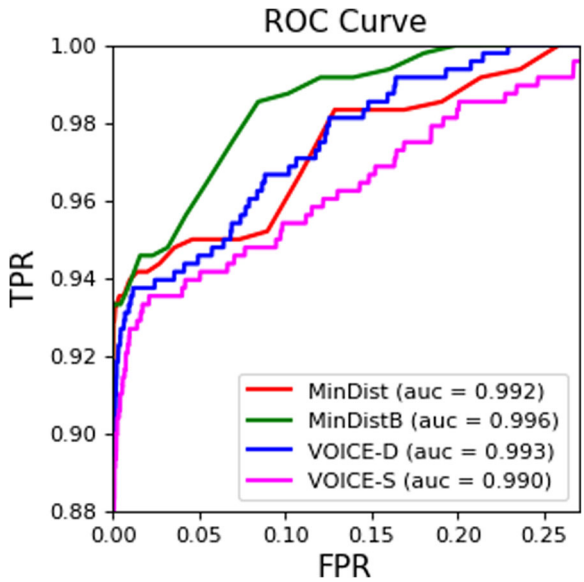

Fig. 6 ROC curve for pairs relatedness detection

Transmission direction

Among tested algorithms, only *ReD* and *VOICE* allows for detection of transmission directions. For that algorithms, percentages of correctly predicted pairs source-recipient were calculated (Table 1). Here the highest accuracy of 87.1% was achieved by *ReD* and $VOICE - S$.

Running time

All tests were performed on PC with DDR3-1333MHz 4 GBx12 RAM and 2 Intel Xeon-X5550 2.67 GHz processors. The fastest algorithms were MinDist and MinDistB, with running times 9 ms for a pair of samples in our dataset. *ReD* requires $\sim 0.1s$ per pair of samples, While the running time of *VOICE* is $\sim 35\ s$ per pair.

Conclusions

Currently, a molecular viral analysis is one of the major approaches used for investigations of outbreaks and inference of transmission networks. Although modern sequencing technologies significantly facilitated molecular analysis, providing unprecedented access to intra-host viral populations, they generated novel bioinformatics challenges.

This work proposed three novel algorithms for the investigation of viral transmissions based on analysis of the intra-host viral populations, which allow clustering genetically related samples, infer transmission directions and predict sources of outbreaks. Evaluation of the algorithms on experimental data from HCV outbreaks demonstrated their ability to accurately reconstruct various transmission characteristics. It should be noted, that although *ReD* was proved to be accurate in estimation of transmission clusters, directions and sources, its accuracy of relatedness detection is lower than for other evaluated methods. However, the advantage of this method over other methods is its non-parametricity (i.e. independence from virus-specific and genomic region-specific thresholds), which makes it more universally applicable and extremely useful in situations, when the lack of training data does not allow to establish reliable relatedness thresholds.

The clustering-based *ReD* approach may be further improved using a more scalable clustering similar to the algorithm proposed in [17]. The simulation-based approach *VOICE* presented here may be further improved by incorporating more complex viral evolution models taking into account cell proliferation rate and immune responses against viral variants.

All algorithms are planned to be integrated into the pipeline of cloud-based web-system "Global Hepatitis Outbreak and Surveillance Technology" (GHOST), which is currently being developed by US Centers for Disease Control and Prevention (https://webappx.cdc.gov/GHOST/).

Funding
AZ was partially supported by NSF Grant CCF-16119110 and NIH Grant 1R01EB025022-01; PS was partially supported by NIH Grant 1R01EB025022-01; OG, SK, AM, and AA were partially supported by GSU Molecular Basis of Disease Fellowship. The publication costs were funded by NSF Grant CCF-1611911.

Availability of data and materials
ReD and *VOICE* are freely available at https://bitbucket.org/osaofgsu/red and https://bitbucket.org/osaofgsu/voicerep, respectively.

About this supplement
This article has been published as part of *BMC Genomics* Volume 18 Supplement 10, 2017: Selected articles from the 6th IEEE International Conference on Computational Advances in Bio and Medical Sciences (ICCABS): genomics. The full contents of the supplement are available online at https://bmcgenomics.biomedcentral.com/articles/supplements/volume-18-supplement-10.

Authors' contributions
OG and SK designed, implemented and tested the algorithms; AM and AA implemented and tested the algorithms; YK designed the algorithms and analyzed the algorithms' results; AZ and PS designed and implemented the algorithms, analyzed the results and supervised the research. All authors read and approved the final manuscript.

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Publisher's Note

Author details
[1]Computer Science Department, Georgia State University, 25 Park Place NE, 30303 Atlanta, GA, USA. [2]Centers for Disease Control and Prevention, 1600 Clifton Rd, 30329 Atlanta, GA, USA.

Published: 6 December 2017

References

1. Drake JW, Holland JJ. Mutation rates among rna viruses. Proc Natl Acad Sci. 1999;96(24):13910–3.
2. Domingo E, Holland J. Rna virus mutations and fitness for survival. Annu Rev Microbiol. 1997;51(1):151–78.
3. Domingo E, Sheldon J, Perales C. Viral quasispecies evolution. Microbiol Mol Biol Rev. 2012;76(2):159–216.
4. Eriksson N, Pachter L, Mitsuya Y, Rhee SY, Wang C, Gharizadeh B, Ronaghi M, Shafer RW, Beerenwinkel N. Viral population estimation using pyrosequencing. PLoS Comput Biol. 2008;4(5):1000074.
5. Archer J, Braverman MS, Taillon BE, Desany B, James I, Harrigan PR, Lewis M, Robertson DL. Detection of low-frequency pretherapy chemokine (cxc motif) receptor 4-using hiv-1 with ultra-deep pyrosequencing. AIDS (London, England). 2009;23(10):1209.
6. Hoffmann C, Minkah N, Leipzig J, Wang G, Arens MQ, Tebas P, Bushman FD. Dna bar coding and pyrosequencing to identify rare hiv drug resistance mutations. Nucleic Acids Res. 2007;35(13):91.
7. Wang W, Zhang X, Xu Y, Weinstock GM, Di Bisceglie AM, Fan X. High-resolution quantification of hepatitis c virus genome-wide mutation load and its correlation with the outcome of peginterferon-alpha2a and ribavirin combination therapy. PLoS ONE. 2014;9(6):100131.
8. Skums P, Campo DS, Dimitrova Z, Vaughan G, Lau DT, Khudyakov Y. Numerical detection, measuring and analysis of differential interferon resistance for individual hcv intra-host variants and its influence on the therapy response. Silico Biol. 2011;11(5):263–9.
9. Campo DS, Skums P, Dimitrova Z, Vaughan G, Forbi JC, Teo CG, Khudyakov Y, Lau DT. Drug resistance of a viral population and its individual intrahost variants during the first 48 h of therapy. Clin Pharmacol Ther. 2014;95(6):627–35.
10. Wertheim JO, Brown AJL, Hepler NL, Pond SLK. The global transmission network of hiv-1. J Infect Dis. 2014;209(2):304–13.
11. Wertheim JO, Pond SLK, Forgione LA, Mehta SR, Murrell B, Shah S, Smith DM, Scheffler K, Torian LV. Social and genetic networks of hiv-1 transmission in new york city. PLoS Pathog. 2017;13(1):1006000.
12. Campo DS, Xia GL, Dimitrova Z, Lin Y, Forbi JC, Ganova-Raeva L, Punkova L, Ramachandran S, Thai H, Skums P, et al. Accurate genetic detection of hepatitis c virus transmissions in outbreak settings. J Infect Dis. 2016;213(6):957–65.
13. Romero-Severson EO, Bulla I, Leitner T. Phylogenetically resolving epidemiologic linkage. Proc Natl Acad Sci. 2016;113(10):2690–5. doi:10.1073/pnas.1522930113. http://www.pnas.org/content/113/10/2690.full.pdf.
14. De Maio N, Wu CH, Wilson DJ. Scotti: Efficient reconstruction of transmission within outbreaks with the structured coalescent. PLoS Comput Biol. 2016;12(9):1005130.
15. Fischer GE, Schaefer MK, Labus BJ, Sands L, Rowley P, Azzam IA, Armour P, Khudyakov YE, Lin Y, Xia G. Hepatitis c virus infections from unsafe injection practices at an endoscopy clinic in las vegas, nevada, 2007–2008. Clin Infect Dis. 2010;51(3):267–73.
16. Apostolou A, Bartholomew ML, Greeley R, Guilfoyle SM, Gordon M, Genese C, Davis JP, Montana B, Borlaug G. Transmission of hepatitis c virus associated with surgical procedures-new jersey 2010 and wisconsin 2011. MMWR Morb Mortal Wkly Rep. 2015;64(7):165–70.
17. Skums P, Artyomenko A, Glebova O, Ramachandran S, Mandoiu I, Campo DS, Dimitrova Z, Zelikovsky A, Khudyakov Y. Computational framework for next-generation sequencing of heterogeneous viral populations using combinatorial pooling. Bioinformatics. 2015;31(5):682–90. doi:10.1093/bioinformatics/btu726. http://bioinformatics.oxfordjournals.org/content/31/5/682.full.pdf+html.
18. Astrakhantseva IV, Campo DS, Araujo A, Teo CG, Khudyakov Y, Kamili S. Differences in variability of hypervariable region 1 of hepatitis c virus (hcv) between acute and chronic stages of hcv infection. Silico Biol. 2011;11(5):163–73.
19. Quirin A, Cordón O, Guerrero-Bote VP, Vargas-Quesada B, Moya-Anegón F. A quick mst-based algorithm to obtain pathfinder networks. J Am Soc Inf Sci Technol. 2008;59(12):1912–24.
20. Campo DS, Dimitrova Z, Yamasaki L, Skums P, Lau DT, Vaughan G, Forbi JC, Teo CG, Khudyakov Y. Next-generation sequencing reveals large connected networks of intra-host hcv variants. BMC Genomics. 2014;15(Suppl 5):4.
21. Deza MM, Deza E. Encyclopedia of Distances. Springer-Verlag Berlin Heidelberg; 2009.