BMC Bioinformatics

SOFTWARE

Open Access

Primary case inference in viral outbreaks through analysis of intra-host variant population

J. Walker Gussler[1,2], David S. Campo[1*], Zoya Dimitrova[1], Pavel Skums[2] and Yury Khudyakov[1]

*Correspondence:
fyv6@cdc.gov
[1] Centers for Disease Control and Prevention, 1600 Clifton Rd, Atlanta, GA 30333, USA
Full list of author information is available at the end of the article

Abstract

Background: Investigation of outbreaks to identify the primary case is crucial for the interruption and prevention of transmission of infectious diseases. These individuals may have a higher risk of participating in near future transmission events when compared to the other patients in the outbreak, so directing more transmission prevention resources towards these individuals is a priority. Although the genetic characterization of intra-host viral populations can aid the identification of transmission clusters, it is not trivial to determine the directionality of transmissions during outbreaks, owing to complexity of viral evolution. Here, we present a new computational framework, PYCIVO: primary case inference in viral outbreaks. This framework expands upon our earlier work in development of QUENTIN, which builds a probabilistic disease transmission tree based on simulation of evolution of intra-host hepatitis C virus (HCV) variants between cases involved in direct transmission during an outbreak. PYCIVO improves upon QUENTIN by also adding a custom heterogeneity index and identifying the scenario when the primary case may have not been sampled.

Results: These approaches were validated using a set of 105 sequence samples from 11 distinct HCV transmission clusters identified during outbreak investigations, in which the primary case was epidemiologically verified. Both models can detect the correct primary case in 9 out of 11 transmission clusters (81.8%). However, while QUENTIN issues erroneous predictions on the remaining 2 transmission clusters, PYCIVO issues a null output for these clusters, giving it an effective prediction accuracy of 100%. To further evaluate accuracy of the inference, we created 10 modified transmission clusters in which the primary case had been removed. In this scenario, PYCIVO was able to correctly identify that there was no primary case in 8/10 (80%) of these modified clusters. This model was validated with HCV; however, this approach may be applicable to other microbial pathogens.

Conclusions: PYCIVO improves upon QUENTIN by also implementing a custom heterogeneity index which empowers PYCIVO to make the important 'No primary case' prediction. One or more samples, possibly including the primary case, may have not been sampled, and this designation is meant to account for these scenarios.

Background

Hepatitis C virus (HCV) infection affects nearly 3% of the world's population and is a major cause of liver disease worldwide [1]. In the United States, HCV infection is an important public health problem, being the most common chronic blood-borne infection as well as the leading cause for liver transplantation [2]. Since 2007, HCV surpasses HIV as a cause of death in the US [3]. Outbreaks of HCV infections are associated with unsafe injection practices, drug diversion, and other exposures to blood and blood products.

RNA viruses such as HCV exist as a heterogeneous population of closely related but genetically distinct variants, known as quasispecies [4]. When a transmission event occurs between an infected person and a susceptible person, the target patient will not receive the entire intra-host HCV population from the infected individual, but only some of the variants. These variants may or may not be representative of the infected individual's HCV population as a whole [5].

Owing to this, methodologies which rely on one consensus sequence per patient are not effective in detection of the primary case. A better alternative is to obtain a large sample of viral variants that adequately represent intra-host viral sub-populations and thus improve accuracy of genetic detection of transmissions. This is achieved through sampling intra-host variants using amplicon-based deep sequencing technology of a highly variable region of the viral genome [6].

We previously developed the Global Hepatitis Outbreak and Surveillance Technology (GHOST), a cloud-based bioinformatics suite that can infer HCV transmission clusters by analyzing intra-host HCV populations [7]. However, given a transmission cluster, GHOST lacked the functionality to identify the likely primary case: the individual who was infected the longest and has the greatest chance to have infected other persons in the outbreak. Given that the identification of the primary case can help in the interruption and prevention of outbreaks, we present PYCIVO, a model that evaluates the strength of the molecular evidence available to identify the potential primary case for any given recent transmission cluster.

Methods

Given a transmission cluster of several samples with available sequence, PYCIVO utilizes a consensus methodology from 2 independent models to identify a potential primary case. All HCV sequences are in FASTA format. Two primary case *indication values* are computed for each patient in the outbreak from each model. We call the first indication value V_1; this is estimated using data from a population-based evolutionary simulator. The second indication value, named V_2, is a composite heterogeneity score which is generated using an ensemble of different metrics. Identification of the likely primary outbreak case in a cluster improves with application of both values over only one of them.

Evolutionary simulations

To generate V_1, an evolutionary distance is determined between every pair of samples in the transmission cluster and stored in a matrix M_T. In order to obtain these distances, we

utilize an evolutionary simulation method proposed in our earlier paper [8]. Briefly, evolutionary-based random processes are simulated over a graph $G=(V, E)$, in which the nodes are haplotypes (unique or distinct sequences), and edges exist between two haplotypes at hamming or edit distance equal to 1. In each simulation, one population is designated as a potential infector and the other as a possible infectee. The distance measure between populations is defined as the analogue of the cover time for the random process.

For this evolutionary simulator to function, it is important to reconstruct unsampled variants that were likely present between the time of transmission and the time of sampling and were linking both samples. This reconstruction is done using maximum parsimony, assuming the minimal number of mutation steps to reach from primary case to target sequences. To achieve this, we first construct a *k-step network* [6, 9–12], i.e. the union of all minimum spanning trees of the complete graph on the set of haplotypes with edge weights corresponding to hamming distance between two given sequences. Next, each edge of length r is subdivided by $r-1$ vertices. These vertices represent unsampled haplotypes and form "bridges" between connected components of observed haplotypes. This procedure is outlined in Fig. 1. The result of the procedure is a connected graph whose edges represent single mutations and whose nodes are haplotypes of three types: those present in the potential infector, those present in the possible infectee, and simulated intermediate sequences which likely were present at some point between the time of transmission and the time of sampling. Evolutionary simulations are conducted on this constructed graph.

Simulations are performed using a model from [8], which is essentially a quasispecies model with logistic growth, depicted in Fig. 2. The model is described by the following equation:

$$x^t = \left(\left(1 - \sum_{i=1}^{n} \frac{x_i^{t-1}}{K}\right)(1+r)E + qA\right)x^{t-1}$$

Here $x^t=(x_1{}^t, \ldots, x_n{}^t)^T$ is a vector of abundances of haplotypes corresponding to the nodes of the graph G at time t, K is the maximum population size, E is an identity matrix and A is an adjacency matrix of G, and replication probability r and mutation probability q are defined as functions of the position-wise mutation rate ε and genome length L:

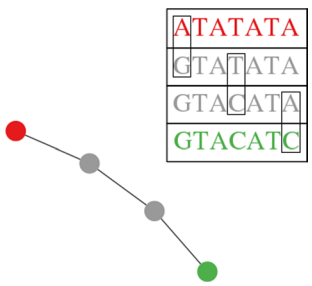

Fig. 1 Example of a bridge reconstruction using maximum parsimony. If the red variant A were to mutate to the green variant B the gray intermediate variants would represent the most parsimonious path between these two variants

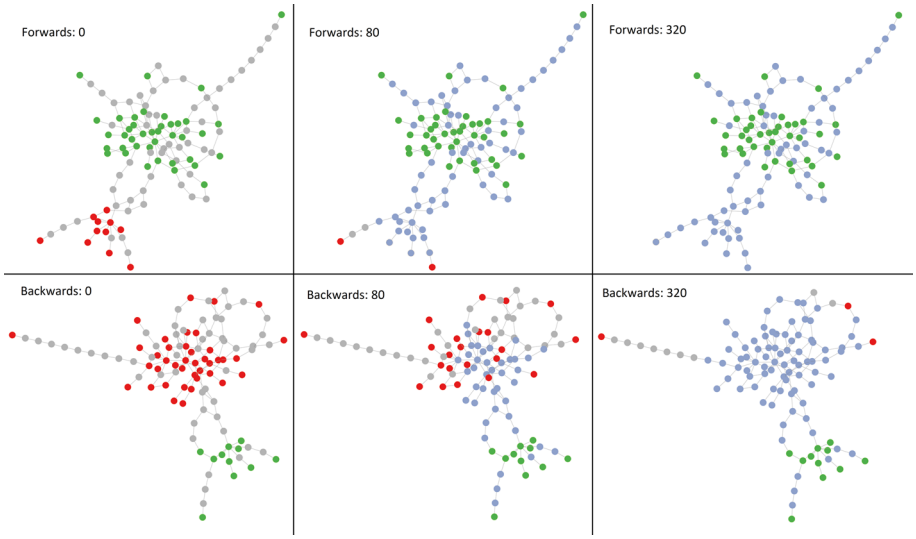

Fig. 2 Visualization of evolutionary simulation. Dots represent unique variants. The simulation is a random walk that starts from the green potential infector variants with the gray intermediate variant dots and red potential infectee variant dots being unexplored. As the simulation carries out, dots turn blue as they are explored. The simulation is complete when all potential infectee variant dots are reached, or when all the red dots have turned blue. The top 3 frames show sample A evolving into sample B. The bottom 3 frames show sample B evolving into sample A. We can see that it is easier for A to evolve into B as it finishes faster, so we would estimate that A transmitted the virus to B rather than vice versa

$$r = (1-\varepsilon)^L, q = \frac{\varepsilon}{3}(1-\varepsilon)^{L-1}$$

The evolutionary distance measure for a pair of populations (i, j) is defined as the time needed by the process from equation 1 to cover all nodes from the target j when starting only from the vertices of the potential infector i. This distance is considered to be infinite if the simulation takes longer than a threshold T_{max}. The simulations are carried out for all pairs of patients and the calculated distances are stored in the distance matrix M_T. Figure 3 depicts the relationship between this evolutionary simulation distance T and the minimum hamming distance between samples which has been shown to be a robust measure of relatedness [6].

The matrix M_T is used to construct a directed *genetic relatedness graph*, G_T, with nodes corresponding to patients, edges corresponding to pairs of patients with genetically related viral populations and their directions representing possible transmission directions. Using the minimal evolution principle, patients i and j are connected by an arc whenever $M_T[i, j] = min(M_T[i, j], M_T[j, i]) < \infty$.

If G_T does not form a connected component with all members of the input, we cannot make any prediction about the primary case. A recent outbreak in which the entire transmission network was sampled should yield a connected component with the PYCIVO evolutionary distance metric [8]. Among all input in this component, the candidate primary case a is selected based on the parameter $V_1(a)$ defined as follows:

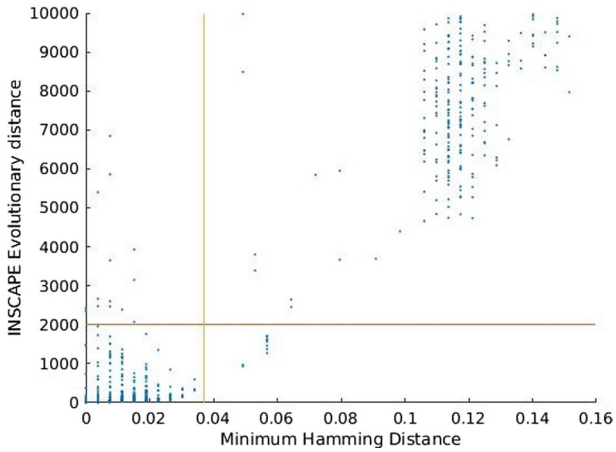

Fig. 3 Distribution of PYCIVO distance measurements among our dataset. Most sample pairs fall either well below or above the empirically derived T_{max} value of 2000. Evolutionary distances used in PYCIVO versus the minimum hamming distances between samples. The threshold for transmission is indicated by vertical and horizontal lines. Only 827 out of 5460 cases in our study were linked by the PYCIVO distance metric, 817 of these are also linked by minimum hamming distance

$$V_1(a) = \frac{\sum_{i=1}^{n} M_T[a,i]}{\sum_{i=1}^{n} M_T[i,a]}$$

where n is the number of samples present, the numerator represents the 'in-evolution' times, or the total time it took for all the samples to evolve to this one. The denominator represents the 'out-evolution' times, or the time it took for this sample to evolve into all the others. If this ratio is high, then that is indicative of a patient being a likely primary case candidate, as they were able to evolve to others more successfully than the other way around in simulations. If the ratio is low, that sample is not considered as a candidate for the primary case. For each sample, this value is the output for this portion of the algorithm. These values average to 1, with the best primary case candidates having higher values.

Measures of heterogeneity

The second method for primary case candidate determination within a transmission cluster works by selecting the infected individual with the highest amount of heterogeneity according to a custom index. We use several different measures to generate a composite score. Genetic heterogeneity has been shown to be higher for primary outbreak cases than other individuals sampled in an outbreak [6]. For each sample, 8 values are measured as shown below. A denotes the clinical sample of interest, where $F(A_i)$ and $H(A_i, A_j)$ are shorthand for frequency and hamming distance, respectively. The chosen heterogeneity values are as follows:

(1) Maximum intra-patient Hamming distance: the largest Hamming difference between two of the haplotypes

$$\max\left(H\left(A_i, A_j\right)\right) \forall i, j \in A$$

(2) Mean intra-patient Hamming distance: the arithmetic mean of all the pairwise Hamming distance between pairs of sequences within the sample.

$$\text{mean}\left(H\left(A_i, A_j\right)\right) \forall i, j \in A$$

(3) Nucleotide diversity: degree of polymorphism in a sample.

$$\sum F(A_i) F\left(A_j\right) H\left(A_i, A_j\right) \forall i, j \in A$$

(4) K-mer entropy: Shannon entropy over the set of 8-mers in the sample. Let $A_k = \{k_1 \ldots, k_n\}$ be the set of k-mers.

$$-1 \sum_{i=1}^{n} F(k_i) \log_2 F(k_i)$$

(5) Haplotype entropy: Shannon entropy over the set of haplotypes in the sample. Let $A_h = \{h_1 \ldots, h_n\}$

$$-1 \sum_{i=1}^{n} F(h_i) \log_2 F(h_i)$$

(6) Average nucleotide entropy: Shannon entropy over the set of nucleotides in the sample. If we consider the alphabet $N = \{A, T, C, G\}$ and amplicon sequence length L, where individual nucleotides are n_i

$$\frac{-1}{L} \sum_{i=1}^{L} \sum_{i=1}^{|N|} F(n_i) \log_2 F(n_i)$$

(7) One step component entropy: Let $A_0 = \{o_1 \ldots o_n\}$ be the set of one-step hamming distance of connected components present in the sample

$$-1 \sum_{i=1}^{n} F(o_i) \log_2 F(o_i)$$

(8) Mutation Frequency: Let M be the viral variant with the highest frequency

$$\sum H(M, A_i) F(A_i) \forall i \in A$$

We chose these eight values because they were empirically shown to be higher for chronically infected individuals than acute ones, as shown in Table 3. We combine them into a single scalar composite index for each sample. Using a matrix in which rows represent patients and columns represent each of the 8 features, each feature is ranked among patients in such a way that a more heterogeneous sample receives a lower rank. To obtain V_2 from this matrix, we compute the ranks in descending order along each index, then calculate the reciprocal of the harmonic mean of these ranks for each sample.

We choose this method rather than sorting in ascending order and using the arithmetic mean because we value a sample that is the most heterogeneous over most of the indices rather than being second or third among all indices. Using the reciprocal

of the harmonic mean rewards samples which score the highest among a given index disproportionately.

Implementation

PYCIVO uses data generated by the GHOST system, which processes genomic data generated by the Illumina MiSeq sequencer using several specially developed procedures [7]. There are some pre-processing requirements associated with this model. The input data should be a multiple sequence alignment in FASTA format, split into one file for each patient. Additionally, all reads in analysis must be same in length; length normalization can optionally be performed by PYCIVO using MAFFT with the -a option. Additionally, each file should not contain duplicate sequences. If there are two reads which have the same sequence, they are collapsed together into one haplotype with an associated frequency. This frequency should be at the end of the sequence ID following an underscore character.

Results

PYCIVO was validated using data from 11 transmission clusters containing 105 samples and 1936 unique sequences [13–19]. In all cases, the primary case was epidemiologically identified. In addition, we used 10 modified transmission clusters in which the known primary case sample was removed in order to test PYCIVO specificity. One transmission cluster included only 2 cases and could not be used effectively as a modified cluster.

The cover times of simulated evolution from intra-host viral variants sampled from patient A to intra-host variants from patient B (out-evolution time) and from patient B to patient A (in-evolution time) are asymmetric; i.e., the time of evolution from A to B and from B to A are different. A good primary case candidate would have overall high in-evolution and low out-evolution times. This concept is illustrated in Fig. 2. In contrast, a simpler metric such as minimal or average hamming distance does not inform on which direction is more likely. The evolutionary time, however, correlates positively ($r = 0.65$, $p < 10^{-200}$) with hamming distances, as shown in Fig. 3.

Both primary case indication values were calculated to issue primary case predictions for each outbreak. We aim to achieve accurate detection of the primary case only when one was actually sampled in an outbreak investigation. We recognize that the sample size for validation fo this software is very small. However, epidemiologically labeled complete outbreak clusters are exceedingly difficult to find. We aim to make our software specific at the cost of sensitivity. With this in mind, PYCIVO issues a "No Primary Case" (NPC) prediction when the primary case was not sampled or was removed from the data set.

This category of output is enabled by discrepancy between the two independent prediction methods, or when the genetic relatedness graph G_T is not connected. In this case, PYCIVO informs the user that there is not a good candidate for the primary case within this group, claiming that they were likely not present in the samples given.

If there is no discrepancy between prediction methods and G_T forms a connected component, then the V_1 and V_2 vectors were used to determine the confidence of the prediction. We get a distribution over the input samples for these two metrics: one from V_1 and another from V_2. Two levels of PYCIVO prediction were empirically established based off these distributions: High Confidence (HC) when a sample

Table 1 Results per outbreak

Id	n of cases	n of features	Primary present	Primary absent
1	33	0	NPC	NPC
2	19	7	HC	NPC
3	15	8	HC	NPC
4	9	0	NPC	NPC
5	7	8	LC	NPC
6	6	7	HC	NPC
7	4	8	LC	NPC
8	4	8	LC	NPC
9	3	8	LC	LC
10	3	7	LC	LC
11	2	8	LC	N/A

The last two columns show the predicted label

Table 2 Results per label

Label	Primary present	Primary absent
HC	3/11	0/10
LC	6/11	2/10
NPC	2/11	8/10

has the highest value along both identification metrics, and both of their z-scores are above 2; Low Confidence (LC) when a sample has the highest value for both primary case indication values but one or both z-scores are below 2. We chose the z-score of 2 because this is analogous to being in the 95th percentile for each metric. This is a very stringent criteria by design, as we care more about avoiding false positives than false negatives. Table 1 shows that LC predictions are much more common than NPC and HC predictions. This is an intentional feature of the software to give more veracity to NPC and HC predictions. This model is still in its early stages and was validated on a small amount of data, so LC predictions at this point are prone to error and have little value without supporting epidemiological information.

Tables 1 and 2 summarize the PYCIVO results obtained using the data from 11 known transmission clusters. The main results are as follows: (1) PYCIVO made primary case predictions in 9/11 clusters while the others were classified as NPC; (2) HC predictions (n = 3) were all accompanied by NPC when the primary case was removed; (3) LC predictions (n = 5) were accompanied by 3 NPC predictions as well as 2 erroneous LC predictions in the modified dataset, and the remaining outbreak did not have enough cases to conduct this analysis; (4) the 2 cases in the modified dataset which issued LC predictions were outbreaks of only 2 patients; (5) NPC predictions were always accompanied by NPC predictions in the modified dataset.

Table 3 shows how each feature performed for the collection of outbreaks for which we had access. This table also lists 5 metrics which were removed due to low positive predictive value as well as high correlation to other metrics within the dataset. Many of the features correctly highlight the intended sample despite having distinct biological meaning, except for the indices at the bottom of Table 3 which were excluded due

Table 3 Feature performance

K-mer entroppy	9/11
Average nucleotide entropy	9/11
Mean hamming distance	9/11
Max hamming distance	9/11
Nucleotide diversity	8/11
Haplotype entropy	8/11
Mutation frequency	8/11
1-step component entropy	8/11
Epistasis coefficient	6/11
Frequency entropy	3/11
Hill numbers	3/11
Mean consensus	2/11
Simpson index	2/11

to their lack of accuracy and high correlation to more efficacious metrics within the set.

Using this multitude of indices rather than just one of them produces a system which is more robust. This is so because artifacts in HCV evolution which reduce the primary case's heterogeneity along one index will not be equally reflected in the other indices.

Discussion

Genetic characterization of intra-host viral populations is often used for the detection of transmission clusters and for the investigation of outbreaks [6]. However, these genetic approaches, although highly effective in uncovering transmissions, are infrequently applied to tracking transmissions, owing to complexity of intra-host [20] and inter-host [5] viral evolution. Here, we present a new computational framework for the identification of the likely primary case in HCV outbreaks. The presented approach is based on simulation of evolution of intra-host HCV variants between cases involved in direct transmission during an outbreak, assessment of 2 identification values and statistical evaluation.

These methods enable PYCIVO to make the important NPC distinction, as well as preventing from making HC predictions on small outbreaks. This is so because the maximum z-score among a small set of random numbers will likely be less than 2, producing a LC prediction. It is difficult to determine the veracity of a source prediction for smaller outbreaks.

Three of the analyzed clusters are made up of three or less patients, which tend to produce LC predictions. In contrast, larger outbreaks make HC or NPC predictions more often. In agreement with this observation, there is a very high correlation between the minimum z-score used for prediction and the number of samples in the cluster ($r^2 = 0.9322$, $p = 2.4e^{-5}$). This is advantageous for most outbreaks as majority of them have more than three patients, which steers PYCIVO towards the more conclusive NPC and HC prediction classes.

PYCIVO evaluates the strength of the molecular evidence available to identify the likely primary case of a recent transmission cluster. Our method considers the information provided by both inter-host and intra-host evolution, showing an accuracy

of 81.8% when the primary case is present. The part of the method that deals with inter-host evolution uses the cover times of evolutionary simulations as an analog of evolutionary distance. If intra-host viral variants in one patient have higher out-evolution times, that means that they are unlikely to evolve to variants in other patients. If intra-host viral variants in a patient have higher in-evolution times, that means that they are unlikely to be evolved to. Intra-host variants of a likely primary case candidate should have a high in-evolution time and low out-evolution time. The part of the method that deals with intra-host evolution uses a composite heterogeneity score, which is generated via an ensemble of different metrics. The rationale for this second value is that, in general, HCV accumulates mutations during intra-host evolution and becomes more genetically heterogeneous over time [20]. Given that the primary case must have been infected for a longer time than all other incident cases, we can use this difference in duration of infection to infer the transmission direction. Indeed, our previous HCV analyses showed that the primary case is infected with a much more diverse HCV population than any incident case from the corresponding transmission cluster [5, 6]. This finding is supported with our earlier observation that, on average, the intra-host HVR1 nucleotide diversity is 1.8 times greater in patients with chronic than acute HCV infection [21].

Identification of the primary case is only possible with these methods if genetic samples from that patient are available. Otherwise, the incident case with the highest value may be erroneously classified as a source of infection in a transmission cluster. These problems of transmission-direction detection have been noted earlier [22, 23]. Therefore, our goal in developing this model was twofold: in addition to being able to predict the primary case, we aimed to be sensitive to input in which the primary case is neither known nor sampled. To achieve this goal, output was suppressed if there was uncertainty about who the primary case may have been, or if it appeared that they were not present among the samples. Output suppression thresholds were determined in via empirical thresholds outlined in the previous section.

PYCIVO is based on knowledge of intra-host and inter-host evolutionary dynamics of HCV shortly after a transmission event. Thus, the method's performance will likely be less reliable if all the sampling times were not both close to the transmission event and close to each other. Over time, intra-host HCV populations from infected cases will both evolve away from one other and the population will grow more heterogeneous over time. The performance of this type of algorithm is impacted when one or more individuals experience superinfection, inflating the intra-host heterogeneity, which results in loss of association between the heterogeneity and duration of infection.

The previously published QUENTIN model [8] showed the same performance as PYCIVO when the transmission cluster does include the primary case. However, QUENTIN always chooses one sample among the input members as the primary case regardless of whether they are actually present in the input group of samples. In contrast, PYCIVO can issue the important NPC prediction on outbreaks, which do not have a clear primary case. The ability for PYCIVO to suppress erroneous output via this mechanism represents the major advantage of this model, as we cannot expect sequences from the primary case to always be available. Another difference is that

in QUENTIN the network linking all the sequences of every pair of samples was a median-joining network rather than the k-step network used here, a change motivated by obtaining almost identical results with a considerably lower computational burden.

Our work has two major limitations. First, the datasets used in this study are limited in size. However, data on well-defined transmission clusters is exceedingly difficult to obtain. These data become available only after comprehensive epidemiological and molecular investigations. Second, the genomic data used here to devise and evaluate PYCIVO were not all obtained by deep sequencing but mainly by an older technology based on end-point limiting-dilution PCR and Sanger sequencing [20]. Both methods produce a population of variants, however, deep sequencing yields orders of magnitude more sequences. Although we previously found that statistical comparisons between these two methods on the same set of samples showed equivalent inter- and intra-host levels of heterogeneity [24], the use of PYCIVO with deep sequencing data needs to be further validated and updated. The results obtained in our study warrant research to further improve accuracy of the model to increase potential benefits of its application in the field. Utilization of differences in inter- and intra-host variability will likely continue as a core component of the model, but there are several avenues for modification in the future. The simplest way to improve upon the model is to accrue additional data from epidemiologically characterized outbreak investigations, which will either further validate the model or present opportunity to adjust parameters accordingly. We expect that our GHOST platform [7] developed to assist in identification of transmission clusters during outbreak investigation or similar technologies will provide ample data to improve the accuracy of this model in the future. Additionally, we plan further improvement of the model by updating the current strategy for calculating V_2 to include a feature vector based on predictions of infection duration by PHACELIA [25]. We may introduce new methodologies to the way in which V_1 is computed as well when novel data become available.

The epidemiological identification of outbreak sources is a very complex task. The genetic detection of a likely primary case can greatly facilitate investigation of outbreaks, assisting in identification of new cases and routes of transmission in specific epidemiological settings, and, thus, in guiding public health interventions for interruption and prevention of disease spread. However, it should be noted that although genetic testing can help detect the source of infection, it does not reveal the actual mechanism of transmission operating during outbreaks. Identification of such mechanisms and routes of transmission as well as the primary case role in an outbreak can be accomplished only through epidemiological investigation. For example, inadequate infection control, unsafe injection practices or drug diversion may be responsible for HCV transmission in healthcare settings rather than actions of source cases [26].

Understanding transmission of infections is crucial for effective public health interventions. The utility of transmission networks to public health interventions was demonstrated in simulation experiments [27–29]. Further improvement of accuracy will enhance potential benefits of the PYCIVO application to prevention and to development of more targeted interventions. However, genetic contact tracing and identification of primary cases present complex ethical issues associated with legal and social

implications [30, 31]. Resolution of these issues is fundamentally related to data security. In this respect, it is important to note that the model reported here cannot collect, use, or produce personally identifiable information and is based exclusively on using genetic viral data.

Conclusion

Here we present PYCIVO, a model that evaluates the strength of the molecular evidence available to identify the primary case for a transmission cluster. Our method takes into account the information provided by both inter-host and intra-host evolution, showing an accuracy of 81.8% when the primary case is present and with the important ability to issue 'No Primary Case' predictions in 80% of modified transmission clusters in which the epidemiologically verified primary case was removed.

Availability and requirements

Project name: PYCIVO.
Project home page: https://www.github.com/walkergussler/PYCIVO.
Operating system(s): Platform independent.
Programming language: Python 2.7.
Other requirements: No.
License: GNU General Public License.
Any restrictions to use by non-academics: No.

Abbreviations
HCV: Hepatitis C Virus; HVR1: Hyper-Variable region 1; PYCIVO: Primary case inference in viral outbreaks; GHOST: Global Hepatitis Outbreak and Surveillance Technology; HC: High confidence; LC: Low confidence; NPC: No primary case.

Acknowledgements
The authors would like to acknowledge Seth Sims for support in software development. The authors would additionally like to thank Bailey Gussler for proposing the name PYCIVO.

Authors' contributions
DSC, PS and YK designed the study. DSC and ZD processed the sequence data. PS and JWG developed the evolutionary distances. DSC, ZD and JWG developed the networks and heterogeneity measures. JWG performed developed the composite measures and performed all calculations. JWG, DSC and YK wrote the manuscript. All authors evaluated the final draft. All authors read and approved the final manuscript.

Funding
This work was supported by intramural funding from Centers for Disease Control and Prevention, Atlanta, GA and by the National Institutes of Health Grant 1R01EB025022-01. The funding bodies did not play any roles in the design of the study, analysis, interpretation of data and in writing the manuscript.

Availability of data and materials
Data available upon request.

Declarations

Ethics approval and consent to participate
Research was conducted as approved by the Institutional Review Board of the Centers for Disease Control and Prevention, Atlanta GA (protocol 7270.0). The findings and conclusions in this report are those of the authors and do not necessarily represent the official opinion of the U.S. Centers for Disease Control and Prevention (CDC).

Consent for publication
Consent for publication was obtained through the internal CDC document clearance system.

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

Author details
[1]Centers for Disease Control and Prevention, 1600 Clifton Rd, Atlanta, GA 30333, USA. [2]Department of Computer Science, Georgia State University, 1 Park Place NE, Atlanta, GA 30303, USA.

Received: 25 November 2020 Accepted: 25 January 2022
Published online: 08 February 2022

References

1. Mohd Hanafiah K, Groeger J, Flaxman AD, Wiersma ST. Global epidemiology of hepatitis C virus infection: new estimates of age-specific antibody to HCV seroprevalence. Hepatology. 2013;57(4):1333–42. https://doi.org/10.1002/hep.26141.
2. Alter MJ. Epidemiology of hepatitis C virus infection. World J Gastroenterol. 2007;13(17):2436–41. https://doi.org/10.3748/wjg.v13.i17.2436.
3. Ly KN, Xing J, Klevens RM, Jiles RB, Ward JW, Holmberg SD. The increasing burden of mortality from viral hepatitis in the United States between 1999 and 2007. Ann Intern Med. 2012;156(4):271–8. https://doi.org/10.7326/0003-4819-156-4-201202210-00004.
4. Domingo E, Sheldon J, Perales C. Viral quasispecies evolution. Microbiol Mol Biol Rev. 2012;76(2):159–216. https://doi.org/10.1128/MMBR.05023-11.
5. Campo DS, Zhang J, Ramachandran S, Khudyakov Y. Transmissibility of intra-host hepatitis C virus variants. BMC Genomics. 2017;18(Suppl 10):881. https://doi.org/10.1186/s12864-017-4267-4.
6. Campo DS, Xia GL, Dimitrova Z, Lin Y, Forbi JC, Ganova-Raeva L, et al. Accurate genetic detection of hepatitis C virus transmissions in outbreak settings. J Infect Dis. 2016;213(6):957–65. https://doi.org/10.1093/infdis/jiv542.
7. Longmire AG, Sims S, Rytsareva I, Campo DS, Skums P, Dimitrova Z, et al. GHOST: global hepatitis outbreak and surveillance technology. BMC Genomics. 2017;18(Suppl 10):916. https://doi.org/10.1186/s12864-017-4268-3.
8. Skums P, Zelikovsky A, Singh R, Gussler W, Dimitrova Z, Knyazev S, et al. QUENTIN: reconstruction of disease transmissions from viral quasispecies genomic data. Bioinformatics. 2018;34(1):163–70. https://doi.org/10.1093/bioinformatics/btx402.
9. Campo DS, Dimitrova Z, Yamasaki L, Skums P, Lau DT, Vaughan G, et al. Next-generation sequencing reveals large connected networks of intra-host HCV variants. BMC Genomics. 2014;15(Suppl 5):S4. https://doi.org/10.1186/1471-2164-15-S5-S4.
10. Campo DS, Roh HJ, Pearlman BL, Fierer DS, Ramachandran S, Vaughan G, et al. Increased mitochondrial genetic diversity in persons infected with hepatitis C virus. Cell Mol Gastroenterol Hepatol. 2016;2(5):676–84. https://doi.org/10.1016/j.jcmgh.2016.05.012.
11. Campbell EM, Jia H, Shankar A, Hanson D, Luo W, Masciotra S, et al. Detailed transmission network analysis of a large opiate-driven outbreak of HIV infection in the United States. J Infect Dis. 2017;216(9):1053–62. https://doi.org/10.1093/infdis/jix307.
12. Palmer BA, Schmidt-Martin D, Dimitrova Z, Skums P, Crosbie O, Kenny-Walsh E, et al. Network analysis of the chronic hepatitis C virome defines hypervariable region 1 evolutionary phenotypes in the context of humoral immune responses. J Virol. 2015;90(7):3318–29. https://doi.org/10.1128/JVI.02995-15.
13. Zeuzem S, Hezode C, Bronowicki JP, Loustaud-Ratti V, Gea F, Buti M, et al. Daclatasvir plus simeprevir with or without ribavirin for the treatment of chronic hepatitis C virus genotype 1 infection. J Hepatol. 2016;64(2):292–300. https://doi.org/10.1016/j.jhep.2015.09.024.
14. Tugwell BD, Patel PR, Williams IT, Hedberg K, Chai F, Nainan OV, et al. Transmission of hepatitis C virus to several organ and tissue recipients from an antibody-negative donor. Ann Intern Med. 2005;143(9):648–54. https://doi.org/10.7326/0003-4819-143-9-200511010-00008.
15. Petruccelli B, Chai F, Williams I, Perz J, Lee K, Harris M, et al. Outbreak of acute hepatitis C virus (HCV) infections of two different genotypes associated with an HCV-infected anesthetist. 2005.
16. Thompson ND, Novak RT, White-Comstock MB, Xia G-l, Ganova-Raeva L, Ramach S, et al. Patient-to-patient hepatitis C virus transmissions associated with infection control breaches in a hemodialysis unit. J Nephrol Therap. 2011;2013:1–5.
17. Fischer GE, Schaefer MK, Labus BJ, Sands L, Rowley P, Azzam IA, et al. 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. https://doi.org/10.1086/653937.
18. Moore ZS, Schaefer MK, Hoffmann KK, Thompson SC, Xia GL, Lin Y, et al. Transmission of hepatitis C virus during myocardial perfusion imaging in an outpatient clinic. Am J Cardiol. 2011;108(1):126–32. https://doi.org/10.1016/j.amjcard.2011.03.010.
19. Warner AE, Schaefer MK, Patel PR, Drobeniuc J, Xia G, Lin Y, et al. Outbreak of hepatitis C virus infection associated with narcotics diversion by an hepatitis C virus-infected surgical technician. Am J Infect Control. 2015;43(1):53–8. https://doi.org/10.1016/j.ajic.2014.09.012.
20. Ramachandran S, Campo DS, Dimitrova ZE, Xia GL, Purdy MA, Khudyakov YE. Temporal variations in the hepatitis C virus intrahost population during chronic infection. J Virol. 2011;85(13):6369–80. https://doi.org/10.1128/JVI.02204-10.
21. 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. In Silico Biol. 2011;11(5–6):163–73. https://doi.org/10.3233/ISB-2012-0451.

22. Bernard EJ, Azad Y, Vandamme AM, Weait M, Geretti AM. HIV forensics: pitfalls and acceptable standards in the use of phylogenetic analysis as evidence in criminal investigations of HIV transmission. HIV Med. 2007;8(6):382–7. https://doi.org/10.1111/j.1468-1293.2007.00486.x.
23. Scaduto DI, Brown JM, Haaland WC, Zwickl DJ, Hillis DM, Metzker ML. Source identification in two criminal cases using phylogenetic analysis of HIV-1 DNA sequences. Proc Natl Acad Sci USA. 2010;107(50):21242–7. https://doi.org/10.1073/pnas.1015673107.
24. Dimitrova Z, Campo DS, Ramachandran S, Vaughan G, Ganova-Raeva L, Lin Y, et al. Evaluation of viral heterogeneity using next-generation sequencing, end-point limiting-dilution and mass spectrometry. In Silico Biol. 2011;11(5–6):183–92. https://doi.org/10.3233/ISB-2012-0453.
25. Lara J, Teka M, Khudyakov Y. Identification of recent cases of hepatitis C virus infection using physical-chemical properties of hypervariable region 1 and a radial basis function neural network classifier. BMC Genomics. 2017;18(Suppl 10):880. https://doi.org/10.1186/s12864-017-4269-2.
26. Defendorf CM, Paul S, Scott GJ. Iatrogenic hepatitis C virus transmission and safe injection practices. J Am Osteopath Assoc. 2018;118(5):311–20. https://doi.org/10.7556/jaoa.2018.062.
27. Campo DS, Khudyakov Y. Intelligent network DisRuption analysis (INDRA): a targeted strategy for efficient interruption of hepatitis C transmissions. Infect Genet Evol. 2018;63:204–15. https://doi.org/10.1016/j.meegid.2018.05.028.
28. Little SJ, Kosakovsky Pond SL, Anderson CM, Young JA, Wertheim JO, Mehta SR, et al. Using HIV networks to inform real time prevention interventions. PLoS ONE. 2014;9(6): e98443. https://doi.org/10.1371/journal.pone.0098443.
29. Liu A, Glidden DV, Anderson PL, Amico KR, McMahan V, Mehrotra M, et al. Patterns and correlates of PrEP drug detection among MSM and transgender women in the Global iPrEx Study. J Acquir Immune Defic Syndr. 2014;67(5):528–37. https://doi.org/10.1097/QAI.0000000000000351.
30. Mehta SR, Schairer C, Little S. Ethical issues in HIV phylogenetics and molecular epidemiology. Curr Opin HIV AIDS. 2019;14(3):221–6. https://doi.org/10.1097/COH.0000000000000538.
31. Dawson L, Benbow N, Fletcher FE, Kassaye S, Killelea A, Latham SR, et al. Addressing ethical challenges in US-based HIV phylogenetic research. J Infect Dis. 2020. https://doi.org/10.1093/infdis/jiaa107.

Publisher's Note