



HHS Public Access

Author manuscript

J Infect Dis. Author manuscript; available in PMC 2016 November 22.

Published in final edited form as:

J Infect Dis. 2016 March 15; 213(6): 957–965. doi:10.1093/infdis/jiv542.

Accurate genetic detection of hepatitis C virus transmissions in outbreak settings

DS Campo^a, GL Xia^a, Z Dimitrova^a, Y Lin^a, JC Forbi^a, L Ganova-Raeva^a, L Punkova^a, S Ramachandran^a, H Thai^a, P Skums^a, S Sims^a, I Rytsareva^a, G Vaughan^a, HJ Roh^a, MA Purdy^a, A Sue^a, and Y Khudyakov^a

^aMolecular Epidemiology and Bioinformatics Laboratory, Division of Viral hepatitis, Centers for Disease Control and Prevention, Atlanta, 30333, USA

Abstract

Hepatitis C is a major public health problem in the United States and worldwide. Outbreaks of hepatitis C virus (HCV) infections are associated with unsafe injection practices, drug diversion, and other exposures to blood, being difficult to detect and investigate. Here, we developed and validated a simple approach for molecular detection of HCV transmissions in outbreak settings. We obtained sequences from the HCV hypervariable region 1 (HVR1) using End-Point Limiting-Dilution (EPLD) from 127 cases involved in 32 epidemiologically defined HCV outbreaks and 193 individuals with unrelated HCV strains. We compared several types of genetic distances and calculated a threshold using minimal Hamming distances that identifies transmission clusters in all tested outbreaks with 100% accuracy. The approach was also validated on sequences from 239 individuals obtained using next-generation sequencing, showing the same accuracy as EPLD. In average, nucleotide diversity of the intra-host population was 6.2-times greater in the source than in any incident case, allowing the correct detection of transmission direction in 8 outbreaks for which source cases were known. A simple and accurate distance-based approach for detecting HCV transmissions developed here streamlines molecular investigation of outbreaks, thus improving the public health capacity for rapid and effective control of hepatitis C.

Keywords

HCV; Outbreak; Threshold; NGS; Nucleotide diversity; Phylogenetic analysis; Hamming distance; Transmission networks

Introduction

Hepatitis C virus (HCV) infects nearly 3% of the world's population and is a major cause of liver disease worldwide [1]. HCV infection is an important US public health problem because it is the most common chronic blood-borne infection and the leading cause for liver transplantation [2]. Since 2007, HCV surpasses HIV as a cause of death in the US [3]. It is estimated that 2.7–3.9 million people in the US have chronic HCV infection and that

Correspondence and requests for reprints should be addressed to DS Campo; Molecular Epidemiology and Bioinformatics Laboratory, Division of Viral Hepatitis, Centers for Disease Control and Prevention; fyv6@cdc.gov.

>15,000 die each year from HCV-related disease, with mortality expected to rise in the coming years [4]. Approximately 80% of patients who become infected with HCV develop chronic infections and are at risk for advanced liver disease; 15%–30% of these patients progress to liver fibrosis and cirrhosis and up to 5% die from liver failure due to cirrhosis or hepatocellular carcinoma[2].

Outbreaks of HCV infections are associated with unsafe injection practices, drug diversion, and other exposures to blood and blood products. During 2005–2013 in the US, a total of 18 healthcare-associated outbreaks were detected, involving 228 outbreak-associated cases and >92,550 at-risk persons notified for screening. Of these, 9 outbreaks occurred in outpatient facilities, 7 occurred in hemodialysis settings, and 2 were caused by HCV-infected health care providers who were diverting drugs [5]. Considering a long incubation period of up to 6 months and an asymptomatic course of HCV infections in >70% of infected persons, the detected hepatitis C outbreaks most likely account for a fraction of outbreaks of HCV infections, resulting in the under-identification of recent transmissions and new cases of infections.

Several molecular approaches have been developed for tracking viral infections [6–15]. Over the last decade, our laboratory investigated numerous outbreaks of viral hepatitis in the US using molecular analysis (Supporting Information) [16–22]. Molecular detection of viral transmissions is usually aided by phylogenetic analysis, for which a small viral genomic region is amplified, sequenced directly and a single sequence per case is used to construct a phylogenetic tree. The identification of transmission clusters from phylogenetic trees is achieved using several criteria. For example, a phylogenetic cluster of sequences can be interpreted as representing a single viral strain shared by cases involved in an outbreak if: (i) genetic or patristic distances among sequences from the cluster are below a certain threshold, and (ii) the ancestral node for the suspected transmission cluster in the tree has a high statistical significance calculated using the Bayesian statistics or bootstrap analysis. If available, data on sharing known association factors such as a close geographical location and high-risk behavior among members of the cluster can be used to evaluate the phylogenetic inferences [14]. Such approaches are commonly used in HIV forensics [23].

Although a single consensus viral sequence from each infected case is usually used for the molecular detection of outbreaks [14], the use of such sequences is not satisfactory for the accurate identification of viral strains because viruses, especially RNA viruses such as HCV, exist as a heterogeneous population of closely related but genetically distinct variants, known as quasispecies, in each infected person [24]. Such a population cannot be adequately represented with a single sequence. Moreover, considering that HCV infections are frequently established by minority variants transmitted from the source [20, 25], a consensus sequence cannot reliably capture such transmissions.

The limitations of consensus sequencing were solved using sequencing of intra-host viral variants isolated by End-Point Limiting-Dilution (EPLD)[26] and, recently, using next-generation sequencing (NGS)[27, 28]. A large sample of intra-host viral variants obtained using NGS represents more adequately intra-host viral subpopulations and their minority variants, thus, improving accuracy of genetic detection of transmissions [27, 29, 30].

However, sampling thousands of intra-host HCV variants complicates phylogenetic analysis, straining computational resources, being time-consuming and difficult to interpret. Given the need for a rapid response during HCV outbreaks, a simple and accurate approach to the detection of HCV transmissions by NGS data would be very useful.

Here, we have developed and validated a methodology for the molecular detection of HCV outbreaks. A simple, accurate and fast distance-based method for the detection of cases linked by transmission was devised using sequence data obtained from cases identified during epidemiologically curated HCV outbreaks. In addition, we found that in 8 outbreaks with known HCV infection source, this source can be accurately identified using the population-level nucleotide diversity.

Methods

Related samples

Sequences of hypervariable region 1 (HVR1) were obtained from 127 cases collected during 32 epidemiologically defined outbreaks (see Supporting information for a complete list, and [16–22]). HCV from 76 cases belonged to genotype 1a and from 51 cases to genotype 1b. All outbreak-associated HCV infections were serologically confirmed and epidemiologically defined as described in [5]. Outbreak-associated HCV infections are defined as those with serological, clinical and epidemiologic evidences supporting transmission and include cases identified with acute infection, or previously undiagnosed chronic infections with epidemiologic evidence indicating that these were likely outbreak-related incident cases that progressed from the acute to chronic stage [5]. The source case of HCV infections was known for 8 outbreaks [16–22]. The HCV strain identified in one of the outbreaks (AW) [22] was additionally sequenced from all 18 cases using NGS. For more details of each outbreak and sequence access, the authors of each cited article should be contacted.

Unrelated samples

Two sets of HCV sequences obtained from epidemiologically unrelated individuals were used in the study. Sequences from 193 HCV infected individuals were obtained by EPLD-PCR using serum specimens from national collections [31, 32] and other surveillance projects conducted in the laboratory. HCV from 118 individuals belonged to subtype 1a and from 75 to subtype 1b. A second set of sequences was obtained by NGS from 221 HCV infected patients, with 81 being infected with HCV subtype 1a, 41 with subtype 1b, 58 with genotype 2, 39 with genotype 3, and 2 with genotype 4. Some of these samples were described in [28] and gathered from other surveillance projects conducted in the laboratory.

End-Point Limiting-Dilution Real-Time PCR

Total nucleic acids from the specimens were extracted from serum, then RNA was precipitated and reverse-transcribed using both random and specific primers as previously described [26]. We used the EPLD-PCR protocol for sequencing multiple clones of HVR1, as previously described [26, 33].

Next Generation Sequencing (NGS)

PCR products were pooled and subjected to pyrosequencing using Junior GS or GS FLX Titanium Sequencing Kit (454 Life Sciences, Roche, Branford, CT). The NGS files were processed using the error correction algorithms KEC and ET[34].

Data analysis

For each sample of HCV sequences, a Multiple Sequence Alignment (MSA) was created using MAFFT 7.221 [35]. The primer sequences were removed and the final sequences were 264 nucleotides in length. The level of genetic heterogeneity of each sample was estimated by the Nucleotide diversity (π), following [36, 37] using MATLAB R2014a [38]. The sequences of every pair of samples were aligned and used to calculate genetic distances with MATLAB [38]. A Maximum likelihood tree using the Nei-Tamura substitution model was constructed using MEGA 6 [39]. Patristic distances were calculated over the phylogenetic tree using MATLAB[38]. For each distance type, we measured the overlap between the distributions of values among related or unrelated samples using Bhattacharyya Coefficient (B_c), which is equal to 0 when two distributions do not overlap and equal to 1 when they completely overlap[40].

Transmission network

A transmission network that represents the genetic relatedness between samples was built using MATLAB[38] and drawn with GEPHI[41].

k-step network

For one of the outbreaks (AW), we built a k-step network as we previously described in [28]. The k-step network contains all possible Minimum Spanning Trees and allows to efficiently visualize the genetic relatedness among all haplotypes present in the sample. The networks were drawn with GEPHI[41].

Results

Intra-host HCV populations

A total of 12987 HCV clones obtained using EPLD-PCR were analysed in this study. These HCV clones were obtained from 320 HCV-infected cases. On average, each sample contains 40.58 clones, with 21.52 of them being different (henceforth, HCV variants), a major variant representing 34% of all clones and hamming distance of 2.58%.

Genetic distances among samples of HCV sequences

There are 374 pairwise comparisons among samples belonging to the same transmission cluster from 32 epidemiologically confirmed outbreaks, of which 78 (20.86%) comparisons are between identical HCV variants and 73 (19.52%) between HCV variants that differ at a single nucleotide position. Although the sharing of identical variants among samples is a strong evidence of direct transmission, not all epidemiologically linked samples share HCV variants, indicating a need in a less stringent threshold to accurately define transmission clusters. In order to establish this threshold, we studied the distribution of pairwise distances

among samples belonging to the same outbreak (related) and samples without any epidemiological linkage (unrelated). Three forms of genetic distances were considered: (1) Hamming distance, which is the number of mismatches; (2) Nei-Tamura distance, which takes into account transitions and transversions; and (3) patristic distance, which is calculated over the branches of the maximum likelihood phylogenetic tree. Given that each sample includes a population of variants, three different measures were used for each distance: (1) minimal distance among all possible pairs of variants between two samples; (2) average distance; and (3) distance between major variants from each sample. The last measure was included to simulate the situation where a single sequence per sample was obtained, as is often the case in several molecular epidemiology studies.

For each distance type, we measured the overlap between the distributions of values among related or unrelated samples (Fig. 1) using the Bhattacharyya Coefficient (Bc), which is equal to 0 when two distributions do not overlap and equal to 1 when they completely overlap. Among the distances, the minimal distance was superior, which is in agreement with the observation that minority variants are frequently responsible for transmission [20, 25]. The overlap between the distributions was ~2.8 times greater than using the minimal distance. Overall, the minimal Hamming and minimal patristic distances were equal in performance, both showing the same overlap values (Bc = 0.043) and being highly correlated ($r = 0.992$).

A relatedness threshold

Considering performance and simplicity, the minimal Hamming distance was selected to find a threshold for the most accurate separation of the related and unrelated samples (Table 1). Fig. 2 shows that the distributions of minimal Hamming distances among related and unrelated samples are very distinct. The threshold of 3.77% was calculated as the average minimum distance among the related samples (1.02%) plus 2 standard deviations (1.37%). This threshold is lower than any of the 18528 unrelated pairwise distances, and greater than all but only 8 (2.14%) of the related pairwise distances. Although some distances between related cases are greater than this threshold, all 127 outbreak samples were found linked to 1 case from the corresponding transmission clusters. Thus, the relatedness threshold yields 100% sensitivity and 100% specificity in the detection of cases involved in these outbreaks.

Application of the threshold to NGS data

The threshold was established using the EPLD-PCR generated data. However, NGS provides an opportunity to sample many more intra-host viral variants, which may affect the threshold accuracy. In order to assess applicability of the threshold to NGS data, we studied a second independent set of HCV HVR1 sequences obtained by NGS from 221 unrelated cases. Among all possible 24310 pairwise comparisons, not a single distance was below the established relatedness threshold, indicating a strong specificity of the threshold applied to the NGS data. In addition, the NGS data were obtained for 18 cases involved in the AW outbreak (See Figs. 3A and 3B). Analysis of the data identified the same transmission cluster, but the application of NGS resulted in ~4-fold increase in the number of links among related cases sharing identical sequences as compared to EPLD-PCR (60.13% vs. 14.38%). This increase indicates that a greater sampling of intra-host HCV variants improves

sensitivity of the transmission detection by increasing the probability of finding more closely related or even identical HCV variants from outbreak-associated cases using the relatedness threshold.

Transmission graphs

A usual representation of the relatedness among sequences is a phylogenetic tree, which often requires considerable expertise for its construction and interpretation (Fig. 4A). However, when the size of the data increases, the tree becomes unwieldy and may obscure a visual identification of transmission clusters. Here the calculated threshold allows for a graphical representation that captures the public health information in the more intuitive form of transmission networks. Fig. 4B shows the HCV HVR1 genetic relatedness data obtained from the AW outbreak as a network, taking the same dataset used to generate the phylogenetic tree shown in Fig. 4A. In this network, each node is an HCV sample; a link is drawn if the minimal Hamming distance between the two samples is smaller than the relatedness threshold (3.77%) and the size of the node is proportional to the nucleotide diversity of the intra-host viral population.

Source identification

The epidemiologically identified source of HCV infections was known for 8 outbreaks, which were studied further in detail. Intra-host HCV populations sampled from the known source were found to be the most genetically heterogeneous among all cases involved in the corresponding transmission cluster. The level of genetic heterogeneity of each sample was measured by calculating the nucleotide diversity, as shown in Fig. 5 for the AW outbreak. In average over all 8 outbreaks, the source HCV population has 6.2-times greater nucleotide diversity than its associated incident cases and 4.3-times greater than the incident case with the highest nucleotide diversity (Fig. 6).

Discussion

The presented approach identified accurately all transmission clusters in the tested outbreaks and separated these clusters from epidemiologically unrelated HCV strains and did not link any unrelated case to each other by transmission. The genetic linkage identified using the threshold approach developed here cannot be unambiguously interpreted as a direct transmission event without epidemiological data supporting such direct transmissions. Rather, our approach detects if two patients share same HCV strain and it should be interpreted as sampling from members of transmission chains or networks. Outbreaks are usually identified as a cluster of cases in certain epidemiological settings. Thus, identification of strain sharing among cases involved in such a cluster is a strong indication of linkage by transmission. However, this approach should be used with caution in investigation of transmissions that have occurred in a distant past because of the uncertainty of HCV evolution in a succession of hosts or over a long time.

Availability of extensive genetic data from outbreaks, as reported here, is unusual. However, certain information on genetic strain identity can be obtained from analyzing intra-host viral heterogeneity [42, 43]. The essential limitation of this approach, however, is focus on intra-

host rather than inter-host viral evolution. Although both play an important role in defining viral genetic identity, transmission is largely an inter-host process, and genetic heterogeneity associated with the transmitted viral strain cannot be accurately measured without consideration of intra- as well as inter-host genetic changes. The approach implemented here takes into consideration both, measuring minimal genetic distances among viral populations of patients infected from a common source. The average Hamming distance among intra-host variants measured in our samples is 2.58%, which is lower than the relatedness threshold (3.77%) identified here. Application of the average value as a threshold would allow for separating all unrelated cases but also would lead to misclassification of 7.48% of related cases, thus reducing sensitivity of transmission detection.

Although simple and efficient in outbreak settings, this threshold approach is very specific to the virus and genomic region used for the detection of transmissions, and needs to be experimentally established for each pathogen or genomic region. A generalized approach that could have been readily applicable to different pathogens and genomic regions of the same pathogen would have significant advantages in detection of transmissions over the targeted threshold approach. The development of such approaches exploiting different clustering techniques promises a more universal detection of transmissions [43, 44].

Simplicity of identification of a single sequence per specimen prompted application of consensus sequences to the detection of transmissions by phylogenetic analysis [14]. However, this study shows that genetic distances among HCV strains using a single sequence result in less accurate separation of related and unrelated cases. Moreover, consensus sequences obtained by direct sequencing are rarely identical to the major variant and frequently different from any sequence variant sampled from a specimen [29, 33], indicating that consensus sequences result from amalgamation of a heterogeneous viral population and should be used for the detection of transmission with caution.

The accuracy of the detection depends significantly on sampling of the sufficient number of intra-host viral variants in order to capture minority populations, which can be achieved with NGS [27, 28, 45]. In general, it can be expected that increases in the number of variants sampled from intra-host viral populations should not change significantly minimal Hamming distances among HCV variants from epidemiologically unrelated cases, whereas minimal genetic distances among epidemiologically related HCV variants may become shorter due to the greater probability of sampling minority variants. Indeed, as we observed in this study, the threshold approach developed using the EPLD-PCR data performed equally well on NGS data obtained from unrelated HCV cases but NGS data showed an improved sensitivity in the detection of transmission links among members of the transmission cluster in the AW outbreak.

Several types of distances were studied here, with the minimal Hamming distance being identified as one of the most accurate for the detection of transmissions and convenient to calculate. Although identical in accuracy of transmission detection to the patristic distance, as shown here, Hamming distance has an advantage in its simplicity. Hamming distance is computationally less intensive than the others and can be calculated rapidly even for large NGS datasets. Minimal distances are most suitable for application to the detection of

transmissions using threshold since recently transmitted viral populations are generally in genetic proximity to the source, even in the case of transmission of minority variants [20].

With the predominance of phylogenetic analyses for identifying transmission clusters, the usual graphical representation of a transmission cluster uses a phylogenetic tree. However, a network representation is more suited for the threshold analysis developed here, showing potential transmissions among cases directly and streamlining interpretation of results. The graph is simple and can be easily modified if additional cases are added without the need to recalculate distances among all previously studied cases, considerably reducing computational time. This is in contrast with phylogenetic reconstructions, which need to be recalculated after adding new sequences, a significant burden in the current NGS period. Further, such graphs can be constructed using any type of validated genetic distances; e.g., distances among mass-spectrometric profiles as was shown earlier in our laboratory [46]. Finally, the current approach allows the implementation of very efficient computational algorithms to remove patient-pairs that can't have sequences with a distance below the threshold, reducing considerably demand on computational resources [47].

Identification of the source of infection is crucial for the interruption and prevention of outbreaks. In general, HCV accumulates mutations during intra-host evolution and becomes more genetically heterogeneous [32, 48]. Thus, the difference in duration of infection between the source and incident cases can be explored for the detection of the transmission direction. Indeed, analysis of HCV cases from 8 outbreaks with known sources of infection showed that the source is infected with a much more diverse HCV population than any incident case from the corresponding transmission cluster. This finding is supported with our earlier observation that the intra-host HVR1 nucleotide diversity is 1.8 times greater in patients with chronic than acute HCV infection {Astrakhantseva, 2011 #2262[49]}. However, the difference in the genetic diversity allows for the accurate identification of the transmission direction only when the source was sampled. Otherwise, the incident case with the most heterogeneous HVR1 population may be classified as a source of infection in a transmission cluster. One possible way to resolve this issue is to establish a threshold but its definition requires a greater number of outbreaks with known source. These problems of transmission-direction detection have been noted earlier [23, 50] and warrant further investigation.

This simple and accurate distance-based approach for detecting HCV transmissions developed here streamlines molecular investigation of outbreaks, thus improving the public health capacity for a rapid and effective control of hepatitis C. Currently, the approach is one of the tools of the Global Hepatitis Outbreak and Surveillance Technology (GHOST), which enables molecular outbreak investigation by an automated analysis of HCV sequences and graphical presentation of results (to be described elsewhere in detail).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors are deeply indebted to all researchers who contributed over the last decade to outbreak investigations referenced here, especially the members of the Epidemiology & Surveillance branch, (DVH, CDC). The authors are also very grateful to Chris Lynberg (ITSO R&D, CDC), Max Mirabito, Silver Wang, Robin Tracy and Thom Sukalac (NCHHSTP IO, CDC) for their constant computational help.¹

The work was partially supported by the Advanced Molecular Detection program (Office of Infectious Diseases, Centers for Disease Control and Prevention).

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; 57:1333–42. [PubMed: 23172780]
2. Alter M. Epidemiology of hepatitis C virus infection. *World J Gastroenterol*. 2007; 13:2436–41. [PubMed: 17552026]
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. *Annals of internal medicine*. 2012; 156:271–8. [PubMed: 22351712]
4. Ward JW. The hidden epidemic of hepatitis C virus infection in the United States: occult transmission and burden of disease. *Topics in antiviral medicine*. 2013; 21:15–9. [PubMed: 23596274]
5. DVH. Healthcare-Associated Hepatitis B and C Outbreaks Reported to CDC in 2008–2013. Available at: <http://www.cdc.gov/hepatitis/Outbreaks/HealthcareHepOutbreakTable.htm>
6. Spada E, Abbate I, Sicurezza E, et al. Molecular epidemiology of a hepatitis C virus outbreak in a hemodialysis unit in Italy. *J Med Virol*. 2008; 80:261–7. [PubMed: 18098132]
7. Spada E, Saggiocca L, Sourdis J, et al. Use of the minimum spanning tree model for molecular epidemiological investigation of a nosocomial outbreak of hepatitis C virus infection. *J Clin Microbiol*. 2004; 42:4230–6. [PubMed: 15365016]
8. Ou CY, Ciesielski CA, Myers G, et al. Molecular epidemiology of HIV transmission in a dental practice. *Science*. 1992; 256:1165–71. [PubMed: 1589796]
9. Esteban JI, Gomez J, Martell M, et al. Transmission of hepatitis C virus by a cardiac surgeon. *N Engl J Med*. 1996; 334:555–60. [PubMed: 8569822]
10. Birch CJ, McCaw RF, Bulach DM, et al. Molecular analysis of human immunodeficiency virus strains associated with a case of criminal transmission of the virus. *J Infect Dis*. 2000; 182:941–4. [PubMed: 10950794]
11. Metzker ML, Mindell DP, Liu XM, Ptak RG, Gibbs RA, Hillis DM. Molecular evidence of HIV-1 transmission in a criminal case. *Proc Natl Acad Sci U S A*. 2002; 99:14292–7. [PubMed: 12388776]
12. Bracho MA, Gosalbes MJ, Blasco D, Moya A, Gonzalez-Candelas F. Molecular epidemiology of a hepatitis C virus outbreak in a hemodialysis unit. *J Clin Microbiol*. 2005; 43:2750–5. [PubMed: 15956393]
13. Gonzalez-Candelas F, Bracho MA, Wrobel B, Moya A. Molecular evolution in court: analysis of a large hepatitis C virus outbreak from an evolving source. *BMC biology*. 2013; 11:76. [PubMed: 23870105]
14. Prosperi MC, De Luca A, Di Giambenedetto S, et al. The threshold bootstrap clustering: a new approach to find families or transmission clusters within molecular quasispecies. *PLoS One*. 2010; 5:e13619. [PubMed: 21049051]
15. Feray C, Bouscaillou J, Falissard B, et al. A novel method to identify routes of hepatitis C virus transmission. *PLoS One*. 2014; 9:e86098. [PubMed: 24465895]

¹The authors state they do not have any conflict of interest.

16. Noviello, S.; Smith, P.; Chai, F., et al. Hepatitis C Virus Transmission by a Cardiac Surgeon in the United States. 2015.
17. Chai, F.; Xia, G.; Williams, I., et al. Transmission of Hepatitis C Virus at a Pain Remediation Clinic - San Diego, California 2003. 43rd Annual Meeting of the Infectious Diseases Society of America (IDSA); San Francisco, California.
18. Lee, K.; Scoville, S.; Taylor, R., et al. Outbreak of Acute Hepatitis C Virus (HCV) Infections of Two Different Genotypes Associated with an HCV-Infected Anesthetist. 47th Annual Meeting of the Infectious Diseases Society of America (IDSA); Philadelphia, Pennsylvania.
19. Thompson N, Novak R, White-Comstock M, et al. Patient-to-Patient Hepatitis C Virus Transmissions Associated with Infection Control Breaches in a Hemodialysis Unit. *J Nephrol Therapeutics*. 2012; S10:002.
20. Fischer GE, Schaefer MK, Labus BJ, et al. Hepatitis C virus infections from unsafe injection practices at an endoscopy clinic in Las Vegas, Nevada, 2007–2008. *Clin Infect Dis*. 2010; 51:267–73. [PubMed: 20575663]
21. Moore ZS, Schaefer MK, Hoffmann KK, et al. Transmission of hepatitis C virus during myocardial perfusion imaging in an outpatient clinic. *The American journal of cardiology*. 2011; 108:126–32. [PubMed: 21529725]
22. Warner AE, Schaefer MK, Patel PR, et al. Outbreak of hepatitis C virus infection associated with narcotics diversion by an hepatitis C virus-infected surgical technician. *American journal of infection control*. 2015; 43:53–8. [PubMed: 25442395]
23. 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 medicine*. 2007; 8:382–7. [PubMed: 17661846]
24. Domingo E, Sheldon J, Perales C. Viral quasispecies evolution. *Microbiol Mol Biol Rev*. 2012; 76:159–216. [PubMed: 22688811]
25. Apostolou A, Bartholomew ML, Greeley R, et al. Transmission of hepatitis C virus associated with surgical procedures - new jersey 2010 and wisconsin 2011. *MMWR Morbidity and mortality weekly report*. 2015; 64:165–70. [PubMed: 25719676]
26. Ramachandran S, Xia GL, Ganova-Raeva LM, Nainan OV, Khudyakov Y. End-point limiting-dilution real-time PCR assay for evaluation of hepatitis C virus quasispecies in serum: performance under optimal and suboptimal conditions. *J Virol Methods*. 2008; 151:217–24. [PubMed: 18571738]
27. Dimitrova Z, Campo DS, Ramachandran S, et al. Evaluation of viral heterogeneity using next-generation sequencing, end-point limiting-dilution and mass spectrometry. *In silico biology*. 2011; 11:183–92. [PubMed: 23202420]
28. Campo DS, Dimitrova Z, Yamasaki L, et al. Next-generation sequencing reveals large connected networks of intra-host HCV variants. *BMC genomics*. 2014; 15(Suppl 5):S4.
29. Forbi JC, Campo DS, Purdy MA, et al. Intra-host diversity and evolution of hepatitis C virus endemic to Cote d'Ivoire. *J Med Virol*. 2014; 86:765–71. [PubMed: 24519518]
30. Forbi JC, Purdy MA, Campo DS, et al. Epidemic history of hepatitis C virus infection in two remote communities in Nigeria, West Africa. *J Gen Virol*. 2012; 93:1410–21. [PubMed: 22456613]
31. Williams I. Epidemiology of hepatitis C in the United States. *The American journal of medicine*. 1999; 107:2S–9S.
32. 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 biology*. 2011; 11:163–73. [PubMed: 23202418]
33. Ramachandran S, Zhai X, Thai H, et al. Evaluation of intra-host variants of the entire hepatitis B virus genome. *PLoS One*. 2011; 6:e25232. [PubMed: 21949887]
34. Skums P, Dimitrova Z, Campo DS, et al. Efficient error correction for next-generation sequencing of viral amplicons. *BMC Bioinformatics*. 2012; 13(Suppl 10):S6.
35. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol*. 2013; 30:772–80. [PubMed: 23329690]

36. Schneider, S.; Roessli, D.; Excoffier, L. ARLEQUIN version 2000: Software for population genetic data analysis. Geneva: Genetics and Biometry Laboratory, University of Geneva; 2000. p. 120
37. Tajima F. Evolutionary relationship of DNA sequences in finite populations. *Genetics*. 1983; 105:437–60. [PubMed: 6628982]
38. Mathworks. Matlab; Natick, MA: 2010.
39. Kumar STK, Nei M. MEGA3: Integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment. *Brief Bioinformatic*. 2004; 5:150–63.
40. Comaniciu, D.; Ramesh, V.; Meer, P. Real-Time Tracking of Non-Rigid Objects using Mean Shift. *IEEE Conference in Computer Vision and Pattern Recognition (CVPR'00)*; Hilton Head Island, SC. 2000. p. 142-9.
41. Bastian, M.; Heymann, S.; Jacomy, M. Gephi: an open source software for exploring and manipulating networks. *International AAAI Conference on Weblogs and Social Media*; San Jose, CA, USA.
42. Olmstead, A.; Montoya, V.; Joy, J., et al. Characterizing hepatitis C virus transmission dynamics using molecular phylogenetic based methods. 2015.
43. Poon AF, Joy JB, Woods CK, et al. The Impact of Clinical, Demographic and Risk Factors on Rates of HIV Transmission: A Population-based Phylogenetic Analysis in British Columbia, Canada. *J Infect Dis*. 2015; 211:926–35. [PubMed: 25312037]
44. Skums, P.; Artyomenko, A.; Glebova, O., et al. Detection of genetic relatedness between viral samples using EM-based clustering of next-generation sequencing data. *Workshop on Computational Advances in Molecular Epidemiology (CAME2014) of the IEEE 4th International Conference on Computational Advances in Bio and Medical Sciences (ICCABS)*; Miami, FL. June 2–4;
45. Wang G, Sherrill-Mix S, Chang K, Quince C, Bushman F. Hepatitis C virus transmission bottlenecks analyzed by deep sequencing. *J Virol*. 2010; 84:6218–28. [PubMed: 20375170]
46. Ganova-Raeva LM, Dimitrova ZE, Campo DS, Khudyakov Y. Application of mass spectrometry to molecular surveillance of hepatitis B and C viral infections. *Antivir Ther*. 2012; 17:1477–82. [PubMed: 23322623]
47. Rytsareva, I.; Campo, D.; Zheng, Y., et al. Efficient Detection of Viral Transmission with Threshold-based methods. *5th IEEE International Conference on Computational Advances in Bio and Medical Sciences (ICCABS)*; Miami, FL, USA.
48. 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; 85:6369–80. [PubMed: 21525348]
49. Campo DS, Dimitrova Z, Yokosawa J, et al. Hepatitis C virus antigenic convergence. *Sci Rep*. 2012:267–77. [PubMed: 22355779]
50. 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 U S A*. 2010; 107:21242–7. [PubMed: 21078965]

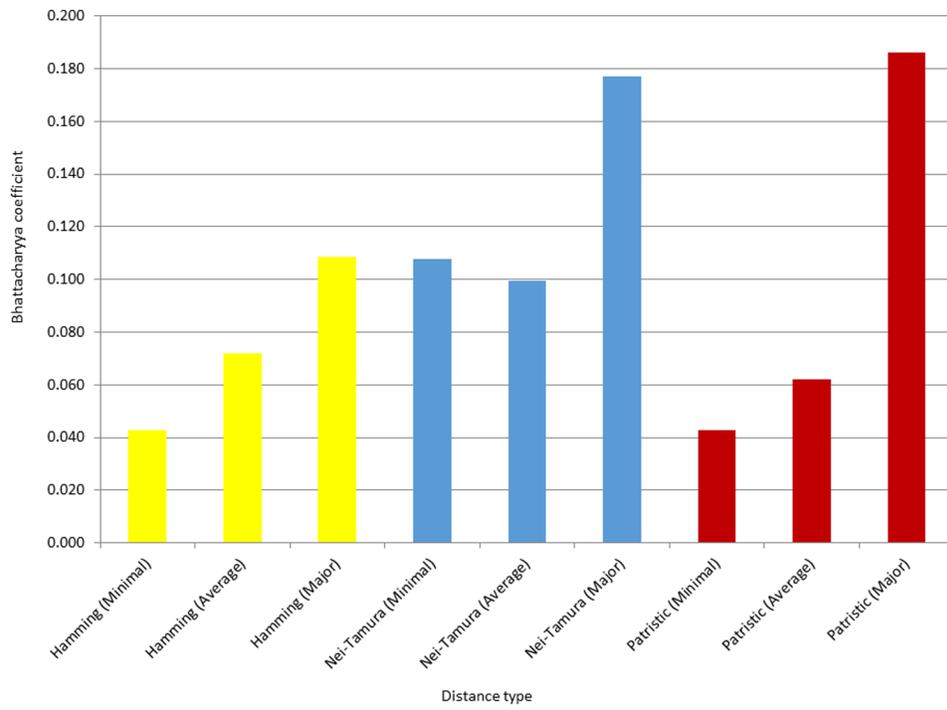


Figure 1. Overlap between the distributions of minimal distances among related or unrelated samples.

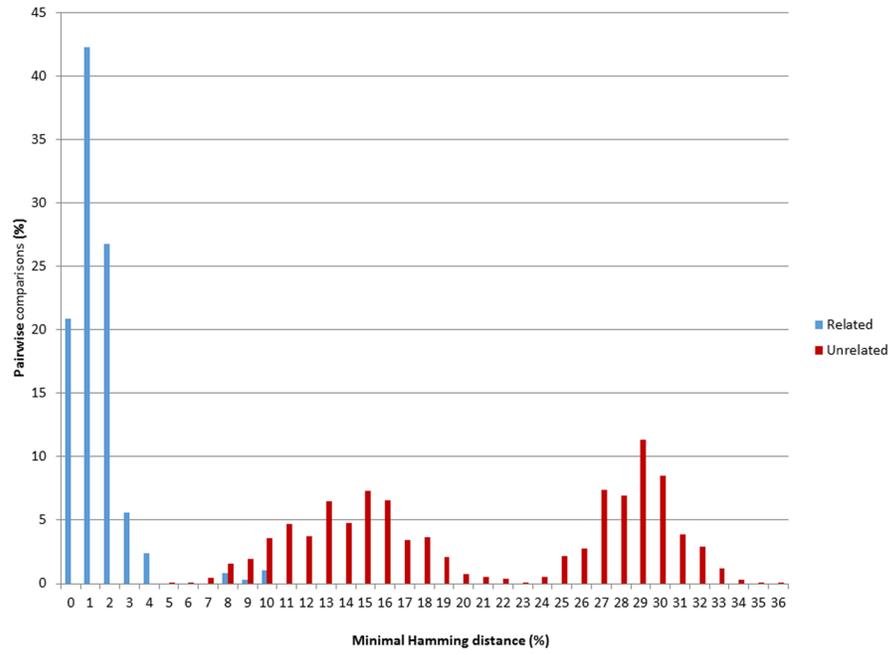


Figure 2. Distribution of pairwise distances. Percentage of pairwise comparisons for each category of Minimal Hamming distance.

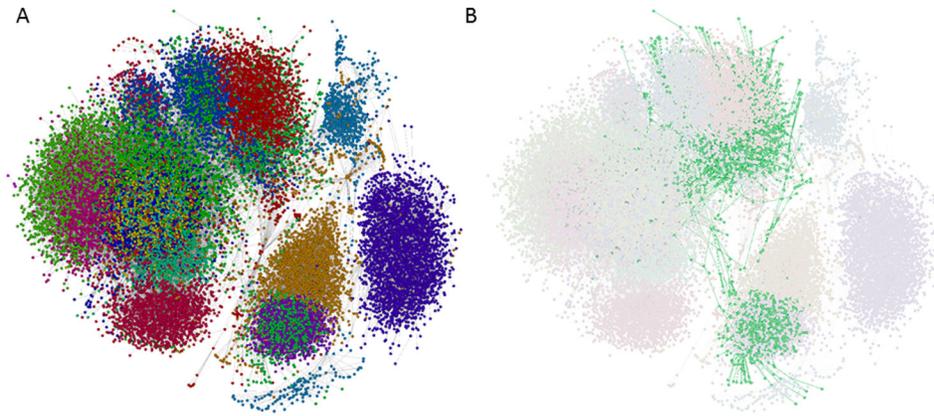


Figure 3.

A) k-step network of the HCV variants sampled during a single outbreak (AW). The k-step network contains all possible Minimum Spanning Trees and allows to efficiently visualize the genetic relatedness among all variants. Each node is an HCV variant and different cases are shown in different colors. B) The same k-step network but highlighting in green the HCV variants present in the source.

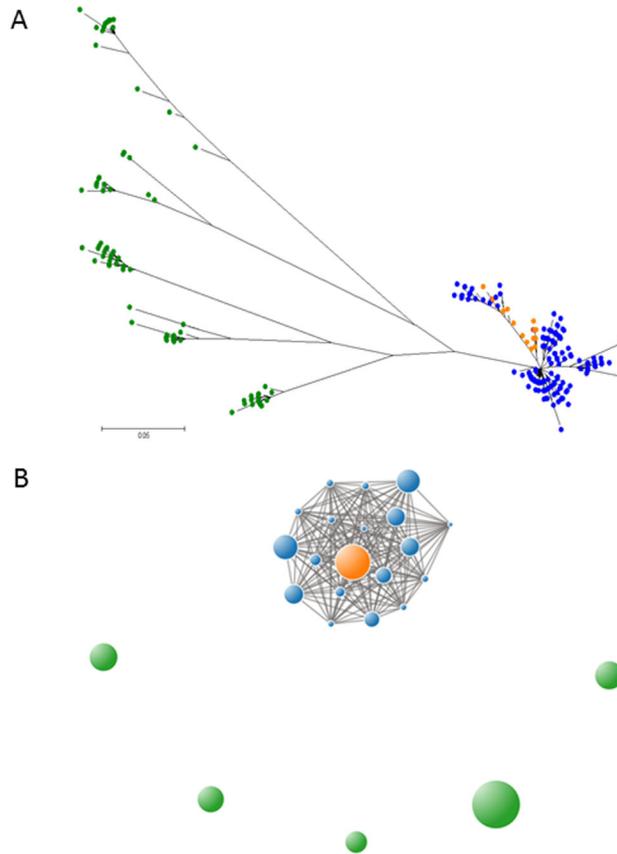


Figure 4. A) Phylogenetic tree of HCV variants sampled during a single outbreak (AW) [46]. Each leaf is an HCV variant. Green nodes represent sequences obtained from 5 unrelated cases, blue from 18 incident cases, and orange from the known source of the outbreak, a drug-diverting, HCV-infected surgical technician. B) Transmission network of the same outbreak (AW). Each node is an HCV sample, a link is drawn if the minimal Hamming distance between the two samples is smaller than the relatedness threshold (3.77%) and the size of the node is proportional to the sample nucleotide diversity. Green nodes represent unrelated cases, blue represents related cases and orange represents the known source of the outbreak.

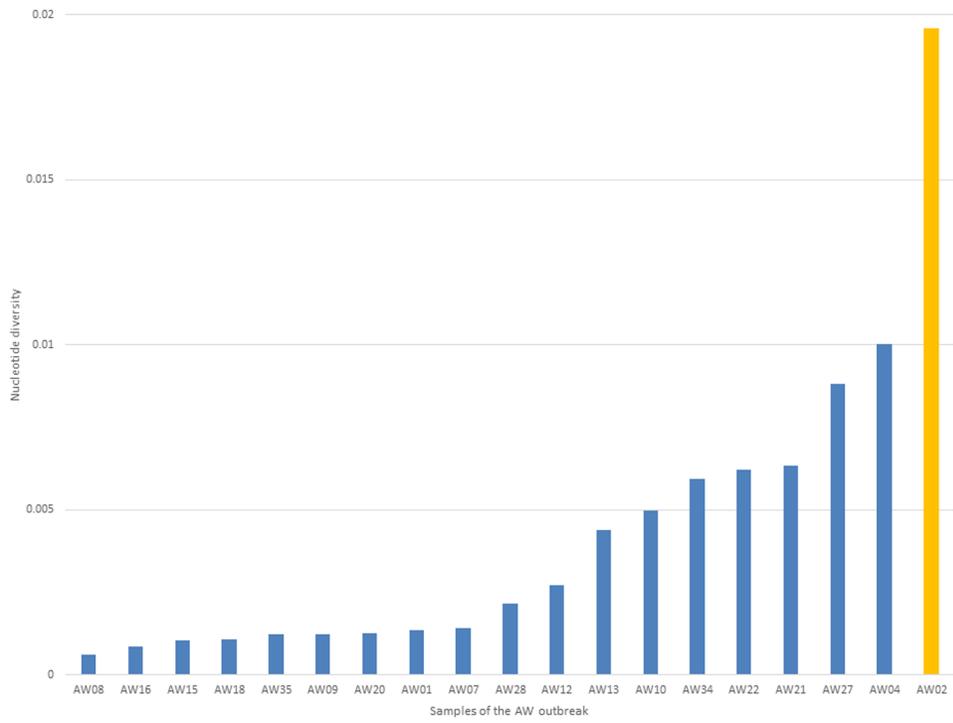


Figure 5. HCV nucleotide diversity of each sample from the AW outbreak. The source of the outbreak (a drug-diverting, HCV-infected surgical technician) is shown in orange.

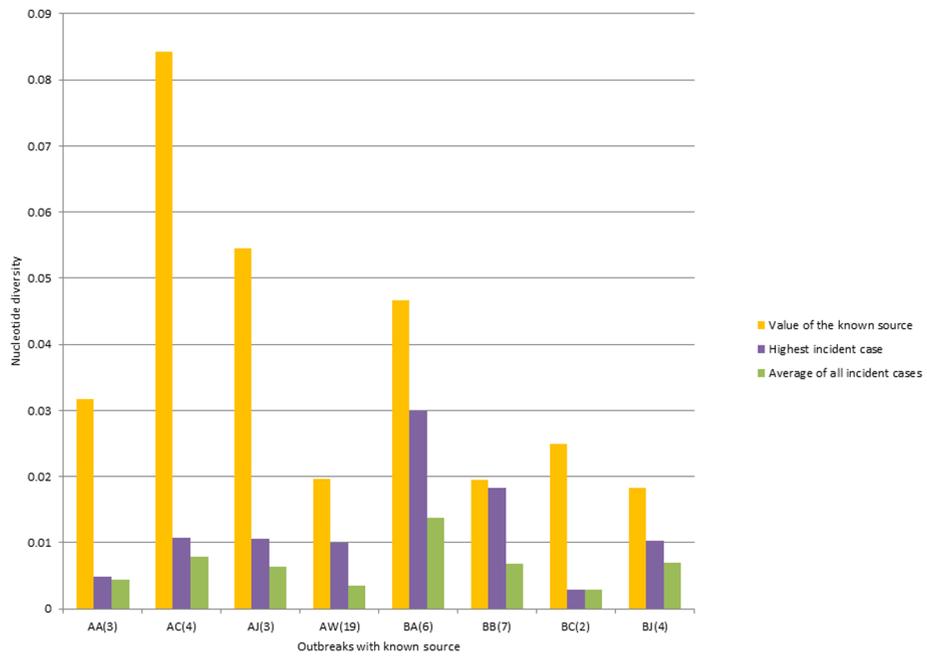


Figure 6. Nucleotide diversity of HCV HVR1 from the source and incident cases. The diversity of HCV HVR1 from the known source in each outbreak is shown in orange. The highest value of HVR1 nucleotide diversity found among incident cases is shown in purple. The average HVR1 nucleotide diversity from incident cases is shown in green. The total number of cases in each outbreak is shown in parenthesis.

Table 1

Descriptive statistics of the studied samples.

	1a	1b	All
n samples	194	126	320
n outbreaks	20	12	32
n related samples	76	51	127
n unrelated samples	118	75	193
n pairwise related	165	209	374
n pairwise unrelated	6903	2775	18528
Average of related distances	0.87	1.14	1.02
S. D. of related distances	1.21	1.48	1.37
Average of unrelated distances	12.97	15.04	20.60
Ratio of mean unrelated/related	14.91	13.19	20.20
n pairwise related > threshold	2	6	8
Pairwise unrelated < threshold	0	0	0
n of missed outbreak samples	0	0	0

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript