Infection, Genetics and Evolution 58 (2018) 77–82

Contents lists available at ScienceDirect

Infection, Genetics and Evolution

journal homepage: www.elsevier.com/locate/meegid

Research paper

Complex patterns of Hepatitis-C virus longitudinal clustering in a high-risk population

Rebecca Rose[a], Susanna L. Lamers[a], Guido Massaccesi[b], William Osburn[b], Stuart C. Ray[b], David L. Thomas[b], Andrea L. Cox[b], Oliver Laeyendecker[b,c,*]

[a] BioInfoExperts LLC, Thibodaux, LA, United States
[b] Department of Medicine, Johns Hopkins University, Baltimore, MD, United States
[c] NIAID, NIH, Baltimore, MD, United States

ARTICLE INFO

Keywords:
Linkage
Clustering
Epidemiology
Phylogenetic
HCV

ABSTRACT

We investigated longitudinal viral clustering among and within subjects in a highly networked cohort of people who inject drugs (PWID). All subjects had estimated dates of infection and two or more E1 sequences (bp 943–1288 relative to H77) with 1 to 14 years of follow up. Two methods (HIV-TRACE and PhyloPart) were used to determine clusters. Genetic distance thresholds were determined by comparing intra-and inter-host distances. Additional phylogenetic analysis was performed on subjects with complicated viral histories. At the optimal threshold of 3.9%, HIV-TRACE found 77 clusters and PhyloPart found 63 clusters, of which 27 and 32 contained multiple subjects, respectively. Furthermore, 1/3 of the subjects had sequences in different clusters over the course of the study, including some cases in which a later-sampled sequence matched a cluster detected much earlier in the infection, despite being separated by RNA-negative lab visit and detection of sequences in different clusters. A detailed phylogenetic analysis of four subjects with such patterns showed that in all four cases, the earlier and later variants grouped closely on the tree, and did not group with concurrent sequences from any other subject. These observations suggest that subjects are either experiencing rapid and recurring infection-clearance-reinfection cycles from the same source, or a single transmission event produces a chronic infection that may go undetected and/or co-circulate with different viruses from separate transmission events. Furthermore, our results show the utility of using longitudinal sampling to obtain a more comprehensive view of the viral linkages in high-risk populations.

1. Introduction

Hepatitis C virus (HCV) causes more mortality in the United States than 59 other reportable infectious diseases combined (Ly et al., 2016). Moreover, there is an epidemic of acute HCV infection linked to the nationwide opioid epidemic (Zibbell et al., 2015). Thus, despite widely publicized new effective treatments, HCV remains a major public health problem (Thrift et al., 2017). The epidemic is particularly devastating in populations of people who inject drugs (PWID), in which an estimated 2/3 of individuals are HCV positive (Nelson et al., 2011), although extensive research is lacking (Larney et al., 2015).

According to a recent report from the National Academies of Science, Engineering, and Medicine, assessing HCV epidemiological dynamics in populations is important to both design and evaluate HCV control strategies (Romano et al., 2010). The identification of transmission clusters can support self-reported transmission events, identify putative transmission chains, and reveal mixing between key risk groups and geographic sub-populations (Aldous et al., 2012). Such approaches are commonly used in molecular epidemiological studies of HIV (Leigh Brown et al., 2011; Hughes et al., 2009; Wertheim et al., 2016; Wertheim et al., 2014; Wertheim et al., 2017) and HCV (Jacka et al., 2016; Olmstead et al., 2015; Jacka et al., 2014; Sacks-Davis et al., 2012; Matthews et al., 2011; Aitken et al., 2004; Bartlett et al., 2016; Bretaña et al., 2015).

However, tracing HCV transmission events, particularly in high-risk population such as PWID, is more complicated than for HIV for several reasons. One, intra-host HCV evolves in highly structured populations where multiple lineages persist through time (Rose et al., 2012) although may only be sporadically detected (Raghwani et al., 2016; Ho et al., 2017). Thus, the viral variants taken from a single time point do not necessarily reflect the true composition of the population (Rose et al., 2012), even when using high-throughput sequencing (Ho et al.,

Abbreviations: HCV, hepatitis-C viruses; ROC, Receiver operator curve; BBAASH, Baltimore Before and After Acute Study of Hepatitis; PWID, people who inject drugs

* Corresponding author at: 855 North Wolfe St., Rangos Building, room 538A, Baltimore, MD 21205, United States.
E-mail address: olaeyen1@jhmi.edu (O. Laeyendecker).

https://doi.org/10.1016/j.meegid.2017.12.015
Received 8 September 2017; Received in revised form 10 November 2017; Accepted 14 December 2017
Available online 16 December 2017
1567-1348/

2017). Second, unlike HIV, individuals can spontaneously clear HCV infection. Individuals can then become re-infected, either with a different strain, or with a highly similar strain as the initial infection. Either scenario is likely among a highly-networked group participating in high risk behaviors. Concurrent infections with different strains are also possible, although they may be transiently detected. Additionally, in PWID populations, individuals may cease the high-risk behavior and/or experience re-lapse, so that the risk of re-infection is not constant (Larney et al., 2015).

Some studies may attempt to avoid such complications by using a single time point per individual taken soon after initial seroconversion (Jacka et al., 2014; Matthews et al., 2011; Bartlett et al., 2016), and longitudinal studies typically only track infection only over a short time frame (e.g. < 2 years) (Olmstead et al., 2015; Bretaña et al., 2015). However, these approaches limit the understanding of the viral dynamics over time within PWID populations.

In this study, we investigated viral patterns among a highly-networked cohort of PWID followed from July 1997 to July 2016. All subjects had at least 2 samples spaced at least one year apart. Our goal was to determine the appropriate definitions of linkage and describe the complex nature of HCV viral evolution in this highly networked population.

2. Methods

2.1. Ethics statement

This study was approved by the Institutional Review Board of the Johns Hopkins University (IRB# NA_00046368, Title: CD8 + T cells and the Outcome of viral hepatitis- Monthly Blood Sampling). This study of stored samples and previously collected data and the parent cohort study was conducted according to the ethical standards set forth by the institutional review board of Johns Hopkins University and the Helsinki Declaration of the World Medical Association. All participants were adults who provided written informed consent.

2.2. Subjects

HCV uninfected subjects were enrolled in the "*Baltimore Before and After Acute Study of Hepatitis*" (BBAASH) cohort, as previously described (Cox et al., 2005a). All were classified as people who inject drugs (PWID) and had estimated dates of infection after entry into the study. At each visit, subjects were counseled to reduce the risk of injection drug use and tested for HCV infection. The study followed hundreds of individuals who were tested for HCV in a protocol designed for monthly follow up and approved by the Johns Hopkins IRB. At time points where HCV RNA testing was positive, HCV sequence data was obtained. The cohort included chronically-infected individuals, as well as individuals who cleared infection, some of whom were subsequently became re-infected. Summary statistics for the subjects analyzed in this study is shown in Table 1.

2.3. Genetic sequence data and analysis

For samples that tested RNA positive in the BBASH cohort, bulk sequencing was performed on the HCV E1 gene (H77 nt 943 to 1288) (Ray et al., 2000). E1 was chosen as it has high variability and is a commonly sampled region. Furthermore, the primer sets have a high amplification efficiency and efficiently amplify across genotypes 1 to 6 (https://hcv.lanl.gov/content/sequence/HCV/primers/primers.html). For this study, we included 89 subjects who had two or more sequences (range 2–31, median = 7 sequences/person) sampled over a 1–14-year period (average time span = 4.7 years). The final alignment comprised 743 sequences which were generated in real time over a near two-decade period (1997–2016). At time points when a subject had a sequence > 5% different from the previous time point, the sample was re-extracted and re-sequenced. All subjects had estimated dates of infection measured as the midpoint between the last negative and first positive HCV serological test.

Table 1
Description of population evaluated.

Characteristic	Number of subjects
	89
% Female	42
Race (%)	
African American	16 (17%)
Caucasian	70 (78%)
Other	3 (3%)
Mean age at seroconversion (IQR[a]) in years	27 (Cox et al., 2005a; Anisimova and Gascuel, 2006)
Mean year of HCV seroconversion (IQR)	2003 (2001, 2006)
Follow up post HCV seroconversion Av. (range) in years	4.7 (Ly et al., 2016; Olmstead et al., 2015)
Total C/E1 sequences	743
Average number of sequences/subject	8.3

[a] IQR inter quartile range.

Clusters were assigned using two methods: HIV-TRACE (Wertheim et al., 2014) and PhyloPart (Prosperi et al., 2011). Pairwise distances for HIV-TRACE were generated using the tn93 genetic distance model. Clusters were assigned using a range of genetic distance thresholds (1–10%). Phylogenetic trees were inferred using the program PhyML v.3.0 (Guindon et al., 2009) using the general-time reversible nucleotide substitution model with gamma-distributed rate variation among sites. Branch support was assessed using the aLRT method (Anisimova and Gascuel, 2006). Clusters were assigned with PhyloPart using a range of patristic distance thresholds (1–10%) and > 70 bootstrap branch support.

The ROC curve and threshold analysis were performed in R (code available on request). Cluster visualizations were performed with the Tableau software (version 10).

3. Results

3.1. Genetic distance

From the 89 subjects, there were 743 sequences from which a total of 254,106 bases were confidently represented. Overall, the average intra-subject distance ranged from 0%–47%. We then compared the average intra- and inter-subject pairwise distance (Fig. 1). The distributions were largely overlapping, with peaks in both distributions occurring at 10%, 35%, 50%, and 65%. The major difference between the two distributions was that < 5% distance was observed for 50% of the intra-subject comparisons but only 0.2% of inter-subject comparisons. We then calculated pairwise distances within HCV genotypes or subtypes (Fig. 1). Within a genotype or subtype, the average genetic distance was 10% for the inter-subject comparisons and < 5% for the intra-subject comparisons. For subtypes 1a vs. 1b, average genetic distance was 35%; 1a/1b vs. 3a was 50%, and 2b vs. any other genotype was 65%. These values correspond to the 4 peaks in the distance distributions. Within the intra-host dataset, there were different HCV genotypes/subtypes detected that, importantly, reflected co- or re-infection.

3.2. Establishing a genetic distance threshold

To determine viral clusters, we used two approaches: HIV-TRACE, which considers only genetic distance, and PhyloPart, which uses both tree-topology and patristic (branch) distances to assign thresholds. We evaluated the behavior of the two approaches by using a series of genetic distance threshold values (1%–10%). For HIV-TRACE, at the 1% threshold, 90% of sequences fell into a cluster, which quickly increased