Plasma virome and the risk of blood-borne infection in persons with substance use disorder

Abraham J. Kandathil¹, Andrea L. Cox¹, Kimberly Page², David Mohr³, Roham Razaghi⁴, Khalil G. Ghanem¹, Susan A. Tuddenham¹, Yu-Hsiang Hsieh⁵, Jennifer L. Evans⁶, Kelly E. Coller⁷, Winston Timp¹,⁴, David D. Celentano⁸, Stuart C. Ray¹ & David L. Thomas¹

There is an urgent need for innovative methods to reduce transmission of bloodborne pathogens like HIV and HCV among people who inject drugs (PWID). We investigate if PWID who acquire non-pathogenic bloodborne viruses like anelloviruses and pegiviruses might be at greater risk of acquiring a bloodborne pathogen. PWID who later acquire HCV accumulate more non-pathogenic viruses in plasma than matched controls who do not acquire HCV infection. Additionally, phylogenetic analysis of those non-pathogenic virus sequences reveals drug use networks. Here we find first in Baltimore and confirm in San Francisco that the accumulation of non-pathogenic viruses in PWID is a harbinger for subsequent acquisition of pathogenic viruses, knowledge that may guide the prioritization of the public health resources to combat HIV and HCV.
High hepatitis C virus (HCV) incidences among PWID are reported worldwide and threaten the United States Department of Health and Human Services' and World Health Organization's 2030 hepatitis elimination goal to reduce new infections by 90%. Because of the opioid epidemic, the incidence of HCV infections per 100,000 in the United States more than tripled from 0.29 to 1.04 between 2010 and 2017, after a decade of steady decline. Likewise, the opioid epidemic has spread HIV-1 including to populations not previously affected.

In addition to HIV, HCV, and other viral pathogens, there are other viruses found in blood that are not known to cause disease in humans (non-pathogenic viruses) and other viruses found in blood that are not known to cause disease in humans (non-pathogenic viruses) and non-pathogenic Anelloviridae members. Collectively, these viruses are referred to as the virome which also includes virus-derived elements in chromosomes and viruses that infect other types of organisms that inhabit humans.

The risk of transmission of bloodborne pathogens is typically measured by their incidences, while their spread is inferred by the phylogenetic relationships of their sequences. Transmission studies have shown that HCV precedes HIV infection in PWID, and hence new HCV infections serve as strong predictors of communities at risk for HIV. We hypothesized that in that same way, the dynamics and abundance of non-pathogenic bloodborne viruses might serve as sentinels to predict the risk of acquiring pathogenic viruses such as HIV-1 and HCV and to reveal insights into transmission networks.

We have extensively characterized the blood virome of PWID. After previously studying plasma from PWID in Baltimore, Maryland including using unbiased, metagenomic approaches, we focused on frequently observed non-pathogenic components of the DNA (alpha-, beta-, and gamma torque viruses) and RNA (pegivirus C and pegivirus H) virome in blood. In this study to test our hypothesis, we first assessed the plasma of PWID enrolled in a unique Baltimore cohort that recruited participants (see methods) when they were still HIV-1 and HCV uninfected and monitored them longitudinally for new bloodborne infections. Our assessment revealed an expansion of the non-pathogenic virome prior to acquisition of HCV, which we also confirmed in a similar PWID cohort based in San Francisco. We also demonstrate the utility of non-pathogenic virome sequences in transmission studies.

### Results

**Expansion of non-pathogenic plasma virome prior to acquisition of HCV.** We studied 20 participants (Table 1) from a Baltimore based PWID cohort before and after HCV seroconversion (HCVneg to pos PWID). Participants were assessed at baseline, a median of 140 days (IQR:168.5) before their first seroconversion visit and then followed-up a median of 1184 days (IQR:213) after baseline assessment. Identification of non-pathogenic virome components was done using PCR (see methods). We similarly tested at two time points, 20 age- and gender-matched PWID from the same cohort who never acquired HCV or HIV (HCVneg to neg PWID) after a similar period of observation (median:1096 days, IQR:152.5). Consistent with our hypothesis, at baseline more non-pathogenic virome components were detected in those who would later acquire HCV than those who would not ($P = 0.0031$). The same assessments performed on plasma from 20 sociodemographic and age matched controls from urban Baltimore with no history of injection drug use revealed even fewer non-pathogenic viruses (Fig. 1a, $P = 0.0046$).

### Table 1 Subject characteristics.

| City, Country | Non-injection drug users ($n = 20$) |
|---------------|-----------------------------------|
|               | HCVneg to neg PWID ($n = 19$)     |
|               | HCVneg to pos PWID ($n = 13$)     |
|               | HCVneg to pos PWID ($n = 19$)     |
|               | HCVneg to neg PWID ($n = 20$)     |
|               | HCVneg to pos PWID ($n = 20$)     |
| Female (%)    | 6 (30)                            |
|               | 27 (4)                            |
|               | 2 (2)                             |
|               | 2 (2)                             |
|               | 2 (2)                             |
| Median (IQR) age years | 26 (6.25) | 26 (6.25) | 26 (6.25) | 26 (6.25) | 26 (6.25) |
| No. of time points tested | 2 | 2 | 2 | 2 | 2 |
| Median (IQR) days tested | 1184 (213) | 1184 (213) | 1184 (213) | 1184 (213) | 1184 (213) |
| Median (IQR) days before HCV seroconversion visit | 140 (168.5) | 140 (168.5) | 140 (168.5) | 140 (168.5) | 140 (168.5) |
| Median (IQR) duration in years of injection drug use | >6 months | >6 months | >6 months | >6 months | >6 months |

### Table 1 Subject characteristics.

| City, Country | Female (%) | Median (IQR) age years | No. of time points tested | Median (IQR) days tested | Median (IQR) days before HCV seroconversion visit | Median (IQR) duration in years of injection drug use |
|---------------|------------|-----------------------|--------------------------|--------------------------|-----------------------------------------------|-----------------------------------------------|
|               |            |                       |                          |                          |                                               |                                               |
|               | HCVneg to pos PWID ($n = 20$)     | HCVneg to neg PWID ($n = 19$)     | HCVneg to pos PWID ($n = 13$)     | HCVneg to neg PWID ($n = 20$)     | HCVneg to pos PWID ($n = 20$)     |
| Female (%)    | 6 (30) | 27 (4) | 2 | 2 | 2 | 2 |
| Median (IQR) age years | 26 (6.25) | 26 (6.25) | 26 (6.25) | 26 (6.25) | 26 (6.25) |
| No. of time points tested | 2 | 2 | 2 | 2 | 2 |
| Median (IQR) days tested | 1184 (213) | 1184 (213) | 1184 (213) | 1184 (213) | 1184 (213) |
| Median (IQR) days before HCV seroconversion visit | 140 (168.5) | 140 (168.5) | 140 (168.5) | 140 (168.5) | 140 (168.5) |
| Median (IQR) duration in years of injection drug use | >6 months | >6 months | >6 months | >6 months | >6 months |
At follow-up, HCV\textsubscript{neg to pos} PWID group had a median of two non-pathogenic viruses compared to one in the HCV\textsubscript{neg to neg} PWID group (Fig. 1b, \(P < 0.001\)). We reasoned that additional pegivirus C infections not detected would have occurred and subsequently cleared before baseline, and hence also performed serological testing to assess for previous pegivirus C exposure\textsuperscript{13,14}. By baseline, antibodies to pegivirus C were already detected in 35% and 30% among HCV\textsubscript{neg to pos} PWID and HCV\textsubscript{neg to neg} PWID, respectively. Accounting for resolved pegivirus C infection at baseline, further confirmed higher viral infections prior to HCV seroconversion in the HCV\textsubscript{neg to pos} PWID group compared to the HCV\textsubscript{neg to neg} PWID group (\(P = 0.0041\)). Pegivirus H antibodies were only detected in one sample from an HCV\textsubscript{neg to pos} PWID at follow-up. We also observed a greater circulating plasma DNA virome richness among HCV\textsubscript{neg to pos} PWID compared to HCV\textsubscript{neg to neg} PWID both at baseline (\(P = 0.0038\)) and follow-up (\(P = 0.0016\)). Similar differences were not observed in the RNA compartment.

To confirm these observations in the discovery panel (Baltimore) regarding expansion of the virome prior to acquisition of bloodborne viral infection among PWID, we performed the same assessments on a confirmatory panel composed of PWID from San Francisco\textsuperscript{17}. Like the Baltimore cohort, this cohort also recruited participants (see methods) when they were HIV-1 and HCV uninfected and monitored them longitudinally for new bloodborne infections. We studied 19 age and gender-matched participants from the HCV\textsubscript{neg to pos} PWID and HCV\textsubscript{neg to neg} PWID group at two time-points (Table 1). The HCV\textsubscript{neg to pos} PWID were tested a median of 140 days before their first HCV seropositive visit. Each group was composed of 20 biologically independent specimens. At baseline (\(p = 0.0316\)) and follow-up (\(p = 0.0384\)), HCV\textsubscript{neg to pos} PWID had significantly greater virome richness in the confirmatory (San Francisco) panel (median: 91 days prior to seroconversion) For panels c and d each group was composed of 19 biologically independent specimens. The differences were also observed on combining study participants (\(n = 39\) biologically independent specimens in each group) from both panels \(c\) and \(d\) (\(p < 0.001\)). The differences were also observed on combining study participants (\(n = 39\) biologically independent specimens in each group) from both panels \(c\) and \(d\) (\(p < 0.001\)). Error bar denotes median. \(^*\) Two tailed \(P\) and \(^*\) one tailed \(P\) value were calculated using Mann–Whitney or Kruskal–Wallis test. HCV was not included in the analysis. List of non-pathogenic viruses identified are shown in Supplementary Table 1. Source data are provided as a Source Data file.
Utility of non-pathogenic plasma virome sequences in identification of drug use networks. It can be important to understand the structure of the drug-use networks that transmit bloodborne infections. Yet, limitations in ascertainment of personal, illegal practices leave most transmission events uncharacterized even with post-infection analysis of HCV sequences. We asked if the phylogenetic relationships can be revealed by sequencing the non-pathogenic virome components. We initially performed Sanger sequencing on Baltimore and Chang Mai samples with pegivirus C infection, including one transmission pair, containing pegivirus C RNA and HCV RNA (Supplementary Fig. 2). As expected, sequences from Baltimore clustered distinctly from those recovered from Thailand, and pegivirus C sequences from persons in known HCV-defined groups also clustered as did samples from same study participants at first and second time points (except when the initial virus sequence was cleared). In addition, pegivirus C sequences revealed new linkages among PWID previously not recognized by HCV sequences and/or self-report.

Based on our observation of the plasma DNA virome, we characterized the predominant non-pathogenic components of the DNA virome, *Anelloviridae*. Seven plasma samples from four HCV-infected PWID partners identified by epidemiological data with *Anelloviridae* viremia were subjected to next generation sequencing (NGS) using nanopore (see methods). Following NGS, reads were taxonomically classified using Centrifuge, a microbial classification engine, that is part of 'What's in my Pot' workflow used for real-time analysis of nanopore sequences. Classified sequences were then used for construction of full-length de novo assemblies using Canu (v1.9). To independently validate multi-species infection, Sanger sequencing was done to confirm the presence of four distinct alphatorquevirus species identified in a study participant. Reads classified as TTV7, TTV8, TTV9, and TTV18 were used to design PCR primers targeting a 500 bp region of ORF1 to confirm their presence in the plasma. The ORF-1 region was targeted since ICTV guidelines use a cut-off value of 35% nucleotide sequence identity in the ORF-1 as a demarcation criterion for species identification.

We then designed primers based on reads classified as TTV8 to amplify positions 275–400 of ORF-1. This sequence contained a relatively conserved region of ORF-1 in the 5' end followed by the first 100 amino acids of the HVR-1 region located in ORF-1. For a single participant, the same set of primers were also used to detect TTV8 at time point 29 months later. Using study participant sequences (n = 6) and four non-Baltimore sequences that included ICTV reference sequences for TTV7 and TTV8, amino acid sequence divergence was calculated using MegaX (Fig. 2). Pairwise distance confirmed relatedness of alphatorquevirus sequences observed between a drug use pair previously identified using HCV core-E1 sequences. In addition, we also observed drug use pairs linked based on alphatorquevirus sequences but not identified using HCV sequences, suggesting that alphatorquevirus sequences might reveal information about networks prior to HCV infection and unique from HCV sequences.

Discussion

These data confirm that the blood virome can reveal the risk of bloodborne infectious diseases and that the phylodynamic structure of the virome may provide unique insights previously hidden among PWID.

Characterization of non-pathogenic components of the virome has clear advantages over traditional epidemiological approaches of viral pathogen sequencing. By using non-pathogenic viruses as sentinels, risk can be ascertained before harm occurs in a population. In principle, this is similar to the use of HCV as a marker of HIV risk. HCV infection was widespread in Scott County Indiana before HIV was introduced and early recognition of HCV might have prevented at least 58 cases of HIV. Since HCV causes more mortality in the United States than HIV, HCV is clearly a suboptimal sentinel. Since we found that non-pathogenic members of the virome were evident before HCV, this approach might be more effective. In addition, identification of sequences from multiple anellovirus species in study participants using nanopore suggests that the PCR based approach likely underestimated the number of members among individual study participants.

There are also important limitations to existing methods of identifying the sources of bloodborne pathogen transmission. Because of stigma and legal issues, self-identification of networks is an insensitive method to detect transmission pairs. Even phylogenetic approaches focused on one virus have rarely been helpful to reveal drug use networks. With HCV, natural clearance, reinfection, superinfection and within-host evolution may render sequences obtained at a particular visit uninterpretable. The addition of even one more virus might exponentially increase the probability of phylodynamic inferences. For example, if an analysis of HCV sequences produced a probability of a shared source of 0.90 and the same pair also shared *Anelloviridae* sequences that suggested the same probability, then the actual probability of sharing rises from 0.90 to 0.99. While this example is simplistic and the assigned probabilities arbitrary, next generation sequencing has already been proposed for the Advanced Molecular Detection Program of the United States Centers for Disease Control and Prevention.

Our data reveal the vast potential applications of full characterization of the blood virome that could transform our understanding of PWID networks at a crucial time when the opioid epidemic is spreading worldwide. While this work provides a virologic foundation, additional work is needed to increase non-pathogenic virome sequences to aid taxonomic identification and for development of suitable phylodynamic methods.

Methods

Study participants. Study samples were drawn from a cohort of injection drug users located in Baltimore, Maryland (USA). At the time of recruitment, participants in the discovery panel (Baltimore) reported injecting drugs in the last 6 months and were negative for both HIV and HCV infections. Participants are then longitudinally followed with plasma and serum collected at each monthly visit. At each visit, participants are tested for HCV and are tested for HIV every 6 months. For this study, we identified 20 participants in whom HCV seroconversion occurred (HCVneg to pos) and there was sufficient stored plasma from before and after HCV infection. The initial HCV negative sample was a control for the ultimate acquisition of HCV. In addition, from the same cohort a total of 20 additional controls were identified with similar time of follow-up but who did not acquire HCV infection (HCVneg to neg). Participants from the Baltimore cohort received a participant compensation of $US25 for all study visits. All 40 participants from the HCVneg to pos PWID and HCVneg to neg PWID groups were tested at two-time points, separated by a median of 1112.5 days. For the HCVneg to pos PWID, the first time point was a median of 126 days before HCV seroconversion. The median age in the HCVneg to pos PWID group was 26 years (IQR:6.25) and in the HCVneg to neg PWID group it was 27 years (IQR:4). Each group had a gender ratio of 2.33:1. For the four transmission pairs, each pair was tested at two visits that were a median of 1085 days (IQR:180) apart.

Since so little was known about the plasma virome of young adults, we also added sociodemographic controls who denied injection drug use:10 men from the Johns Hopkins Emergency Department and 10 women from the gynecological clinic. The control group had a median age of 26.5 years (IQR:7.7) and a gender ratio of 1:1. Finally, for confirmation, a geographically distinct cohort of HIV negative PWID from San Francisco, California (USA) was used. Similar to the discovery panel, participants in this confirmatory panel reported injecting drugs in the last 6 months and were negative for both HIV and HCV infections. We included 19 HCVneg to pos PWID and HCVneg to neg participants. Each total 38 participants were tested at two timepoints. The median age in the HCVneg to pos PWID group was 26 years (IQR:4) and in the HCVneg to neg PWID group it was 24 years (IQR:4.5). Each group had a gender ratio 3.75:1. HCVneg to pos were tested a median of 91 days (IQR:32) before HCV seroconversion and a median of 427 days
detected using primers described by Berg et al. (Supplementary Table 2)\textsuperscript{30}. The superscript IV (Thermo Fisher Scientific, Cat# D7003) was used to extract DNA and RNA from 200 µL of plasma.

Pre-extraction steps included spinning the samples at 1600 g for 15 min at 4 °C to remove debris (e.g., insoluble complexes) followed by clinical review. Pegivirus C detection was done using a one-step SuperScript\textsuperscript{TM} cDNA synthesis using superscript IV (Thermo Fisher Scientific Cat#18090010), pegivirus H infection was detected using primers described by Berg et al. (Supplementary Table 2)\textsuperscript{30}. The single round PCR targeted a 160 bp region in the NS2-3 junction of pegivirus H. Pegivirus C detection was done using a one-step SuperScript\textsuperscript{TM} IV One-Step RT-PCR (Thermo Fisher Scientific Cat#12594100) protocol followed by second round PCR using Platinum Taq DNA polymerase High Fidelity (Thermo Fisher Scientific Cat#11304-011). A hemi-nested PCR targeted nucleotide positions 979–1316 for pegivirus C on the ARCHITECT immunoassay platform using previously described antibody assays\textsuperscript{34}. Briefly, separate anti-IgG assays were developed that use pegivirus H recombinant antigens, NS4AB or E2, or pegivirus C recombinant antigen, E2, to detect IgG specific responses. Provisional assay cut-offs were based on screening a volunteer donor population that was negative for pegivirus H and pegivirus C RNA or antibodies. A signal to cut-off (S/CO) value greater than or equal to 1.0 was considered reactive in the pegivirus H or pegivirus C assays.

Serology testing was performed to detect IgG antibodies specific for pegivirus H and pegivirus C on the ARCHITECT immunoassay platform using previously described antibody assays\textsuperscript{34}. Briefly, separate anti-IgG assays were developed that use pegivirus H recombinant antigens, NS4AB or E2, or pegivirus C recombinant antigen, E2, to detect IgG specific responses. Provisional assay cut-offs were based on screening a volunteer donor population that was negative for pegivirus H and pegivirus C RNA or antibodies. A signal to cut-off (S/CO) value greater than or equal to 1.0 was considered reactive in the pegivirus H or pegivirus C assays.

Nanopore sequencing. Sequencing of Anelloviridae was done using Oxford Nanopore Sequencing on the GridION platform. The concamers generated using RCA were digested using T7 endonuclease to debranch and generate fragments of varying sizes. The fragmented DNA was subjected to a magnetic bead-based size exclusion cleanup protocol to keep fragments size between 500 and 10,000 bp. Following the size clean up, samples that had a concentration greater than 500 ng were selected for library preparation using the using SQK-LSK109 kit with native barcoding EXP-NBD104. Libraries were pooled in equimolar amounts and loaded onto a single R9.4 flowcell. Sequencing and basecalling was performed on the GridION running 19.12.2 according to protocol.

The reads were analyzed using What’s In My Pot (WIMP), a quantitative analysis tool for nanopore reads\textsuperscript{19}. As part of the WIMP workflow, taxonomic classification of the reads was done using Centrifuge\textsuperscript{35}. For classification of reads, Centrifuge uses a database generated from sequences in RefSeq\textsuperscript{38}. Classified reads for each alphatorquevirus species were used with Canu v1.9\textsuperscript{39} (parameters: genomeSize=3k and -nanopore-raw) to assemble species-specific viral genomes. When primers were needed for Sanger sequencing of specific alphatorquevirus

### Table 1: Pair Subject IDs and Time between sample collection

| TTV Pair | Pair Subject IDs | Time between sample collection (months) |
|----------|-----------------|----------------------------------------|
| 1        | Subject A(v1) – Subject A(v2) | 29 |
| 2        | Subject A(v1) – Subject D | 8 |
| 3        | Subject A(v2) – Subject D | 28 |
| 4        | Subject B – Subject A(v1) | 5 |
| 5        | Subject B – Subject D | 13 |
| 6        | Subject B – Subject A(v2) | 41 |

Fig. 2 Amino acid sequence divergence between alphatorquevirus pairs confirms known drug use pair while also revealing newly recognized drug use pairs. The number of amino acid substitutions per site between alphatorquevirus ORF-1 (positions: 263-404) sequences are shown. Error bar denotes standard error estimate obtained by a bootstrap procedure (1000 replicates). Analyses were conducted using the Poisson correction model. This analysis involved 10 amino acid (6 Baltimore and 4 non-Baltimore) sequences from biologically independent samples and 97 positions. All positions containing gaps and missing data were eliminated. Evolutionary analyses were conducted in MEGA X. Table shows drug use pairs identified using alphatorquevirus sequences and their temporal relation. (Gen Bank ID: AB054648, JN980171, AB054647, AF261761, MZ889122-MZ889127). All other pairs denote sequence divergence between Baltimore and non-Baltimore alphatorquevirus.

Molecular detection and serology. Extraction of plasma DNA and RNA was as previously described\textsuperscript{32}. Briefly, the Quick-DNA/RNA miniprep plus kit (Zymo Research, Cat# D7003) was used to extract DNA and RNA from 200 µL of plasma. Pre-extraction steps included spinning the samples at 1600 g for 15 min at 4 °C to remove debris (e.g., insoluble complexes) followed by filtration using 0.2 µM syringe filters (Fisher Scientific, Cat# 6778-1302). Following cDNA synthesis using superscript IV (Thermo Fisher Scientific Cat#18090010), pegivirus H infection was detected using primers described by Berg et al. (Supplementary Table 2)\textsuperscript{30}. The first round PCR primers amplified a conserved region found in the 5’ end of all three torquevirus genera using forward primers NG779/NG780 and reverse primers NG781/NG782. Genus specific primers for alpha-(NG779/NG780 and NG785), beta- (NG792/NG793/NG794 and NG791), and gamma (NG795 and NG796) torqueviruses allowed for differential detection of each genus. The size of the resulting amplicons was used to differentiate the three genera – alpha– (112–117 bp), beta– (70–72 bp), and gamma– (88 bp) torquevirus. The PCR for alpha torquevirus detection was semi-nested in design while a nested PCR for beta- and gamma torquevirus was used. Additionally, we enhanced sensitivity of the PCR by incorporating a rolling circle amplification (RCA) based preamplification step. We also observed increased sensitivity of RCA + PCR without the initial 95 °C denaturation step for RCA. Enrichment of circular templates using RCA was done using TempliPhi™ (Sigma Aldrich GE25-6400-10). RCA was carried out at 30 °C for 18 h with 2 nM of DNA as input. For the PCR, a 1:5 dilution of RCA concatamers was taken for subsequent PCR reactions.

Serology testing was performed to detect IgG antibodies specific for pegivirus H and pegivirus C on the ARCHITECT immunoassay platform using previously described antibody assays\textsuperscript{34}. Briefly, separate anti-IgG assays were developed that use pegivirus H recombinant antigens, NS4AB or E2, or pegivirus C recombinant antigen, E2, to detect IgG specific responses. Provisional assay cut-offs were based on screening a volunteer donor population that was negative for pegivirus H and pegivirus C RNA or antibodies. A signal to cut-off (S/CO) value greater than or equal to 1.0 was considered reactive in the pegivirus H or pegivirus C assays.

Nanopore sequencing. Sequencing of Anelloviridae was done using Oxford Nanopore Sequencing on the GridION platform. The concamers generated using RCA were digested using T7 endonuclease to debranch and generate fragments of varying sizes. The fragmented DNA was subjected to a magnetic bead-based size exclusion cleanup protocol to keep fragments size between 500 and 10,000 bp. Following the size clean up, samples that had a concentration greater than 500 ng were selected for library preparation using the using SQK-LSK109 kit with native barcoding EXP-NBD104. Libraries were pooled in equimolar amounts and loaded onto a single R9.4 flowcell. Sequencing and basecalling was performed on the GridION running 19.12.2 according to protocol.

The reads were analyzed using What’s In My Pot (WIMP), a quantitative analysis tool for nanopore reads\textsuperscript{19}. As part of the WIMP workflow, taxonomic classification of the reads was done using Cen
sequences, samples were aligned to reference ICTV species, indexed used SAM tools and visualized on integrative genomics viewer. Sanger sequencing was used to confirm the presence of multiple alphatorquevirus sequences in a single study participant and to obtain sequences for phylogenetic analysis.

### Statistical analysis

Since data were not normally distributed, nonparametric statistical tests were used to characterize the probability of the data were explained by the null hypothesis, that there was no difference in the viromes of persons who acquired HCV or did not. The number tested was determined to provide more than 90% power to detect a difference of two additional viruses in the virome of the HCV

### Phylogenetic analysis

**Method:**

**Data:**

The data used in the analysis were obtained from the GenBank database under accession code MN938857-MN938890 (pegivirus-C), MN954475-MN954517 (HCV), and M2889122-M2889127 (alphatorquevirus). The metagenomic dataset was generated and analyzed during the study and were deposited in the National Center for Biotechnology Information (NCBI) database.

**Phylogenetic tools:**

The phylogenetic relationships of HCV and pegivirus C nucleotide sequences were analyzed using the MUSCLE algorithm followed by analysis using Poisson correction model with a bootstrap value of 1000. The phylogenetic tree was inferred using the maximum likelihood method based on the Tamura-Nei model. The sequences were also analyzed using the maximum likelihood method based on the Tamura-Nei model.

**Results:**

The phylogenetic tree was constructed using the Maximum Likelihood method. The tree is rooted using the midpoint rooting method. The tree was inferred using the Maximum Likelihood method based on the Tamura-Nei model. The sequences were also analyzed using the Maximum Likelihood method based on the Tamura-Nei model.

**Conclusion:**

The phylogenetic analysis of the sequences revealed a consistent temporal pattern of genetic variation of HCV among high-risk subjects may impact inference of transmission networks. Hepatitis C infection is a major public health concern and affects millions of people worldwide. The occurrence of new viruses poses a significant challenge to public health officials. The study findings can be used to inform public health policies and interventions to control the spread of HCV.

### Data availability

The data used in the analysis were obtained from the GenBank database under accession code MN938857-MN938890 (pegivirus-C), MN954475-MN954517 (HCV), and M2889122-M2889127 (alphatorquevirus). The metagenomic dataset was generated and analyzed during the study and were deposited in the National Center for Biotechnology Information (NCBI) database.

**Acknowledgements**

This work was supported in part by National Institute on Drug Abuse (NIDA) (R01DA0380 for D.LT., and A.J.K., R31DA041657 for K.A.P.), and National Institute of Allergy and Infectious Diseases (NIAID) (U19AI159822 for A.L.C.). The funders had no role in study design, data analysis, decision to publish, or preparation of the manuscript.
**Author contributions**

D.L.T. and A.J.K.: conception and design. A.J.K. and J.L.E.: collection and assembly of data. A.J.K., R.E., W.T., and D.M.: computational analysis. A.L.C., K.A.P., K.G.G., S.A.T., D.D.C., Y.H., and K.E.C.: provision of the samples and study materials. D.L.T., A.J.K., and S.C.R.: analysis and interpretation of data. A.J.K. and D.L.T.: drafting of the article that was reviewed and edited by all the authors.

**Competing interests**

S.T. has been a consultant for Biofire Diagnostics, Roche Molecular Diagnostics and Luca Biologics, and has received speaker honoraria from Roche Molecular Diagnostics. All other authors declare no competing interests.

**Additional information**

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-021-26980-8.

Correspondence and requests for materials should be addressed to David L. Thomas.

Peer review information Nature Communications thanks the anonymous reviewers for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2021