Published in final edited form as:

J Viral Hepat. 2017 January ; 24(1): 43–52. doi:10.1111/jvh.12616.

Phylogenetic analysis of full-length, early infection, hepatitis C virus genomes among people with intravenous drug use: the InC³ Study

Chaturaka Rodrigo¹, Auda A. Eltahla¹, Rowena A. Bull¹, Fabio Luciani¹, Jason Grebely², Gregory J. Dore², Tanya Applegate², Kimberly Page³, Julie Bruneau⁴, Meghan D. Morris⁵, Andrea L. Cox⁶, William Osburn⁶, Arthur Y. Kim⁷, Naglaa H. Shoukry⁴, Georg M. Lauer⁷, Lisa Maher², Janke Schinkel^{8,9}, Maria Prins^{8,9}, Margaret Hellard^{10,11,12}, and Andrew R. Lloyd¹ on behalf of the InC³ Collaborative

¹School of Medical Sciences, UNSW, NSW, Australia ²The Kirby Institute, UNSW, NSW, Australia ³University of New Mexico, Albuquerque, New Mexico, USA, CRCHUM ⁴Université de Montréal, Montreal, Canada ⁵University of California, San Francisco, USA ⁶Johns Hopkins Medical Institutions, Baltimore, MD, USA ⁷Harvard Medical School, Boston, MA, USA ⁸Academic Medical Center, Amsterdam, The Netherlands ⁹GGD Public Health Service of Amsterdam ¹⁰Burnet Institute, Melbourne, VIC, Australia ¹¹Monash University, Australia ¹²Alfred Hospital, Melbourne, Australia

Abstract

Cross-continental phylogenetic analysis is important to understand subtle molecular differences of currently circulating hepatitis C virus (HCV) subtypes. Existence of such differences can be crucial in pursuing a universal hepatitis C vaccine. We characterised molecular epidemiology of early HCV infections identified across nine cohorts [North America (n=4), Australia (n=4) and Europe (n=1)] in the International Collaborative of Incident HIV and Hepatitis C in Injecting Cohorts (InC³). One hundred and ninety two full-length HCV genomes were amplified from plasma of incident infections and subjected to next generation sequencing to establish the largest cross continental, full-length acute HCV genomic dataset available to-date. Genomes from the most common subtypes (1a: n=94, 2b: n=15 and 3a: n=68) were used in phylogenetic analysis. Using full genome trees, 78 sequences (44%) were found to lie within 29 phylogenetic clusters/pairs defined on the basis of molecular similarity of consensus sequences. Of these, 26 each had exclusively Australian or North American sequences indicating a strong geographical bias for molecular similarity. On further analysis of behavioural and demographic associations, binary logistic regression analysis showed that older age and non-Caucasian ethnicity were significantly associated with clustering. HCV probably evolves in micro-epidemics within geographically isolated communities.

Correspondence: Chaturaka Rodrigo MBBS MRCP(UK) MD, Infection and Inflammation Research Centre, School of Medical Sciences, University of New South Wales, 2052, Sydney, NSW, Australia, c.rodrido@unsw.edu.au, Tel:+61 452095648, Fax:+61 2 9385 1389.

Disclosures: The authors have nothing to disclose in relation to this work.

Keywords

People who inject drugs; molecular epidemiology; NGS; Hepatitis C virus; Phylogenetics

Hepatitis C virus (HCV) remains one of the most prevalent and incident blood borne viral infections globally, particularly among people who inject drugs (PWID) (1–4). In USA, incident cases of HCV are increasing since 2010 and approximately 7.8%–26.7% of PWIDs are estimated to be infected annually (5–11). The “International Collaborative of Incident HIV and HCV in Injecting Cohorts” (InC3) has successfully created the first and largest merged international multi-cohort project of pooled biological and behavioural data from nine prospective observational studies (in Australia, Canada, USA and the Netherlands) of incident HIV and HCV infections in high-risk cohorts, principally people who inject drugs (PWID) (12). This has allowed in-depth, cross-continental epidemiological analysis of incident HCV infections.

Viral sequence data have been increasingly used to complement traditional epidemiological studies examining the incidence and risk factors for infections. This molecular epidemiology approach has largely superseded the more limited analysis of simple viral traits such as genotype or viral load (13). RNA viruses, such as HIV and HCV, feature rapid mutations across the genome, both within infected hosts and across infected populations. This rapid viral evolution allows tracking of related infections by identification of viral genomic similarities in different hosts (phylogenetics).

However, traditional phylogenetic studies on HCV have several limitations. Firstly, the molecular footprint identified from an individual host at a later time point in infection can be different from that of an early time point as the characteristics of the transmitted/founder sequence can change due to viral adaptation and host immune selection pressures (13). Accordingly, viral sequences obtained from late infection time points (beyond six months in HCV infection) may be less accurate in representation of the transmission strain(s) (14). Secondly, all phylogenetic studies of HCV published to date have concentrated on selected segments of the genome for analysis, most commonly Core-E2 or NS5B (15). The phylogenetic signal from these segments had never been compared to that of the full genome probably due to the technical difficulties in amplifying the full genome (15, 16). Thirdly, first generation (Sanger) sequencing used in many studies does not allow an in-depth appreciation of the quasi-species diversity of the virus within a host. Finally many molecular epidemiological studies on HCV were carried out within a limited geographical area and logistical difficulties in obtaining samples had prevented a detailed cross-continental molecular epidemiological analysis.

This study aimed to overcome these limitations by using our recently developed technique (17) for amplification of full length HCV genomes from serum samples stored during acute infection across nine prospective cohorts in four countries. Many samples were from North America (USA and Canada) and Australia allowing a cross continental phylogenetic analysis of sequences with regard to demographic and behavioural factors associated with phylogenetic clustering. Comparison of phylogenetic resolution when using full length

genome and segments of the same genome (the standard practice for many previous studies) was another aim of this study.

Methods

Sample selection

Samples were selected from the individual cohorts of the InC³ collaboration (North America (n=4), Australia (n=4) and Europe (n=1)), (Table 1). The InC³ Collaborative is described in detail elsewhere (12). The dataset was searched for plasma samples collected within 180 days of the estimated date of infection and with a viral load of greater than 1,000 IUml⁻¹ or a positive qualitative viral RNA result. The estimated date of infection was calculated as described previously (18). When there were multiple eligible samples from the same subject, a preferred sample was selected based on two criteria in the following order of priority: a) sampling date within 100 days of estimated date of infection, and b) an estimated viral load >20,000 IUml⁻¹.

Generation of full length hepatitis C amplicons and sequencing

Near full-length HCV amplicons were generated from plasma or serum samples using the methods described by Bull et al. (17). In brief, viral RNA was extracted from 140µl of serum or plasma using the QIAmp Viral RNA kit according to manufacturers' instructions (Qiagen, Chadstone Centre, Vic, Australia). Reverse transcription of viral RNA was carried out using SuperScript III First-Strand Synthesis System (Life Technologies, Australia) and a pan-genotype primer (oligo dA₂₀). The cDNA from this step was subjected to a nested PCR with a combination of genotype specific and universal hepatitis C primers using KlenTaq LA enzyme (Clontech, USA). PCR products were subjected to gel electrophoresis (0.8% agarose) to guide selection of samples to be taken directly into library preparation for NGS, based on identification of an appropriate (9Kb) sized band. All PCR products were purified with magnetic beads (Agencourt AMPure XP, Beckman Coulter, USA, Cat: A63881) according to the manufacturer's instructions and stored at (-80 °C). Plasma samples from one cohort (the Networks/SuperMIX study, Melbourne, Australia) had been collected in Lithium heparin, which is a PCR inhibitor. The RNA extracts from these samples were decontaminated using heparinase as described by Johnson et al (19).

If the full-length amplicon generation was apparently unsuccessful (no band on the gel), remaining RNA from the same extraction was used to attempt amplification of full genome in three separate fragments with a different protocol and primer sets (14). This hemi-genomic amplification method was optimized for genotypes 1a and 3a, which were the most common genotypes in the sample. When both methods were unsuccessful, a fresh RNA extraction was undertaken from the remainder of the plasma with ultracentrifugation to concentrate the virus. For this step, up to 1ml of plasma was centrifuged at 120,000g for 90 minutes and the supernatant discarded. The pellet was resuspended in 140µl of phosphate buffered saline (PBS; Life Technologies, NSW, Australia) before being subjected to RNA extraction and full length amplicon generation as described above.

All samples with a visible band at 9Kb on agarose gel were subjected to NGS on a MiSeq sequencer (Illumina, San Diego, California, USA) using an optimized in-house library preparation (supplementary material 1). For samples without a visible band, the PCR product was diluted by 1:1000 and subjected to a quantitative PCR (qPCR) using two universal HCV primers (hep21b and hep22) to amplify a 150–200bp segment in the Core region (17). Samples with at least 10^7 copies per ml in the qPCR were carried forward in the protocol for NGS (based on pilot experiments which revealed that approximately 50% of samples above this cut-off which did not show a band on gel, were ultimately successfully sequenced).

The raw data from NGS were analysed to generate a consensus sequences for each sample using a bioinformatics pipeline implemented within the Geneious package (Biomatters Inc, Version 8.0.5, Auckland, New Zealand) (Supplementary material 1).

Phylogenetic analysis

Only near full length sequences (Core to NS5B, approximate length 8900) were considered for the phylogenetic analyses described here. The consensus sequences were categorised according to genotype and aligned using the MUSCLE (Multiple Sequence Comparison by Log-Expectation) algorithm, excluding the non-coding region. Each genotype specific alignment was used to generate a maximum likelihood consensus tree of 1,000 bootstrap replicates using the RAxML (Randomized Axelerated Maximum Likelihood) plugin in Geneious (Nucleotide model: GTR GAMMA, Algorithm: rapid bootstrapping). The bootstrap consensus tree was exported in newick format to Phylopart software for cluster analysis based on the mean patristic distance (MPD). Cluster definition using MPD is dependent on the total variability of the phylogenetic tree (20). Firstly an MPD distribution for all pairs on the tree are calculated and any pair with an MPD that falls below defined cut-off are classified as a cluster. A mean patristic distance of 0.01 (below 1st percentile of the total variation of the tree) was used as a cut-off for cluster definition with a bootstrap support of 95%. Due to the stochastic nature of phylopart algorithm, the maximum reported number of clusters per run was selected after running the programme 10 times for each dataset. A sensitivity analysis for clustering was done by relaxing the MPD up to 0.1(13).

Pairwise time differences since the estimated dates of infection were compared between case-sequences that clustered and those sequences which did not. Geographical location of participants at the time of infection was also recorded (State, district or postcode) to explore evidence of proximity at the time of infection.

The phylogenetic trees from the above step were redrawn by introducing additional subtype specific sequences from public domain databases (Los Alamos HCV database and Genbank), collected across a wide time range and geographical locations not exclusive to locations of InC³ sequences. Preservation of the clusters after introduction of these extraneous sequences was taken as a measure of robustness (strong molecular relatedness) of the identified clusters.

Finally, segments coding for separate structural and non-structural proteins (Core, E1, E2, P7-NS2, NS3, NS4, NS5) were extracted from the final alignments (for subtypes with more

than 50 sequences in the final alignment) and phylogenetic trees were reconstructed using the same algorithm to compare with that of the full genome. In addition, two commonly used concatenated regions of the genome [Core – partial E2 without HVR1, Core – partial E2 (without HVR1) + NS5B] were also used for this comparison (15). Without a gold standard for a valid cluster, only those appearing in at least two of the trees (full genome or segment trees) were counted.

Demographic data

The demographic, behavioural and clinical data for subjects in the nine individual cohorts has been merged into a common dataset for shared use within the InC3 Collaborative. These were used to explore associations with clustering. The dataset used for this study was updated in February 2014. The variables used for this analysis include; age, gender, continent of origin, ethnicity, past history of incarceration, incarceration at baseline (at time of recruitment to the study), level of education, sexual orientation, type of drug injected at baseline, being on opioid substitution therapy and HIV co-infection. These selections were based on study hypotheses (continent of origin) and previously demonstrated associations for clustering.

Statistical analysis

Chi square and independent T tests were used initially to identify significant demographic and behavioural factors associated with clustering. Descriptive statistics were summarized into proportions and averages based on the scales of measurement. SPSS statistical software (version 15 IBM, New York, USA) was used for data analysis. Statistical significance was set at $p < 0.05$. Binary logistic regression including all factors with $p < 0.1$ in bivariate analyses was conducted to identify independent associations of clustering.

Ethical considerations

Ethics approval for this study was granted by the University of New South Wales, Australia (No. 14201). In addition, each of the individual cohorts had ethical approval for data and specimen collections.

Results

Samples

Three hundred and sixty eight samples collected between 1987 and 2012 met the inclusion criteria and were available from the individual cohorts. The majority of these (299, 81%) had a quantitative viral RNA load $> 1,000\text{IU/ml}$ (recorded at the time of sampling, prior to freezing for storage), while the remaining 69 (19%) only had a positive qualitative viral RNA test.

Generation of HCV amplicons and sequencing

Near full length HCV amplicons were successfully generated and sequenced for 190 (52%) samples. Partial sequences were generated from another 23 (6%) samples (at least one third of the genome) and 153 (42%) could not be sequenced. The average coverage from deep

sequencing was approximately 17,000 nucleotides per position. Two cohorts, namely Amsterdam Cohort Studies (ACSD) and the SuperMIX study (NET) had relatively poor results with regard to successful amplicon generation (Table 1). The samples from the former cohort had predominantly been collected in the 1980s and had likely degradation of viral RNA; and the latter had stored samples in Lithium heparin requiring decontamination with heparinase, which was partially successful with three full and five partial genomes recovered from the 18 samples received. The most common genotypes in the InC³ sample set were 1a, 3a and 2b respectively (Table 2).

Phylogenetic clusters recognized by the full length genomes

Of the 187 full length sequences used in the analysis, only genotypes 1a, 2b and 3a were present in adequate numbers to construct phylogenetic trees (n=94, n=15 and n=68, respectively) and for these subtypes 14, 4 and 11 clusters or pairs were identified respectively. Of the total dataset (n=177), a total of 78 sequences (44%) were grouped in either a pair (n=34) or a cluster (n=44).

A flow chart on sample selection is shown in supplementary material 2. The phylogenetic trees and identified clusters for genotypes 1a and 3a are shown in Figure 1 and supplementary material in Figure 3 respectively (bootstrap support is indicated at each node). Two additional clusters identified by the software for genotype 1a were later classified as non-clusters. In one pair, the envelope region of both sequences was missing, rendering the phylogenetic signal weak. The other pair did not appear as a cluster in any of the trees drawn using limited segments of the genome, casting doubt on the validity of the cluster. The separation of clusters by the geographic origin of the samples was striking. Of the 29 total clusters identified, 26 (90%) were with exclusively North American or Australian sequences, all with good bootstrap support (> 95% on most occasions). Within the North American sequences, the Canadian sequences (HEP) formed unique sub-clusters for both genotypes 1a and 3a from the samples from USA.

In the sensitivity analysis, when the MPD cut-off of 0.01 was relaxed up to 0.05 (5th percentile), there were no major changes in the clusters. For both genotypes, a few clusters identified as separate at MPD of 0.01 coalesced to form a single cluster at 0.05 (while still maintaining their geographical identity). When the MPD cut-off was raised to 0.1, cluster definitions became less reliable.

Clusters described here are only defined according to molecular similarity (more similar sequences within a cluster than compared to two random sequences within the tree). There was no circumstantial evidence for them to be recognized as transmission clusters (temporal or spatial proximity at time of infection). On the comparison of pairwise age differences since the time of infection within and outside clusters, sequences within clusters did not show a significant temporal proximity in infection. Detailed spatial proximity at the time of infection for those within a cluster could not be established as exact geographical locations could not be shared due to privacy laws in the United States, and inadequate spatial resolution from the data available for Canadian and Australian cohorts (one entire cohort was from a highly mobile population of individuals within a network of correctional centres).

To explore the robustness of the identified clusters, additional 47 full length genotype 3a sequences (all available full length 3a sequences with geographical metadata in the Los Alamos HCV database) and 244 genotype 1a sequences were introduced into the analysis. The selection for genotype 1a had 171 sequences from the United States and 73 sequences from other countries combined (Germany, Brazil, Switzerland, Japan). Most clusters recognized in the InC³ sequence analysis remained without any changes (9/14 and 8/11 for genotypes 1a and 3a respectively). None of the clusters were lost, thereby supporting the strong molecular relatedness of the clusters compared to overall variability within the tree. The clusters that varied had addition or removal of one to two sequences from a similar geographical region. A sequence from The Netherlands belonging to the InC³ dataset that did not cluster previously entered a cluster of German sequences from the Los Alamos database (there were no full length 1a sequences from The Netherlands in the public database and Germany was the nearest neighbour with sequences). Interestingly, despite the inclusion of sequences from a wide time interval (1991 – 2012), country of origin influenced clustering more than the age of sequences. Finally, after inclusion of the additional sequences, 55 extra clusters/pairs were identified of which 45 had sequences from the same country, while two had sequences from neighbouring regions (USA/Canada and Germany/Switzerland) (Figure 2).

Demographic and behavioural associations with clustering

This analysis was restricted to sequences from Australia and North America as only two sequences were available from Europe. Another three sequences were excluded due to non-availability of demographic data, making the total sample size 172. The characteristics of these subjects are summarized in Table 2.

Bivariate analyses for associations with clustering was performed: a) for the entire sample of 172 sequences, b) Australian and North American sequences separately, c) patients from East and West coast of North America, d) for individual cohorts with a sample size > 15 (HITS-p, HEP, UFO, BBA, ATAHHC).

In the analysis of the entire dataset, there were significant associations between clustering and older age (mean 29.1 years versus 24.8 years; mean difference 4.24 years, $p < 0.001$), continent of origin being Australia, ethnicity (Caucasians were less likely to be in a cluster, $p < 0.046$) and incarceration at baseline compared to recruitment from the community at baseline ($p < 0.028$) (Table 3). Other variables including: sex, level of education, most commonly injected drug at baseline, receiving opioid substitution therapy, HIV co-infection, sexual orientation (self-reported) and past history of incarceration, were not associated with clustering. Cohort-based analyses revealed a significant association for clustering and ethnicity (Caucasians were less likely than non-Caucasians to be in a cluster) in the HITS-p cohort ($p = 0.027$). Age, non-Caucasian ethnicity, continent of origin being Australia, incarceration at baseline and HIV co-infection were included in binary logistic regression analysis. The predictive capacity of the model increased from 55% (non-model) to 68% when these variables were included in the equation. The contribution of age ($p = 0.015$) and ethnicity ($p = 0.019$) remained significant when adjusted for other variables.

Comparison of full genome versus segments in phylogenetics

The phylogenetic trees and clusters of full-length sequences for genotypes 1a and 3a were compared to that of regions and concatenated regions as mentioned in methods to see if the full length phylogenetic signal had better resolution compared to segments of the genome. Most studies on HCV phylogenetics had used segments of the genome so far due to technical difficulties in generating full length genomes (see below). Some full-length genomes (9 of genotype 1a and 6 of genotype 3a) in our dataset had regions of ambiguous or absent reads (extending to up to 1000 nucleotides), due to incomplete sequence coverage. These were excluded from the analysis. The full genome trees were also re-drawn from this subset.

Except for one cluster in genotype 1a, all clusters identified by at least one of the sub-genomic regions were also present in the full genome tree (supplementary material 4). The number of clusters predicted by individual segments of the genome varied and was consistently less than that predicted by the full genome tree. The Core was the most restrictive region in this regard. However, the concatenated Core-partial (p) E2 (without HVR1) plus NS5B region had an equal capacity to predict clusters compared to the full genome for genotype 1a (supplementary material 4). By contrast, for genotype 3a this approach fell well short of the clusters identified using the full genome tree, as several clusters predicted by the NS3 and NS5A regions (plus the full genome) were not identified with the Core-pE2+NS5B tree. The phylogenetic relationships were observed with better resolution when the segment of the genome used for analysis was longer and closer to the full genome (Figure 3).

The Genbank accession numbers of sequences used in this manuscript are listed in supplementary material 5.

Discussion

This study utilised full length HCV sequences from early infection time points and geographically dispersed sources to infer molecular relatedness by phylogenetics. Three novel features of this analysis are noteworthy: firstly near full length HCV sequences were generated with an optimised high-throughput protocol for NGS; secondly, the analysis was restricted to viral sequences from incident infections thereby enabling detection of early viral traits; and thirdly, the project was a multi-national effort allowing analysis of a large number of sequences from nine separate prospective cohorts in four countries. The sequencing output of the project has more than doubled the number of publicly available, full length genotype 3a sequences globally, and established a large repository of 1a and 3a full length sequences from Australia.

In this study, the geographical origin of the sample strongly influenced clustering. Within the InC³ sequence dataset, the North American and Australian sequences sat apart from each other in clusters. This geographical bias in HCV phylogenetics has been proposed by previous studies that utilised selected segments of the genome (21–24). The findings here consolidate this observation by analysing full length genomes to incorporate the complete

phylogenetic signal. It is plausible that the virus is evolving separately in micro-epidemics between communities that have restricted population mixing.

There are no previous cross-continent studies for comparison of demographic associations for clustering noted in this study (between Australia and North America). However, studies within limited geographical areas have revealed some demographic associations. For instance, in Vancouver, Canada, an analysis using the Core-E2 region, demonstrated that out of 501 participants infected with genotypes 1a or 3a, 156 (31%) belonged to a cluster or a pair (clusters were defined with a MPD cut-off of 0.05 and bootstrap support of 90%) (25), with younger age (<40 years), HIV co-infection and syringe borrowing being associated with clustering. Another analysis from one of the participant cohorts (Networks study, NET) identified 26 clusters (of 2–7 subjects) from 244 HCV infections. Acute infection and genotype 3a were significantly associated with clustering (26). Finally, a detailed spatial-temporal analysis for transmission pair or clusters conducted in one of the cohorts in InC³, the HITS-p cohort, identified four likely transmission events of which three were associated with recent sharing and borrowing of needles (13). Unfortunately, several of these selected subjects/samples from both NET and HITS-p studies did not meet the inclusion criteria for the current study.

After adjusting for other variables, older age and non-Caucasian ethnicity remained both significant predictors for clustering in the current dataset. The mean age of participants (in both the clustering group and the remainder) was less than 30 years, but those in clusters were comparatively older – by 4.2 years on average, which contrasts with the finding in Vancouver described above. Although the multivariate analysis sought to identify independent associations with clustering, residual confounding with other variables remains plausible. For instance, the term “non-Caucasian ethnicity” encompassed a wide variety of ethnicities, including African American, Hispanic, Asian and Indigenous Australians. Analysis of each ethnic group separately was not feasible due to the limited numbers in each category. In the analysis of the individual cohorts, the association with ethnicity was significant only in the HITS-p cohort which had a relatively large number of Indigenous Australians who are over-represented among PWID and within prisons in Australia (27, 28). These trends are consistent with recent Australian epidemiological estimates (29, 30). There were only 8 HIV co-infections in the InC³ dataset analysed here, which limited the potential to observe any relationship of this co-factor with HCV clustering.

The availability of a large number of full length genomes made it possible to carry out a comparison against utilising one or more segments of the genome (as was the practice in many previous studies on HCV phylogenetics) (31). The most popular regions in this regard are the Core, E1 and E2 glycoproteins (with or without HVR1 of E2), and the NS5B region (15, 25, 32, 33). It has been recently shown that tree topology and clustering results of phylogenetic analysis change with the segment of the genome used for analysis (15). These authors used 11 different combinations of Core, E1, E2 (with or without HVR1) and NS5B regions (either as single regions or concatenated combinations) for a comparative analysis. The Core was considered unsuitable for clustering analysis due to the fewer number of clusters identified, while concatenated regions of Core-E2 without HVR1 and Core-pE2 (without HVR1) plus NS5B identified a larger number of clusters. This study showed that,

the concatenated Core-E2+NS5B region was comparable to the full genome for genotype 1a, but not for genotype 3a.

The clinical relevance of these findings of the divergence between continents lies in relation to both neutralising antibody and T cell epitopes of HCV. If the mutations that define the uniqueness of clusters based on geographical region abrogate recognition by host T or B cell responses then this will have implications for the worldwide success of a prophylactic vaccine. The analysis presented here indicates that the genetic variations driving clustering are mostly accounted for by the Envelope (E1 and E2) and NS5B regions taken together (plus NS3 for genotype 3a). This may be important as all broadly neutralizing antibody epitopes discovered to-date in the HCV genome are located in the E1 and E2 glycoproteins, whereas T cell epitopes predominate in the non-structural proteins. It will also be important to understand whether resistance-associated variations relevant to the new directly acting antiviral (DAA) therapies are selected at the population level and would thereby influence antiviral treatment decisions, as well as treatment as prevention strategies. The NGS data generated from this project is being examined to answer these questions.

The observations in this study are applicable to North America and Australia as sequences could not be extracted from many European samples. The samples that failed in genome amplification are unlikely to bias the conclusions as they were uniformly distributed across the Australian and North American cohorts. PCR bias is a recognized limitation of this study. Methodologies that eliminate this bias (e.g. RNAseq approach) were only successful in making full length HCV amplicons when the viral load was above one million IUml⁻¹ (results not shown). Therefore using this approach (which was abandoned after a pilot test) for this study would have increased the failure rate as many early infection samples had a low viral load. The nested PCR method used in this study made amplicons from samples with a viral load as low as 14,000 IUml⁻¹ (17). While this manuscript details a consensus level analysis, deep sequencing data from the project had been used for Bayesian evolutionary analysis, identifying frequencies of resistant association variants within quasi-species and predicting epistatic relationships of the full HCV genome. These results are beyond the scope of the aims of this manuscript.

In conclusion, this is the largest molecular epidemiological and phylogenetic study of incident HCV infections conducted worldwide. Pooling of viraemic samples, as well as demographic and behavioural data from two continents (Australia and North America), combined with the technological advantage of NGS analysis of full length HCV genomes, has revealed that phylogenetic clustering was strongly influenced by the geographical origin of the sequences. This suggests that the HCV epidemic is probably evolving independently in micro-epidemics in geographically separated populations.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors acknowledge the support of the Ramaciotti Centre for Genomics, UNSW, Australia in conducting the next generation sequencing. We are grateful to the study participants and to field and laboratory staff (Geng Zang, Rachel Bouchard, Elisabeth Deschenes, Marie-Eve Turcotte, Yanie Nan, Mélanie Gagnon, Isabelle Boisvert, Nathalie Bédard, Rachel Sack Davis, Paul Dietze, Peter Higgs, Campbell Aitken) and key co-investigators (Elise Roy and Didier Jutras-Aswad).

Financial support: The infrastructure for sharing of data and specimens in InC³ was funded by the National Institute on Drug Abuse (NIDA) R01DA031056. The content of this manuscript is solely the responsibility of the authors and does not necessarily represent the official views of NIDA or the National Institutes of Health (NIH). Research support for the InC³ cohorts includes: the Netherlands National Institute for Public Health and the Environment to the Amsterdam Cohort Study (ACS); Baltimore Before and After Study (BBA) - National Institutes of Health (NIH U19 AI088791); Boston Acute HCV Study: transmission, immunity, outcomes network (BAHSTION) is funded under NIH NIAID U19 AI066345; ATACH - NIDA R01 DA 15999-01; HITS-p - National Health and Medical Research Council of Australia (NHMRC) - Project No. 222887, Partnership No. 1016351, Program Nos. 510488 and 1053206; HITS-c - UNSW Hepatitis C Vaccine Initiative and NHMRC Project Grant No. 630483; Networks/MIX - NHMRC Project Grants Nos. 331312 and 545891 and the Victorian Operational Infrastructure Support Programme (Department of Health, Victoria, Australia); HepCo - the Canadian Institutes of Health Research (MOP-103138 and MOP-106468; Réseau SIDA Maladies Infectieuses, Fonds de la Recherche du Québec-Santé); UFO - NIH R01 DA016017. The following InC³ investigators are funded by research fellowships: NHMRC, Australia - JG, MH, LM, GD, ARL; Fonds de la Recherche du Québec, Canada - JB and NHS.

References

1. Hajarizadeh B, Grebely J, Dore GJ. Epidemiology and natural history of HCV infection. *Nat Rev Gastroenterol Hepatol*. 2013; 10(9):553–62. [PubMed: 23817321]
2. Fittig J, Swaminathan M, Murrill CS, Kaplan JE. Global epidemiology of HIV. *Infect Dis Clin North Am*. 2014; 28(3):323–37. [PubMed: 25151559]
3. Murray CJ, Ortblad KF, Guinovart C, et al. Global, regional, and national incidence and mortality for HIV, tuberculosis, and malaria during 1990–2013: a systematic analysis for the Global Burden of Disease Study 2013. *Lancet*. 2014; 384(9947):1005–70. [PubMed: 25059949]
4. Gower E, Estes C, Blach S, Razavi-Shearer K, Razavi H. Global epidemiology and genotype distribution of the hepatitis C virus infection. *J Hepatol*. 2014; 61(1 Suppl):S45–57. [PubMed: 25086286]
5. Mehta SH, Astemborski J, Kirk GD, et al. Changes in blood-borne infection risk among injection drug users. *J Infect Dis*. 2011; 203(5):587–94. [PubMed: 21282191]
6. Page K, Hahn JA, Evans J, et al. Acute hepatitis C virus infection in young adult injection drug users: a prospective study of incident infection, resolution, and reinfection. *J Infect Dis*. 2009; 200(8):1216–26. [PubMed: 19764883]
7. Cox AL, Netski DM, Mosbruger T, et al. Prospective evaluation of community-acquired acute-phase hepatitis C virus infection. *Clinical infectious diseases: an official publication of the Infectious Diseases Society of America*. 2005; 40(7):951–8. [PubMed: 15824985]
8. Hagan H, Thiede H, Des Jarlais DC. Hepatitis C virus infection among injection drug users: survival analysis of time to seroconversion. *Epidemiology*. 2004; 15(5):543–9. [PubMed: 15308953]
9. Maher L, Jalaludin B, Chant KG, et al. Incidence and risk factors for hepatitis C seroconversion in injecting drug users in Australia. *Addiction*. 2006; 101(10):1499–508. [PubMed: 16968352]
10. Hahn JA, Page-Shafer K, Lum PJ, et al. Hepatitis C virus seroconversion among young injection drug users: relationships and risks. *J Infect Dis*. 2002; 186(11):1558–64. [PubMed: 12447730]
11. Roy E, Alary M, Morissette C, et al. High hepatitis C virus prevalence and incidence among Canadian intravenous drug users. *Int J STD AIDS*. 2007; 18(1):23–7. [PubMed: 17326858]
12. Grebely J, Morris MD, Rice TM, et al. Cohort profile: the international collaboration of incident HIV and hepatitis C in injecting cohorts (InC3) study. *International journal of epidemiology*. 2013; 42(6):1649–59. [PubMed: 23203695]
13. Bretana NA, Boelen L, Bull R, et al. Transmission of Hepatitis C Virus among Prisoners, Australia, 2005–2012. *Emerging infectious diseases*. 2015; 21(5):765–74. [PubMed: 25897788]

14. Bull RA, Luciani F, McElroy K, et al. Sequential bottlenecks drive viral evolution in early acute hepatitis C virus infection. *PLoS pathogens*. 2011; 7(9):e1002243. [PubMed: 21912520]
15. Lamoury FM, Jacka B, Bartlett S, et al. The Influence of Hepatitis C Virus Genetic Region on Phylogenetic Clustering Analysis. *PloS one*. 2015; 10(7):e0131437. [PubMed: 26192190]
16. Liu L, Fisher BE, Dowd KA, Astemborski J, Cox AL, Ray SC. Acceleration of hepatitis C virus envelope evolution in humans is consistent with progressive humoral immune selection during the transition from acute to chronic infection. *Journal of virology*. 2010; 84(10):5067–77. [PubMed: 20200239]
17. Bull RA, Eltahla AA, Rodrigo C, et al. A method for near full-length amplification and sequencing for six hepatitis C virus genotypes. *BMC genomics*. 2016; 17(1):247. [PubMed: 26988550]
18. Hajarizadeh B, Grady B, Page K, et al. Patterns of hepatitis C virus RNA levels during acute infection: the InC3 study. *PloS one*. 2015; 10(4):e0122232. [PubMed: 25837807]
19. Johnson ML, Navanukraw C, Grazul-Bilska AT, Reynolds LP, Redmer DA. Heparinase treatment of RNA before quantitative real-time RT-PCR. *BioTechniques*. 2003; 35(6):1140–2. 4. [PubMed: 14682046]
20. Prosperi MC, Ciccozzi M, Fanti I, et al. A novel methodology for large-scale phylogeny partition. *Nature communications*. 2011; 2:321.
21. Fu Y, Qin W, Cao H, et al. HCV 6a prevalence in Guangdong province had the origin from Vietnam and recent dissemination to other regions of China: phylogeographic analyses. *PloS one*. 2012; 7(1):e28006. [PubMed: 22253686]
22. Akkarathamrongsin S, Hacharoen P, Tangkijvanich P, et al. Molecular epidemiology and genetic history of hepatitis C virus subtype 3a infection in Thailand. *Intervirology*. 2013; 56(5):284–94. [PubMed: 23838334]
23. Iles JC, Raghwanji J, Harrison GL, et al. Phylogeography and epidemic history of hepatitis C virus genotype 4 in Africa. *Virology*. 2014; 464–465:233–43.
24. Li C, Yuan M, Lu L, et al. The genetic diversity and evolutionary history of hepatitis C virus in Vietnam. *Virology*. 2014; 468–470:197–206.
25. Jacka B, Applegate T, Krajden M, et al. Phylogenetic clustering of hepatitis C virus among people who inject drugs in Vancouver, Canada. *Hepatology*. 2014; 60(5):1571–80. [PubMed: 25042607]
26. Sacks-Davis R, Daraganova G, Aitken C, et al. Hepatitis C virus phylogenetic clustering is associated with the social-injecting network in a cohort of people who inject drugs. *PloS one*. 2012; 7(10):e47335. [PubMed: 23110068]
27. Butler, T.; Lim, D.; Callander, D. National Prison Entrants' Bloodborne Virus and Risk Behaviour Survey Report 2004, 2007, and 2010: National Report. Sydney: Kirby Institute; 2011. Prevalence of HIV, HBV, HCV, and Risk Behaviours among Australian Prison Entrants.
28. Iversen, J.; Maher, L. Australian Needle and Syringe Program National Data Report 1995–2014. Sydney: The Kirby Institute, the University of New South Wales; 2015.
29. The Kirby Institute. Bloodborne Viral and Sexually Transmitted Infections in Aboriginal and Torres Strait Islander People: Surveillance and Evaluation Report 2012. Sydney: Kirby Institute, the University of New South Wales; 2012.
30. Treloar C, Jackson C, Gray R, et al. Care and treatment of hepatitis C among Aboriginal people in New South Wales, Australia: implications for the implementation of new treatments. *Ethnicity & health*. 2015; 1–19. [PubMed: 25372242]
31. Messina JP, Humphreys I, Flaxman A, et al. Global distribution and prevalence of hepatitis C virus genotypes. *Hepatology*. 2015; 61(1):77–87. [PubMed: 25069599]
32. Sato M, Maekawa S, Komatsu N, et al. Deep sequencing and phylogenetic analysis of variants resistant to interferon-based protease inhibitor therapy in chronic hepatitis induced by genotype 1b hepatitis C virus. *Journal of virology*. 2015; 89(11):6105–16. [PubMed: 25810555]
33. Choudhary MC, Natarajan V, Pandey P, et al. Identification of Indian sub-continent as hotspot for HCV genotype 3a origin by Bayesian evolutionary reconstruction. *Infection, genetics and evolution: journal of molecular epidemiology and evolutionary genetics in infectious diseases*. 2014; 28:87–94.

Appendix: List of Abbreviations

HCV

Hepatitis C virus

InC³

International Collaboration of Incident HIV and Hepatitis C in Injecting Cohorts

HIV

Human immunodeficiency virus

PWID

People who inject drugs

RNA

Ribonucleic acid

IU

International Units

EDTA

Ethylenediaminetetraacetic acid

ACD

Acid citrate dextrose

PCR

Polymerase chain reaction

PBS

Phosphate saline buffer

NGS

Next generation sequencing

BWA

Burrows-Wheeler Alignment tool

MUSCLE

MUltiple Sequence Comparison by Log-Expectation

RAxML

Randomized Axelerated Maximum Likelihood

GTR

Generalized time reversible

MPD

Mean Patristic Distance

ACSD

Amsterdam cohort studies

NET

Networks and SuperMIX study, Melbourne Australia

ATAHC/AUS

Australian Trial in Acute Hepatitis C, Sydney Australia

BAHSTION (BASMGGH, BASLSH)

Boston Acute HCV Study: transmission, immunity, outcomes network (BAHSTION) community and prison studies

BBAASH (BBA)

Baltimore Before and After Acute Study in Hepatitis

HITS-p

Hepatitis C Incidence and Transmission – prison study, Sydney, Australia

HITS-c

Hepatitis C Incidence and Transmission – community study, Sydney, Australia

HEP

HEPco study, Montreal, Canada

UFO

UFO study, San Francisco, California, USA

PWID

People who inject drugs

Figure 1. Maximum likelihood consensus tree (of 1000 bootstrap replicates) of genotype 1a full length genomes from the InC³ study (n=94)

Each subject is name has following components a) cohort abbreviation, b) subject code, c) HCV subtype. Cohort abbreviations: ACSD; Amsterdam Cohort Studies, AUS; Australian Trial in Acute Hepatitis C, BBAASH; Baltimore Before and After Acute Study in Hepatitis, BASMGH/LSH; Boston Acute HCV Study, HEP; HEPco Study (Canada), HITS-p; Hepatitis C Incidence and Transmission in Prison Study (Australia), HITS-c: Hepatitis C Incidence and Transmission in Community Study (Australia), NET: SuperMIX Study (Australia), UFO; UFO Study, (San Francisco, United States). Colour code; green – clusters with sequences exclusively from North America, blue - clusters with sequences exclusively from Australia, red – clusters with sequences from both continents.

Figure 2. Maximum likelihood consensus tree (of 1000 bootstrap replicates) of genotype 1a full-length genomes from InC³ study and public databases (n=341). Sequences that formed clusters are indicated by coloured tips with each colour corresponding a geographical region/country (The tree is collapsed to show only nodes with a bootstrap support over 70%).

Figure 3.

a. Maximum likelihood consensus tree (of 1000 bootstrap replicates) of genotype 1a: Panel A. core region (n=85), Panel B. E2 region (n=85), Panel C. Core-E2 (without HVR1) + NS5B concatenated region (n=85), Panel D. full genome (n=85). The trees are collapsed to show only nodes with bootstrap support over 70%. Each subject is identified via the following components: a) cohort abbreviation, b) subject code, c) HCV subtype. Cohort abbreviations: ACSD; Amsterdam cohort studies, AUS; Australian Trial in Acute Hepatitis C, BBAASH; Baltimore Before and After Acute Study in Hepatitis, BASMGH/LSH; Boston Acute HCV Study, HEP; HEPco study (Canada), HITS-p; Hepatitis C Incidence and Transmission in Prison study (Australia), HITS-c: Hepatitis C Incidence and Transmission in Community Study (Australia), NET: SuperMIX study (Australia), UFO; UFO study, (San Francisco, United States).

Table 1

Number of received and sequenced samples from individual cohorts

Cohort	Country	Samples received	Successful FG amplification *	Successful PG amplification **
Amsterdam Cohort studies	Netherlands	21	2	1
Australian Trial in Acute Hepatitis C	Australia	50	21	6
Boston Acute HCV Study: transmission, immunity, outcomes network (BAHSTION) community and prison studies	USA	22	8	3
Baltimore Before and After Acute Study in Hepatitis	USA	69	40	3
Hepatitis C Incidence and Transmission Study - prison	Australia	49	38	1
Hepatitis C Incidence and Transmission Study - community	Australia	12	11	0
HepCo Study	Canada	24	17	0
SuperMIX Study	Australia	18	3	5
UFO Study	USA	103	50	4
Total		368	190	23

* FG-Full Genome,

** PG – Partial Genome

Table 2

Characteristics of participants from whom the full length virus was successfully sequenced and used for cluster analysis (n=172)

Characteristic	Number/Mean $\pm$ SD (%)
Continent of origin	
Australia	63 (36.6)
North America	109 (63.4)
Age (years)	26.7 $\pm$ 8.05
Genotype	
1a	91 (52.9)
2b	14 (8.0)
3a	67 (38.3)
Other	17 (8.8)
Male (vs. female)	110 (64.0)
History of injecting drug use	171 (97.7)
Caucasian ethnicity (vs. others)	138 (80.0)
Incarceration at baseline	39 (22.7)
Past history of incarceration (n=132)	59 (44.7)
Level of education (n=124)	
Did not complete high school	55 (44.3)
Completed high school	41 (33.1)
Some post high school education	28 (22.6)
Sexual orientation (n=59)	
Heterosexual	51 (86.4)
Gay or bisexual	8 (13.6)
Most commonly injected drug at recruitment (n = 86)	
Heroin	46 (53.5)
Methamphetamine	24 (27.9)
Cocaine	6 (6.9)
Other opiates	7 (8.1)
Heroin and cocaine	3 (3.5)
Opioid substitution therapy at recruitment (n 98)	27 (27.6)
HIV co-infection	8 (4.6)

Table 3

Demographic and behavioural associations with phylogenetic clustering*

Characteristic**	Total (%) (n - 172)	Within pair/cluster (n-75)	Not within pair/cluster (n-97)	Unadjusted Odd ratio (95% CI)	P value	Adjusted Odd ratio (95% CI)	P value
Age (odd ratio calculated per 1 year increment)	NA	NA	NA	1.07 (1.02–1.12)	0.001	1.06 (1.01–1.12)	0.015
Male vs. female	111 (64%)	50 (67%)	61 (63%)	1.18 (0.63–2.22)	0.607	-	-
Australia vs. North America	63 (37%)	39 (52%)	24 (25%)	3.29 (1.73–6.29)	< 0.001	0.57 (0.23–1.41)	0.220
Non Caucasian vs. Caucasian	34 (20%)	20 (27%)	14 (14%)	2.23 (1.04–4.79)	0.046	2.94 (1.19–7.26)	0.019
Incarceration at baseline (yes vs. no)	39 (23%)	23 (31%)	16 (16%)	2.24 (1.08–4.63)	0.028	0.81 (0.31–2.16)	0.678
Genotype 1 vs others	86 (50%)	38 (51%)	48 (49%)	1.05 (0.57–1.92)	0.878	-	-
HIV co-infection (yes vs. no)	8 (5%)	6 (8%)	2 (2%)	4.13 (0.81–21.1)	0.067	0.64 (0.1–3.98)	0.633

Percentages indicate column percentages.

** Characteristics with > 10% of missing data are not shown, none of them showed significant associations.

* phylogenetic analysis completed with full genomes, n – 175, 82 (45.5%) sequences were found to be in clusters, NA–Not applicable