Historical trends in the hepatitis C virus epidemics in North America and Australia

Chaturaka Rodrigo¹, Auda A. Eltahla¹, Rowena A. Bull¹, Jason Grebely², Gregory J. Dore², Tanya Applegate², Kimberly Page³, Julie Bruneau⁴, Meghan D. Morris⁵, Andrea L. Cox⁶, William Osburn⁶, Arthur Y. Kim⁷, Janke Schinkel⁸, Naglaa H. Shoukry⁴, Georg M. Lauer⁷, Lisa Maher⁷, Margaret Hellard⁹, Maria Prins⁸¹⁰, Chris Estes¹¹, Homie Razavi¹¹, Andrew R. Lloyd¹, and Fabio Luciani¹, on behalf of the InC3 Study Group

¹School of Medical Sciences, Faculty of Medicine, UNSW Australia, Sydney, NSW, Australia

²The Kirby Institute, UNSW Australia, Sydney, NSW, Australia

³University of New Mexico, Albuquerque, New Mexico, USA

⁴CRCHUM, Université de Montréal, Montreal, QC, Canada

⁵Department of Epidemiology and Biostatistics, University of California, San Francisco, San Francisco, CA, USA

⁶Department of Medicine, Johns Hopkins Medical Institutions, Baltimore, MD, USA

⁷Harvard Medical School, Boston, MA, USA

⁸Department of Internal Medicine, Division of Infectious Diseases, Tropical Medicine and AIDS, Center for Infection and Immunity Amsterdam, Academic Medical Center, Meibergdreef, Amsterdam, The Netherlands

⁹Burnet Institute, Melbourne, VIC, Australia

¹⁰GGD Public Health Service of Amsterdam, Amsterdam, The Netherlands

© The Author 2016. Published by Oxford University Press for the Infectious Diseases Society of America. All rights reserved. For permissions, e-mail journals.permissions@oup.com.
Center for Disease Analysis, Louisville, CO, USA

Correspondence: Andrew Lloyd AM, MBBS, MD, FRACP, Infection and Inflammation Research Centre, School of Medical Sciences, University of New South Wales, 2052, Sydney, NSW, Australia, Email: a.lloyd@unsw.edu.au, Tel:+61 2 9385 2534, Fax:+61 2 9385 1389

Footnote Page

Financial support:

The infrastructure for sharing of data and specimens in InC\(^2\) 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 InC\(^2\) 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; ATAHC – 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 InC\(^2\) investigators are funded by research fellowships: NHMRC, Australia – JG, MH, LM, GD, ARL; Fonds de la Recherche du Québec-Santé, Canada – JB and NHS.

Conflicts of interest: The authors have nothing to disclose in relation to this work
Abstract

Background

Bayesian evolutionary analysis (coalescent analysis) based on genetic sequences has been used to describe the origins and spread of rapidly mutating RNA viruses such as Influenza, Ebola, HIV and hepatitis C virus (HCV).

Methods

Full length subtype 1a and 3a sequences from early HCV infections from the International Collaborative of Incident HIV and Hepatitis C in Injecting Cohorts (InC3), as well as from public databases from a time window of 1977–2012, were used in a coalescent analysis with BEAST software to estimate the origin and progression of the HCV epidemic in Australia and North America. Convergent temporal trends were sought via independent epidemiological modelling.

Results

The epidemic of subtype 3a had more recent origins (around 1950) than subtype 1a (around 1920) in both continents. In both modelling approaches and in both continents, the epidemics underwent exponential growth between 1955 and 1975, which then stabilized in the late 20th century.

Conclusions

Historical events that fuelled the emergence and spread of injecting drug use, such as the advent of intravenous medical therapies and devices, and growth in the heroin trade, as well as population mixing during armed conflicts, are likely drivers for the cross-continental spread of the HCV epidemics.
Background (3487 words - allowed 3500)

Hepatitis C (HCV) is a chronic viral infection predominantly spread via blood-to-blood contamination during injecting drug use, with outbreaks also attributed to non-sterile medical practices, unscreened blood transfusions and sexual transmission amongst men-who-have-sex-with-men (MSM) [1-5]. The HCV genera is divided into seven genotypes and 67 subtypes [6]. The subtypes that have a global presence are; 1a, 1b, 2a, 2b and 3a[7]. Given the limited repertoire of modes of transmission, HCV provides an ideal opportunity to study how historical trends in public health measures, risk behaviours (e.g. injecting drug use), and major socioeconomic changes such as wars and migration patterns may have shaped the global epidemic.

Viral sequences and phylogenetic analysis have been used previously to infer evolutionary dynamics of Ebola and HIV, as well as HCV epidemics [1, 8]. As these RNA viruses feature high mutation rates, the evolutionary and epidemiological timescales are similar. These viruses are therefore termed “measurably evolving populations” (MEP) [9]. In evolutionary analysis, the extent to which the sampled genealogy (i.e the ancestral lineage of the virus) represents the true genealogy of the population (and hence the accuracy of estimated parameters) depends on several factors: firstly, the similarity between an experimentally-reconstructed and the true genealogy; secondly, the length of the viral sequences (phylogenetic signal); thirdly the model of evolution used; and finally the phylogenetic techniques used to reconstruct the genealogy [9].

We have recently established the largest repository of early infection, full length, HCV genome sequences available globally[10]. The contributions to this repository came from nine prospective cohorts constituting the International Collaboration of Incident HIV and Hepatitis C in Injecting Cohorts (InC3) [11]. Phylogenetic analysis of these sequences has shown that North American and Australian sequences clustered separately from each other, suggesting that HCV largely evolves
separately in micro-epidemics in geographically separate communities [10]. Here, we explore this theory further using Bayesian evolutionary analysis to model the spread of the subtypes 1a and 3a HCV epidemics in North America and Australia in the 20th century to understand the similarities and differences in the epidemics. This is the first description of historical trends in the HCV epidemic for subtypes 1a and 3a in Australia, and for subtype 3a in North America using full length genomes to maximise sensitivity in phylogenetic analyses.

**Methods**

**Selection of sequences**

From 368 early incident samples from InC3, 190 full length and 23 partial length HCV sequences were generated using a protocol of reverse transcription and nested PCR [12]. All amplicons were subjected to next generation sequencing (Miseq, Illumina, California, USA). The sequencing output had an average coverage of 17,000 reads per site. The consensus sequences for each sample were generated using the bioinformatics pipeline previously described[10]. The current analysis utilized all of the consensus sequences originating from samples collected in Australia and North America (Canada and United States) (n=161) for subtypes 1a and 3a (4 datasets). These incident infection samples had been collected between 1998 and 2012. Public databases (Los Alamos HCV sequence database and Genbank) were also searched for full length HCV sequences originating from North America or Australia with a known date of sampling. For subtype 3a, there were no additional full length sequences for Australia, but an additional six sequences were found for North America (United States – 2, Canada – 4). By contrast, there were more than 700 full length sequences were available for subtype 1a from the United States. From these additional sequences, a randomly selected subset was included to represent the time window from 1977 – 1998 to supplement the more recent InC3 sequences. The composition of the final datasets is shown in table 1 and in supplementary material (S1). Each dataset was aligned with MUSCLE (MUltiple Sequence Comparison by Log-Expectation) algorithm.
Ethics approval for the InC3 sequencing project was granted by the UNSW Australia Human Research Ethics Committee (No. 14201), and included local ethical approval at each contributing site.

**Coalescent analysis**

The analysis was completed using Bayesian evolutionary methods with BEAST version 1.8 [13]. The alignments were analysed in MEGA (version 6.0) to identify the best substitution model to explain the observed mutations [14]. The generalised time reversible model (GTR+G+I) was selected as the best model. The most suitable clock model for the analysis was selected after running the analysis separately using strict, lognormal relaxed, and exponential relaxed models for two datasets of genotype 1a and 3a. The lognormal relaxed clock was selected as the best model from marginal likelihood analysis conducted by calculating posterior simulation-based analogue of Akaike's Information Criterion (AICM) (data not shown)[15, 16]. Previously estimated substitution rates for the full genome by Gray et al. and Choudhary et al. for subtypes 1a and 3a respectively, were used as external priors in this analysis [15, 17]. The SRD06 model was used to estimate relative substitution rates, α parameter and the proportion of invariant sites for the first two codon positions and the third codon position separately. The Bayesian skyline plot (BSP) approach was used to infer population dynamics of the epidemic [18].

The simulations were run with a Monte Carlo Markov Chain (MCMC) until the chain entered an area of high probability for each parameter. For each dataset, the MCMC was run 50 – 500 million times for convergence. All estimates were drawn from a pool with an ESS greater than 200. The MCMC chains were visualized by TRACER software[19]. The tree log file was annotated using the Tree Annotator software [18] and the maximum clade credibility tree was visualized with the Figtree application (version 1.4.2). Bayesian skyline plots were generated from the BEAST output to visualize the changes in the effective population size since the time of the most recent common ancestor (tMRCA) to the most recent sampling time point.
Epidemiological model

A previously developed epidemiological model was applied to estimate the time course of changes in the incidence and prevalence of HCV infection [20, 21]. The Markov model (implemented in Microsoft Excel (Redmond, WA)) was used to predict the HCV-infected population, and the disease progression from 1950 to 2012. The average number of new infections per year was estimated using the known prevalence estimates in the United States [22, 23] and Australia (see below) after taking into account the trends in all-cause mortality, excess mortality among people who inject drugs, estimated liver related deaths and the impact of treatment. The change in incidence was estimated by distributing the average number of new infections between 1950 and the year of known prevalence to match the age distribution of the known prevalence. Uncertainty and sensitivity analyses were also performed utilising Crystal Ball, an Excel add-in by Oracle. Beta-PERT distributions were used to model uncertainty associated with all inputs. The model was simulated using a Monte-Carlo approach to determine the 95% uncertainty interval (UI) for prevalence [20, 21].

The epidemiological estimates of HCV prevalence by age and gender [22-24] for the United States were extracted from the National Health and Nutrition Examination Survey (NHANES), the US Centers for Disease Control and Prevention (CDC) surveillance data between 1997 and 2007, and the number of patients cured by antiviral therapy reported by publications between 1997 and 2007 [20]. The model was also calibrated to account for populations not captured by the NHANES in United States, such as incarcerated and homeless people. For Canada, data reported by Public Health Agency Canada (PHAC) on HCV notifications between 1998 and 2011 were used to calibrate the model [25]. For Australia, data from the National Health Surveillance system between 1995 and 2013 were used [26]. These epidemiological estimates of prevalence were then compared to the changes of effective population size over time inferred by the Bayesian skyline model. The epidemiological model estimates the prevalence of HCV as the total infected population over time in a given region, while the Bayesian estimates represent the changes in effective population size.
which is a simulated population based on the available sequences. These two quantities are not identical but the trends in these two estimates may be compared [9].

Results

Coalescent analysis

The evolutionary analysis aimed to identify the origins of the subtype 1a and 3a epidemics in both North America and Australia and the historical progression since the origin. Results are summarized in Table 2.

The subtype 1a epidemic was estimated to have originated at the beginning of the 20th century, around 1915 (95% high posterior density (HPD): 1863-1961) and 1925 (95% HPD: 1914 – 1936) for Australia and North America respectively. HPD interval for the 1a epidemic in Australia was wider, due to the lack of availability of sequences dated closer to the origin of the epidemic compared to other datasets. The subtype 3a epidemic had more recent origins than the subtype 1a epidemic in both continents. The Australian 3a epidemic was estimated to have originated around 1953 (95% HPD: 1934 – 1968) and the North American epidemic was estimated to have originated around 1949 (95% HPD: 1937-1960). The estimated substitution rates for genotypes 1a and 3a in both continents (Australia: 1.13 and 1.19 X 10^{-3}, North America: 1.07 and 1.16 X 10^{-3}) were similar, despite the difference in genotype-specific prior rates (1.48 X 10^{-3} for genotype 1a and 1.07 X 10^{-3} for genotype 3a).

History of the epidemics in Australia and North America

Reconstructions of the epidemics for each subtype and continent are shown on the Bayesian skyline plots in Figure 1. Following the origin of the epidemic of subtype 1a in the early 20th century in North America, the effective population size remained low until 1955. Thereafter, two phases of rapid growth occurred between 1955 and 1975. The effective population size stabilized thereafter and declined slightly after the 1990s. The subtype 1a epidemic for Australia also showed a period of
rapid growth between 1955 and 1975. However, the effective population size was in decline from the 1980s, with a faster drop observed after year 2000.

By contrast, the subtype 3a epidemics in both continents had more recent origins compared to subtype 1a. The 3a epidemic in North America showed an exponential increase in the period between 1955 and 1970. The numbers started to decline slightly after 1995. In Australia, the rapid rise in infections happened between 1960 and 1975, before stabilization through to the recent era. The maximum clade credibility trees for each dataset are given in supplementary material (S2).

**Comparison of coalescent projections with epidemiological model projections**

The BEAST skyline output for each dataset was superimposed on data derived from the epidemiological model estimates derived using the previously published system dynamic modelling framework [20, 21]. Comparison of the trends reveals considerable similarities with regard to the timing of the exponential increases in the effective population size with the Bayesian phylogenetic analysis and the size of the HCV-infected population from epidemiological back projections for both genotypes and continents (Figure 1).

**Discussion**

These convergent virological and epidemiological analyses have allowed tracing of historical trends in the cross-continental spread of HCV subtype 1a and 3a epidemics in North America and Australia, and identified dramatic increases in the relative size of the epidemics in the mid-20th century.

As the HCV genome features widely varied evolutionary rates, both within infected hosts, across infected populations, and between different regions within the genome, assumptions regarding these rates have the potential to influence the outcome of the complex statistical methods applied in the coalescent analysis described here. Accordingly, for the present analysis, the substitution rates for the full genome estimated by Gray et al. were utilised as priors [17]. It should be noted that since most sequences used in our study were from acute infections, they could plausibly have a higher
substitution rate in the Envelope region, as the virus is recognised to mutate frequently in this region as it adapts for efficient entry and to the immune responses within the new host. However, as the full length genome was used, the averaged substitution rate across the genome is not likely to vary significantly from the estimates used in the previous study [17].

The tMRCA estimates for both North American and Australian 1a sequences were in the early 20th century, suggesting the virus was present in both continents at the beginning of the century.

Previous studies have estimated the tMRCA for subtype 1a in other parts of the world using limited segments of the genome. For instance, a previous global evolutionary analysis of subtype 1a using the NS5B region of sequences obtained from 21 countries, placed the tMRCA at a similar time point and with a Bayesian skyline plot showing a similar pattern to that reported here for North America [1]. However this study included a disproportionate number of North American sequences (399 sequences vs. 443 sequences from 20 other countries combined). Regarding regional studies, using Core-E1 sequences from Vietnam, Li et al. placed the tMRCA of subtype 1a around 1928 in South-East Asia. By contrast, two studies from South America using 1a NS5B sequences placed the tMRCA for Brazilian sequences between 1950 and 1979. Even after adjusting for the slower rate of evolution of NS5B, this estimate is more recent than estimates of the origin of the virus in other continents. Taken together, these data suggest that in Australia, North America and South-East Asia, subtype 1a originated at the beginning of the 20th century with a more recent introduction into South America.

The epidemic of genotype 3a has also been examined in previous studies [4, 15, 27, 28], including two studies which utilized full length genomes [15, 28]. However, for both of these previous studies, less than 50 full length 3a sequences were available from across the globe with very few, or no, sequences from United States and Australia. With the InC3 sequencing project, the number of full length subtype 3a sequences available has been more than doubled (67 new sequences added). The current dataset now allows tracing of the global spread of subtype 3a infections by provision of previously missing information from North America and Australia and has placed the tMRCA around
1950 in both continents. Li et al. used 42 sequences of eight different genotype 3 subtypes and estimated the tMRCA to be around 1934 (95% HPD: 1915 – 1949) for subtype 3a [28]. Choudhry et al., also constructed a time scaled phylogeny of 55 subtype 3a full length sequences obtained from public databases with a global distribution, and estimated the tMRCA to be approximately 1913 (95% HPD: 1859 – 1944) [15]. Interestingly, in both of these previous studies, the first sequences to branch out from the root were from India. In the latter study, the only sequence from Australia, and three sequences from United States, all branched out around 1950, as did other sequences from Denmark, Italy, China and New Zealand. Two other studies from South America that used partial 3a sequences from Brazil, placed the tMRCA around 1967 – 1979, which is considerably more recent [4, 27].

Comparison of the trends predicted by the coalescent analysis of phylogenetic data and the epidemiological model estimates were remarkably similar with regard to the rapid rise of infections from the late 1950s onwards and the later stabilization of the epidemics which serves as validation for the predictions by two entirely different methods. It should be noted however, that both datasets are limited by the first available raw data points (1977 and 1989 for the 1a and 3a viral sequences respectively, and 1997 for the epidemiological data). Having more sequences spread across the time period of interest allows better estimation of parameters including substitution rate. For example, the subtype 1a analysis in Australia, where the distance between the predicted root and the oldest available sequence is the largest, the HPD intervals around tMRCA are also large. The robustness of coalescent analyses in predicting HCV epidemics has already been established [18], and a recent paper using this method on subtype 1a data in North America had a very similar prediction to that reported here regarding the timing of the epidemic [29]. Regarding the epidemiological modelling, the data is dependent on reported numbers in national surveys and notification systems, which may be affected by under-reporting.
The genetic diversity between HCV subtypes is immense with some subtypes thought to have existed much earlier than other subtypes (as endemic infections in restricted geographical locations). For example, the origins of subtype 6 in East Asia have been traced back to approximately 1000 AD by evolutionary analysis of sequence data [30]. Similarly, the origin of the genotype 4 epidemic in Africa has been traced back to the 17th century [31]. However, evolutionary analysis of the more common subtypes with a global distribution (1a, 1b and 3a), which are frequently referred to as “epidemic” subtypes, have consistently traced the ancestor to within the last century in many locations around the world [32, 33]. The original endemic viral population sources for these emergent subtypes have not been identified.

The origins of subtypes 1a and 3a in North America and Australia generally coincide with the First and Second World Wars (WW) respectively and are likely to be closely linked to developments in medical technology and practice in that era. The United States mobilized more than one million soldiers to battlefronts in Europe in WWI, and 16 million in WWII [34]. There were 330,000 Australians who enlisted for military service and fought in Europe in WWI, constituting approximately 7% of the total population at that time [35], and one million served abroad in WWII [35, 36]. Similarly, the Korean War (1950-53) allowed the mixing of large numbers of young Australians and Americans [37, 38]. In conjunction with these conflicts blood transfusion for management of major trauma was first undertaken in 1916 [39], followed by establishment of the first blood donation service [39]. In the same time period, there was rapid growth in the use of hypodermic needles and syringes initially for subcutaneous, and subsequently intravenous, administration of medications. There was a dramatic fall in prices for these devices between 1920 and 1950 [40]. In addition, there had been a preceding rise in the routine administration of opium and morphine by physicians, which is cited as the leading cause of opiate dependency at that time [41].
The Bayesian skyline plots for subtypes 1a and 3a reveal a period of exponential growth of infections in both North America and Australia between 1955 and 1975. Soon after WWII, injecting drug use became popular in United States and was accompanied by a post-War surge in heroin use [42]. Armstrong et al. modelled lifetime probability of injecting drug use and the prevalence of anti-HCV positivity by birth year for United States [22, 23]. This analysis shows that in both cases, numbers reached a peak for cohorts born from 1948 to 1962. The age of initiation of injecting drug use is predominantly in the late teens or early twenties [43, 44], which suggests that the peak HCV prevalence may follow 15-20 years later. This correlates well with the Bayesian skyline projections described here (supplementary S3). In Australia, there is no data on drug use prior to 1960 [45]. Recent models estimating injecting drug use patterns in Australia assume that the prevalence increased by approximately 8% per year between 1970 – 1997 [45]. The simulated rise in the number of HCV infected individuals for both subtypes in North America and Australia identified here thus correlates with the surge in injecting drug use.

In both continents since the mid-1970s, the projections show stabilization of the epidemic. Since the early 1980s, the risks of blood borne virus transmission, notably for HIV and subsequently HCV, have become well known to the public, and led to harm minimisation strategies for injecting drug users. In the United States, the HIV incidence in people who inject drugs (PWID) has dropped by 50% from 1997 – 2009 (80% since 1980) [46]. Surveillance data indicate that the incidence of acute hepatitis C also declined from 2000-2004 and remained stable until 2010 after which the incidence has been on an increasing trend probably due to increasing injecting drug use [47]. The findings of this study are consistent with these observations though it failed to show the later surge of infections as the most recent samples were collected in 2011 from USA. In Australia, epidemiological estimates and surveillance data suggest that injecting drug use and incident HCV infections have declined since the early 2000s [45, 48]. There is a stable or falling trend in the prevalence of hepatitis C among attendees of needle and syringe programmes, prison entrants, and blood donors [49, 50]. These data are consistent with the trends in the projections described here.
Conclusions

Using coalescent analysis of full length HCV genomes, the historical trends in the HCV epidemics in Australia and North America have been mapped with convergent data derived from epidemiological modelling. The subtype 3a epidemic has more recent origins than subtype 1a. All projections showed an exponential increase in infections from 1955 to 1975 coincident with the increase in injecting drug use. The epidemics subsequently stabilized up to the time window examined, potentially explained by improved harm minimisation measures following the discovery of HIV and HCV.
References


16. Yuan M, Lu T, Li C, Lu L. The evolutionary rates of HCV estimated with subtype 1a and 1b sequences over the ORF length and in different genomic regions. PLoS One 2013; 8:e64698.


### Tables

Table 1. Selection of sequences for the coalescent analysis by continent and subtype

<table>
<thead>
<tr>
<th>Continent</th>
<th>Subtype</th>
<th>Source</th>
<th>Time window</th>
<th>Number of sequences</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>North America</strong></td>
<td>1a</td>
<td>InC3</td>
<td>1998-2011</td>
<td>58</td>
<td>72</td>
</tr>
<tr>
<td></td>
<td>Public</td>
<td></td>
<td>1977-1998</td>
<td>14</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3a</td>
<td>InC3</td>
<td>2000-2011</td>
<td>35</td>
<td>42</td>
</tr>
<tr>
<td></td>
<td>Public</td>
<td></td>
<td>1991-2000</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td><strong>Australia</strong></td>
<td>1a</td>
<td>InC3</td>
<td>2004-12</td>
<td>30</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td>Public</td>
<td></td>
<td>Not available</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3a</td>
<td>InC3</td>
<td>2005-12</td>
<td>35</td>
<td>35</td>
</tr>
<tr>
<td></td>
<td>Public</td>
<td></td>
<td>Not available</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>
Table 2. Summary of coalescent analysis for subtypes 1a and 3a for North America (NA) and Australia (AUS)

<table>
<thead>
<tr>
<th>Variable</th>
<th>1a AUS</th>
<th>1a NA</th>
<th>3a AUS</th>
<th>3a NA</th>
</tr>
</thead>
<tbody>
<tr>
<td>tMRCA (Years) ± 95% HPD</td>
<td>97.12 (50.88-149.38)</td>
<td>85.68 (75.11-96.32)</td>
<td>59.69 (43.31-77.53)</td>
<td>61.59 (50.65-73.51)</td>
</tr>
<tr>
<td>Ucl mean</td>
<td>1.13x10^{-3}</td>
<td>1.07x10^{-3}</td>
<td>1.19x10^{-3}</td>
<td>1.16x10^{-3}</td>
</tr>
<tr>
<td>COV</td>
<td>0.586</td>
<td>0.123</td>
<td>0.073</td>
<td>0.111</td>
</tr>
<tr>
<td>α C(1+2)</td>
<td>0.529</td>
<td>0.384</td>
<td>0.450</td>
<td>0.435</td>
</tr>
<tr>
<td>α C(3)</td>
<td>1.521</td>
<td>1.282</td>
<td>1.358</td>
<td>1.423</td>
</tr>
<tr>
<td>pInV C(1+2)</td>
<td>0.754</td>
<td>0.694</td>
<td>0.732</td>
<td>0.738</td>
</tr>
<tr>
<td>pInV C(3)</td>
<td>0.120</td>
<td>0.075</td>
<td>0.029</td>
<td>0.048</td>
</tr>
<tr>
<td>μ C(1+2)</td>
<td>0.391</td>
<td>0.513</td>
<td>0.479</td>
<td>0.516</td>
</tr>
<tr>
<td>μ C(3)</td>
<td>2.218</td>
<td>1.975</td>
<td>2.041</td>
<td>1.968</td>
</tr>
<tr>
<td>MCMC length</td>
<td>400 million</td>
<td>500 million</td>
<td>50 million</td>
<td>400 million</td>
</tr>
</tbody>
</table>

tMRCA-Time to the most recent common ancestor, HPD – High posterior density, Ucl mean – Estimated substitution rate, COV – Coefficient of variation, PInv – Proportion of invariant sites, μ - relative substitution rate, MCMC – Monte Carlo Markov Chain, AUS – Australia, NA – North America
Figures

Figure 1. Bayesian skyline simulations of the effective population size with hepatitis C infections in North America and Australia compared with epidemiological back projections

The solid line represents the change in median effective population size (left Y axis) with time (X axis, calendar years) as estimated by the Bayesian evolutionary analysis. The effective population size refers to a population that is inferred from the available sequence data and does not refer to real incidence or prevalence values in the population. The 95% high posterior density (HPD) interval for each time point is shown by the error bars. The lines with triangles represent the total infected population as estimated by epidemiological back projections (right Y axis) which are real prevalence estimates for the population under study. Both Y axes are on a logarithmic scale. The epidemiological projections are available only up to 1950 while the Bayesian projections extend to the lower HPD interval of the estimated time to the most recent common ancestor (Table 2). Only the trends of Y axes may be compared, not the absolute numbers due to the intrinsic the differences in methodology and output parameters mentioned above.

The panels are as follows; (A) North America: subtype 1a, (B) North America: subtype 3a (C) Australia: subtype 1a, (D) Australia: subtype 3a.