PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B

rstb.royalsocietypublishing.org

Research

Cite this article: Gray RR, Tanaka Y, Takebe Y, Magiorkinis G, Buskell Z, Seeff L, Alter HJ, Pybus OG. 2013 Evolutionary analysis of hepatitis C virus gene sequences from 1953. Phil Trans R Soc B 368: 20130168. http://dx.doi.org/10.1098/rstb.2013.0168

One contribution of 13 to a Theme Issue 'Paleovirology: insights from the genomic fossil record'.

Subject Areas:
evolution, bioinformatics, health and disease and epidemiology

Keywords:
molecular epidemiology, phylogenetics

Authors for correspondence:
Rebecca R. Gray
e-mail: rebecca.gray@zoo.ox.ac.uk
Oliver G. Pybus
e-mail: oliver.pybus@zoo.ox.ac.uk

Electronic supplementary material is available at http://dx.doi.org/10.1098/rstb.2013.0168 or via http://rstb.royalsocietypublishing.org.

Royal Society Publishing
Informing the science of the future

Evolutionary analysis of hepatitis C virus gene sequences from 1953

Rebecca R. Gray[1], Yasuhito Tanaka[2], Yutaka Takebe[3], Gkikas Magiorkinis[1], Zelma Buskell[4], Leonard Seeff[5], Harvey J. Alter[6] and Oliver G. Pybus[1]

[1]Department of Zoology, University of Oxford, Oxford, UK
[2]Department of Virology and Liver Unit, Nagoya City University Graduate School of Medical Sciences, Kawasumi, Mizuho, Nagoya, Japan
[3]AIDS Research Center, National Institute of Infectious Diseases, Tokyo 162-8640, Japan
[4]Veterans Affairs Medical Center, Washington, DC, USA
[5]The Hill Group, Bethesda, MD, USA
[6]Department of Transfusion Medicine, National Institutes of Health, Bethesda, MD, USA

Reconstructing the transmission history of infectious diseases in the absence of medical or epidemiological records often relies on the evolutionary analysis of pathogen genetic sequences. The precision of evolutionary estimates of epidemic history can be increased by the inclusion of sequences derived from 'archived' samples that are genetically distinct from contemporary strains. Historical sequences are especially valuable for viral pathogens that circulated for many years before being formally identified, including HIV and the hepatitis C virus (HCV). However, surprisingly few HCV isolates sampled before discovery of the virus in 1989 are currently available. Here, we report and analyse two HCV subgenomic sequences obtained from infected individuals in 1953, which represent the oldest genetic evidence of HCV infection. The pairwise genetic diversity between the two sequences indicates a substantial period of HCV transmission prior to the 1950s, and their inclusion in evolutionary analyses provides new estimates of the common ancestor of HCV in the USA. To explore and validate the evolutionary information provided by these sequences, we used a new phylogenetic molecular clock method to estimate the date of sampling of the archived strains, plus the dates of four more contemporary reference genomes. Despite the short fragments available, we conclude that the archived sequences are consistent with a proposed sampling date of 1953, although statistical uncertainty is large. Our cross-validation analyses suggest that the bias and low statistical power observed here likely arise from a combination of high evolutionary rate heterogeneity and an unstructured, star-like phylogeny. We expect that attempts to date other historical viruses under similar circumstances will meet similar problems.

1. Introduction

For some of the most important viral epidemics of the twentieth century, most notably those caused by HIV-1 and the hepatitis C virus (HCV), little reliable epidemiological information is available until the discovery of the infectious agent, in 1983 for HIV-1 and in 1989 for HCV. Reconstruction of the early transmission dynamics of these epidemics therefore relies upon statistical inferences, either through the application of mathematical epidemiological models or by the evolutionary analysis of viral genetic sequences. The latter approach has gained in popularity owing to the widespread and increasing availability of viral genomes and the adoption of analytical methods that can estimate epidemic history from pathogen sequences [1]. The precision and accuracy of such estimates may be improved by inclusion of 'ancient' virus sequences that are evolutionarily distinct from current strains [2] and which, ideally, represent infections that occurred early in an epidemic. Notable viral sequences that have been obtained from archived sera or preserved tissue include the genome of the devastating 1918 'Spanish flu' H1N1 human influenza virus [3] and partial variola virus sequences obtained from a 300-year-old Siberian mummy [4]. Viral sequences such as these can be thought of as an 'archaeological record' of viral diversity,

complementing the much more ancient viral 'fossil record' provided by endogenous viral elements [5].

Historical virus sequences may be uncommon because the pathogen was not identified during early epidemic spread, hence specimens were not kept, or because prevalence was much lower in the past, or because the method of specimen storage did not lend itself to nucleic acid preservation. When historical viral sequences *are* available, they not only prove that an infection must have existed at a certain time in the past, but can also strengthen statistical estimates of parameters of interest, such as rates of molecular evolution, dates and locations of epidemic origin and past rates of transmission [2]. For HIV-1, the utility of historical isolates was demonstrated through the discovery and analysis of two archived strains from the Democratic Republic of Congo, one obtained from a 1959 plasma sample [6,7] and the other from a 1960 paraffin-embedded tissue sample [6]. However, for HCV, there are remarkably few sequences that substantially predate discovery of the virus in 1989. The earliest dated HCV subgenomic sequence published to date represents a 1976 strain from the USA (AB079715; [8]). Of more than 117 000 sequences available in the HCV sequence database [9], only 162 sequences with known dates of sampling (and derived from human infections) were collected before 1986, most of which are short subgenomic fragments.

HCV currently circulates worldwide and is classified into seven genetically divergent genotypes, with each genotype being further subdivided into numerous subtypes [10]. A few subtypes (most notably 1a and 1b) are highly prevalent worldwide and considered to be 'epidemic' subtypes [11,12]. Although HCV now infects an estimated 3% of the world's population, using medical records to reconstruct the history of HCV transmission before 1989 has proved difficult, because the symptoms of acute HCV infection are typically mild or unspecific. From historical evidence we know that during the second half of the twentieth century the number of hepatitis cases in the USA increased dramatically among individuals who had received blood transfusions, particularly among those who received pooled and/or commercial plasma [13]. At the time, many of these infections could not be attributed to hepatitis A virus or hepatitis B virus infection [14]; hence a distinct but unknown infection, termed non-A non-B hepatitis, was proposed to be the disease-causing agent [15]. After many years of research this pathogen was identified in 1989 as HCV, an enveloped single-stranded positive-sense RNA virus classified in the family *Flaviviridae* [16].

While a few studies have used mathematical models to back-calculate historical incidence from contemporary patterns of HCV seroprevalence [17–19], the majority of evidence for HCV transmission before 1989 has come from the evolutionary analysis of virus gene sequences collected in the last two decades of the twentieth century. Using phylogenetic, molecular clock and coalescent methods, these studies have revealed a long and complex history of transmission that began substantially before the mid-twentieth century (e.g. [8,11,12]).

To address the paucity of HCV genetic data from the decades preceding its identification, we here report and analyse two subgenomic HCV subtype 1b sequences that were recovered from human sera sampled in the USA in 1953. These sequences represent the earliest genetic evidence of HCV infection to date. We use a new phylogenetic 'tip-dating' method to investigate the possibility of contamination with modern HCV sequences. Our analyses produced surprising results that highlight the need to carefully interpret and validate the results of molecular clock methods commonly used to date historical pathogen sequences. We find that HCV subtype 1b diversity was already high by the mid-twentieth century, and we provide new estimates for the likely time of entry of the virus into the USA.

2. Material and methods

(a) Archived hepatitis C virus isolates

In previous work, Seeff *et al.* [20] screened for antibodies against HCV more than 8000 archived blood samples that had been obtained between 1948 and 1955 from US military recruits. They reported that 17 samples were positive for anti-HCV antibodies, of which 11 were positive for HCV RNA using PCR. Ten of the 11 PCR-positive samples were classified as subtype 1b but no HCV sequences were reported [20]. Here, we provide partial NS5B sequences from two of these 10 isolates, named US1953a and US1953b (genome positions 7939–8274 relative to the H77 reference strain). The amplification and sequencing protocol undertaken to obtained these sequences was the same as that described in [8]. Sequences have been deposited in GenBank under accession nos. KF261594 and KF2615945. Alignments are available from the authors on request.

(b) Whole genome sequence alignment

We downloaded all available whole genome sequences of HCV subtype 1b with known dates of sampling from the Los Alamos HCV Database ($n = 170$; as of February 2012) [9]. Sampling dates ranged from 1988 to 2008. Sequences were aligned manually using BioEdit. All isolate names and accession numbers are provided in the electronic supplementary material, table S1.

(c) Pairwise diversity

Pairwise genetic distances for the whole genome dataset were calculated using the maximum composite likelihood model with pairwise deletions and gamma-distributed among-site rate variation, as implemented in MEGA v. 1.5 [21].

(d) Phylogenetic analysis

We estimated two maximum-likelihood (ML) phylogenies that represent (i) the 170 sequences in the whole genome dataset and (ii) the 170 sequences plus the two new isolates from 1953. Both trees were estimated using a general-time reversible nucleotide substitution model with gamma-distributed among-site rate variation, using the ML method implemented in MEGA v. 1.5 [21]. Phylogenetic support was evaluated using a bootstrap approach: ML trees were estimated from 200 bootstrap replicate alignments using PhyML v. 3.0 [22]. To evaluate the temporal evolutionary signal in the dataset prior to molecular clock analysis, we used Path-O-Gen v. 1.2 to calculate regressions of root-to-tip genetic distance against sampling time (available from http://tree.bio.ed.ac.uk). Phylogenies were visualized and annotated in FigTree (available from http://tree.bio.ed.ac.uk).

(e) Bayesian estimation of sequence sampling time

Shapiro *et al.* [23] recently developed a new phylogenetic approach that estimates the date of sampling of a historical molecular sequence. Given a set of temporally sampled (heterochronous) sequences whose sampling dates are known, plus a target sequence whose sampling time is unknown, molecular clock models can be used to estimate the unknown date with statistically appropriate confidence intervals. Here, we applied this 'tip-dating' method to

assess the evolutionary information contained in our archived HCV sequences and to explore the possibility that they represent contamination by modern HCV sequences. To estimate the sampling date of a sequence, we set the isolation date of that sequence to be a random variable, then used Bayesian Markov chain Monte Carlo (MCMC) sampling to estimate the posterior probability distribution of the unknown date [23]. All analyses were performed using BEAST v. 1.7 [24]. MCMC sampling was performed for at least 1×10^7 generations, with states sampled every 1000th generation. MCMC convergence and effective sample sizes were evaluated using Tracer (http://tree.bio.ed.ac.uk/software/tracer/).

For computational tractability, in each tip-dating analysis, we used a smaller dataset of 59 HCV genomes with known sequence sampling dates selected from the whole genome dataset, which is hereafter referred to as the 'tip-dating whole genome' alignment. Sequences were selected for removal in a manner that preserved a wide and even range of sampling dates. We performed tip-dating analyses separately for each of the two 1953 sequences. Additionally, we undertook cross-validation of the tip-dating approach by repeating it on four reference strains whose year of sampling was unambiguous (accession numbers: EU256088, 2003; HQ110091, 1996; EU155336, 1992; EU482849, 1989). All isolate names and accession numbers are provided in the electronic supplementary material, table S2.

To determine the effect of using short subgenomic fragments during tip-dating analysis, the four above-mentioned reference strains were tested both as full genomes and as subgenomic sequences cropped to the same 336 nt region as that available for the 1953 isolates. Analyses were performed using both strict and relaxed (lognormal) relaxed molecular clock models, the HKY + Γ model of nucleotide substitution, empirical base frequencies and the Bayesian skyline coalescent model [25]. We imposed a constraint on phylogenetic topology during MCMC sampling to prevent the 1953 strain from being erroneously placed as a monophyletic outgroup to the remainder of the taxa. Preliminary analyses indicated a lack of MCMC convergence in several runs. We posited that this was because the data contain limited information about the phylogenetic time-scale (i.e. the variance in taxa sampling time is small compared with the age of the tree), as suggested by the bootstrap results in figure 1*c*. To resolve this, we placed a normal prior distribution (mean = 1920, s.d. = 23 years) on the date of the phylogeny root. This information on root age was obtained from the most comprehensive molecular clock analysis of HCV subtype 1b undertaken to date [26].

(f) Epidemic history of the US epidemic

Of the 170 whole HCV subtype 1b genomes obtained for analysis, 106 were sampled from the USA. These 106 genomes, referred hereafter as the 'USA whole genome' alignment, plus the two 1953 sequences, were used to estimate the date of origin and epidemic history of HCV subtype 1b in the USA (isolate names and accession numbers provided in the electronic supplementary material, table S3). The time of the most recent common ancestor of these taxa was estimated using BEAST v. 1.7 [24] using both constant size and Bayesian skyline coalescent models. In addition, a second set of analyses was performed on the 'USA whole genome' alignment only (i.e. without the 1953 sequences) in order to determine the impact of inclusion of the archived strains on the estimation of evolutionary parameters.

3. Results

(a) Pairwise diversity

Pairwise genetic diversities were calculated only for the 336 nt region of the HCV NS5B gene that was available for the historical sequences US1953a and US1953b. Within this subgenomic region, the mean pairwise nucleotide diversity among the 170 subtype 1b reference strains was 0.071 substitutions per site. In comparison, the pairwise nucleotide diversity between the two sequences from 1953 was 0.066 substitutions per site, close to that observed among all strains (figure 1*a*).

(b) Phylogenetic analysis

The ML tree for subtype 1b estimated from the whole genome alignment ($n = 170$) plus the two 1953 sequences is shown in figure 1*b* (a fully annotated tree is provided in the electronic supplementary material, figure S1). The phylogeny shows some spatial structure, with two well-supported clades containing the majority of the Chinese and Japanese sequences, as noted previously [27]. Several well-supported clades of US sequences are also evident. However, clades containing mainly European sequences did not have high bootstrap scores. The two new archived isolates (US1953a and US1953b) are shown in red and did not cluster together. The genetic distances of US1953a and US1953b from the root of the tree appeared to be smaller than those for other taxa.

(c) Root-to-tip divergence

To explore the temporal signal in our dataset, we calculated a plot of root-to-tip genetic distance versus taxon sampling date (figure 1*c*, red) from the ML phylogeny estimated above (figure 1*b*). To visualize the variation arising from phylogenetic uncertainty in this plot, we superimposed the results from 25 randomly selected ML bootstrap phylogenies (figure 1*c*, black). Despite a high variation in estimated root-to-tip distances for the 1953 isolates, there is a tendency for these strains to be closer to the root than the sequences sampled after 1988 (figure 1*c*). The gradient of this plot, which represents evolutionary rate, was 0.00107 substitutions per site per year, which is marginally lower than the rate estimated previously for subtype 1b whole genomes using a Bayesian relaxed molecular clock approach (0.00118–0.00125 substitutions per site per year) [26]. The *x*-intercept, which represents an estimate of the date of the phylogeny root, was 1894. Confidence limits and *p*-values for this regression are not statistically valid [28] and therefore not reported here.

(d) Bayesian estimation of sequence sampling time

Tip-dating analyses were performed separately for four target sequences with unambiguous years of sampling (EU256088, 2003; HQ110091, 1996; EU155336, 1992; EU482849, 1989) as well as for the two 1953 sequences. Preliminary analyses showed that a lognormally distributed relaxed molecular clock model fitted the data significantly better than the strict molecular clock and thus the former was used in all subsequent analyses (data not shown). As expected, using a short subgenomic region for the four target sequences considerably increased the variance of estimated sampling dates, and in some cases led to a failure of MCMC convergence (data not shown). Figure 2 shows the results of the tip-dating analyses for the four whole genome target sequences, plus the two 1953 sequences.

Figure 2*a* shows the posterior probability density of the estimated sampling date for each of the six sequences investigated. Each distribution is truncated at 2008, the date of the most recent sequence in the alignment. As the true

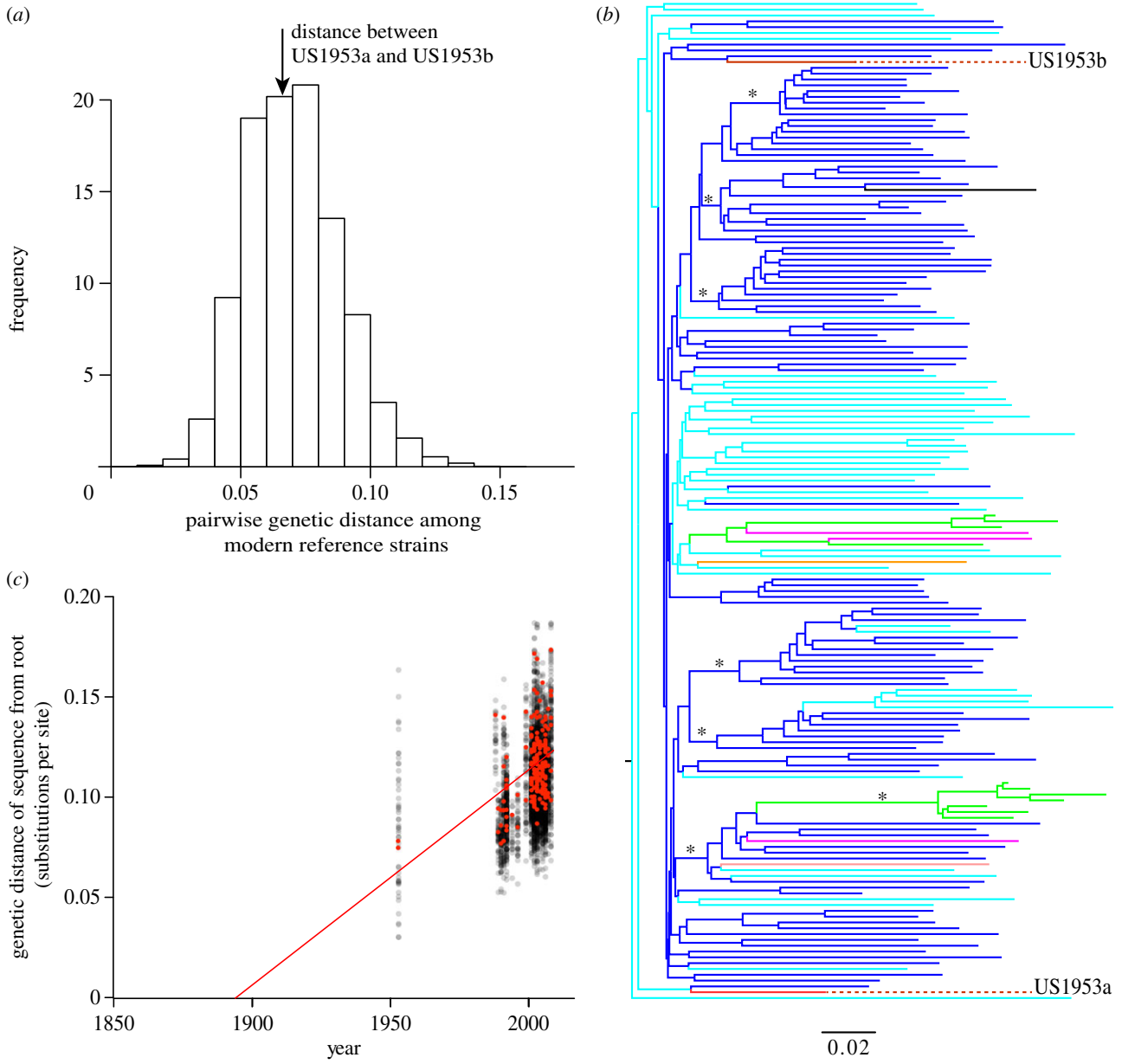

Figure 1. Evolutionary analysis of HCV subgenomic sequence fragments obtained from blood sampled in 1953. (*a*) Pairwise nucleotide diversity. Histogram of all pairwise genetic distances among 170 subtype 1b reference sequences, calculated using a 336 nt region of the NS5b gene. The pairwise genetic distance between the two 1953 sequences is noted with an arrow. (*b*) ML tree of HCV subtype 1b. All non-1953 sequences were full genomes, and the tree was mid-point rooted. Sampling locations are indicated by the colour of the branch as follows: USA (blue branches), Europe (cyan), Brazil (orange), Japan (magenta) and China (green). The two USA isolates from 1953 are shown in red and highlighted with a dotted line. Branches represent number of substitutions per site according to the scale at the bottom. Selected branches with bootstrap values greater than 70% are noted with an asterisk. (*c*) Plot of root-to-tip genetic distances against sampling time. Each *y*-axis value represents the genetic distance from a given tip (sampled sequence) to the root, and the *x*-axis value represents the corresponding sampling date of that tip. The points and regression line shown in red were obtained from the ML tree presented in (*b*). The points in black were obtained from 25 ML bootstrap replicate trees, each of which was mid-point rooted.

sequence age becomes more recent, the probability density shifts to the right (towards the present) and the variance of the posterior distribution decreases. Further, we can use these distributions to directly calculate the probability that the 1953 isolates were sampled before 1989, which is the year that PCR amplification of HCV in diagnostic laboratories began. This probability was 0.67 for US1953a and 0.81 for US1953b. By contrast, the probabilities that the 2003, 1996, 1992 and 1989 sequences were sampled prior to 1989 were lower (0.18, 0.33, 0.25 and 0.53, respectively).

A variety of statistics could be drawn from the posterior probability distributions in figure 2*a*, and we can use our cross-validation results to explore which might be the most statistically informative. In figure 2*b*, we plot the true (or proposed) year of sampling against three values from each of the six posterior distributions shown in figure 2*a*: (i) the mean of the posterior distribution, (ii) the upper 95% credible interval (CI) of the posterior distribution and (iii) the lower 95% CI of the posterior distribution. For a perfect estimator of sequence age, the gradient of the estimated sampling date against the true date should be 1.0. A regression (black line) drawn through the mean values (squares) for the four target sequences has a gradient of only 0.55 (figure 2*b*), indicating that this value increasingly underestimates the true sampling date as historical sequences get older. Remarkably, the corresponding gradient for the lower 95% CI values (triangles) is zero, demonstrating almost no statistical power to reject recent sampling dates, even when whole genomes are

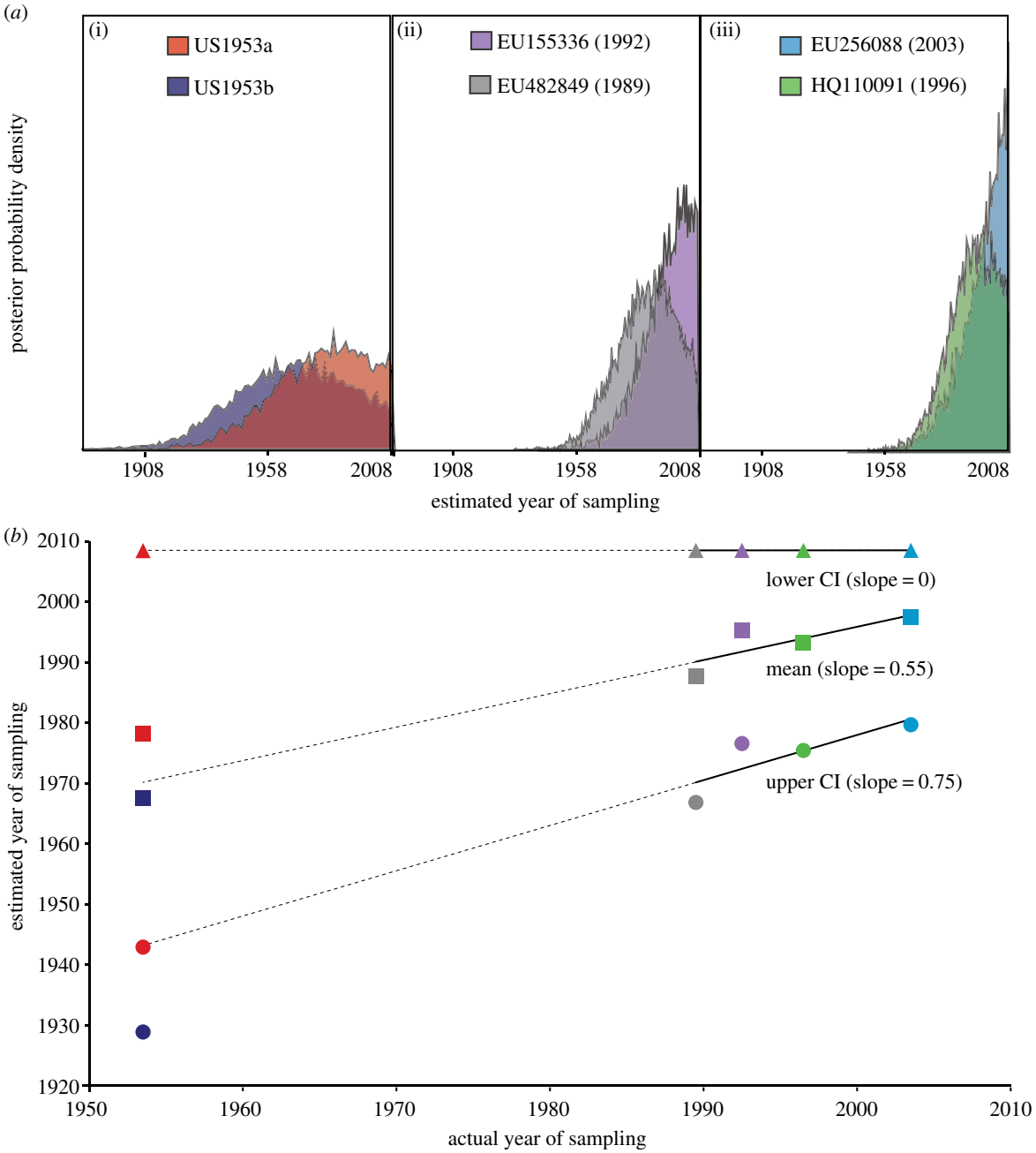

Figure 2. Results of the tip-dating analyses for the two 1953 sequences and four target reference sequences. (*a*) Marginal posterior probability distributions of sequence sampling dates. The posterior probability densities of the estimated sampling dates for six sequences are shown: (i) sequence US1953a (red) and US1953b (blue); (ii) EU155336 (purple) and EU482849 (grey) and (iii) EU256088 (blue) and HQ110091 (green). (*b*) Comparison of true and estimated sampling times. Three values are shown for each of the six sequences analysed (the two 1953 sequences plus the four target reference sequences): (i) the mean of the corresponding posterior distribution (squares), (ii) the lower 95% CI of the corresponding posterior distribution (triangles) and (iii) the upper 95% CI of the corresponding posterior distribution (circles). Colours match those used above in (*a*). The black lines show the best-fit regression for each value, calculated using the four target reference sequences, and subsequently extrapolated back to 1953 (dotted line).

used and the true date of sampling is approximately 20 years into the past (figure 2*b*). The gradient for the upper 95% CI values (circles) is 0.75, which means that the left-hand tail of each posterior distribution tracks changes in true sampling date better than the mean (each upper 95% CI is approximately 15–25 years earlier than the true date; figure 2). We extrapolated these three regression lines (dotted lines) from the four unambiguously dated target sequences to the two new 1953 isolates (blue and red). In each case, the mean and 95% CI values of the 1953 isolates approximately match the extrapolated values (figure 2*b*). In other words, although the mean sampling date estimates of the 1953 sequences are 1968 and 1978, respectively, this level of underestimation should be expected in this dataset for sequences more than 50 years old.

(e) History of the US epidemic

In order to reconstruct HCV epidemic history in the USA and to explore the effects of including the 1953 sequences on such estimates, we analysed 106 whole genome sequences with and without the two sequences from 1953. Bayesian MCMC analyses were performed under three models: (i) a strict molecular clock with a constant size coalescent model, (ii) a relaxed molecular clock with a constant size coalescent model and (iii) a relaxed molecular clock with a Bayesian

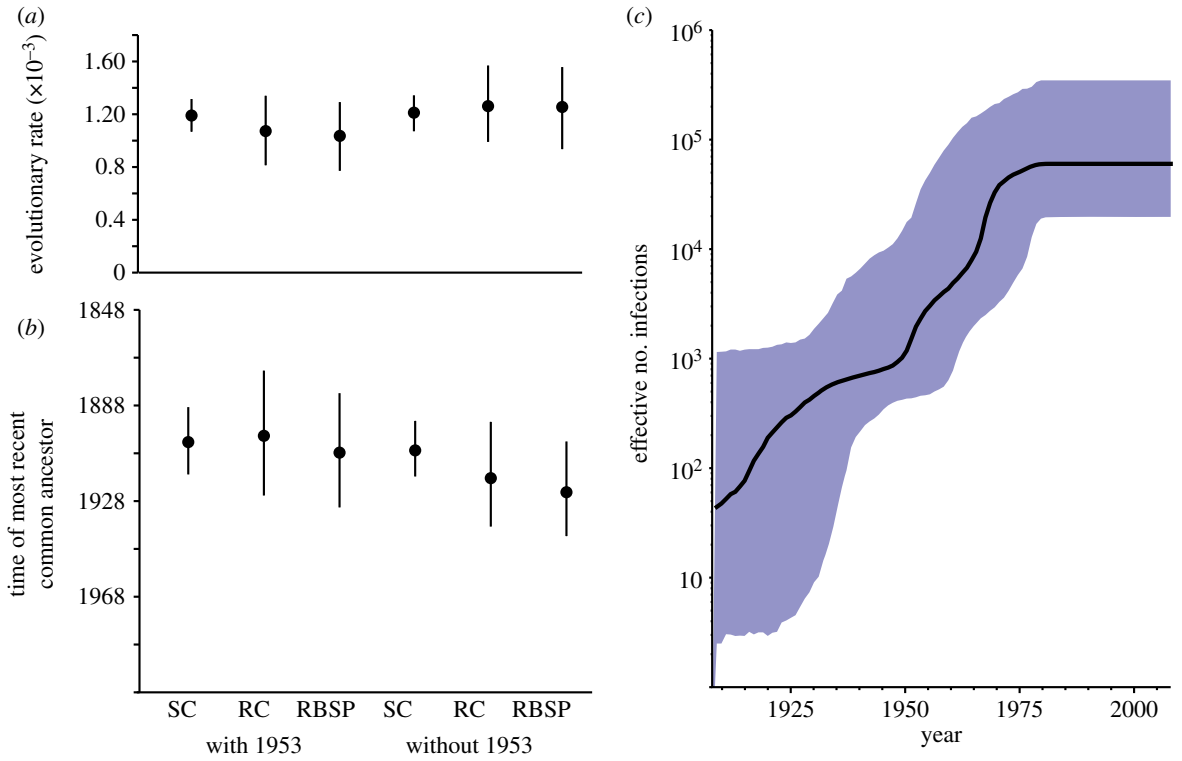

Figure 3. Estimates of the epidemic history of HCV subtype 1b in the USA. (*a*) Estimated mean evolutionary rates with 95% CIs (vertical bars). (*b*) Estimated dates of the most recent common ancestor of US HCV1b sequences with 95% CIs (vertical bars). Three different molecular clock and coalescent model combinations were tested for each dataset: SC, strict molecular clock and constant size coalescent model; RC, relaxed molecular clock and constant size coalescent model; RBSP, relaxed molecular clock and Bayesian skyline coalescent model. (*c*) Bayesian skyline plot for HCV subtype 1b in the USA. Estimates include the two 1953 sequences and used a relaxed molecular clock model.

skyline plot coalescent model. The relaxed clock analyses (lognormal model) significantly outperformed the strict clock analysis (Bayes factor greater than 400). When analyses (ii) and (iii) were compared, the Bayesian skyline plot analysis was not significantly better than the constant size model (Bayes factor less than 1.0). The estimated HCV substitution rates were similar under each model but were slightly higher when the 1953 sequences were excluded (figure 3*a*). Consequently, the estimated date of the most recent common ancestor of HCV in the US was slightly more recent when the 1953 sequences were excluded, although this change was not significant (figure 3*b*). When the 1953 isolates were included, the estimated date of the most common ancestor of the US epidemic was 1901 (95% CIs: 1874–1926). Figure 3*c* shows the estimated Bayesian skyline plot, which matches previous reconstructions [12]. The skyline plot shows exponential growth in the effective number of infections between approximately 1900 and the 1980s, albeit within very wide CIs.

4. Discussion

Here, we report and analyse two subgenomic HCV subtype 1b sequences that were recovered from sera sampled in the USA in 1953. These sequences represent the earliest genetic evidence of HCV infection to date and precede all previously available HCV sequences by several decades. Such samples have the potential to provide information about the dynamics of the HCV epidemic prior to the discovery of the virus in 1989. However, data associated with historical samples (such as the date of sampling) could become lost or incorrectly transposed, and there is the possibility of contamination with modern virus sequences during sample handling. Thus it is important to have a statistical technique that can validate the date of a historical sequence directly from sequence data itself.

In this study, we used a new Bayesian phylogenetic 'tip-dating' method [23] to estimate the isolation dates of two historical sequences, as well as four reference strains whose sampling dates were known. Although in each case, the CIs contained the true (or proposed) date of sampling, our results were in some ways surprising: estimated sampling dates were associated with statistical bias and uncertainty, both of which increased for older sequences even when whole HCV genomes were used. This contrasts with the tight CIs and low bias reported previously when dengue virus sequences were analysed using the same technique [23]. Three differences relating to phylogenetic structure and sampling may explain these divergent results. First, in the dengue virus analysis, sequences were sampled during almost the entire course of the epidemic, including historical strains close to the phylogeny root. By contrast, for the HCV subtype considered here, all available sequences (with the exception of the 1953 strains) were sampled during the last third of the epidemic and none were sampled near the phylogeny root. Second, HCV is known to exhibit substantial among-lineage evolutionary rate variation, which can add considerable uncertainty to molecular clock estimates. Third, the dengue virus phylogeny was more structured (i.e. contained several well-supported internal branches) than the phylogeny of HCV subtype 1b, which is much more star-like, with long terminal branches and few well-supported internal nodes (figure 1).

These factors likely combine in our study to explain the lack of statistical power. Specifically, high among-lineage

variation in HCV evolutionary rates means that a long terminal branch in the HCV phylogeny can take a range of plausible rate/date combinations. Since there are few well-supported internal nodes to help constrain these branches, their lengths (and associated sampling dates) are only weakly bounded, in one direction by the present, and in the other by the numerous divergence events that cluster around the root of the phylogeny. The latter bound is clearly stronger than the former, as there is consistently more statistical power to reject sampling dates that are too old compared with those that are too recent (figure 2). Analyses of the four target sequences using a strict molecular clock support this explanation: when among-branch rate variation is absent, there is much more statistical power to reject recent sampling dates (data not shown). Consequently, in our study, the mean, median and mode of the posterior distribution of sampling date are all uninformative estimators of the true date. These results demonstrate that the temporal sampling of sequences, phylogeny structure and evolutionary rate variation should be taken into account when using tip-dating methods, and that the results of such analyses must be interpreted carefully. The statistical performance of tip-dating methods clearly needs to be investigated under a wider variety of evolutionary scenarios than those previously considered [23].

The 1953 sequences investigated here were obtained from sera stored in glass vials at −20°C and rarely aliquoted [20]. The discovery of antibodies against HCV in the same collection [20] testifies to their good preservation. The sequences are therefore unlikely to have accumulated a significant amount of 'post-mortem' DNA damage, a problem more commonly encountered in the analysis of ancient DNA sequences obtained from environmental material [29]. Further, adding a statistical model of sequence damage to our molecular clock analyses [29] had no notable effect on estimated sequence sampling times (data not shown). Thus, there is no evidence to suggest that the estimation bias we encountered arose from sequence damage during sample storage.

Interestingly, the pairwise nucleotide diversity between the two 1953 sequences was close to that observed among modern reference strains, suggesting that transmission of HCV subtype 1b had been underway for some time before 1953. Such transmission may have occurred at a low rate, and previous evolutionary analyses have indicated that rapid exponential growth of subtype 1b did not begin until the 1950s [12]. This observation is consistent with an epidemiological study that reconstructed the HCV epidemic in US haemophiliacs, which found increased HCV incidence for individuals with severe or moderate haemophilia concurrent with the introduction of fresh-frozen plasma (1940s) and cryoprecipitate (1960s) [30]. A high pairwise genetic distance among historical isolates was also reported for HIV-1; two early sequences recovered from clinical material obtained in 1959–1960 were genetically divergent, despite being sampled more than 20 years before the discovery of HIV-1 [6]. These results highlight the importance of the earliest phases of pathogen emergence in a population, during which a substantial amount of viral diversity may accumulate even though transmission is slow and population prevalence may be low.

The subsequent global dissemination of HCV subtypes 1a and 1b probably commenced after World War II. Phylogenetic analyses have suggested that the direction of spread was mainly from developed countries to the developing world, with the USA as a likely source location [12]. It is hypothesized that the global dispersal of subtypes 1a and 1b occurred in two steps, starting with widespread export of the virus from the USA to the rest of the world around 1940–1950, followed by local expansion of the exported lineages [12]. The initial step coincided with the international transportation of dried plasma units, a massive operation that involved the pooling and shipping of at least 10 million independent donations to sites of military operations over 5–10 years [13]. Our observation of high HCV sequence diversity in the USA during the 1950s is in agreement with this scenario.

We estimate that the time of the most recent common ancestor of subtype 1b in the USA to be 1901 (1874–1926), consistent with previous estimates [11,12]. However, the present analysis provides tighter CIs than those reported previously for subtype 1b using two subgenomic regions (1905–1965 and 1806–1959; [12]), reflecting the increased information gained from using whole-genome reference sequences and from the inclusion of the two 1953 sequences. While direct amplification of viral genomes from archived biological samples may prove challenging using long-range PCR approaches, owing to nucleic acid degradation, modern high-throughput sequencing technologies may alleviate such difficulties, as they have done in the field of ancient DNA research [31]. The generation of further historical viral sequences has great potential to enhance our incomplete understanding of the epidemic history of HCV as well as that of many other established and newly emergent infections of humans.

Acknowledgements. The authors very gratefully acknowledge the work of Prof. Edward Kaplan, University of Minnesota, who first highlighted and surveyed the Warren AFB sera collection and without whom this research would not have been possible. We thank Marc Suchard and Philippe Lemey for helpful discussions.

Funding statement. R.R.G. was supported by the UK Medical Research Council. G.M. was supported by Marie Curie Actions.

References

1. Pybus OG, Rambaut A. 2009 Evolutionary analysis of the dynamics of viral infectious disease. *Nat. Rev. Genet.* **10**, 540–550. (doi:10.1038/nrg2583)
2. Drummond AJ, Pybus OG, Rambaut A, Forsberg R, Rodrigo AG. 2003 Measurably evolving populations. *Trends Ecol. Evol.* **18**, 481–488. (doi:10.1016/S0169-5347(03)00216-7)
3. Taubenberger JK, Reid AH, Krafft AE, Bijwaard KE, Fanning TG. 1997 Initial genetic characterization of the 1918 'Spanish' influenza virus. *Science* **275**, 1793–1796. (doi:10.1126/science.275.5307.1793)
4. Biagini P *et al*. 2012 Variola virus in a 300-year-old Siberian mummy. *N. Engl. J. Med.* **367**, 2057–2059. (doi:10.1056/NEJMc1208124)
5. Katzourakis A, Gifford RJ. 2010 Endogenous viral elements in animal genomes. *PLoS Genet.* **6**, e1001191. (doi:10.1371/journal.pgen.1001191)
6. Worobey M *et al*. 2008 Direct evidence of extensive diversity of HIV-1 in Kinshasa by 1960. *Nature* **455**, 661–664. (doi:10.1038/nature07390)
7. Zhu T, Korber BT, Nahmias AJ, Hooper E, Sharp PM, Ho DD. 1998 An African HIV-1 sequence from 1959 and implications for the origin of the epidemic. *Nature* **391**, 594–597. (doi:10.1038/35400)
8. Tanaka Y, Hanada K, Mizokami M, Yeo A, Shih J, Gojobori T, Alter HJ. 2002 A comparison of the

molecular clock of hepatitis C virus in the United States and Japan predicts that hepatocellular carcinoma incidence in the United States will increase over the next two decades. *Proc. Natl Acad. Sci. USA* **99**, 15 584–15 589. (doi:10.1073/pnas.242608099)

9. Yusim K *et al*. 2005 Los Alamos hepatitis C immunology database. *Appl. Bioinform.* **4**, 217–225. (doi:10.2165/00822942-200504040-00002)
10. Simmonds P. 2004 Genetic diversity and evolution of hepatitis C virus—15 years on. *J. Gen. Virol.* **85**, 3173–3188. (doi:10.1099/vir.0.80401-0)
11. Pybus OG, Charleston M, Gupta S, Rambaut A, Holmes E, Harvey P. 2001 The epidemic behavior of the hepatitis C virus. *Science* **292**, 2323–2325. (doi:10.1126/science.1058321)
12. Magiorkinis G, Magiorkinis E, Paraskevis D, Ho S, Shapiro B, Pybus OG, Allain J-P, Hatzakis A. 2009 The global spread of hepatitis C virus 1a and 1b: a phylodynamic and phylogeographic analysis. *PLoS Med.* **6**, e1000198. (doi:10.1371/journal.pmed.1000198)
13. Kendrick D. 1964 *Blood program in World War II. Supplemented by experiences in the Korean War*. Washington, DC: Office of the Surgeon General Department of the Army.
14. Alter HJ, Holland PV, Morrow AG, Purcell RH, Feinstone SM, Moritsugu Y. 1975 Clinical and serological analysis of transfusion-associated hepatitis. *Lancet* **2**, 838–841. (doi:10.1016/S0140-6736(75)90234-2)
15. Alter HJ, Purcell RH, Holland PV, Popper H. 1978 Transmissible agent in non-A, non-B hepatitis. *Lancet* **1**, 459–463. (doi:10.1016/S0140-6736(78)90131-9)
16. Alter HJ, Purcell RH, Shih JW, Melpolder JC, Houghton M, Choo QL, Kuo G. 1989 Detection of antibody to hepatitis C virus in prospectively followed transfusion recipients with acute and chronic non-A, non-B hepatitis. *New Engl. J. Med.* **321**, 1494–500. (doi:10.1056/NEJM198911303212202)
17. Armstrong GL, Alter MJ, McQuillan GM, Margolis HS. 2000 The past incidence of hepatitis C virus infection: implications for the future burden of chronic liver disease in the United States. *Hepatology* **31**, 777–82. (doi:10.1002/hep.510310332)
18. Sypsa V *et al*. 2004 Reconstructing and predicting the hepatitis C virus epidemic in Greece: increasing trends of cirrhosis and hepatocellular carcinoma despite the decline in incidence of HCV infection. *J. Viral Hepat.* **11**, 366–374. (doi:10.1111/j.1365-2893.2004.00517.x)
19. Deuffic S, Buffat L, Poynard T, Valleron AJ. 1999 Modeling the hepatitis C virus epidemic in France. *Hepatology* **29**, 1596–601. (doi:10.1002/hep.510290528)
20. Seeff LB, Miller RN, Rabkin CS, Buskell-Bales Z, Straley-Eason KD, Smoak BL, Johnson LD, Lee SR, Kalpan EL. 2000 45-year follow-up of hepatitis C virus infection in healthy young adults. *Ann. Intern. Med.* **132**, 105–11. (doi:10.7326/0003-4819-132-2-200001180-00003)
21. Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S. 2011 MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol. Biol. Evol.* **28**, 2731–2739. (doi:10.1093/molbev/msr121)
22. Guindon S, Delsuc F, Dufayard J, Gascuel O. 2009 Estimating maximum likelihood phylogenies with PhyML. *Methods Mol. Biol.* **537**, 113–137. (doi:10.1007/978-1-59745-251-9_6)
23. Shapiro B, Ho SY, Drummond AJ, Suchard MA, Pybus OG, Rambaut A. 2011 A Bayesian phylogenetic method to estimate unknown sequence ages. *Mol. Biol. Evol.* **28**, 879–887. (doi:10.1093/molbev/msq262)
24. Drummond AJ, Rambaut A. 2007 BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol. Biol.* **7**, 214. (doi:10.1186/1471-2148-7-214)
25. Drummond AJ, Rambaut A, Shapiro B, Pybus OG. 2005 Bayesian coalescent inference of past population dynamics from molecular sequences. *Mol. Biol. Evol.* **22**, 1185–1192. (doi:10.1093/molbev/msi103)
26. Gray RR, Parker J, Lemey P, Salemi M, Katzourakis A, Pybus OG. 2011 The mode and tempo of hepatitis C virus evolution within and among hosts. *BMC Evol. Biol.* **11**, 131. (doi:10.1186/1471-2148-11-131)
27. Nakano T, Lu L, He Y, Fu Y, Robertson BH, Pybus OG. 2006 Population genetic history of hepatitis C virus 1b infection in China. *J. Gen. Virol.* **87**, 73–82. (doi:10.1099/vir.0.81360-0)
28. Drummond A, Pybus OG, Rambaut A. 2003 Inference of viral evolutionary rates from molecular sequences. *Adv. Parasitol.* **54**, 331–358. (doi:10.1016/S0065-308X(03)54008-8)
29. Ho SYW, Heupink TH, Rambaut A, Shapiro B. 2007 Bayesian estimation of sequence damage in ancient DNA. *Mol. Biol. Evol.* **24**, 1416–1422. (doi:10.1093/molbev/msm062)
30. Goedert JJ, Chen BE, Preiss L, Aledort LM, Rosenberg PS. 2007 Reconstruction of the hepatitis C virus epidemic in the US hemophilia population, 1940–1990. *Am. J. Epidemiol.* **165**, 1443–1453. (doi:10.1093/aje/kwm030)
31. Rasmussen M *et al*. 2010 Ancient human genome sequence of an extinct Palaeo-Eskimo. *Nature* **463**, 757–762. (doi:10.1038/nature08835)