Downloaded from http://rsbl.royalsocietypublishing.org/ on March 23, 2018

biology letters

rsbl.royalsocietypublishing.org

Research

Cite this article: Hedge J, Lycett SJ, Rambaut A. 2013 Real-time characterization of the molecular epidemiology of an influenza pandemic. Biol Lett 9: 20130331. http://dx.doi.org/10.1098/rsbl.2013.0331

Received: 12 April 2013
Accepted: 1 July 2013

Subject Areas:
evolution, health and disease and epidemiology

Keywords:
Bayesian phylogenetics, influenza, pandemic, parameter estimation, real-time

Authors for correspondence:
J. Hedge
e-mail: jessica.hedge@ndm.ox.ac.uk
A. Rambaut
e-mail: a.rambaut@ed.ac.uk

Electronic supplementary material is available at http://dx.doi.org/10.1098/rsbl.2013.0331 or via http://rsbl.royalsocietypublishing.org.

Royal Society Publishing
Informing the science of the future

Molecular evolution

Real-time characterization of the molecular epidemiology of an influenza pandemic

J. Hedge[1], S. J. Lycett[1] and A. Rambaut[1,2]

[1]Institute of Evolutionary Biology, University of Edinburgh, Ashworth Laboratories, Edinburgh, UK
[2]Fogarty International Center, National Institutes of Health, Bethesda, MD, USA

Early characterization of the epidemiology and evolution of a pandemic is essential for determining the most appropriate interventions. During the 2009 H1N1 influenza A pandemic, public databases facilitated widespread sharing of genetic sequence data from the outset. We use Bayesian phylogenetics to simulate real-time estimates of the evolutionary rate, date of emergence and intrinsic growth rate (r_0) of the pandemic from whole-genome sequences. We investigate the effects of temporal range of sampling and dataset size on the precision and accuracy of parameter estimation. Parameters can be accurately estimated as early as two months after the first reported case, from 100 genomes and the choice of growth model is important for accurate estimation of r_0. This demonstrates the utility of simple coalescent models to rapidly inform intervention strategies during a pandemic.

1. Introduction

When the swine-origin influenza A virus (A(H1N1)pdm09) was detected in April 2009, rapid characterization of its transmission potential and pathogenicity was urgently required for determination of appropriate interventions [1]. Early estimates of its emergence and transmission using phylogenetic analysis of genetic sequence data were reported within just three months of detection [2,3]. Such analyses are possible owing to rapid accumulation of genetic variation within the virus population, enabling its evolution to be modelled on an epidemiological timescale [4].

Here, we determine the efficiency with which Bayesian phylogenetics based on coalescent processes can estimate the evolutionary rate, date of emergence and intrinsic growth rate, r_0, of A(H1N1)pdm09 using whole-genomes. The evolutionary rate provides an indication of the adaptive potential of a virus in a new host population. Accurate estimation is required for inferring divergence times and population size changes. The time of the most recent common ancestor (TMRCA) of a random sample of viruses provides an upper bound to the date of emergence of an epidemic. We use simple parametric growth models to estimate r_0 as a measure of the relative ease with which A(H1N1)pdm09 spread through a host population.

2. Material and methods

We downloaded all available A(H1N1)pdm09 whole-genome sequences sampled April–December 2009 from the EpiFlu database hosted by the Global Initiative on Sharing All Influenza Data (GISAID; platform.gisaid.org) on 26 April 2010 (see the electronic supplementary material, table S1). We analysed whole-genomes (by concatenating segments) to maximize genetic variation in the dataset and only included North American samples to limit spatial heterogeneity in viral

© 2013 The Authors. Published by the Royal Society under the terms of the Creative Commons Attribution License http://creativecommons.org/licenses/by/3.0/, which permits unrestricted use, provided the original author and source are credited.

Downloaded from http://rsbl.royalsocietypublishing.org/ on March 23, 2018

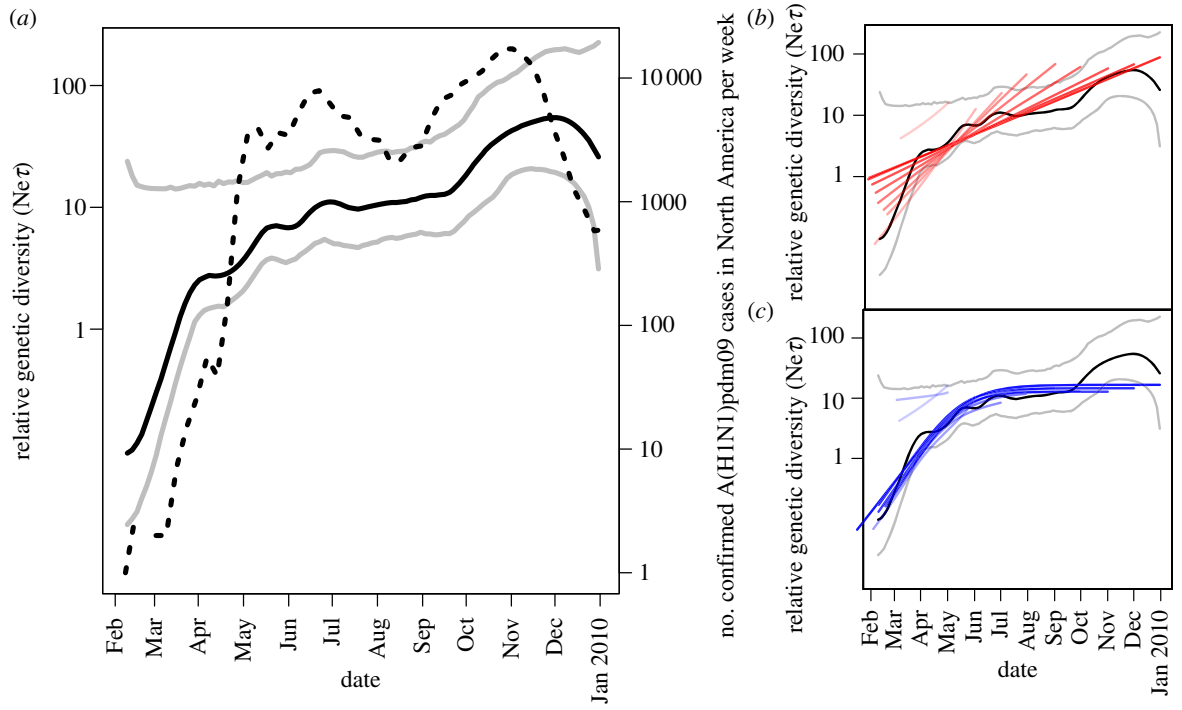

Figure 1. Bayesian skyride reconstruction of the demographic history of A(H1N1)pdm09 in North America until December 2009. Mean genetic diversity (solid black) with corresponding 95% BCI (grey) are shown in (*a*–*c*). Incidence rate (number of new A(H1N1)pdm09 cases confirmed by the WHO/week; dashed) is plotted on secondary axes in (*a*). Similar reconstructions from analysis of the nine cumulative datasets under the (*b*) exponential and (*c*) logistic growth models are plotted with saturation increasing with dataset size in each analysis.

population structure. We removed all isolates sampled from a non-human host, missing an exact sampling date, or with sequencing coverage less than 80% for any genome segment. To minimize the effect of epidemiologically linked cases, which may confound assumptions of the coalescent, we subsampled one isolate/location/day [5,6], resulting in a dataset of 328 sequences. After aligning, we trimmed sequences to 13 158 bp.

We carried out Bayesian phylogenetic analysis of the entire dataset in BEAST v. 1.7.4 [7–9], using the GTR+Γ nucleotide substitution model and uncorrelated lognormal relaxed molecular clock, which had greater Bayes factor support than a strict clock (BF = 3.65) [10,11]. The clock used a gamma-distributed prior on the mean evolutionary rate, with a mean of 1 substitution/site/year ($k = 0.001$, $\theta = 1000$) and exponentially distributed prior on the standard deviation ($\mu = 0.333$). To model the demographic history of the virus population, we used the non-parametric Gaussian Markov random fields Bayesian skyride model [12], which specifies the prior on the TMRCA. We performed four independent Markov chain Monte Carlo runs of 100 million steps to achieve good mixing, sampling trees every 10 000 steps and combining runs after removing 10% burn-in.

To investigate how accurately and precisely Bayesian phylogenetics can estimate the evolutionary rate, date of emergence and r_0 throughout the pandemic, we extracted nine subsets of sequences, each with an increasingly longer temporal range and size. This sampling strategy is akin to carrying out phylogenetic analyses using all genome data available at the end of each month between April and December 2009. Given the increasing capacity with which sequencing can be performed, we included all data available from GISAID on 26 April 2010 to estimate parameters from the maximum amount of sequence data that could have potentially been available if samples were sequenced immediately. We used the same evolutionary models as above but replaced the skyride model with either an exponential or logistic growth model [13]. Here, we use the TMRCA to represent the date of emergence of the virus into the larger human population, assuming a single initial case.

We quantified the relative fit of both growth models by comparing their marginal likelihoods as Bayes factors. The marginal-likelihood measures the average fit of a model to the data and we estimated this using a recently described path sampling procedure [11].

3. Results

The skyride plot in figure 1 shows that the reconstructed past population dynamics of A(H1N1)pdm09 closely follows the number of newly confirmed A(H1N1)pdm09 cases per week (accessed via FluNet; http://who.int/influenza/gisrs_laboratory/flunet/), used here as a measure of incidence rate. This plot captures the exponential growth phase of the first pandemic wave, the plateau in genetic diversity and the growth phase during the second pandemic wave.

The evolutionary rate and date of emergence estimated from the first 34 sampled genomes have wide 95% Bayesian credible intervals (BCI) under both growth models, representing high uncertainty associated with the small sample size (figure 2). Precision increases when the dataset size is increased threefold with the addition of sequences sampled during May, from which the date of emergence is estimated to be 2 February 2009 (95% BCI: 12 January 2009, 2 March 2009) and evolutionary rate 3.93×10^{-3} substitutions/site/year (95% BCI: 2.99, 5.53×10^{-3}) with a standard deviation of 0.24 (95% BCI: 8.9×10^{-6}, 4.5×10^{-1}) under the exponential growth model. The date of emergence remains roughly consistent at later time-points and any further increase in precision is limited by the lack of alternative independent loci. Conversely, the mean evolutionary rate estimates tend to decrease with the addition of data over time, suggesting that many early deleterious/neutral mutations may have later been purged from the population through purifying selection [3,5,14].

Downloaded from http://rsbl.royalsocietypublishing.org/ on March 23, 2018

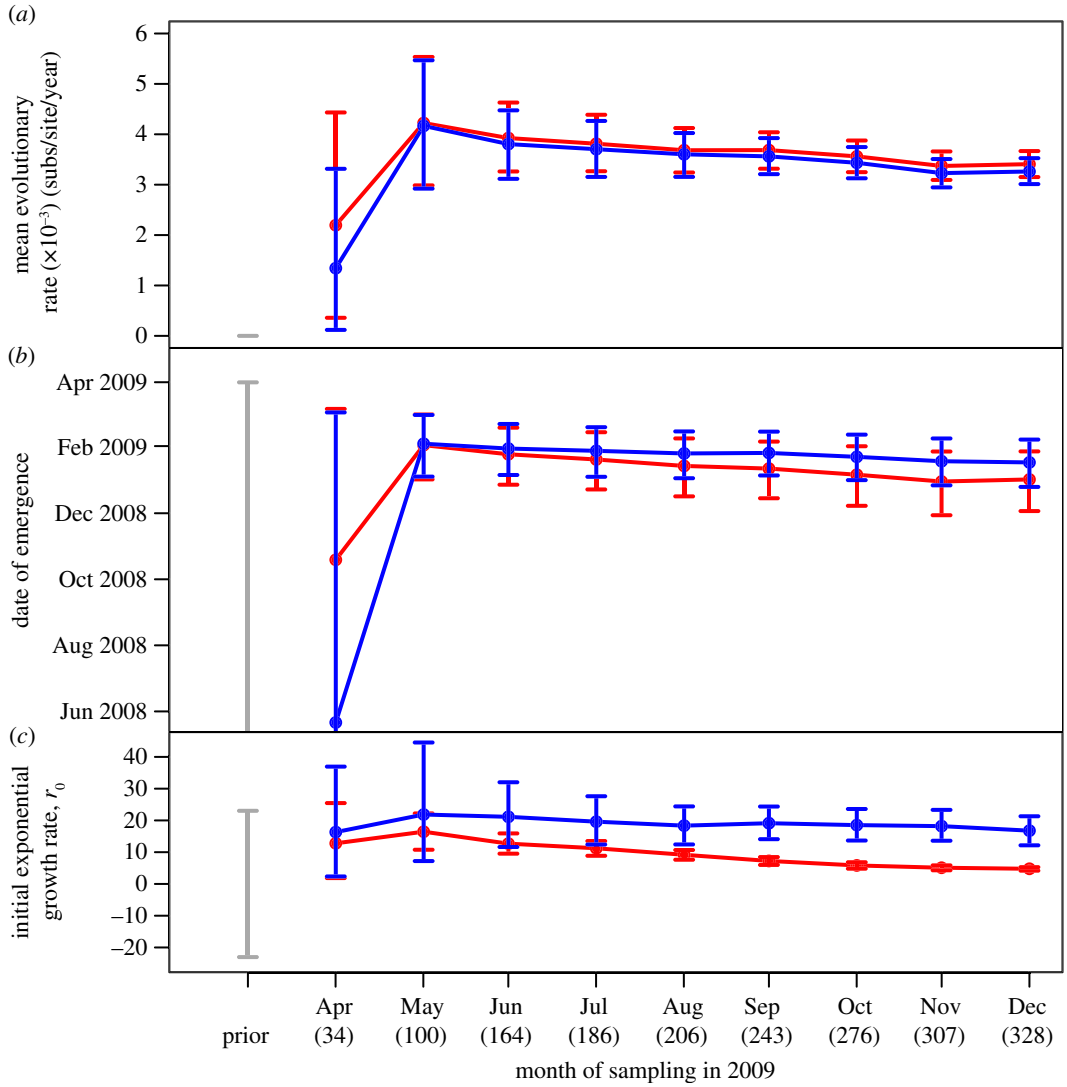

Figure 2. (*a*) Mean evolutionary rate, (*b*) date of emergence and (*c*) r_0 estimates from Bayesian phylogenetic analysis of A(H1N1)pdm09 whole-genomes sampled cumulatively at the end of every month between April and December 2009 across North America. Exponential (red) and logistic (blue) growth models were used in analyses of each dataset. Error bars represent 95% BCI. Dataset size is displayed underneath month names in brackets.

In contrast to either of the other parameters, the choice of growth model has a considerable effect on r_0 estimation (figure 2). By the end of April, both growth models fail to estimate r_0 with sufficient precision to discriminate between slow and rapid epidemic growth because of the small number of sequences sampled. However, uncertainty rapidly reduces by approximately 50% under the exponential growth model with the addition of 66 sequences in May. In comparison, precision remains low during the first three months under the logistic growth model, representing over-parametrization of the model with smaller datasets. The exponential growth model consistently estimates r_0 with greater precision than the logistic model early in the pandemic, once data sampled after May are included in the analysis. Genetic diversity plateaus around June and the exponential growth model inappropriately adjusts for this by lowering the r_0 estimate (figure 1*b*). As the logistic growth model accommodates for this plateau, the accuracy of r_0 estimates remains largely unaffected by the inclusion of data from the second wave (figure 1*c*). A Bayes factor test favours the exponential over the logistic model until June, when support switches and increases as data sampled throughout the following months are included in the analysis (table 1).

Here, r_0 can be used to estimate the basic reproductive ratio (R_0), which describes the average number of secondary infections arising from a primary infection [15]. For example, assuming a gamma-distributed generation time [16,17] with $\mu = 2.6$ and $\sigma = 1.3$ (estimates from household data in the USA [18]), and r_0 estimated from the first two months of data (100 sequences) under an exponential growth model, we estimate an R_0 of 1.12 (95% BCI: 1.07, 1.16). This supports previous estimates from phylogenetic analyses of A(H1N1)pdm09 but is towards the lower end of estimates from incidence data sampled over similar temporal and spatial scales [2,19,20].

We investigated the effect of sample size on parameter estimation by constraining each cumulative dataset to 100 randomly selected genomes (see electronic supplementary material, figure S1). Although variation exists between the means of estimates from different random subsamples at each time-point, their 95% BCIs overlap with one another and those from analysis of the complete dataset. Additionally,

Downloaded from http://rsbl.royalsocietypublishing.org/ on March 23, 2018

Table 1. Log-marginal likelihoods of both growth models used to analyse the nine subsets of sequences with increasing temporal ranges. The preferred model for each dataset (values in italics) was determined using a Bayes factor test, in which the exponential growth model was the null model.

		log-marginal likelihood		
last month sampled in dataset	no. sequences in dataset	exponential growth model	logistic growth model	Bayes factor
April	34	*−19697.73*	−19698.77	−1.03
May	100	*−22630.98*	−22633.16	−2.18
June	164	−26029.90	*−26029.20*	0.70
July	186	−27662.79	*−27650.54*	12.25
August	206	−29552.01	*−29543.42*	8.59
September	243	−33031.06	*−33009.23*	21.83
October	276	−36103.46	*−36078.91*	24.55
November	307	−39147.32	*−39115.72*	31.60
December	328	−41934.96	*−41912.87*	22.09

the complete dataset provides only slightly higher precision of estimates of each parameter.

4. Discussion

Widespread genome sequencing and rapid sharing of data during the 2009 H1N1 pandemic enabled real-time characterization of an influenza pandemic for the first time [2,3,5]. Within approximately two months of the first cases (100 genomes), estimates of evolutionary rate, date of emergence and r_0 from sequence data were in agreement with those from analyses of incidence data, where comparison is available [1,2,19,20]. Over a longer sampling period, parameter estimates from 100 genomes maintain similar accuracy and precision to estimates from more intensively sampled datasets. We discuss potential reasons for the general decrease in evolutionary rate observed over time, although the difference in evolutionary rates is not significant and the datasets are not independent so this result should be interpreted with caution.

The exponential growth model accurately estimates all three parameters during the exponential growth phase, although precision was low with less than 100 sequences. Once growth begins to plateau, this model should be replaced by the logistic growth model to avoid severely underestimating r_0. A demographic reconstruction using a non-parametric coalescent model, such as the skyline or skyride model, can be used to reveal when exponential growth ceases [12,21]. However, these models are unable to estimate the change in relative genetic diversity between the most recent coalescent event and the youngest sample. If this time is large, the demographic plot may appear to flatten prematurely [6]. The exponential growth model is unaffected by an absence of recent coalescence events, estimating r_0 from the density of early coalescent events.

Simple parametric coalescent models are powerful tools for early characterization of an epidemic, even while growth remains exponential. More complex phylogenetic models have been developed to estimate epidemiological parameters that cannot be achieved with parametric coalescent models alone [22–25]. However, the over-parametrization of early A(H1N1)pdm09 data under the logistic growth model highlights the disadvantages of using highly parametrized models during the initial stages of an epidemic. With the increasing capacity of sequencing technologies, the lag between sampling and sequencing viral genomes is expected to decrease, making earlier parameter estimation feasible in future epidemics and before alternative types of data become available.

Acknowledgement. We acknowledge the originating and submitting laboratories of sequences from GISAID's EpiFlu Database (platform.gisaid.org) with which this analysis is carried out (see electronic supplementary material, table S1).

Data accessibility. All models, priors and settings used in this analysis are provided in the XML used as input for BEAST 1.7.4, which are available from the Dryad Digital Repository: doi:10.5061/dryad.jm858. EpiFlu accession numbers of sequences are provided in the acknowledgement of electronic supplementary material, table S1.

Funding statement. The research leading to these results has received funding from the Wellcome Trust (grant no. 092807), the European Research Council under the European Community's Seventh Framework Programme (FP7/2007–2013) under grant agreement no. 278433-PREDEMICS and ERC grant agreement no. 260864 and the Biotechnology Biological Sciences Research Council (UK). We acknowledge funding and resources from the Interdisciplinary Centre for Human and Avian Influenza Research (ICHAIR) and the University of Edinburgh Centre for Infection Immunity and Evolution (CIIE).

References

1. Centers for disease control and prevention. 2009 Swine influenza A(H1N1) infection in two children-Southern California, March–April 2009. *Morb. Mortal. Wkly Rep.* **58**, 400–402.
2. Fraser C *et al*. 2009 Pandemic potential of a strain of influenza A (H1N1): early findings. *Science* **324**, 1557–1561. (doi:10.1126/science.1176062)
3. Smith GJD *et al*. 2009 Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic. *Nature* **459**, 1122–1125. (doi:10.1038/nature08182)
4. Grenfell BT, Pybus OG, Gog JR, Wood JLN, Daly JM, Mumford JA, Holmes EC. 2004 Unifying the epidemiological and evolutionary dynamics of

Downloaded from http://rsbl.royalsocietypublishing.org/ on March 23, 2018

pathogens. *Science* **303**, 327–332. (doi:10.1126/science.1090727)

5. Rambaut A, Holmes E. 2009 The early molecular epidemiology of the swine-origin A/H1N1 human influenza pandemic. *PLoS Curr.* **1**, RRN1003. (doi:10.1371/currents.RRN1003)
6. de Silva E, Ferguson NM, Fraser C. 2012 Inferring pandemic growth rates from sequence data. *J. R. Soc. Interface* **9**, 1797–1808. (doi:10.1098/rsif.2011.0850)
7. Drummond AJ, Rambaut A. 2007 BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol. Biol.* **7**, 214. (doi:10.1186/1471-2148-7-214)
8. Drummond AJ, Suchard MA, Xie D, Rambaut A. 2012 Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Mol. Biol. Evol.* **29**, 1969–1973. (doi:10.1093/molbev/mss075)
9. Ayres DL *et al*. 2012 BEAGLE: an application programming interface and high-performance computing library for statistical phylogenetics. *Syst. Biol.* **61**, 170–173. (doi:10.1093/sysbio/syr100)
10. Drummond AJ, Ho SYW, Phillips MJ, Rambaut A. 2006 Relaxed phylogenetics and dating with confidence. *PLoS Biol.* **4**, e88. (doi:10.1371/journal.pbio.0040088)
11. Baele G, Lemey P, Bedford T, Rambaut A, Suchard MA, Alekseyenko AV. 2012 Improving the accuracy of demographic and molecular clock model comparison while accommodating phylogenetic uncertainty. *Mol. Biol. Evol.* **29**, 2157–2167. (doi:10.1093/molbev/mss084)
12. Minin VN, Bloomquist EW, Suchard MA. 2008 Smooth skyride through a rough skyline: Bayesian coalescent-based inference of population dynamics. *Mol. Biol. Evol.* **25**, 1459–1471. (doi:10.1093/molbev/msn090)
13. Griffiths RC, Tavaré S. 1994 Sampling theory for neutral alleles in a varying environment. *Phil. Trans. R. Soc. Lond. B* **344**, 403–410. (doi:10.1098/rstb.1994.0079)
14. Pybus OG, Rambaut A, Belshaw R, Freckleton RP, Drummond AJ, Holmes EC. 2007 Phylogenetic evidence for deleterious mutation load in RNA viruses and its contribution to viral evolution. *Mol. Biol. Evol.* **24**, 845–852. (doi:10.1093/molbev/msm001)
15. Anderson RM, May RM. 1992 *Infectious diseases of humans: dynamics and control*. Oxford, UK: Oxford University Press.
16. Wallinga J, Lipsitch M. 2007 How generation intervals shape the relationship between growth rates and reproductive numbers. *Proc. R. Soc. B* **274**, 599–604. (doi:10.1098/rspb.2006.3754)
17. Grassly N, Fraser C. 2008 Mathematical models of infectious disease transmission. *Nat. Rev. Microbiol.* **6**, 477–487. (doi:10.1038/nrmicro1845)
18. Cauchemez S, Donnelly CA, Reed C, Ghani AC, Fraser C, Kent CK, Finelli L, Ferguson NM. 2009 Household transmission of 2009 pandemic influenza A (H1N1) virus in the United States. *N. Engl. J. Med.* **361**, 2619–27. (doi:10.1056/NEJMoa0905498)
19. White LF, Wallinga J, Finelli L, Reed C, Riley S, Lipsitch M, Pagano M. 2009 Estimation of the reproductive number and the serial interval in early phase of the 2009 influenza A/H1N1 pandemic in the USA. *Influenza Other Respir. Viruses* **3**, 267–276. (doi:10.1111/j.1750-2659.2009.00106.x)
20. Tuite AR *et al*. 2010 Estimated epidemiologic parameters and morbidity associated with pandemic H1N1 influenza. *CMAJ* **182**, 131–136. (doi:10.1503/cmaj.091807)
21. 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)
22. Volz EM, Kosakovsky Pond SL, Ward MJ, Leigh Brown AJ, Frost SDW. 2009 Phylodynamics of infectious disease epidemics. *Genetics* **183**, 1421–1430. (doi:10.1534/genetics.109.106021)
23. Stadler T, Kühnert D, Bonhoeffer S, Drummond AJ. 2013 Birth–death skyline plot reveals temporal changes of epidemic spread in HIV and hepatitis C virus (HCV). *Proc. Natl Acad. Sci. USA* **110**, 228–233. (doi:10.1073/pnas.1207965110)
24. Rasmussen DA, Ratmann O, Koelle K. 2011 Inference for nonlinear epidemiological models using genealogies and time series. *PLoS Comput. Biol.* **7**, e1002136. (doi:10.1371/journal.pcbi.1002136)
25. Dearlove B, Wilson DJ. 2013 Coalescent inference for infectious disease: meta-analysis of hepatitis C. *Phil. Trans. R. Soc. B* **368**, 20120314. (doi:10.1098/rstb.2012.0314)