Inferring infectious disease phylodynamics with notification data

Sebastián Duchêne[1]*, Francesca Di Giallonardo[2], Edward C. Holmes[3,5], Timothy G. Vaughan[6,7].

[1]Department of Microbiology and Immunology, Peter Doherty Institute for Infection and Immunity, University of Melbourne, Melbourne, VIC, Australia.

[2]The Kirby Institute, University of New South Wale, Randwick, NSW, Australia.

[3]Marie Bashir Institute for Infectious Diseases and Biosecurity, The University of Sydney, Sydney, NSW, Australia.

[4]Charles Perkins Centre, School of Life and Environmental Sciences, The University of Sydney, Sydney, NSW, Australia.

[5]Sydney Medical School, The University of Sydney, Sydney, NSW, Australia.

[6]Department of Biosystems Science and Engineering, ETH Zurich, Basel, Switzerland.

[7]Swiss Institute of Bioinformatics (SIB), Switzerland.

* Corresponding author
Email: sebastian.duchene@unimelb.edu.au

Abstract (150 words)

Genomic surveillance is increasingly common for infectious pathogens. Phylodynamic models can take advantage of pathogen genome sequence data to infer epidemiological dynamics, such as those based on the exponential growth coalescent and the birth-death process. Here we investigate the potential of including case notification data without associated genome sequences in such phylodynamic analyses. Using simulations, we demonstrate that birth-death phylodynamic models can capitalise on notification data to eliminate bias in estimates of the basic reproductive number, R_o, particularly when the sampling rate varies over time. In addition, an analysis of data collected from the 2009 pandemic H1N1 influenza virus demonstrates that using only samples from the prevalence peak results in biased estimates of the reproductive number over time, whereas using case notification data has a comparable accuracy to that achieved when using genome samples throughout the duration of the pandemic.

Keywords

Phylodynamics, Notification data, Bayesian phylogenetics, Birth-death model, Coalescent model, Influenza virus.

Main text (2000 words max.)

Outbreak investigations increasingly rely on genome sequencing of the causative pathogens. For example, it has been estimated that approximately 70% of Ebola cases that occurred in Sierra

Leone during the 2013-2016 West African Ebola virus outbreak have been sequenced (Stadler et al. 2014). Phylodynamic methods can take advantage of these data to infer epidemiological dynamics (Grenfell et al. 2004). Recent sequencing technologies can generate such data very rapidly, such that phylodynamic inferences can be conducted in nearly real-time (Gardy and Loman 2018; Hadfield et al. 2018; Grubaugh et al. 2019). The main appeal of phylodynamics is that the sequence data can inform on epidemiological dynamics for timescales prior to the earliest collected sample. Moreover, because phylodynamic inferences assume an underlying phylogenetic tree, the internal nodes and branches are informative about transmission.

Phylodynamic models describe a branching process. In Bayesian phylogenetic implementations the phylodynamic model is part of the prior and is sometimes referred to as the 'tree prior'. Internal nodes in the tree are assumed to be associated with transmission events while the tree tips represent sampling events, after which an individual is typically not infectious (du Plessis and Stadler 2015). The simplest models posit that the number of infected individuals increases exponentially over time. Although more sophisticated methods now exist (Kühnert et al. 2014; Popinga et al. 2015; Kühnert et al. 2016; Rasmussen et al. 2017; Vaughan et al. 2017; Volz and Siveroni 2018), we focus our simulations on those that assume simple exponential growth that are appropriate for the early stages of an infectious disease outbreak.

Two commonly used phylodynamic models are the coalescent exponential and the birth-death, both of which assume that the infected population size, N, grows at a rate r; $N(t)=e^{rt}$, where t is time after the origin. In the context of a branching process, r, is the difference between the transmission rate, λ, and the become uninfectious rate, δ, ($r= \lambda- \delta$), and where $1/\delta$ is the duration of infection. The basic reproductive number, R_o, is the average number of secondary infections in a fully susceptible population, estimated as $R_o= \lambda/\delta$. The exponential coalescent is a generalisation of the Wright-Fisher model where population size is a deterministic function of time (Griffiths and Tavare 1994; Volz et al. 2009; Volz et al. 2013). The birth-death model typically assumes a stochastic process with sampling through time (Stadler 2010; Stadler et al. 2012; Stadler and Yang 2013), with $\delta=\psi+\mu$, where μ is the recovery rate and ψ is the sampling rate with recovery (the sampling proportion, p, can be calculated as $p= \psi/\psi+\mu$). This model can treat the time of origin of the outbreak as a parameter, which is not the case with the coalescent exponential. The individual parameters, λ, δ, are non-identifiable because the tree likelihood in both models depends on two compound parameters, λ-δ and $\lambda\delta p$, so prior information about any individual parameter is necessary to estimate the rest (Boskova et al. 2014).

Phylodynamic analyses typically require sequence data and sampling times (Rambaut 2000; Drummond et al. 2002; Drummond et al. 2003; Biek et al. 2015; Rieux and Balloux 2016). The number of samples and their times are also informative for the birth-death model because they are explicitly modelled (e.g. they inform ψ) (Boskova et al. 2018). Although the amount of sequence data in outbreak investigations has increased, a key consideration is that sequencing efforts are often conducted only after a large number of cases are reported. For instance, the trees in Fig 1 were simulated under an R_o of 2, a constant sampling effort, and over the course of 1 year. If sequencing was only conducted for samples collected after 0.75 years samples from the deep sections of the tree would be missed (*late sampling* in Fig 1). Such sampling bias can mislead

inferences of epidemiological dynamics because there are no data to inform inferences of the early stages of the outbreak.

Here we investigate bias in epidemiological parameters due to sampling heterogeneity, and we propose some effective approaches to reduce such bias. The first approach involves using a birth-death skyline model (Stadler et al. 2013), that requires an understanding of the sampling effort. For example, if there is knowledge that there was no attempt to collect samples early in the outbreak one can set two intervals for the ψ parameter. However, without knowledge of sampling effort this scenario is indistinguishable from one with a constant sampling effort but where initial prevalence was so low as to preclude obtaining any sequence data in the early stages of the outbreak. The second approach consists of including early case notification data in the analyses, where a notification is a clinically-confirmed case that was not sequenced (*notifications* scenario in Fig 1). Indeed, notifications are an inexpensive source of information traditionally used in epidemiology, such that they could be readily applied to leverage sequence data in outbreak investigations. In a Bayesian phylogenetic framework notification data can be incorporated by assigning a sampling time with no sequence data, and topological uncertainty is naturally incorporated into the analysis. An analogous approach can be used to coherently specify fossil data for molecular clock calibration (Heath et al. 2014; Heath and Moore 2014).

Simulation study

We simulated phylogenetic trees under a birth-death process in MASTER v6.1 (Vaughan and Drummond 2013), with the following parameterisation; R_0=2 or 1.5, δ=91, p=0.05, and an outbreak duration of one year ($1/\delta$ = 0.011 of one year for a duration of infection of about 4 days). The number of tips and their ages are naturally variable (from 100 to 150 tips). We assumed a strict molecular clock with an evolutionary rate of 0.01 substitutions per site per year (subs/site/year) and the HKY+Γ substitution model to produce alignments of 13,000 nucleotides using NELSI (Ho et al. 2015) and Phangorn v2.4 (Schliep 2011). These settings are broadly similar to an influenza virus outbreak (Hedge et al. 2013). We then assumed three sampling scenarios: (i) *constant* sampling with all sequences from the simulation included (e.g. the sequence for every sample in the tree in Fig 1 is included), (ii) *late sampling* only with samples after time T_s (e.g. only sequences for samples after the dashed line in the tree in Fig 1), and (iii) *notifications* in which sequence data are available only after time T_s and the sampling time for samples before T_s are included with no sequence data (i.e. notifications). We set T_s at 0.75 and 0.9 years. For each parameter configuration we simulated 100 sequence data sets which were subsampled according to the three scenarios above. We analysed the data in BEAST v2.5 (Bouckaert et al. 2014; Bouckaert et al. 2018), considering several phylodynamic models; a coalescent exponential and the birth-death. For the *late sampling* scenario we also considered the birth-death skyline with two intervals for the ψ parameter, with the interval time fixed at T_s. We matched the substitution and clock model to those used to generate the data and we used an informative prior on δ using a Γ distribution with mean 91 and standard deviation of 1.

Analyses of data sets with late sampling using a birth-death model produced inaccurate estimates of R_0. In only 11 of 100 simulations with R_0=2 the 95% highest posterior density (HPD) for this parameter included the value used to generate the data (Table 1 and Fig 2). For simulations with R_0=1.5 this model was never able to recover the true R_0 (Table S1 and Fig S1). The birth-death

skyline had much better performance, with 96 of 100 simulations estimating R_o accurately (i.e. the true value was within the HPD). The coalescent exponential had better performance than the birth-death, but it was still less accurate than the birth-death skyline, with 90 simulations producing accurate estimates. In general, for data sets with late sampling we observed that R_o tended to be overestimated with the birth-death and underestimated with the coalescent exponential (Fig 2). Interestingly, estimates of the evolutionary rate displayed a similar pattern to those of R_o, with the birth-death skyline and the birth-death being the most and least accurate, respectively.

As expected, analyses of the data with constant sampling were accurate in a majority of cases, with 97 and 93 of 100 simulations being accurate for R_o, and 98 and 97 for the evolutionary rate, using the birth-death and the coalescent exponential, respectively (the correct model is the birth-death, such that it is expected to perform better than the coalescent). Estimates of R_o including notification data were similarly accurate as those with complete sampling under the birth-death model, where 96 analyses correctly estimated this parameter, but this was not the case for the coalescent exponential, where only 84 analyses included the true value (Table 1). Evolutionary rate estimate with notification data were less accurate than those from the complete data, with 88 accurate analyses using the birth-death and 66 using the coalescent exponential. These results can be attributed to the fact that the birth-death treats sampling times as data, whereas the coalescent is conditioned on the number of samples and their ages (Stadler et al. 2015; Boskova et al. 2018). Notifications improve the accuracy of R_o in the birth-death and they pose an informative prior on the age of the tree height (Boskova et al. 2014), which can also improve the accuracy of the evolutionary rate relative to the coalescent exponential, but this estimate is unlikely to be as accurate as that with the complete sequence data because there is necessarily less molecular information.

The coalescent exponential appears to be more robust to the sampling process, with greater accuracy than the birth-death for the late sampling analyses. This model may be a good alternative when the sampling process is poorly understood, and with no reliable notification data. However, our simulations suggest that this comes at the expense of estimates that are less precise (with higher uncertainty) than those from the birth-death, as measured by the mean estimate divided by the HPD width (Table 1).

Empirical case study: *A/H1N1 Influenza virus from North America*

To illustrate the accuracy of notification data relative to completely sequenced data sets we analysed 639 whole genome sequences sampled from the 2009 A/H1N1 influenza pandemic from North America that were downloaded from GenBank (Supplementary material). The sequences were collected from early April to October 2009. We chose this period of time because it corresponds to a densely sampled clade and captures the peak number of infections as reported in the FluView application (CDC 2019). We considered three data subsets for our analyses; (i) 'complete sampling' with all of the genome sequences, (ii) 'notifications' with 104 sequences only from September and the remaining 535 samples treated as notifications, and (iii) 'late sampling' with only the 104 sequences from September 2009. Because we do not expect constant exponential growth for these data, we used a birth-death skyline model to estimate the reproductive number, R_e (similar to R_o, but not assuming a fully susceptible population), on a monthly basis from January to October. We set the duration of infection at 4 days with a Γ prior on δ with mean 91 and standard

deviation of 1, and we used the HKY+Γ substitution model and a strict molecular clock model. In all cases we set two intervals for ψ, to allow for a low sampling probability before the oldest sample.

The birth-death skyline plot revealed nearly identical trends for the complete data and that using notifications (Fig 3a). For example, the highest R_e was estimated in April, with a mean of 1.52 (HPD: 1.40 – 1.66) for the complete sampling and a mean of 1.54 (HPD: 1.40 – 1.70) for the analysis using notifications. The estimate for the epidemic origin for the late sampling analyses was around late June, such that we cannot estimate R_e before this time. However, for July, August and September we found that late sampling resulted in substantially different estimates to those from the other analyses (Fig 3b-d), particularly in July where R_e with late sampling had a mean of 2.10 (HPD: 1.60 – 2.54) and those with complete sampling and notifications were 0.79 (HPD: 0.72 – 0.85 and 0.72 – 0.86, respectively). Importantly, estimates of R_e using complete sampling or notifications are overall similar in magnitude to those from large-scale epidemiological studies (Fraser et al. 2009; Biggerstaff et al. 2014), and previous estimates using genome sequence data (Hedge et al. 2013).

Notification data in empirical phylodynamic studies

Our simulations and empirical data analyses reveal that notification data are a rich source of information for birth-death models that can dramatically improve the accuracy and precision in estimates of epidemiological parameters. A key consideration is that notifications should represent confirmed cases that would have been sequenced if sequencing effort had been constant. Combining notification and sequence data can be particularly useful in situations where it is unknown whether sequence sampling has been constant over time or where there exist several confirmed cases but a smaller number of sequences. For example, in recently emerging outbreaks combining both sources of data can provide timely insight about the recent evolution of the pathogen in question.

Acknowledgements

SDG was supported by a Discovery Early Career Fellowship from the Australian Research Council (DE190100805) and a McKenzie fellowship from the University of Melbourne.

References

Biek R, Pybus OG, Lloyd-Smith JO, Didelot X. 2015. Measurably evolving pathogens in the genomic era. Trends Ecol. Evol. 30:306–313.

Biggerstaff M, Cauchemez S, Reed C, Gambhir M, Finelli L. 2014. Estimates of the reproduction number for seasonal, pandemic, and zoonotic influenza: a systematic review of the literature. BMC Infect. Dis. 14:480.

Boskova V, Bonhoeffer S, Stadler T. 2014. Inference of epidemiological dynamics based on simulated phylogenies using birth-death and coalescent models. PLOS Comput Biol 10:e1003913.

Boskova V, Stadler T, Magnus C. 2018. The influence of phylodynamic model specifications on parameter estimates of the Zika virus epidemic. Virus Evol. 4:vex044.

Bouckaert R, Heled J, Kühnert D, Vaughan T, Wu C-H, Xie D, Suchard MA, Rambaut A, Drummond AJ. 2014. BEAST 2: a software platform for Bayesian evolutionary analysis. PLOS Comput. Biol. 10:e1003537.

Bouckaert R, Vaughan TG, Barido-Sottani J, Duchene S, Fourment M, Gavryushkina A, Heled J, Jones G, Kuhnert D, de Maio N. 2018. BEAST 2.5: An Advanced Software Platform for Bayesian Evolutionary Analysis. bioRxiv:474296.

CDC. 2019. Centres for Disease Control and Prevention. Available from: https://gis.cdc.gov/grasp/fluview/

Drummond AJ, Nicholls GK, Rodrigo AG, Solomon W. 2002. Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. Genetics 161:1307–1320.

Drummond AJ, Pybus OG, Rambaut A, Forsberg R, Rodrigo AG. 2003. Measurably evolving populations. Trends Ecol. Evol. 18:481–488.

Fraser C, Donnelly CA, Cauchemez S, Hanage WP, Van Kerkhove MD, Hollingsworth TD, Griffin J, Baggaley RF, Jenkins HE, Lyons EJ. 2009. Pandemic potential of a strain of influenza A (H1N1): early findings. Science. 324:1557–1561.

Gardy JL, Loman NJ. 2018. Towards a genomics-informed, real-time, global pathogen surveillance system. Nat. Rev. Genet. 19:9.

Grenfell BT, Pybus OG, Gog JR, Wood JLN, Daly JM, Mumford JA, Holmes EC. 2004. Unifying the epidemiological and evolutionary dynamics of pathogens. Science 303:327–332.

Griffiths RC, Tavare S. 1994. Sampling theory for neutral alleles in a varying environment. Philos. Trans. R. Soc. London. Ser. B Biol. Sci. 344:403–410.

Grubaugh ND, Ladner JT, Lemey P, Pybus OG, Rambaut A, Holmes EC, Andersen KG. 2019. Tracking virus outbreaks in the twenty-first century. Nat. Microbiol. 4:10.

Hadfield J, Megill C, Bell SM, Huddleston J, Potter B, Callender C, Sagulenko P, Bedford T, Neher RA. 2018. Nextstrain: real-time tracking of pathogen evolution. Bioinformatics 34:4121–4123.

Heath TA, Huelsenbeck JP, Stadler T. 2014. The fossilized birth–death process for coherent calibration of divergence-time estimates. Proc. Natl. Acad. Sci. 111:E2957–E2966.

Heath TA, Moore BR. 2014. Bayesian inference of species divergence times. In: Chen M-H, Kuo L, Lewis PO, editors. Bayesian Phylogenetics, Methods, Algorithms, and Applications. Boca Raton, Florida: CRC Press. p. 277–318.

Hedge J, Lycett SJ, Rambaut A. 2013. Real-time characterization of the molecular epidemiology of an influenza pandemic. Biol. Lett. 9:20130331.

Ho SS, Duchêne S, Duchêne DA. 2015. Simulating and detecting autocorrelation of molecular evolutionary rates among lineages. Mol. Ecol. Resour. 15:688–696.

Kühnert D, Stadler T, Vaughan TG, Drummond AJ. 2014. Simultaneous reconstruction of evolutionary history and epidemiological dynamics from viral sequences with the birth–death SIR model. J. R. Soc. Interface 11:20131106.

Kühnert D, Stadler T, Vaughan TG, Drummond AJ. 2016. Phylodynamics with migration: A computational framework to quantify population structure from genomic data. Mol. Biol. Evol. 33:2102–2116.

du Plessis L, Stadler T. 2015. Getting to the root of epidemic spread with phylodynamic analysis of genomic data. Trends Microbiol. 23:383–386.

Popinga A, Vaughan T, Stadler T, Drummond AJ. 2015. Inferring epidemiological dynamics with Bayesian coalescent inference: the merits of deterministic and stochastic models. Genetics 199:595–607.

Rambaut A. 2000. Estimating the rate of molecular evolution: incorporating non-contemporaneous sequences into maximum likelihood phylogenies. Bioinformatics 16:395–399.

Rasmussen DA, Kouyos R, Günthard HF, Stadler T. 2017. Phylodynamics on local sexual contact networks. PLoS Comput. Biol. 13:e1005448.

Rieux A, Balloux F. 2016. Inferences from tip-calibrated phylogenies: a review and a practical guide. Mol. Ecol. 25:1911–1924.

Schliep KP. 2011. phangorn: Phylogenetic analysis in R. Bioinformatics 27:592–593.

Stadler T. 2010. Sampling-through-time in birth-death trees. J. Theor. Biol. 167:696–404.

Stadler T, Kouyos R, von Wyl V, Yerly S, Böni J, Bürgisser P, Klimkait T, Joos B, Rieder P, Xie D. 2012. Estimating the basic reproductive number from viral sequence data. Mol. Biol. Evol. 29:347–357.

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. 110:228–233.

Stadler T, Kühnert D, Rasmussen DA, du Plessis L. 2014. Insights into the early epidemic spread of Ebola in Sierra Leone provided by viral sequence data. PLoS Curr. 6.

Stadler T, Vaughan TG, Gavryushkin A, Guindon S, Kühnert D, Leventhal GE, Drummond AJ. 2015. How well can the exponential-growth coalescent approximate constant-rate birth–death population dynamics? Proc. R. Soc. B Biol. Sci. 282:20150420.

Stadler T, Yang Z. 2013. Dating phylogenies with sequentially sampled tips. Syst. Biol. 62:674–688.

Vaughan TG, Drummond AJ. 2013. A stochastic simulator of birth–death master equations with application to phylodynamics. Mol. Biol. Evol. 30:1480–1493.

Vaughan TG, Leventhal GE, Rasmussen DA, Drummond AJ, Welch D, Stadler T. 2017. Directly Estimating Epidemic Curves From Genomic Data. bioRxiv:142570.

Volz E, Siveroni I. 2018. Bayesian phylodynamic inference with complex models. PLoS Comput. Biol. 14:e1006546.

Volz EM, Koelle K, Bedford T. 2013. Viral phylodynamics. PLOS Comput. Biol. 9:e1002947.

Volz EM, Pond SLK, Ward MJ, Brown AJL, Frost SDW. 2009. Phylodynamics of infectious disease epidemics. Genetics 183:1421–1430.

Figure legends

Fig 1. Example of a phylogenetic trees generated under a birth-death process with a basic reproductive number, R_0, of 2 a become uninfectious rate, δ, of 100 for three analysis scenarios. The solid line denotes the number of samples collected over time. In *constant* sampling samples are collected and sequenced at a rate ψ=5 (i.e. sampling probability, p, of 0.05). In *late sampling* samples are collected and sequenced after time T_s shown with the dashed line. In *notifications* samples are collected constantly over time, but only sequenced after time T_s, such that before T_s only notifications (sampling times with no sequence data) are included. Blue circles represent samples with sequence data, whereas those in orange correspond to notifications. In the *notifications* scenario a Bayesian analysis would integrate over their phylogenetic uncertainty. The solid line represents the number of samples collected over time. In *late sampling* there are no samples collected before T_s, such that assuming constant sampling can produce a bias in estimates of epidemiological dynamics.

Fig 2. Posterior densities for estimates of the basic reproductive number, R_0, and the evolutionary rate for 100 simulations with true R_0 of 2 and an evolutionary rate of 0.01 subs/site/year. The bars

represent the 95% highest posterior density (HPD) and the points are the median. We analysed the data by sampling late in the outbreak only (i.e. after 0.75 of the tree height), with a constant sampling effort (with all samples sequenced), and by including notifications. The colours represent four different models; red for the coalescent exponential, blue for the birth-death skyline, and orange for the birth-death with constant sampling. For the data with sampling late in the outbreak only we use the birth-death skyline model with constant R_o and two intervals for the sampling rate, ψ, before time 0.75. This model is not applicable to analyses with complete sampling or with notifications where sampling is constant. The dashed horizontal lines correspond to the true parameter value used to generate the data.

Fig 3. Estimates of the reproductive number, R_e, for empirical data of the 2009 A/H1N1 influenza pandemic in North America. **a.** Is a birth-death skyline plot where in which R_e is estimated per month from January to October 2009. Each line is a sampled trajectory from the posterior using each analysis, with red for that from the complete sampling (639 sequences), blue for notifications with 104 sequences from September (the month with largest number of infections) and 535 notifications, and green is for only the 104 sequences from September. Note that the analysis using only sequence data from September has a more recent origin parameter, such that it is only possible to estimate R_e from June. The ticks along the x-axis represent the timing of sequences sampled in all analyses in black, and those treated as notifications (with no sequence data) in the 'notifications' analysis. Panels **b.**, **c.** and **d.** show the posterior density for estimates of R_e in July, August, and September, respectively, for each analysis with colours matching those in panel **a.** (i.e. they are the densities for these months shown in **a.**).

Tables

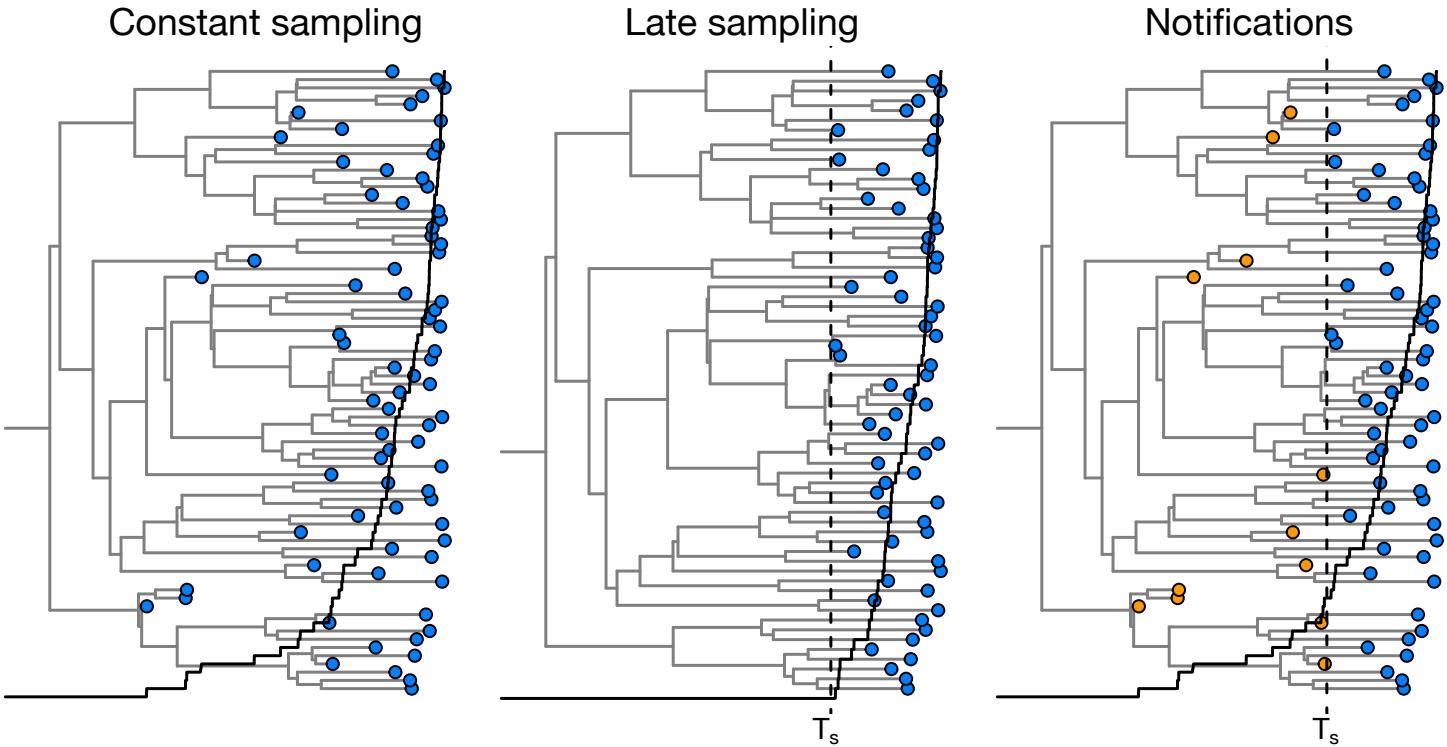
Table 1. Results of the simulation study with R_0 of 1.5 and evolutionary rate of 0.01 subs/site/year. The rows correspond to the seven treatments. The first two columns denote the number of simulations (out of 100) where the value used to generate the data was contained within the 95% highest posterior density (HPD). The last two columns are a measure of precision of the estimates calculated as the estimated mean estimate of R_0 and the evolutionary rate divided by the 95% HPD width, such that large values imply low precision. Here we report the mean value over 100 simulations.

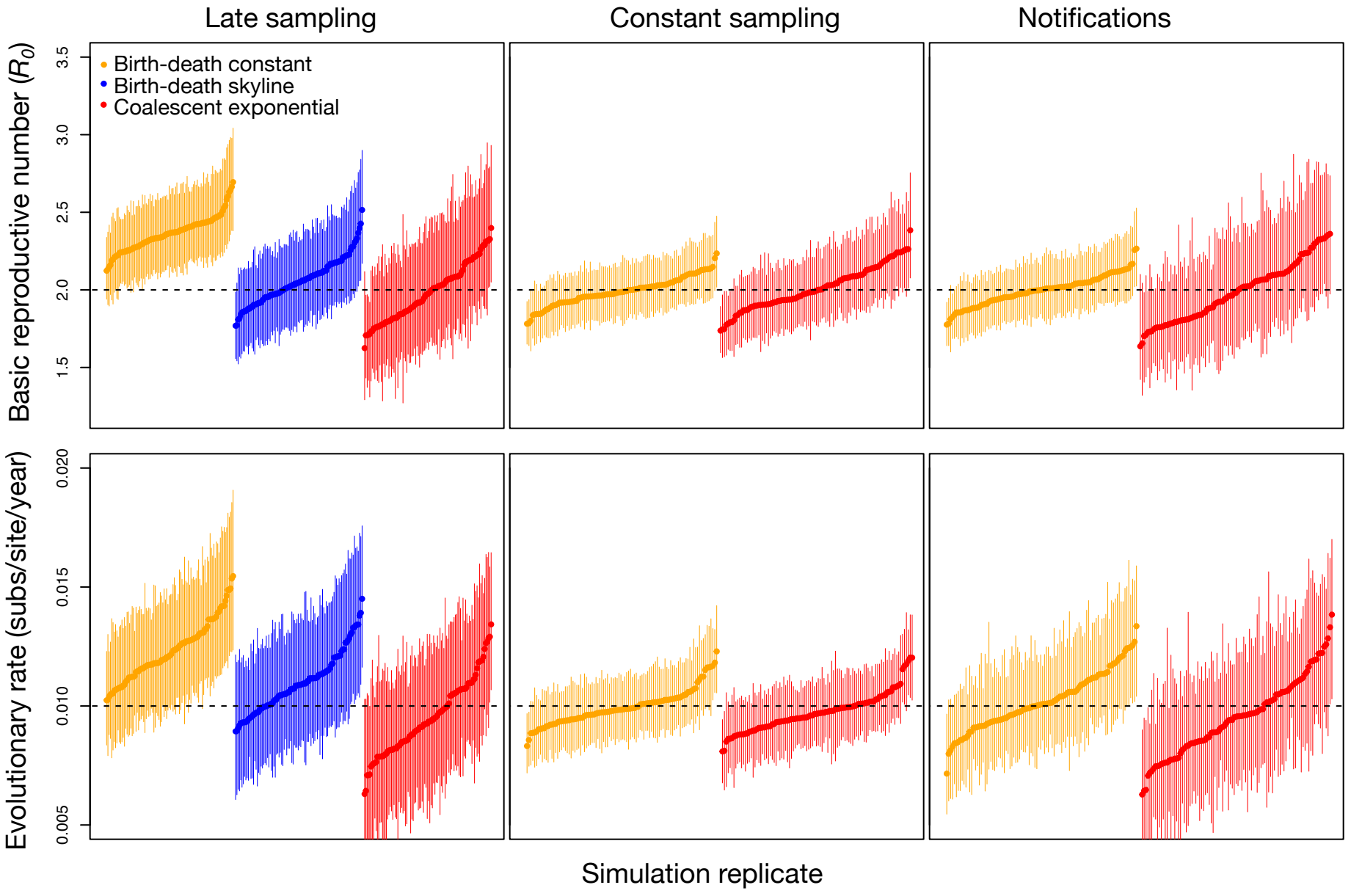
Supplementary material

Table S1. Results of the simulation study with R_o=1.5, evolutionary rate of 0.01 subs/site/year, and late sampling starting at 0.9 years of a total time of 1 year. The rows correspond to the seven treatments. The first two columns denote the number of simulations (out of 100) where the value used to generate the data was contained within the 95% highest posterior density (HPD). The last two columns are a measure of precision of the estimates calculated as the estimated mean estimate of R_o and the evolutionary rate divided by the 95% HPD width, such that large values imply low precision. Here we report the mean value over 100 simulations.

Fig S1. Posterior densities for estimates of the basic reproductive number, R_o, and the evolutionary rate for 100 simulations with true R_o of 1.5 and an evolutionary rate of 0.01 subs/site/year. The bars represent the 95% highest posterior density (HPD) and the points are the median. We analysed the data by sampling late in the outbreak only (i.e. after 0.75 of the tree height), with a constant sampling effort (with all samples sequenced), and by including notifications. The colours represent four different models; red for the coalescent exponential, blue for the birth-death skyline, and orange for the birth-death with constant sampling. For the data with sampling late in the outbreak only we use the birth-death skyline model with constant R_o and two intervals for the sampling rate, ψ, before time 0.75. This model is not applicable to analyses with complete sampling or with notifications where sampling is constant. The dashed horizontal lines correspond to the true parameter value used to generate the data.

Supplementary data. Zip file with input files to generate trees in MASTER and to analyse sequence data in BEAST according to the birth-death skyline, birth-death, and the coalescent exponential models. Accession numbers for empirical A/H1N1 Influenza virus data.


Constant sampling
Late sampling
Notifications
T_s
T_s
Sequenced samples
Notifications (no sequence data available)


Late sampling
Constant sampling
Notifications
Birth-death constant
Birth-death skyline
Coalescent exponential
Basic reproductive number (R_0)
3.5
3.0
2.5
2.0
1.5
Evolutionary rate (subs/site/year)
0.020
0.015
0.010
0.005
Simulation replicate

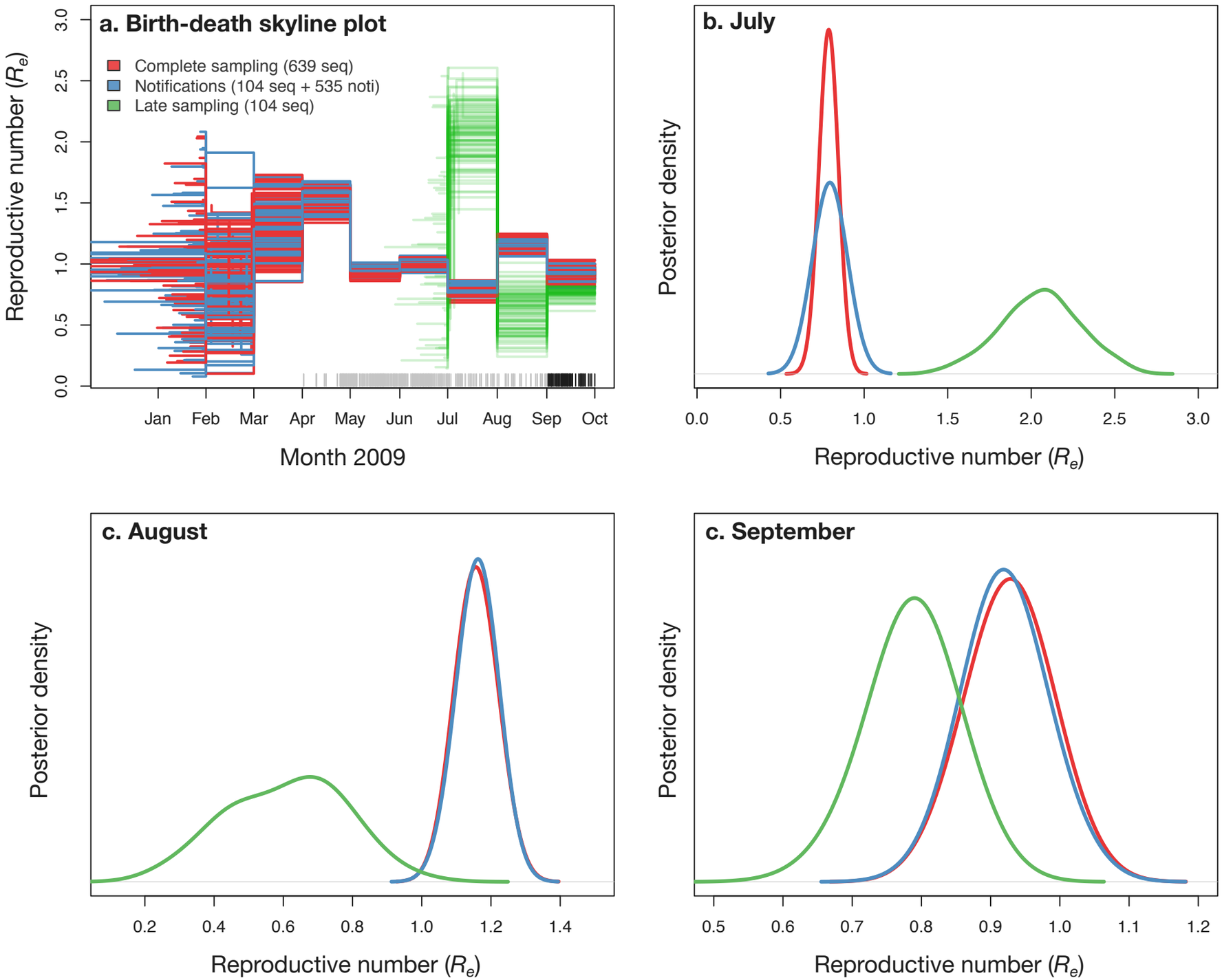

a. Birth-death skyline plot
Complete sampling (639 seq)
Notifications (104 seq + 535 noti)
Late sampling (104 seq)
Reproductive number (R_e)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Month 2009
b. July
Posterior density
Reproductive number (R_e)
c. August
Posterior density
Reproductive number (R_e)
c. September
Posterior density
Reproductive number (R_e)

Table 1. Results of the simulation study with R_0=2, evolutionary rate of 0.01 subs/site/year, and late sampling starting at 0.75 years of a total time of 1 year. The rows correspond to the seven treatments. The first two columns denote the number of simulations (out of 100) where the value used to generate the data was contained within the 95% highest posterior density (HPD). The last two columns are a measure of precision of the estimates calculated as the estimated mean estimate of R_0 and the evolutionary rate divided by the 95% HPD width, such that large values imply low precision. Here we report the mean value over 100 simulations.

	R_0 within 95% HPD	**Evol. rate within 95% HPD**	**Mean R_0 / HPD width**	**Mean evol. rate / HPD width**
Late sampling BD const.	11	58	0.21	0.41
Late sampling BD skyline	96	96	0.28	0.51
Late sampling Coal. exp.	90	93	0.36	0.63
Constant sampling BD const.	97	98	0.18	0.27
Constant sampling Coal. exp.	93	97	0.23	0.29
Notifications BD const.	96	88	0.19	0.37
Notifications Coal. Exp	84	66	0.31	0.49