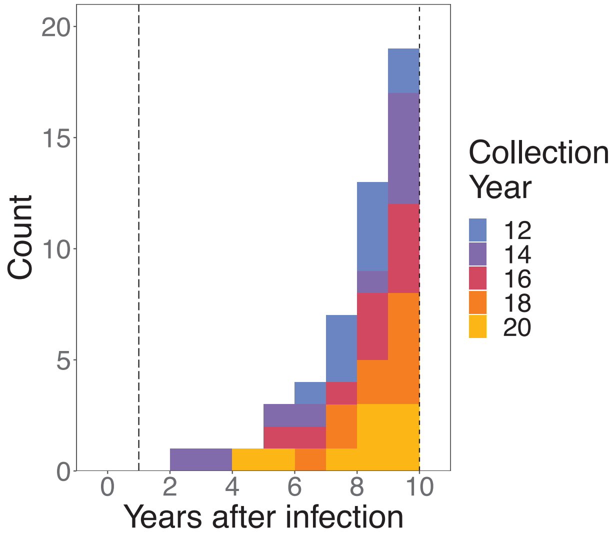
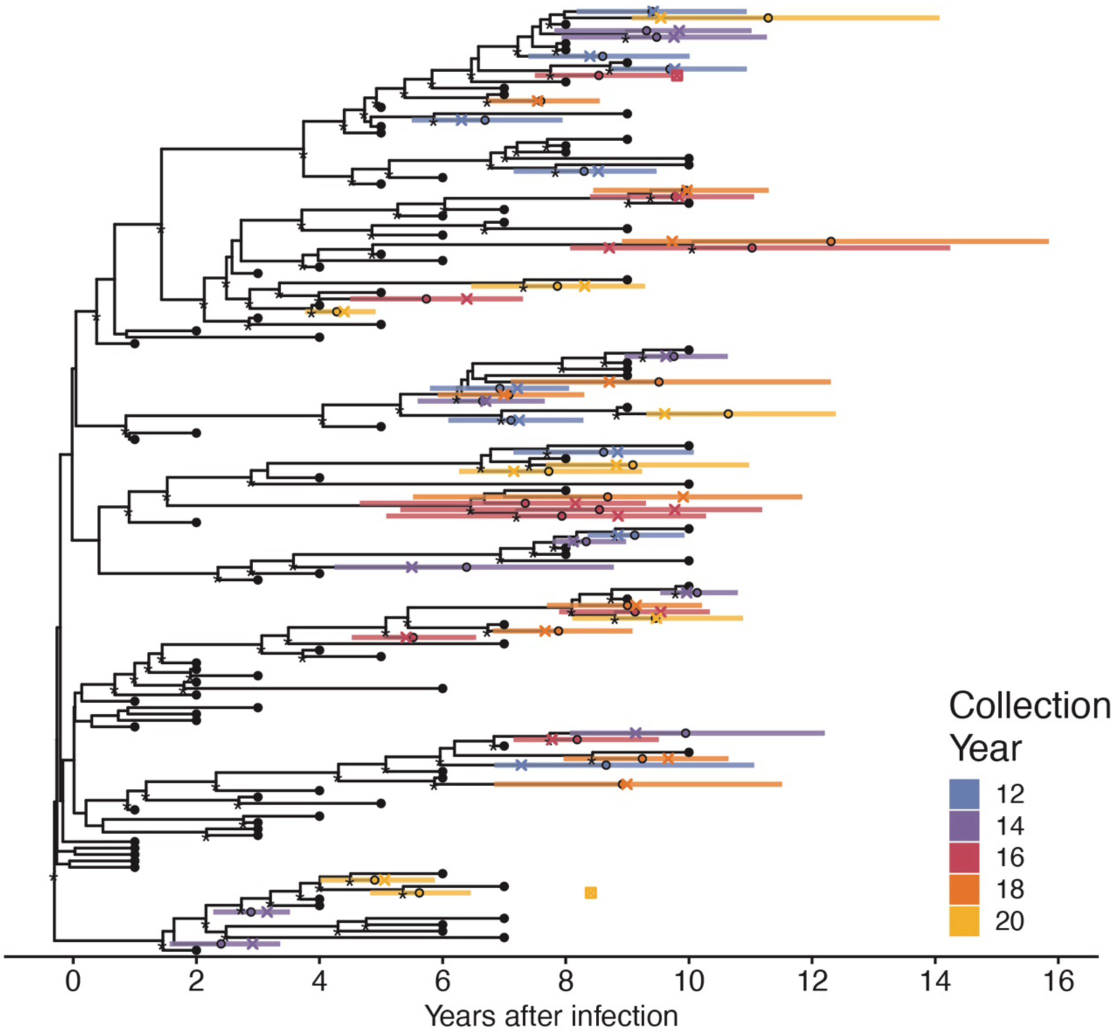
Inferring Human Immunodeficiency Virus 1 Proviral Integration Dates with Bayesian Inference: Supplementary Materials

Bradley R. Jones and Jeffrey B. Joy

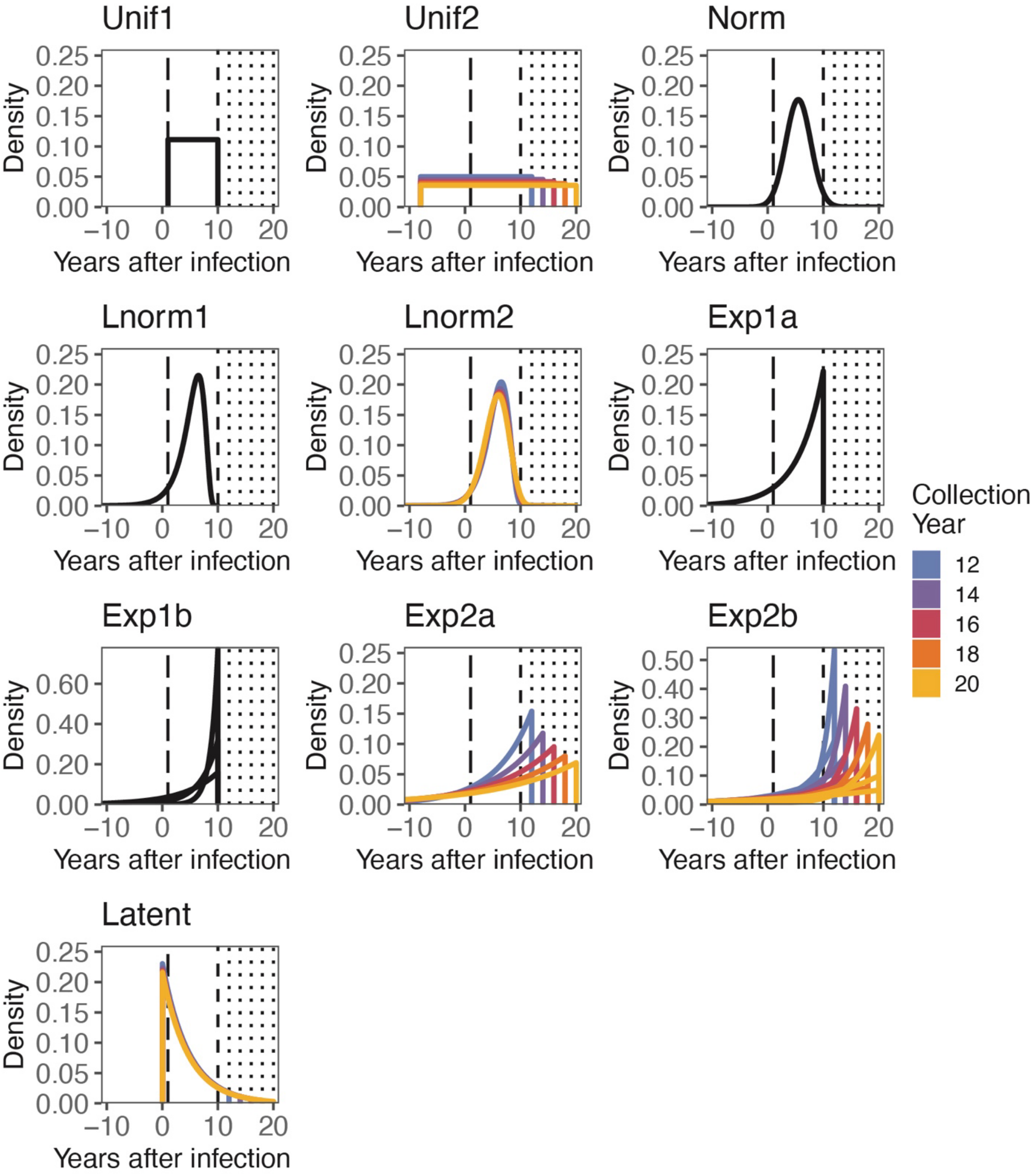
FIGURES



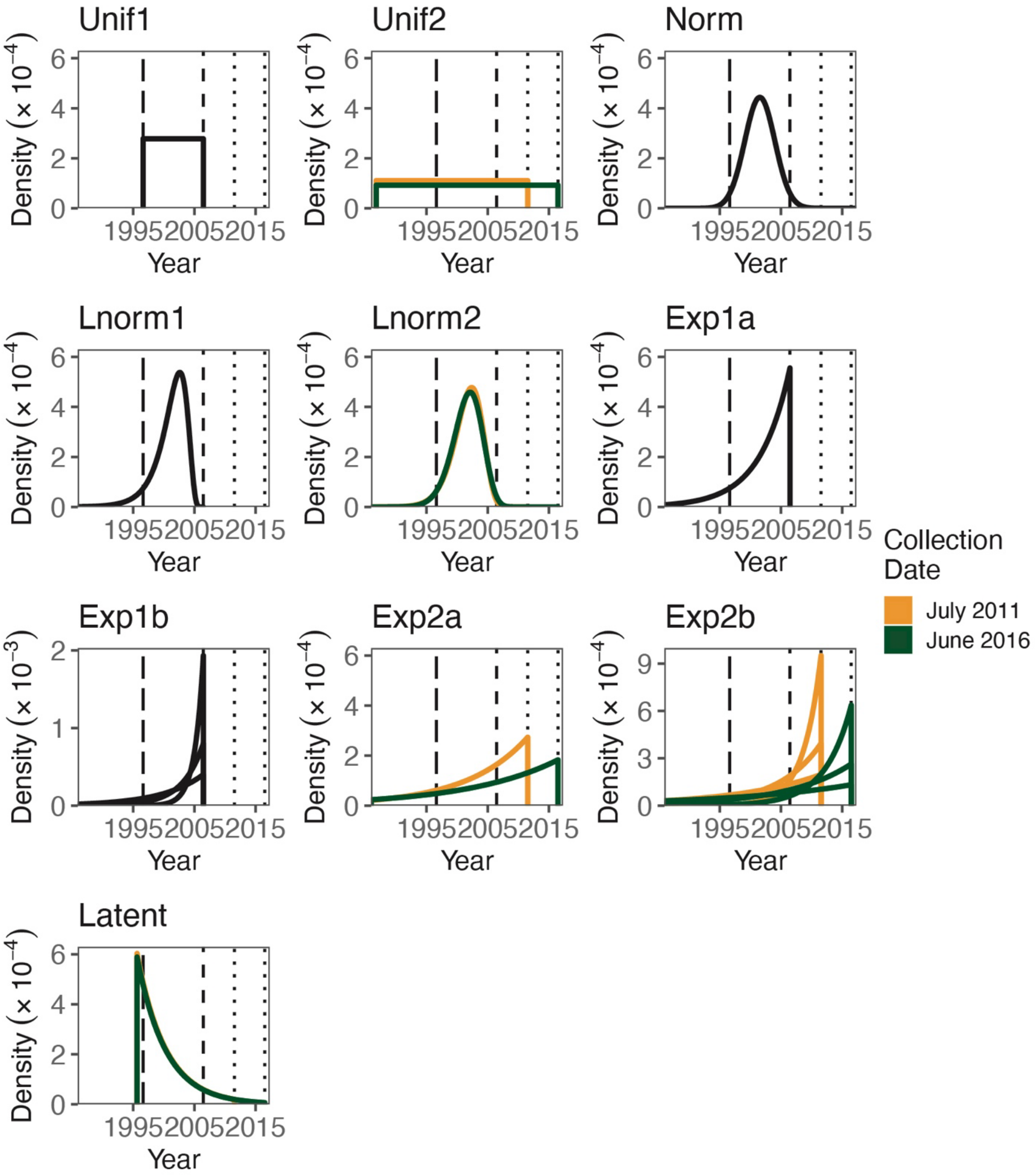
Supplementary Figure 1. Latent sequence integration date distribution of the simulated data set. Colouring shows the collection date of the latent sequences. Dashed lines show the start and end of active sampling.



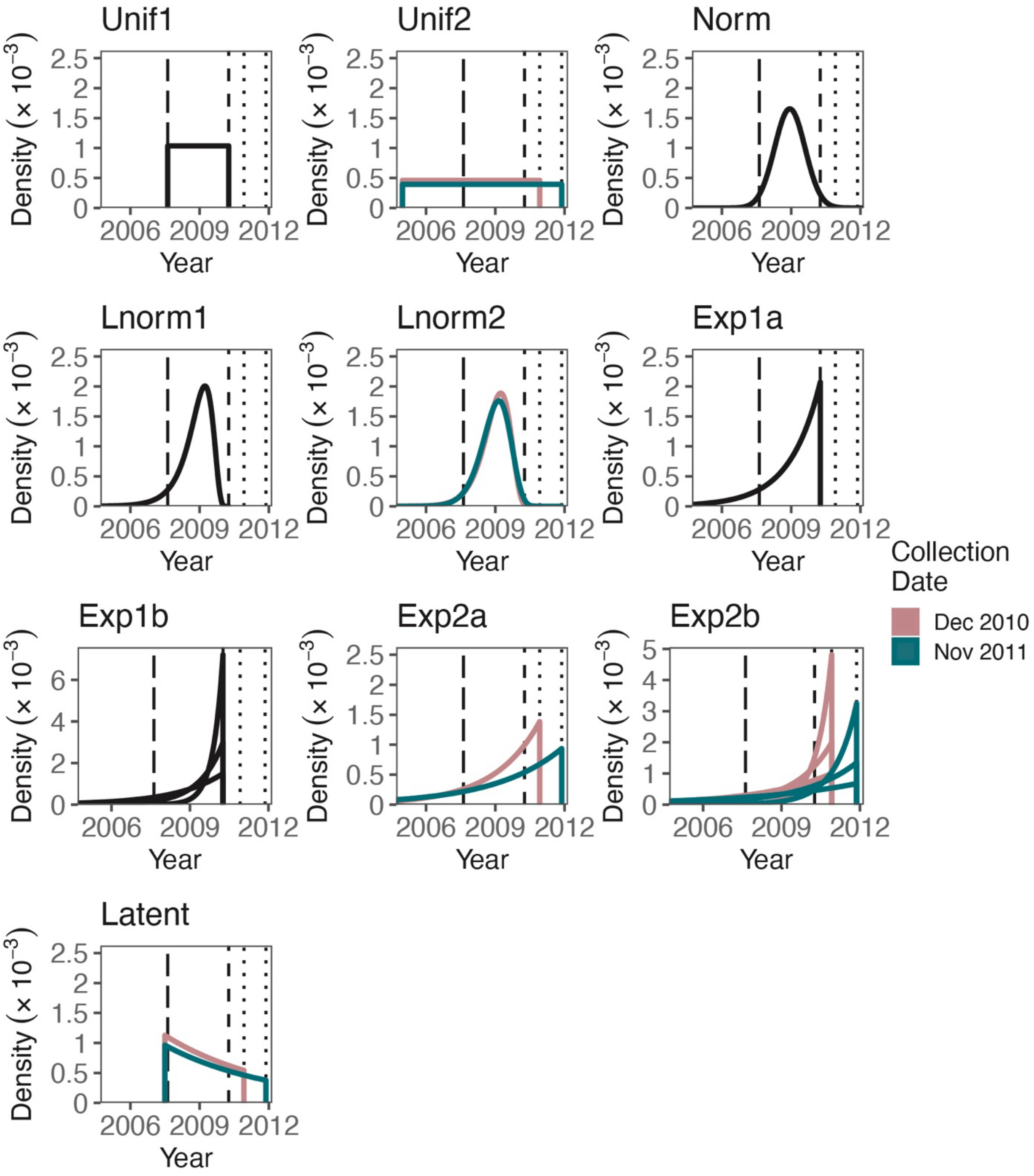
Supplementary Figure 2. Maximum clade credibility tree of date estimation analysis on simulated data with unfixed tree topology. Nodes of the phylogeny are placed at their mean date. Black circles indicate active sequences. Coloured circles indicate the mean integration dates (i.e., estimated dates) and coloured bars indicate 95% highest posterior density intervals for the integration dates. Crosses indicate real integration dates. The small squares indicate the two sequences whose real integration dates fell outside the 95% highest posterior density interval. Nodes with at least 70% posterior clade support are marked with a ‘*’.



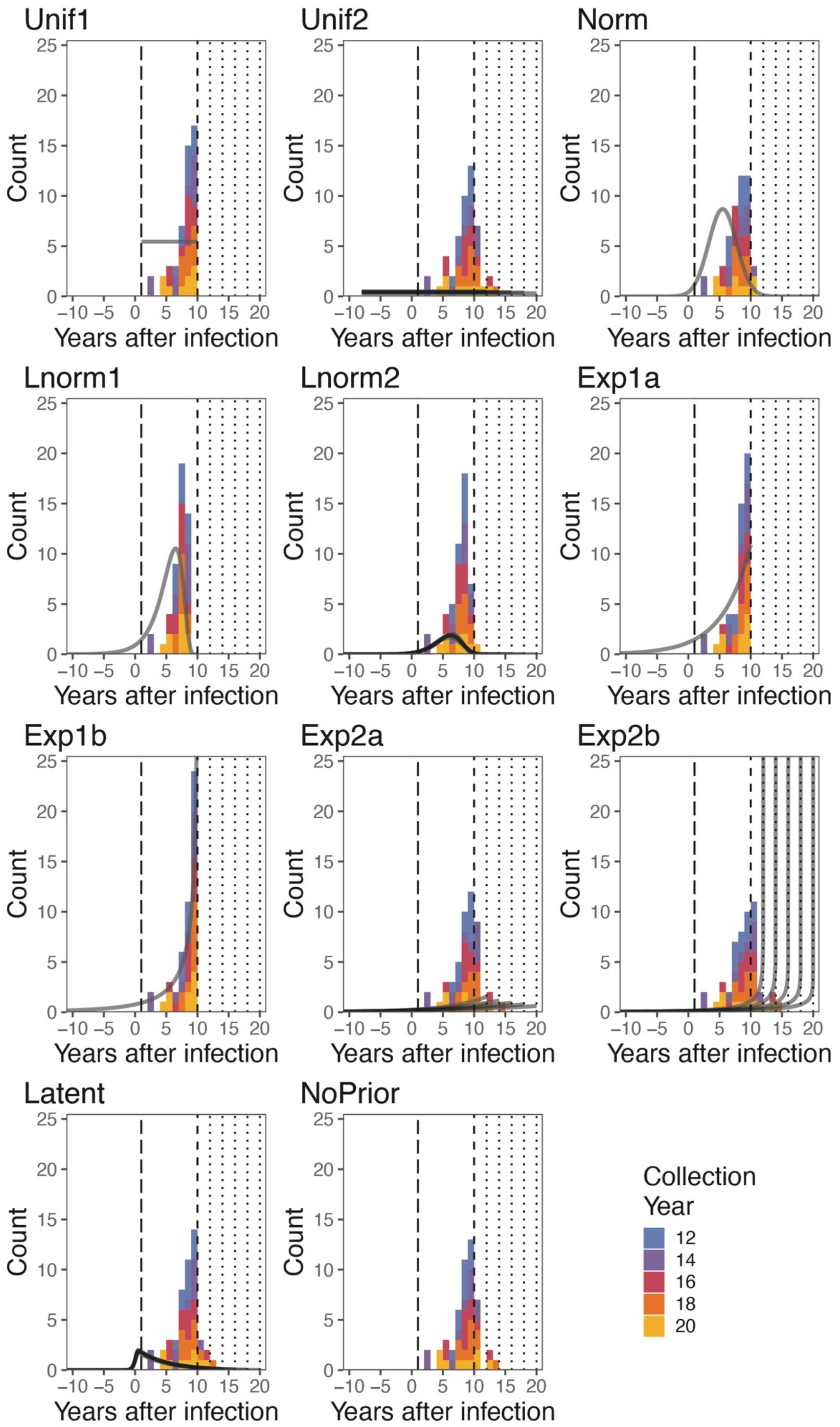
Supplementary Figure 3. BEAST2 tip date priors for simulated data. Dashed vertical lines show start and end of active sampling and dotted lines show latent sampling time points. Colour of outline indicates collection date of latent sequences sampled in years after infection. Black outline means the distribution does not depend on latent sequence collection date. Exp1b and Exp2b plots show prior densities with means at 25%, 50% and 75% quantiles and have an extended y-axis. Latent prior shows distribution for mean root date.



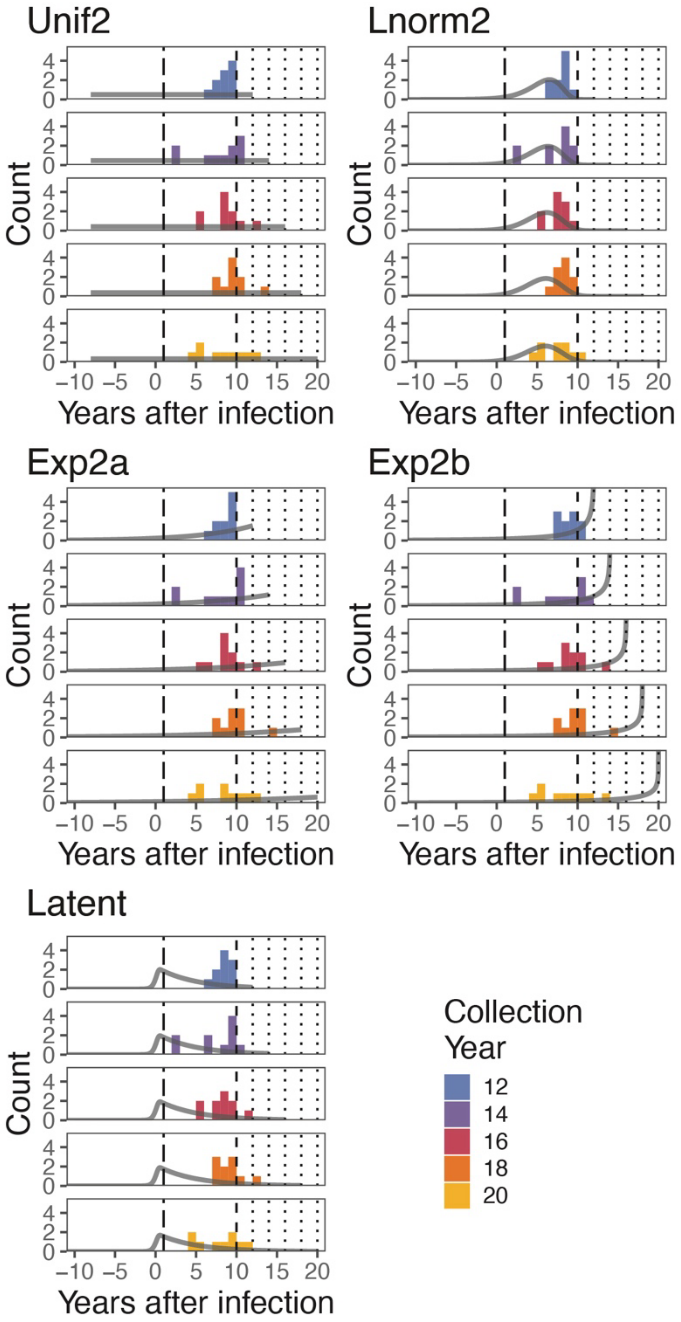
Supplementary Figure 4. BEAST2 tip date priors for P1. Density values were multiplied by 10,000 (1000 for Exp1b). Colour of outline indicates collection date (black outline means distribution does not depend on collection date). Dashed vertical lines show start and end of active sampling and dotted lines show latent sampling time points. Exp1b and Exp2b plots show prior densities with means at 25%, 50% and 75% quantiles and have an extended y-axis. Latent prior shows distribution for mean root date.



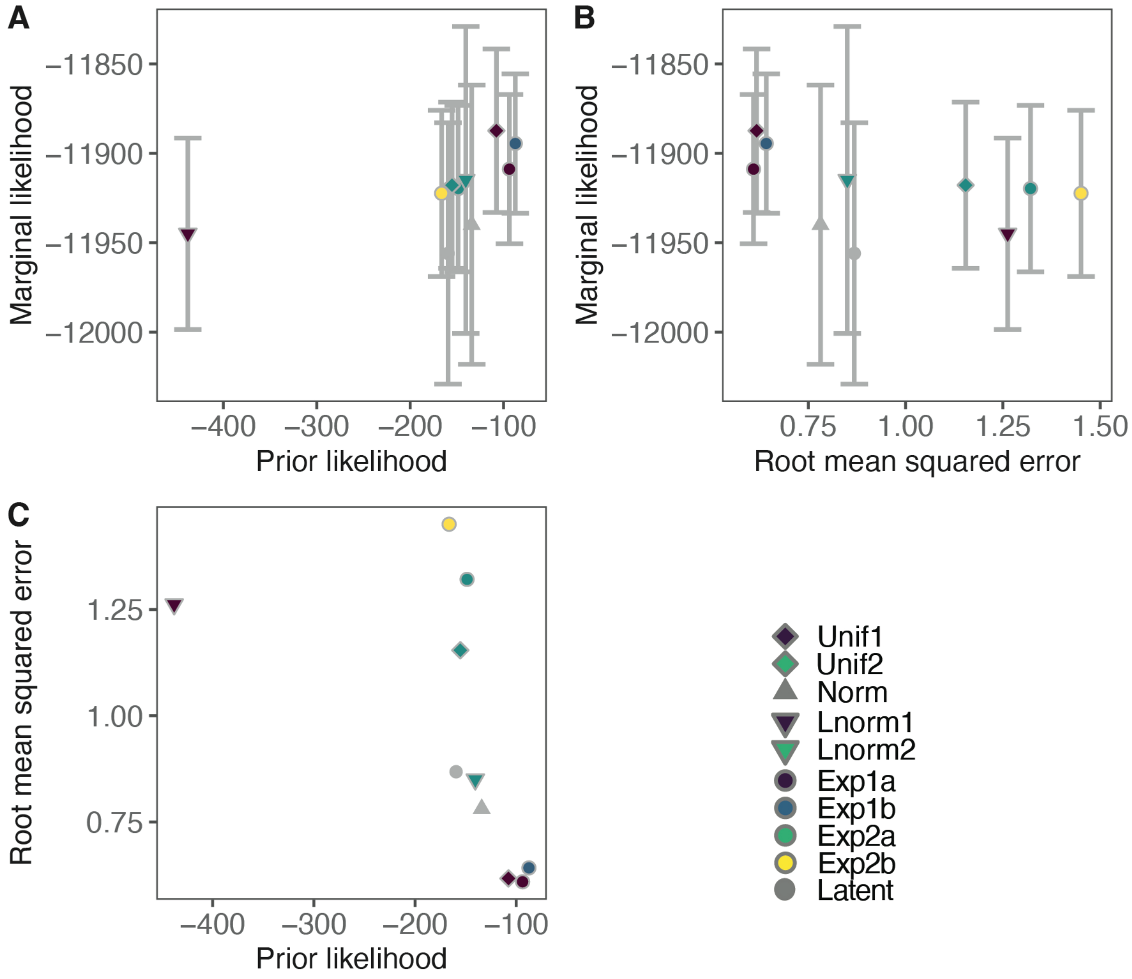
Supplementary Figure 5. BEAST2 tip date priors for N133M. Density values were multiplied by 1000. Colour of outline indicates collection date (black outline means distribution does not depend on collection date). Dashed vertical lines show start and end of active sampling and dotted lines show latent sampling time points. Exp1b and Exp2b plots show prior densities with means at 25%, 50% and 75% quantiles and have an extended y-axis. Latent prior shows distribution for mean root date.


Unif1
Unif2
Norm
Lnorm1
Lnorm2
Exp1a
Exp1b
Exp2a
Exp2b
Latent
NoPrior
Count
0
5
10
15
20
25
Years after infection
−10
−5
0
5
10
15
20
Collection Year
12
14
16
18
20

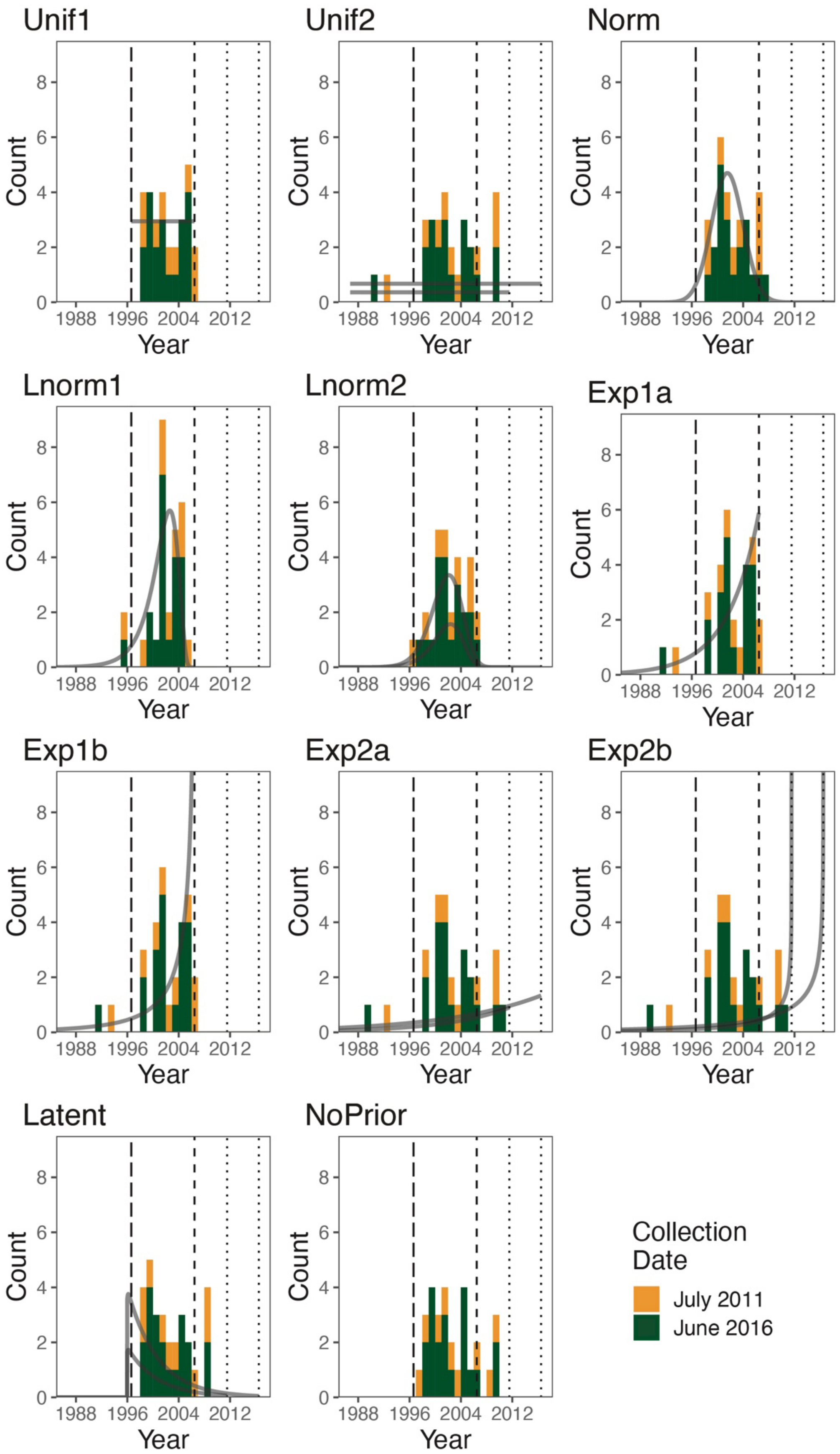
Supplementary Figure 6. Histogram of mean estimated integration dates for simulated data. Colour shows collection date. Dashed lines show the start and end of active sampling and dotted lines show latent sampling times. Grey lines show the prior distribution scaled so that the area under the curve equals the number of latent sequences.



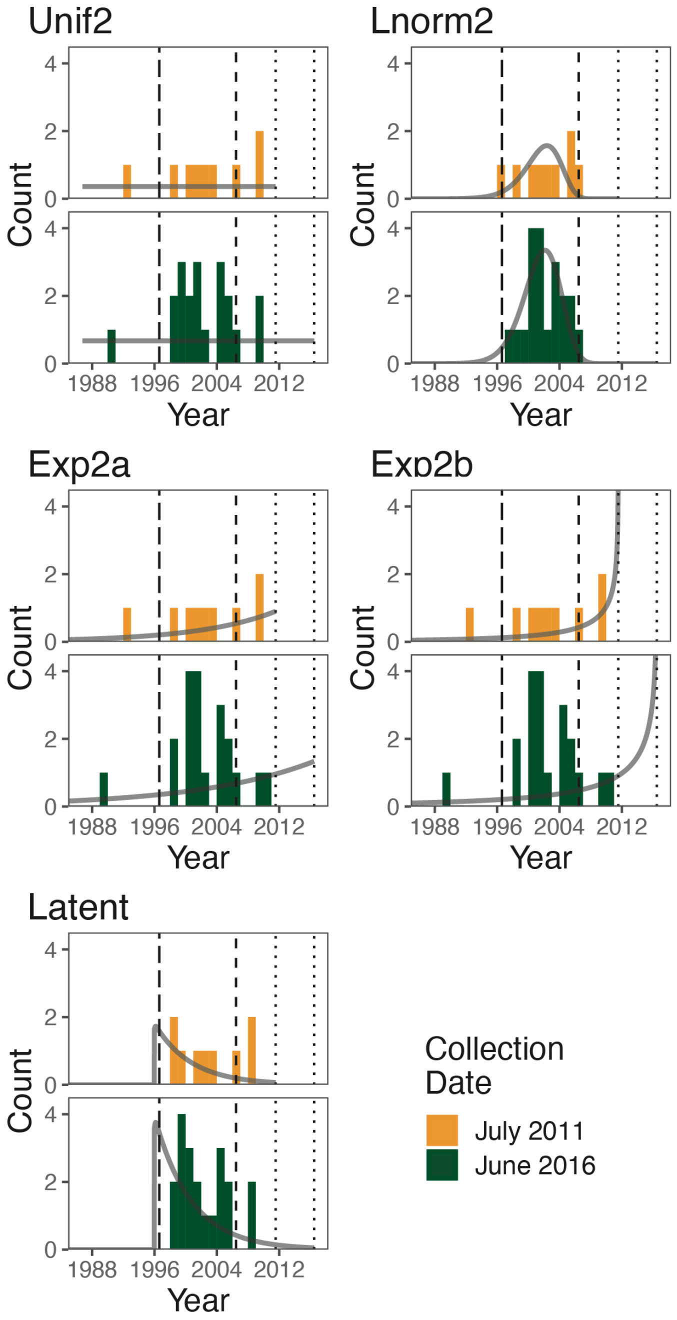
Supplementary Figure 7. Histogram of mean estimated integration dates for simulated data stratified by collection date. Colour shows collection date. Dashed lines show the start and end of active sampling and dotted lines show latent sampling times. Grey lines show the prior distribution scaled so that the area under the curve equals the number of latent sequences.



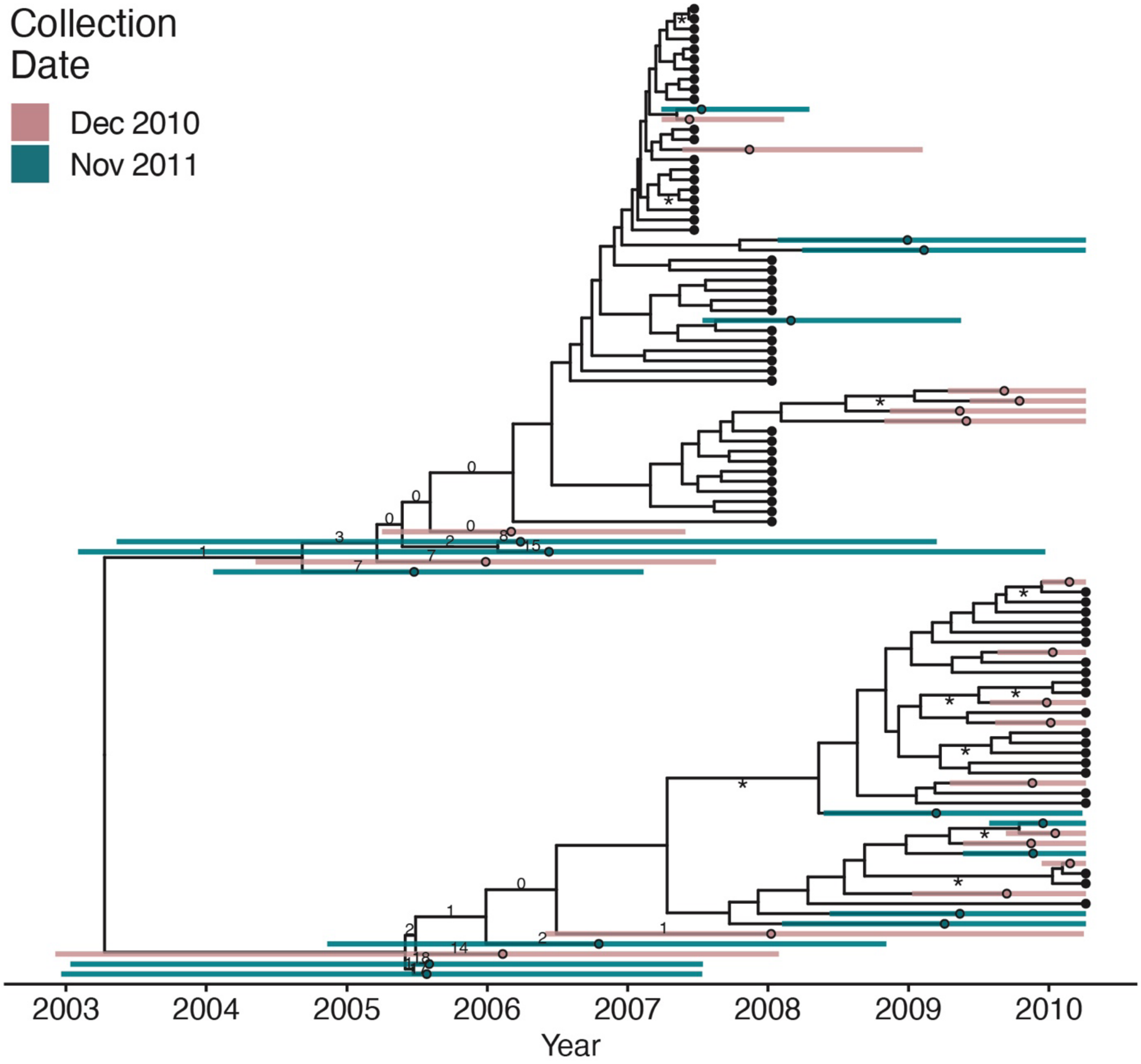
Supplementary Figure 8. Comparison of marginal likelihood, prior likelihood and root mean squared error of simulated data with an unfixed tree. *(A)* The likelihood of the real dates using a specific date prior versus the marginal likelihood of the BEAST2 run using that date prior. Error bars show twice the range of the standard deviation of the estimate of the marginal likelihood. *(B)* Root mean squared error of the date estimation of a BEAST2 run using a specific prior versus the marginal likelihood using that prior. Error bars show twice the range of the standard deviation of the estimate of the marginal likelihood. *(C)* Likelihood of real dates using a specified date prior versus the root mean squared error of the date estimation of the BEAST2 run using that date prior.


Unif1
Unif2
Norm
Lnorm1
Lnorm2
Exp1a
Exp1b
Exp2a
Exp2b
Latent
NoPrior
Count
0
2
4
6
8
Year
1988
1996
2004
2012
Collection Date
July 2011
June 2016

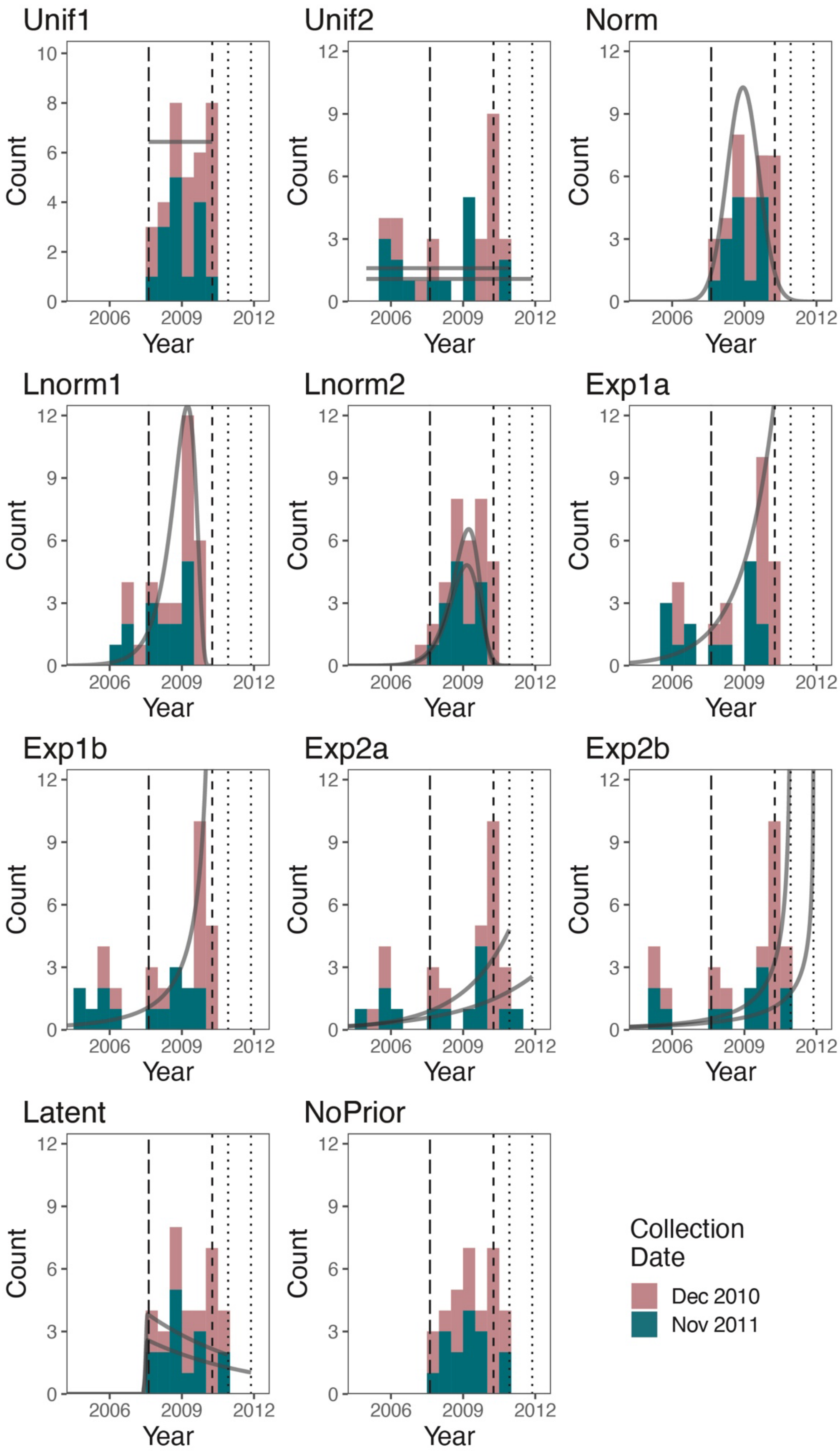
Supplementary Figure 9. Histogram of mean estimated integration dates for P1. Colour shows collection date. Dashed lines show the start and end of active sampling and dotted lines show latent sampling times. Grey lines show the prior distribution scaled so that the area under the curve equals the number of latent sequences.



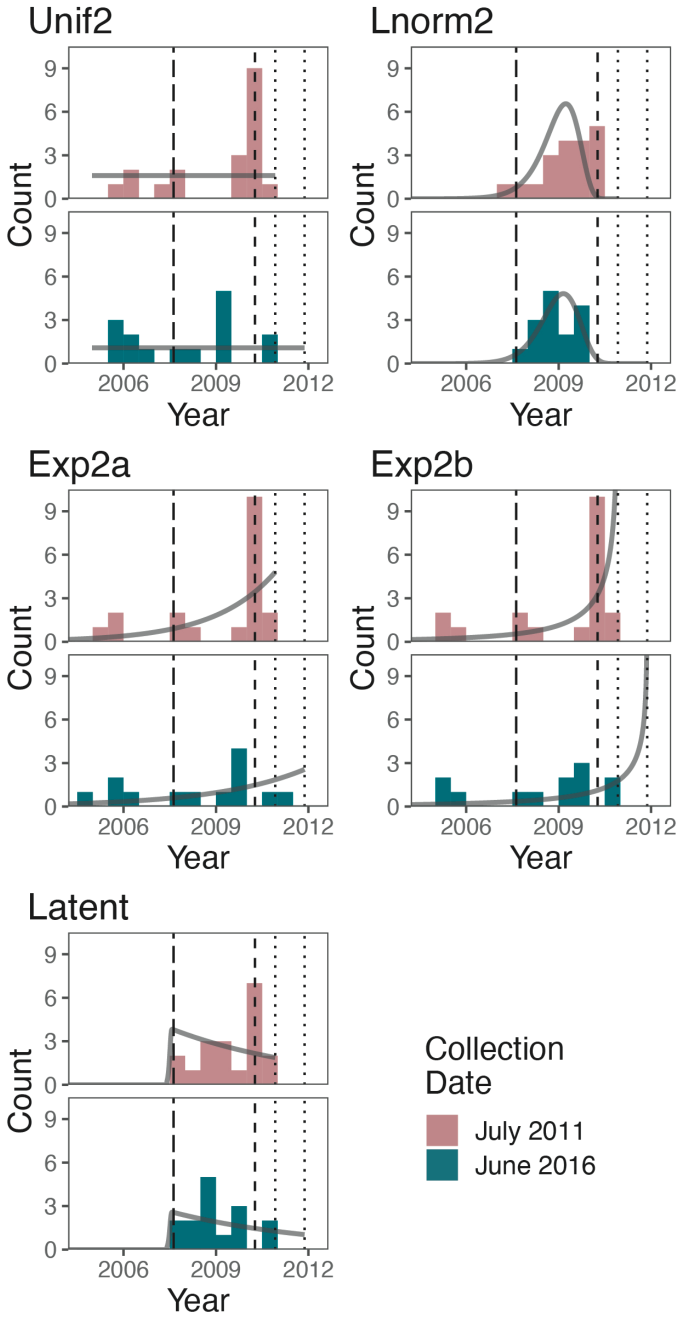
Supplementary Figure 10. Histogram of mean estimated integration dates for empirical data stratified by P1. Colour shows collection date. Dashed lines show the start and end of active sampling and dotted lines show latent sampling times. Grey lines show the prior distribution scaled so that the area under the curve equals the number of latent sequences.



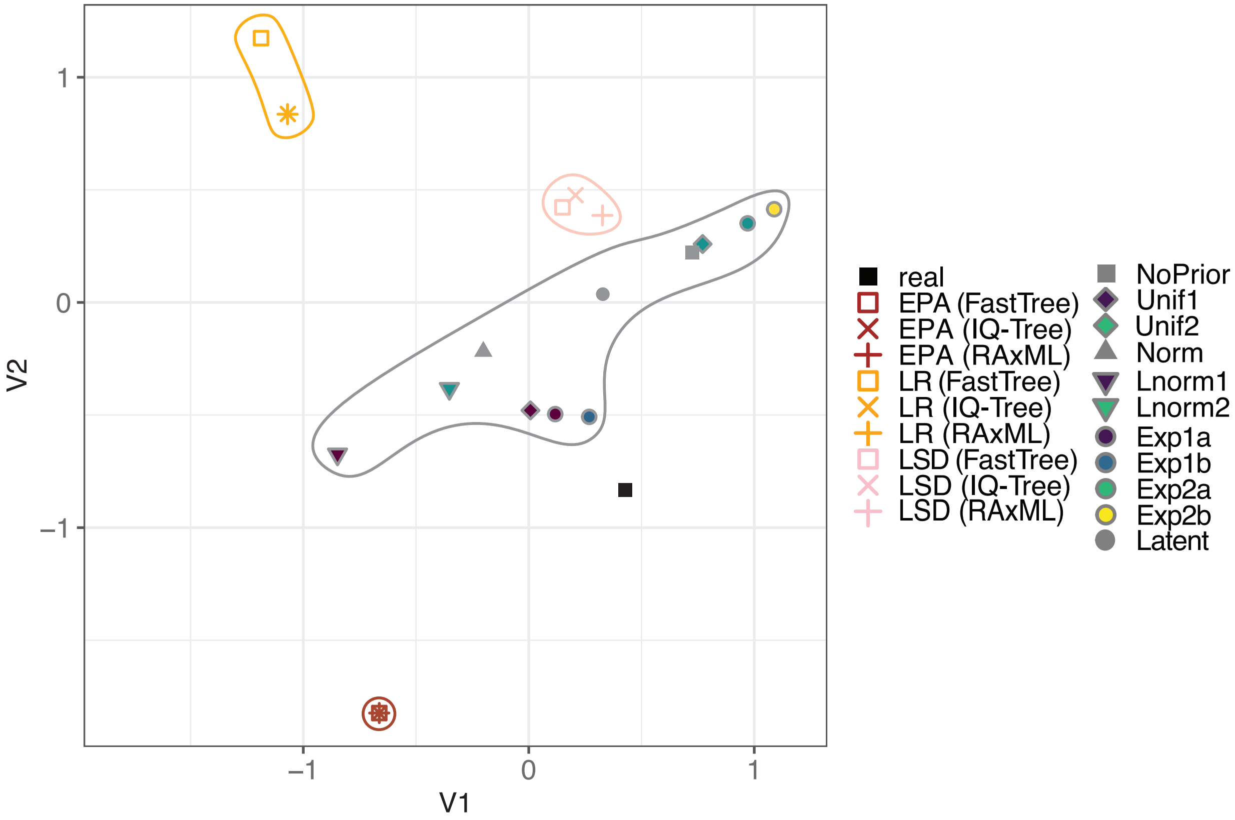
Supplementary Figure 11. Phylogeny of date estimation analysis with the Exp1a prior on N133M. Nodes of the phylogeny are placed at their mean date. Coloured circles indicate mean integration dates (i.e., estimated dates) and coloured bars indicated their 95% highest posterior density intervals for the integration dates. Black circles denote active sequences. Numbers on edges show the percentage of times that edge was sampled as the root after burn-in. Edges without numbers were not sampled as the root. Edges with at least 70% maximum likelihood bootstrap support are marked with a ‘*’.


Unif1
Unif2
Norm
Lnorm1
Lnorm2
Exp1a
Exp1b
Exp2a
Exp2b
Latent
NoPrior
Count
Year
0
2
4
6
8
10
0
3
6
9
12
2006
2009
2012
Collection Date
Dec 2010
Nov 2011

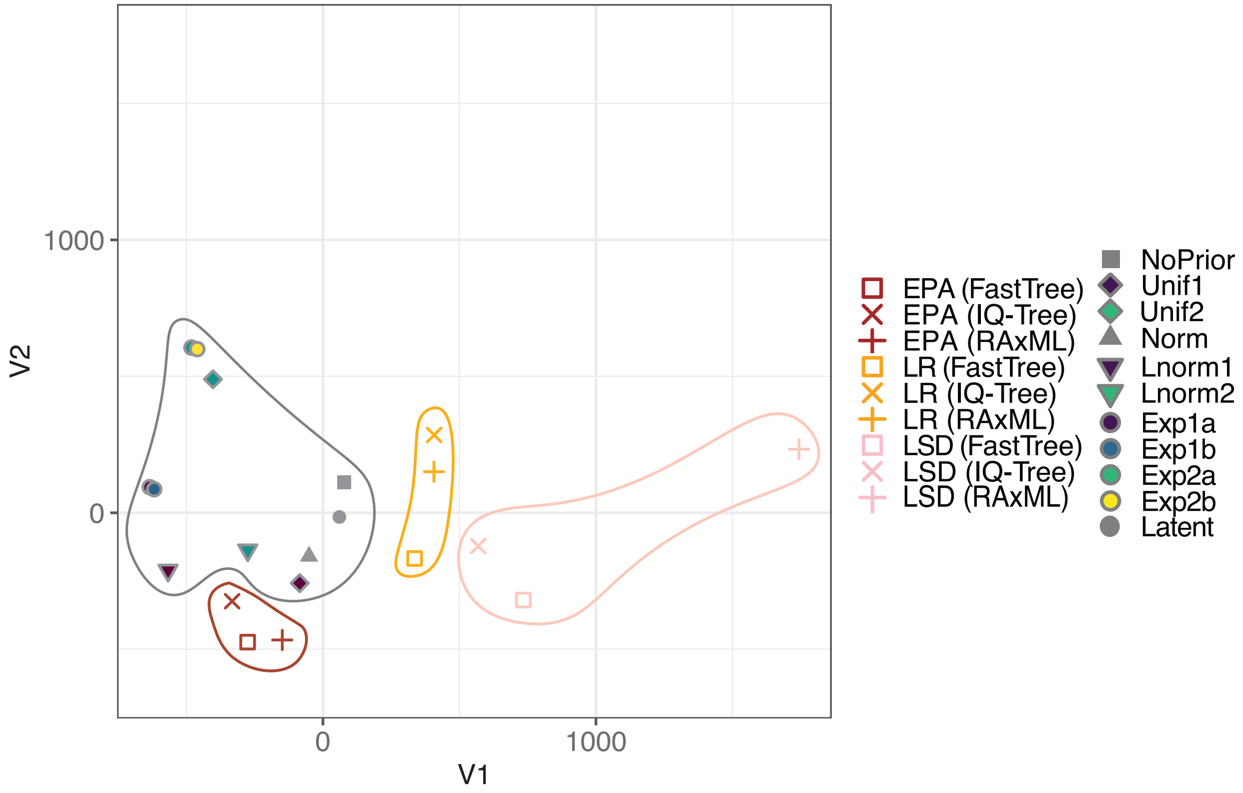
Supplementary Figure 12. Histogram of mean estimated integration dates for N133M. Colour shows collection date. Dashed lines show the start and end of active sampling and dotted lines show latent sampling times. Grey lines show the prior distribution scaled so that the area under the curve equals the number of latent sequences.



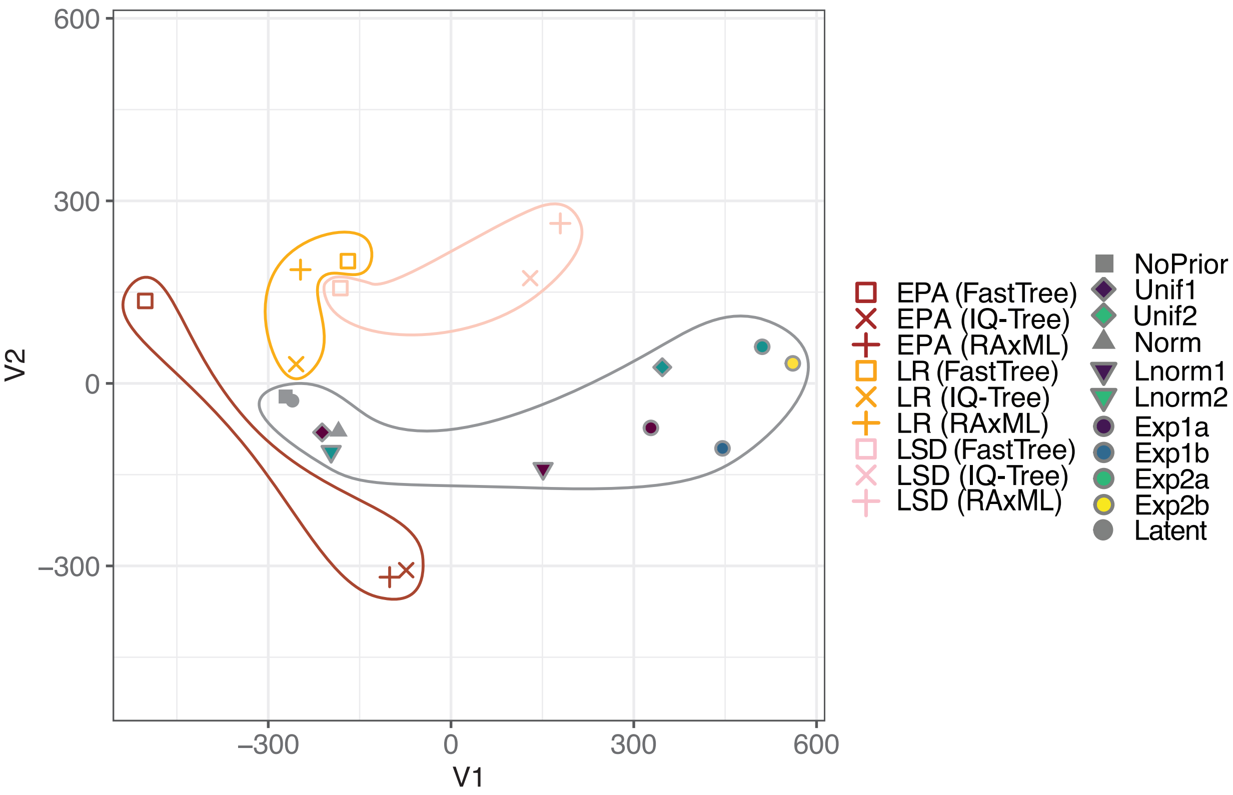
Supplementary Figure 13. Histogram of mean estimated integration dates for empirical data stratified by N133M. Colour shows collection date. Dashed lines show the start and end of active sampling and dotted lines show latent sampling times. Grey lines show the prior distribution scaled so that the area under the curve equals the number of latent sequences.



Supplementary Figure 14. Sammon multidimensional scaling of the methods by the root mean squared deviation of their estimates for simulated data. Root mean squared deviations including EPA were calculated over sequences where the dates were computable. Regions containing each method are circled (EPA: red, LR: orange, LSD: pink, BEAST2: grey). Axes units are years. LR stands for Linear Regression. The red and orange stars correspond to overlapping results of EPA and LR respectively.



Supplementary Figure 15. Sammon multidimensional scaling of the methods by the root mean squared deviation of their estimates for P1. Root mean squared deviations including EPA were calculated over sequences where the dates were computable. Regions containing each method are circled (EPA: red, LR: orange, LSD: pink, BEAST2: grey). Axes units are days. LR stands for Linear Regression.



Supplementary Figure 16. Sammon multidimensional scaling of the methods by the root mean squared deviation of their estimates for N133M. Root mean squared deviations including EPA were calculated over sequences where the dates were computable. Regions containing each method are circled (EPA: red, LR: orange, LSD: pink, BEAST2: grey). Axes units are days. LR stands for Linear Regression.

TABLES

Supplementary Table 1. Substitution model selection. Values show best model according to the given criterion as reported by ModelTest-NG. BIC: Bayesian information criterion, AIC: Akaike information criterion, AICc: Akaike information criterion with correction.

Data set	BIC	AIC	AICc
Simulated	012340[F]+G4	012340[F]+G4	012340[F]
P1	010210[F]+I+G4	012310[F]+I+G4	012310[F]+I+G4
N133M	TIM1+I+G4	012234[F]+I+G4	012234[F]+I+G4

Supplementary Table 2. BEAST2 clock and tree priors considered for model selection.

Type	Model	Priors
Clock	Strict clock	Rate: Lognormal with real mean 0.013 sub/site/year (Cuevas, et al. 2015) or 1.2E-5 sub/site/day (Zanini, et al. 2017) and sigma 5
	Relaxed clock exponential	Mean rate: Lognormal with real mean 0.013 sub/site/year (Cuevas, et al. 2015) or 1.2E-5 sub/site/day (Zanini, et al. 2017) and sigma 5
	Relaxed clock log normal	Mean rate: Lognormal with real mean 0.013 sub/site/year (Cuevas, et al. 2015) or 1.2E-5 sub/site/day (Zanini, et al. 2017) and sigma 5 Standard Deviation: Gamma with alpha 0.5396 and beta 0.3819
Tree	Coalescent constant population	Population size: Lognormal with mean 1 and sigma 4
	Coalescent exponential population	Population size: Lognormal with mean 0 and sigma 5 Growth rate: Laplace with mu 0.001 and scale 30.701135

Supplementary Table 3. BEAST2 prior selection. Entries show the marginal log likelihood plus standard deviation. Bold and italicized text highlights the best model for each data set. Italicized text highlights models where the difference in the marginal log likelihood from the best model is less than twice the sum of their standard deviations.

Data set	Clock model	Tree model	Marginal log likelihood
Simulated	*Strict*	*Coalescent constant population*	*-9523.47 ± 12.31*
	Strict	*Coalescent exponential population*	*-9508.64 ± 13.79*
	Relaxed exponential	Coalescent constant population	-9554.44 ± 13.73
	Relaxed exponential	*Coalescent exponential population*	*-9524.84 ± 13.28*
	Relaxed log normal	*Coalescent constant population*	*-9513.48 ± 12.84*
	Relaxed log normal	*Coalescent exponential population*	***-9492.74 ± 9.33***
P1	Strict	Coalescent constant population	-3227.48 ± 0.76
	Strict	Coalescent exponential population	*-3212.55 ± 0.91*
	Relaxed exponential	Coalescent constant population	-3246.41 ± 1.23
	Relaxed exponential	Coalescent exponential population	-3235.87 ± 1.03
	Relaxed log normal	Coalescent constant population	-3225.45 ± 0.55
	Relaxed log normal	*Coalescent exponential population*	***-3209.68 ± 0.53***
N133M	Strict	Coalescent constant population	-6992.20 ± 0.40
	Strict	Coalescent exponential population	-6989.72 ± 0.33
	Relaxed exponential	Coalescent constant population	-6997.16 ± 0.76
	Relaxed exponential	Coalescent exponential population	-6995.19 ± 0.59
	Relaxed log normal	Coalescent constant population	-6986.30 ± 0.41
	Relaxed log normal	*Coalescent exponential population*	***-6983.56 ± 0.38***

Supplementary Table 4. BEAST2 tip date priors. Name column gives an identifier for the distribution. Prior column shows the distribution. Variable column indicates what variable the distribution is acting on (e.g., date or time before a specific time point). Parameters column describes the values of the parameters for the distribution.

Name	**Prior**	**Variable**	**Parameters**
Unif1	Uniform	Date	Lower bound: earliest active sequence sampling time Upper bound: latest active sampling time
Unif2	Uniform	Date	Lower bound: earliest active sequence sampling time minus active sampling range Upper bound: sampling date
Norm	Normal	Date	Mean: active sampling midpoint Sigma: one fourth of the active sampling range
Lnorm1	Log normal	Years/days before latest active sampling date	Real mean: active sampling midpoint Sigma: chosen to make real standard deviation one fourth of active sampling range
Lnorm2	Log normal	Years/days before sampling date	Real mean: active sampling midpoint Sigma: chosen to make real standard deviation one fourth of active sampling range
Exp1a	Exponential	Years/days before latest active sampling date	Mean: active sampling midpoint
Exp1b	Exponential	Years/days before latest active sampling date	Mean: estimated with an exponential prior whose mean is the active sampling midpoint
Exp2a	Exponential	Years/days before sampling date	Mean: active sampling midpoint
Exp2b	Exponential	Years/days before sampling date	Mean: estimated with an exponential prior whose mean is the active sampling midpoint
Latent	Latency	Date	Become latent rate: 0.364 year^{-1} / 1 × 10^{-3} day^{-1} Reactivation rate: 0.151 year^{-1} / 4.15 × 10^{-4} day^{-1} Log normal prior on root date with mu: 0 log years (simulated), 5.900 log days (P1), 3.829 log days (N133M); sigma: 0.347
NoPrior	None	Date	Log normal prior on root date with mu: 0 log years (simulated), 5.900 log days (P1), 3.829 log days (N133M); sigma: 0.347

Supplementary Table 5. BEAST2 date prior selection (simulated data). Bold text highlights optimal values. Italicized text highlights models where the difference in the marginal log likelihood from the best model is less than twice the sum of their standard deviations. RMSE stands for Root Mean Squared Error and is in units of years. PII stands for Percent real In estimated 95% highest posterior Interval. Mean HPD is the mean width of the 95% highest posterior density interval for the integration dates and is in years.

Prior	Log likelihood	RMSE	Concordance	PII	Mean HPD
NoPrior	—	1.27	0.842	0.939	3.20
Unif1	-11609.20 ± 0.43	0.64	0.933	**0.959**	1.96
Unif2	-11625.27 ± 0.45	1.17	0.834	0.939	3.33
Normal	-11619.10 ± 0.47	0.82	0.893	0.939	2.52
Lnorm1	-11642.18 ± 1.30	1.25	0.730	0.531	2.07
Lnorm2	-11623.24 ± 0.44	0.86	0.876	0.878	2.26
Exp1a	***-11603.52 ± 0.45***	**0.62**	**0.937**	**0.959**	**1.88**
Exp1b	*-11604.27 ± 0.40*	0.65	0.931	0.939	1.85
Exp2a	-11632.56 ± 0.66	1.33	0.800	0.939	3.56
Exp2b	-11642.30 ± 0.77	1.42	0.779	0.919	3.78
Latent	-11614.48 ± 0.47	0.89	0.890	**0.959**	2.97

Supplementary Table 6. BEAST2 date prior selection (simulated data unfixed). Bold text highlights optimal values. Italicized text highlights models where the difference in the marginal log likelihood from the best model is less than twice the sum of their standard deviations. RMSE stands for Root Mean Squared Error and is in units of years. PII stands for Percent real In estimated 95% highest posterior Interval. Mean HPD is the mean width of the 95% highest posterior density intervals for the integration dates and is in years.

Prior	Log likelihood	RMSE	Concordance	PII	Mean HPD
NoPrior	—	1.11	0.846	0.939	3.31
Unif1	*-11887.38 ± 45.66*	0.62	0.937	**0.959**	1.98
Unif2	*-11917.84 ± 46.45*	1.15	0.837	0.939	3.41
Normal	*-11939.91 ± 78.09*	0.78	0.901	**0.959**	2.57
Lnorm1	*-11944.96 ± 53.51*	1.26	0.724	0.531	2.13
Lnorm2	*-11914.91 ± 85.82*	0.85	0.878	0.878	2.29
Exp1a	*-11908.85 ± 41.71*	**0.61**	**0.938**	0.939	1.92
Exp1b	***-11894.53 ± 38.96***	0.64	0.931	**0.959**	**1.89**
Exp2a	*-11919.71 ± 46.53*	1.32	0.801	0.939	3.71
Exp2b	*-11922.38 ± 46.45*	1.45	0.773	0.919	3.92
Latent	*-11955.98 ± 36.52*	0.87	0.893	**0.959**	3.03

Supplementary Table 7. BEAST2 date prior selection (P1). Bold text highlights optimal values. Italicized text highlights models where the difference in the marginal log likelihood from the best model is less than twice the sum of their standard deviations. Mean HPD is the mean width of the 95% highest posterior density intervals for the integration dates and is in days.

Prior	**Log likelihood**	**Mean HPD**
NoPrior	—	1412
Unif1	-3712.47 ± 0.16	**934**
Unif2	-3710.94 ± 0.17	1495
Normal	-3712.29 ± 0.16	1123
Lnorm1	-3714.47 ± 0.18	1141
Lnorm2	-3716.57 ± 0.18	1139
Exp1a	-3709.89 ± 0.16	1174
Exp1b	-3711.04 ± 0.17	1191
Exp2a	-3716.07 ± 0.16	1774
Exp2b	-3716.77 ± 0.16	1820
Latent	***-3705.12 ± 0.16***	1122

Supplementary Table 8. BEAST2 date prior selection (N133M). Bold text highlights optimal values. Italicized text highlights models where the difference in the marginal log likelihood from the best model is less than twice the sum of their standard deviations. Mean HPD is the mean width of the 95% highest posterior density intervals for the integration dates and is in days.

Prior	**Log likelihood**	**Mean HPD**
NoPrior	—	287
Unif1	-9414.53 ± 0.26	**220**
Unif2	-9410.47 ± 0.25	734
Normal	-9420.39 ± 0.26	260
Lnorm1	-9428.00 ± 0.32	755
Lnorm2	-9428.09 ± 0.28	257
Exp1a	***-9408.34 ± 0.22***	765
Exp1b	-9411.53 ± 0.24	895
Exp2a	-9409.44 ± 0.22	1102
Exp2b	-9410.46 ± 0.24	1169
Latent	-9418.91 ± 0.25	276

Supplementary Table 9. Alternate dating methods (simulated data). RMSE stands for Root Mean Squared Error and is in years. Score is the placement probability of the tips for EPA, the difference between the Akaike Information Criterion (AIC) of the null model with a slope of zero and the AIC of the linear model for Linear Regression and the objective function for LSD. Optimal scores are highlighted in bold. Percent Computable shows the percent of latent sequences for which estimates could be computed. [a]EPA RMSE and concordance were calculated over sequences where dates were computable.

Method	Tree	RMSE	Concordance	Score	Percent Computable
EPA	FastTree	1.64 [a]	0.453 [a]	**3.20E-3**	73.5
EPA	IQ-Tree	1.64 [a]	0.453 [a]	3.14E-3	73.5
EPA	RAxML	1.64 [a]	0.453[a]	3.15E-3	73.5
Linear Regression	FastTree	2.15	0.500	181.8	100
Linear Regression	IQ-Tree	1.89	0.572	**189.2**	100
Linear Regression	RAxML	1.89	0.572	189.2	100
LSD	FastTree	1.23	0.802	0.167	100
LSD	IQ-Tree	1.23	0.804	**0.136**	100
LSD	RAxML	1.17	0.822	0.143	100

Supplementary Table 10. Alternate dating methods (P1). Score is the placement probability of the tips for EPA, the difference between the Akaike Information Criterion (AIC) of the null model with a slope of zero and the AIC of the linear model for Linear Regression and the objective function for LSD. Optimal scores are highlighted in bold. Percent Computable shows the percent of latent sequences for which estimates could be computed.

Method	Tree	Score	Percent Computable
EPA	FastTree	6.29E-5	69.0
EPA	IQ-Tree	4.76E-4	75.9
EPA	RAxML	**1.51E-3**	89.7
Linear Regression	FastTree	**196.2**	100
Linear Regression	IQ-Tree	175.7	100
Linear Regression	RAxML	158.4	100
LSD	FastTree	**0.036**	100
LSD	IQ-Tree	0.102	100
LSD	RAxML	0.132	100

Supplementary Table 11. Alternate dating methods (N133M). Score is the placement probability of the tips for EPA, the difference between the Akaike Information Criterion (AIC) of the null model with a slope of zero and the AIC of the linear model for Linear Regression and the objective function for LSD. Optimal scores are highlighted in bold. Percent Computable shows the percent of latent sequences for which estimates could be computed.

Method	Tree	Score	Percent Computable
EPA	FastTree	**9.39E-1**	100
EPA	IQ-Tree	5.81E-1	100
EPA	RAxML	2.37E-1	100
Linear Regression	FastTree	216.3	100
Linear Regression	IQ-Tree	231.2	100
Linear Regression	RAxML	**238.6**	100
LSD	FastTree	**0.015**	100
LSD	IQ-Tree	0.028	100
LSD	RAxML	0.026	100

Supplementary Table 12. BEAST2 date estimation run parameters. Models not explicated shown for the simulated data are represented by a "—".

Data set	Model	Chain length	Operator scheme	Split?
Simulated	—	200,000,000	Normal	No
	NoPrior (both) Latent (both) Norm (unfixed) Lnorm2 (unfixed)	500,000,000	Normal	No
P1	Unif1 Unif2	200,000,000	Random walker operator added for 3 difficult-to-date sequences	No
	Exp1a Exp2a Exp2b Latent	500,000,000	Normal	No
	Norm Exp1b	500,000,000	Random walker operator added for 5 difficult-to-date sequences and increased RootExchange weight	No
	NoPrior	10,000,000,000	Random walker operator added for 5 difficult-to-date sequences and random walker operator added to move two sequences at once	10×
	Lnorm1 Lonrm2	50,000,000,000	Random walker operator added for 5 difficult-to-date sequences and random walker operator added to move two sequences at once	10×
N133M	NoPrior Lnorm2 Exp1a Latent	500,000,000	Normal	No
	Unif1 Norma	500,000,000	DateOp4	No
	Lnorm1 Unif2 Exp1b Exp2a Exp2b	500,000,000 50,000,000,000	DateOp5	No