An efficient coalescent epoch model for Bayesian phylogenetic inference

Remco R. Bouckaert
r.bouckaert@auckland.ac.nz
University of Auckland

June 29, 2021

Abstract

We present a two headed approach called Bayesian Integrated Coalescent Epoch PlotS (BICEPS) for efficient inference of coalescent epoch models. Firstly, we integrate out population size parameters and secondly we introduce a set of more powerful Markov chain Monte Carlo (MCMC) proposals for flexing and stretching trees. Even though population sizes are integrated out and not explicitly sampled through MCMC, we are still able to generate samples from the population size posteriors, which allows demographic reconstruction through time. Altogether, BICEPS can be considered a more muscular version of the popular Bayesian skyline model.

We demonstrate its power and correctness by a well calibrated simulation study. Furthermore, we demonstrate with an application to COVID-19 genomic data that some analyses that have trouble converging with the traditional Bayesian skyline prior and standard MCMC proposals can do well with the BICEPS approach.

BICEPS is available as open source package for BEAST 2 under GPL license and has a user friendly graphical user interface.

1 Introduction

Muscling through Bayesian MCMC analyses is often hampered by convergence issues with the tree prior in our experience. Coalescent tree priors based on Kingman's theory (Kingman, 1982) are popular in phylogenetic analyses. These tree priors are driven by a population function that define the effective population size through time. A population function can be parametric, like exponential or constant (Kuhner et al., 1998), but non-parametric methods that split up the time frame spanning a tree into epochs allow a population function to be constant in an epoch but vary over time. The classic skyline model (Pybus et al., 2000), introduced in a maximum likelihood framework, is based on epochs for every coalescent event and later generalised to epochs grouping coalescent events in the generalised skyline model (Strimmer and Pybus, 2001). The Bayesian skyline plot (Drummond et al., 2005) generalised this to a Bayesian setting, where epochs span multiple coalescent events, and the number of coalescent events as well as population sizes for an epoch are sampled during MCMC. Furthermore, a smoothing

prior is employed that links population sizes in consecutive epochs. Other popular epoch based coalescent models with different smoothing priors include the skyride (Minin et al., 2008), skygrid (Gill et al., 2013; Hill and Baele, 2019), and besp (Parag et al., 2020) models.

All the above Bayesian methods sample the population function parameters. By assuming an inverse gamma prior distribution on population size, we demonstrate that the population size can be integrated out during MCMC. The technique is used in the multispecies coalescent models StarBeast2 (Ogilvie et al., 2017) and STACEY (Jones, 2017), where a constant population size is associated with each branch of the species tree. Here, we generalise this method to the case where we have a single tree, potentially with sampled tip dates, and assume a piecewise constant population for each of epoch under an inverse gamma prior. The mean for the population size of the youngest epoch can be sampled but for consecutive epochs the posterior mean of the previous epoch can be used, providing us with a smoothing prior. Even though population sizes are integrated out, they can still be sampled from the posterior population sizes for the epochs conditioned on the tree. This allows us to reconstruct population size history including uncertainty intervals in a similar fashion as for the Bayesian skyline plot.

Apart from introducing a more efficient way to infer population size histories at different epochs, we also consider a number of new MCMC proposals that can lift a large number of nodes in a tree simultaneously. Observing that tree priors tend to be highly correlated with the length of a tree, we target tree length changes by moving nodes in randomly chosen time intervals (not necessarily the ones used for the tree prior). Furthermore, noting that scaling of trees tend to be hampered by serially sampled tips we design a new scale proposal that moves all nodes in a tree simultaneously but with better exploratory powers than standard scalers. Both proposals move tree length, and since clock rates tend to be inversely correlated with the tree length (Douglas et al., 2021b), we designed proposals that simultaneously move the clock rate to compensate for a changing tree length. We demonstrate the effectiveness of these MCMC proposals for improving mixing of tree lengths, and thus tree priors.

Together, integrating out population sizes and employing more sophisticated MCMC proposals allow us to do inference efficiently, and make it possible to perform larger analyses, as we demonstrate using COVID-19 data. In the next section, we consider the technical details around integrating out parameters and new MCMC proposals. Section 3 continues with validating the method and Section 4 presents results. In the conclusions (Section 5) we consider ways to generalise the approach and in particular point out how to integrate out parameters for an epoch version of the Yule prior (Appendix B and C).

2 Methods

First, we consider integrating out population size parameters, then we design a set of new MCMC proposals.

2.1 Integrating out parameters

Let T be a rooted binary tree with n taxa sampled at r different times.[1] So, $r = 1$ when all taxa are sampled at the same time and $r = n$ when all taxa are sampled at different times. Then there are $n + r - 2$ times $t_1, t_2, \ldots, t_{n+r-2}$ that are either sampling times or coalescent times ordered from youngest tip (t_1) to the the coalescent time at the root (t_{n+r-2}) and let $\Delta t_i = t_i - t_{i-1}$ denote the length of an interval. Let k_i be the number of lineages at event i, so i decreases by one at a coalescent event, but increases at a sampling event. Let $\mathcal{I}_c(i)$ be an indicator function that indicates whether the ith event is a coalescent event ($\mathcal{I}_c(i) = 1$) or a sampling event ($\mathcal{I}_c(i) = 0$).

Consider m epochs defined by groups of coalescent events, and let $A = \{a_1, a_2, \ldots, a_m\}$ be the number of coalescent events in each of the m epochs that cover the whole tree (so $\sum_{i=1}^{m} a_i = n - 1$). Parag and Pybus (2019) show that having a similar number of coalescent events per epoch increases accuracy of population size estimates, so in practice we keep group sizes constant and evenly spread. The number of epochs is a parameter to be provided by the user, but by default 10 epochs will be used unless the epoch sizes become less than 6 ($\lfloor n/6 \rfloor$ groups will be used) or larger than 30 ($\lfloor n/30 \rfloor$ groups will be used).

Let $\Theta = \{\theta_1, \theta_2, \ldots, \theta_m\}$ be the effective population sizes for the m epochs, that define a piecewise constant population function for the m epochs. Let $h(i)$ be a function $1, \ldots, n+s-2 \rightarrow m$ that map the coalescent and sampling events i to epochs (Drummond et al. (2005) Eq(4)). Then the log likelihood $\log p(T|\Theta, A)$ of the tree T given Θ and A is (Drummond et al. (2005) Eq(3)):

$$\log p(T|\Theta, A) = \sum_{i=1}^{n+r-2} \mathcal{I}_c(i) \log \frac{k_i(k_i - 1)}{2\theta_{h(i)}} - \frac{k_i(k_i - 1)\Delta t_i}{2\theta_{h(i)}} \tag{1}$$

Taking the exponent, gives the density

$$p(T|\Theta, A) = \prod_{i=1}^{n+r-2} \exp\{\mathcal{I}_c(i) \log \frac{k_i(k_i - 1)}{2\theta_{h(i)}}\} \exp\{-\frac{k_i(k_i - 1)\Delta t_i}{2\theta_{h(i)}}\} \tag{2}$$

Let $p_j(T|\Theta, A)$ denote the contribution for a single epoch j so $p(T|\Theta, A) = \prod_{j=1}^{m} p_j(T|\Theta, A)$, and let b_j be the index of event i at the start of the jth epoch (so, $h(i) = j$ for $b_i \leq i < b_{i+1}$), then

$$p_j(T|\Theta, A) = \prod_{i=b_j}^{b_{j+1}-1} \exp\{\mathcal{I}_c(i) \log \frac{k_i(k_i - 1)}{2\theta_j}\} \exp\{-\frac{k_i(k_i - 1)\Delta t_i}{2\theta_j}\} \tag{3}$$

which can be simplified to

$$p_j(T|\Theta, A) = \left(\frac{1}{\theta_j}\right)^{Q_j} e^{\frac{-R_j}{\theta_j}} \tag{4}$$

[1] Though the notation of Drummond et al. (2005) is mostly followed here, we use r instead of s since later s will be used to denote scale factors for MCMC proposals.

with $Q_j = \prod_{i=b_j}^{b_{j+1}-1} \exp\{\mathcal{I}_c(i) \log k_i(k_i-1)/2\}$ and $R_j = \sum_{i=b_j}^{b_{j+1}-1} k_i(k_i-1)\Delta t_i/2$. Following Liu et al. (2008), we note that the inverse gamma distribution $f(x;\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{-\alpha-1} e^{-\beta/x}$ is conjugate for θ_j, in other words, the posterior is $f(\theta_j|\alpha+Q_j, \beta+R_j)$ and integrating out θ_j gives

$$\int_0^\infty p_j(T|\Theta, A) f(\theta_j|\alpha,\beta) d\theta_j = \int_0^\infty \left(\frac{1}{\theta_j}\right)^{Q_j} e^- R_j/\theta_j \frac{\beta^\alpha}{\Gamma(\alpha)} \theta_j^{-\alpha-1} e^{-\beta/\theta_j} d\theta_j = \frac{\beta^\alpha}{\Gamma(\alpha)} \frac{\Gamma(\alpha+Q_j)}{(\beta+R_j)^{(\alpha+Q_j)}} \tag{5}$$

thus we get a closed form density for the contribution of epoch j that has the population size θ_j integrated out. Since $\log p(T|\Theta, A) = \prod_{j=1}^m \log p_j(T|\Theta, A)$ and θ_j independent, we can do this for each of the intervals.

This leaves us to choose the parameters for the inverse gamma prior on population sizes. By default, the shape value of $\alpha = 3$ is fixed as suggested elsewhere (Ogilvie et al., 2017; Liu et al., 2008) and the population mean $\mu_1 = \frac{\beta}{\alpha-1}$ estimated during the MCMC run with a lognormal($\mu = 1, \sigma = 1$).

2.1.1 Smoothing priors

Epoch models can show abrupt changes in population size estimates when population sizes for the epochs are assumed to be independent. That is why smoothing priors are applied (Drummond et al., 2005; Minin et al., 2008; Gill et al., 2013) to suppress large fluctuations of population size in consecutive epochs. One way to do this is to sample only the population mean for the first epoch. For consecutive epochs, the posterior mean of the previous epoch $\mu_j = \frac{\beta+R_j}{\alpha+Q_j-1}$ can be used to set $\beta_{j+1} = \mu_j(\alpha-1)$.

2.1.2 Inferring skyline plots

While models that explicitly sample population sizes of each epoch store population sizes and epoch information during MCMC, we do not have population size information available when integrating them out. However, given that for each epoch j we have a posterior distribution $f(\theta_j|\alpha+Q_j, \beta_j+R_j)$ we can just sample a value from that posterior and approximate the population size distribution for each epoch, which allows us to perform demographic reconstruction. A sample from an inverse gamma distribution can be obtained by sampling a gamma distribution with shape $\alpha+Q_j$ and scale $1/(\beta_j+R_j)$ and taking the reciprocal value of the sample.

2.2 New MCMC proposals

To help convergence of the MCMC algorithm, we introduce a number of new proposals that move a large number of heights of internal node in the tree while keeping leaf node heights constant. These proposals have a large effect on the length of a tree, and thus indirectly on the tree prior. Note that the methods introduced are applicable to all phylogenetic tree priors and are not restricted to the skyline model discussed above.

2.2.1 New tree stretch proposal

The standard tree scale proposal in BEAST 2 simply multiplies all internal node heights h_i (for node i) with the same randomly chosen scale factor s, but leaf node heights remain unchanged. This can lead to negative branch lengths if an internal node is scaled down below a tip height of a descendant, at which point the scale proposal is instantly rejected (see node D in Fig. 1a when scaled down). When there are many dated tips over a large time range, and there is little variation in sequence data resulting in short branches to tips, scaling up can make relatively short branches to tip stretch out a lot causing a marked reduction in tree likelihood causing the proposal to be rejected (see node D in Fig. 1a when scaled up).

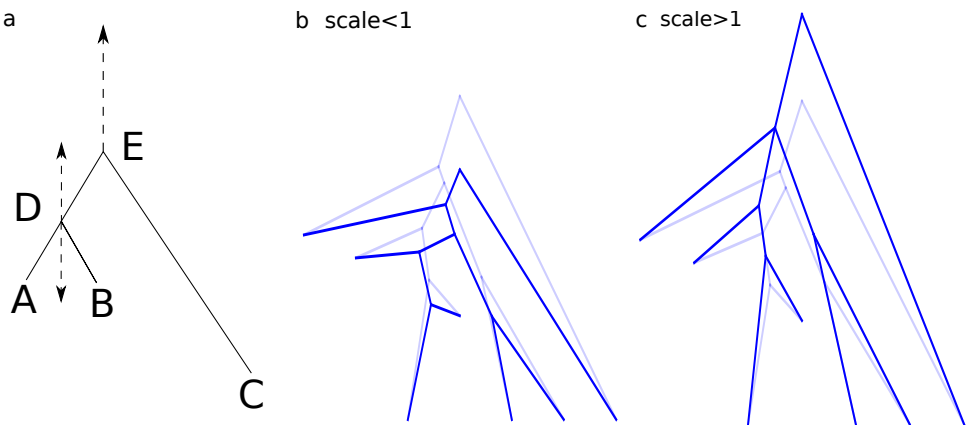


Figure 1: a) The traditional scale operator that gets often rejected when there are many tip dates ($r \gg 1$) because of high probability of negative branch lengths when scaling down or inappropriately stretching short branches into older tips when scaling up. Tree stretch proposal moves nodes near tips (e.g. node D) less far than nodes away from tips (e.g. node E) b) for scale factor less than 1 where lighter trees are the original state and darker trees are proposals and c) for scale factor larger than one.

To remedy such low acceptance, the range from which the scale factor is sampled can be reduced, but that leads to smaller overall changes to the tree. Note that when scaling all nodes in the tree the pruning algorithm (Felsenstein, 1981) for calculating the tree likelihood needs to recalculate all so called partials for internal nodes, which is a computationally expensive task (see (Felsenstein, 1981) for details). So, ideally we would like to make bold proposals to justify this computationally costly operation.

Instead of simply multiplying internal node heights, as the standard scale operator does, we can do a post-order traversal where we scale branch lengths and add them to the height of the left and right child, then take the average of these heights to set the height of the current node in the traversal. Note that down stretching can lead to increased branch lengths, and up stretching to reduced branch lengths, for example in the internal branch below the left branch below the root in Fig.1b and c respectively.

Formally, let s be a randomly chosen scale factor from a Bactrian kernel (Yang and Rodríguez, 2013; Thawornwattana et al., 2018), that is, we randomly sample a value from a standard Gaussian $N(0,1)$ scaled with $c\sqrt{1-m^2}$, and randomly add or subtract m. Here, m determines the shape of the Bactrian distribution and is set to 0.95 by default, and c is a tuning parameter. The tuning parameter is automatically optimised (Drummond and Bouckaert, 2015) during MCMC to obtain optimal balance between better acceptance (at lower values of c) and boldness (at larger values of c). A target acceptance probability of 0.4 suggested in Yang and Rodríguez (2013) appears to give good results. Let b_i be the branch length above node i, so $b_i = h_p - h_i$ when p is the parent of node i. We traverse the tree and do not change leaf node heights, but for a node i with children j and k (assuming they were already visited), we set the new height h'_i of node i to $(h'_j + sb_j + h'_k + sb_k)/2$. When all tips are contemporary, this proposal is the same as the traditional tree scale operator (because $h'_i = (h'_j + sb_j + h'_k + sb_k)/2 = (h'_j + s(h_i - h_j) + h'_k + s(h_i - h_k))/2$ under induction assumption $h'_j = sh_j$ and $h'_k = sh_k$, giving $(sh_j + s(h_i - h_j) + sh_k + s(h_i - h_k))/2 = (sh_i + sh_i)/s = sh_i = h'_i$). But, with dated tips, nodes closer to dated tips move less than nodes farther away from tips.

The Hastings ratio is $\prod_i \frac{h'_i}{h_i}$. Note that it is still possible for node heights to be proposed that result in negative branch lengths, but the probability is much reduced.

2.2.2 New epoch flex proposals

The epoch flex-operator randomly selects a lower bound L and upper bound U in the range between the root height of the tree and the youngest leaf (enforcing $L < U$ by swapping values if $L > U$), then scales the interval with a random scale value s drawn from a Bactrian distribution (Yang and Rodríguez, 2013; Thawornwattana et al., 2018) with respect to the lower bound. Internal nodes above the upper bound U are moved to accommodate the scaled height of the interval. Internal nodes below L and leaf nodes do not have their heights changed, which allows caching of the partial calculations for the tree likelihood for at least the nodes below L (Fig. 2), making it a more time efficient operator that the tree stretch operator.

More formally, for every node i with height h_i the proposed height h'_i is

$$h'_i = \begin{cases} h_i + (s-1)(H-L) & \text{if } U < h_i \\ L + s(h_i - L) & \text{if } L \leq h_i \leq U \\ h_i & \text{if } h_i < L \end{cases} \tag{6}$$

The Hastings ratio requires taking into account selecting L and H and since these are chosen uniform in the interval $[0, h_{root}]$ and we have a new root height h'_{root} after the proposal the contribution is $(h'_{root}/h_{root})^2$ for these two random values. Furthermore, let there be k nodes with heights in between L and U, then the contribution of scaling these k nodes is s^k, making the log Hastings ratio $2\log(\frac{h'_{root}}{h_{root}}) + k\log(s)$.

Like for the tree stretch operator, a tuning parameter c is used for sampling s to obtain an optimal acceptance probability of 0.4.

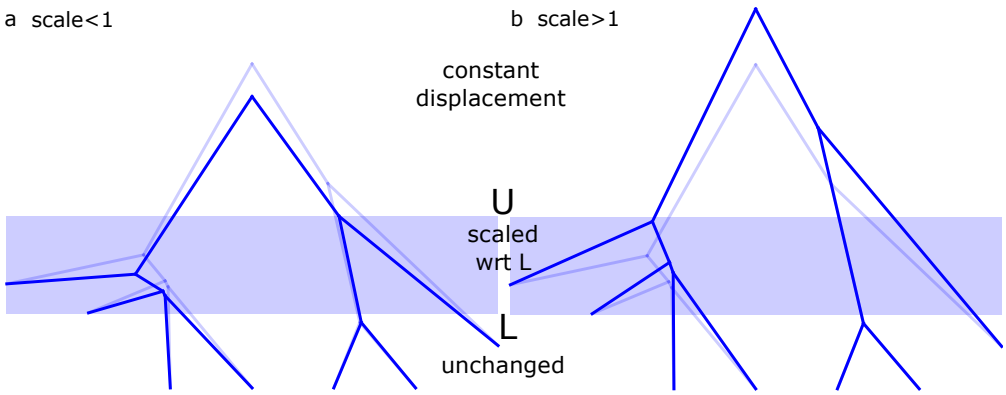


Figure 2: Epoch operator selects lower bound L , upper bound L and scale factor s and scale all nodes between L and U. Nodes above U are moved to make space for the newly scaled epoch. a) applied to light tree giving dark tree when scale factor less than one, and b) when scale factor larger than one.

The proposal can result in direct rejection if any of the scaled nodes are assigned heights below a tip. One way to prevent this from happening is to enforce the lower bound to be older than the oldest tip, so only part of the tree above the oldest tip is scaled. Since that part of the tree tends to be less constrained by tips, bolder proposals are possible, so having both the restricted and unrestricted version of the operator in the mix can lead to better proposals overall.

2.2.3 New up/down proposal

Mean clock rate, tree prior parameters and tree height tend to be highly correlated, so moving them at the same time (but in opposite direction) can help mixing. The so called up/down operator in BEAST randomly picks a scale factor s and scales up the tree with factor s while scaling down the clock by scaling with factor $1/s$. Tree prior parameters like birth rate or population size can be scaled in the appropriate direction at the same time.

The new tree stretch and epoch flex operators also change tree height, so we can use $s = h_{root'}/h_{root}$ as scale factor in a similar fashion as for the up/down operator and scale clock rates and tree prior parameters. For each scaled parameter, a contribution of s when scaling up (or $1/s$ when scaling down) must be added to the Hasting ratio.

3 Validation

We performed a well calibrated simulation study in order to make sure our implementation is correct, and performed an analysis of COVID-19 for community outbreaks in New Zealand.

parameter	epoch flexer	tree stretcher	up/down
tree height	91	95	94
tree length	94	91	92
kappa	96	99	99
gamma shape	99	98	96
population parameter	97	94	93
clock rate	96	99	98
tree prior	98	92	93
frequencies A	95	91	92
frequencies C	97	94	95
frequencies G	95	93	95
frequencies T	96	96	96

Table 1: Coverage of the true value by 95%HPD estimates from 100 independent runs of BICEPS for various parameters in the model and for different operators added to the standard set of operators. All coverage is in the expected 91 to 99 range, providing confidence there are not errors in the operator implementation.

3.1 The implementation is correct

To establish correct implementation of BICEPS, we performed a well calibrated simulation study sampling 50 tip dates randomly from the interval 0 to 1. To establish correctness of the new operators, we use a coalescent tree prior with constant population size (log-normal($\mu = 1, \sigma = 1.25$) distributed), a HKY model with kappa log-normal($\mu = 1, \sigma = 1.25$) distributed and gamma rate heterogeneity with four categories with shape parameter exponentially distributed with mean=1, and frequencies Dirichlet(1,1,1,1) distributed. Further, gamma is lower bounded by 0.1 to give reasonable range of rates (Bouckaert, 2020) and frequencies lower bounded by 0.2 to prevent atypical parameter values. We use a strict clock where the clock rate times tree height has a tight normal($\mu = 1, \sigma = 0.05$) prior. Sampling 100 instances from this distribution using MCMC in BEAST 2 (Bouckaert et al., 2019), we get a range of tree heights from 1.03 to 8.8 with mean 1.6 (note that due to the tips being sampled from 0 to 1, the tree height is lower bounded by 1) and a clock rate range of 0.1 to a fraction over 1 in our study. With these trees, we sample sequences of 1000 sites using the sequence generator in BEAST 2.

Table 1 shows the coverage of true parameter values (and some other statistics) used to simulate the sequence data by the 95% highest probability density (HPD) intervals estimated after running MCMC. With 100 experiments, the 95%HPD of the binomial distribution with p=0.95 ranges from 91 to 99 inclusive. All analyses were run for 20 million samples, which was sufficient to obtain effective sample sizes of at least 200 for each of the parameters shown in Table 1. All coverages observed are in the expected range, suggesting no problems with the implementation.

	Coverage			Average ESS		Minimum ESS	
		linked		linked		linked	
Parameter	unlinked	standard	new	standard	new	standard	new
tree height	97	93	94	1640	**1709**	116	**1219**
tree length	99	97	94	1540	**1669**	113	**1200**
population size	94	97	96	1709	**1736**	783	**1250**
coalescent prior	98	98	94	1547	**1690**	121	**1216**
pop size epoch 1	97	96	98	1721	**1731**	969	**1358**
pop size epoch 2	97	96	93	1704	**1728**	250	**1401**
pop size epoch 3	99	96	95	1711	**1731**	232	**1285**
pop size epoch 4	97	96	97	1695	**1711**	364	**1343**
pop size epoch 5	94	95	96	1693	**1716**	289	**1035**
kappa	97	96	93	1530	**1536**	1114	**1132**
gamma shape	92	99	93	1532	**1543**	644	**741**
frequencies A	91	95	97	**790**	778	213	**342**
frequencies C	94	96	94	750	**752**	**390**	349
frequencies G	91	93	98	**787**	760	**425**	369
frequencies T	92	89	95	**789**	761	**293**	228
clock rate	96	92	94	1622	**1690**	125	**1182**

Table 2: Results for 100 BICEPS analysis with 50 taxa, 250 sites and unlinked and linked population sizes with standard operators and with the new operators added in. Coverage as in Table 1 for unlinked BICEPS with standard operators, linked BICEPS with and without new operators. Effective sample size (ESS) shown compares the standard with new operators, where bold numbers indicate the better ESS. Coverage is in the expected 91 to 99 range for all cases but ESS increase for tree related parameters, in particular the minimum ESS of the 100 runs increases significantly. Coverage of the true value by 95%HPD estimates from 100 independent runs of BICEPS for various parameters in the model and for different operators added to the standard set of operators.

3.2 COVID-19 in New Zealand

We use the 887 full genome sequence data from Douglas et al. (2021a) containing samples from the 11 community outbreaks in New Zealand plus closely related sequences from the rest of the world. Further, we use a subsample of all taxa sampled up to 31 August 2020 consisting of 257 taxa for performance comparison. The data was analysed as follows. Genomic sites were partitioned into the three codon positions, plus non-coding, as described by (Douglas et al., 2020). For each partition we model evolution with an HKY substitution model with log-normal($\mu = 1, \sigma = 1.25$) prior on kappa, frequencies estimated with Dirichlet(1,1,1,1) prior, and relative substitution rates with Dirichlet(1,1,1,1). We use a strict clock model with log-normal($\mu = -7, \sigma = 1.25$) prior on mean clock rate as in Douglas et al. (2020, 2021a), and for tree prior we use a Bayesian skyline model (Drummond et al., 2005) with Markov chain distribution on population sizes and log-normal($\mu = 0, \sigma = 2$) on first population size and compare this with a BICEPS tree prior.

4 Results

4.1 Operator performance analysis

Fig. 3a-c show violin plots for effective sample sizes (ESS) obtained with the 100 runs for the posterior, prior and tree length where the first item was done with standard operators, the second with the epoch operator added and the the third with skewed scale operator added as well. There was some beneficial effect from these operators on the posterior, more so on the prior, as well as the tree length. Site model parameters were practically unaffected by adding these operators, but there was some beneficial effect on the ESS for the clock rate. Note these ESSs were obtained under similar run times, so the plots suggest the operators are moderately beneficial for vanilla simulated data, or at least no detrimental to mixing. However, for empirical data we observed more marked differences (see below).

The adaptable operator sampler (Douglas et al., 2021b) is an operator that selects among a set of operators by keeping track of relevant performance indicators of the various operators, namely amount of change in node heights, amount of time required to calculate the new state, and probability of acceptance. Together, these factors are used by the adaptable operator sampler to re-weigh sets of operators for optimal amount of node height change per unit of time.

Fig. 3d,e show the end weight distribution over the 100 runs of the well calibrated simulation study for the case where the tree scaler, up/down operator and tree stretcher were reweighted by an adaptable operator sampler, and Fig. 3e the case where a new up/down operator was added. In the first case shown in Fig. 3d, an overwhelming amount of weight is distributed towards the tree stretcher. In the second case shown in Fig. 3e, about standard operators hardly get any weight assigned, while most of the weight is distributed almost evenly between tree stretcher and new up/down operator, with a slight preference for the up/down operator.

This illustrates the new tree stretch and up/down variant perform well when balancing the size of change, the time to recalculate the posterior and how often the operator is accepted. Since there is no directly comparable version of the epoch flexer, we omitted it from the mix.

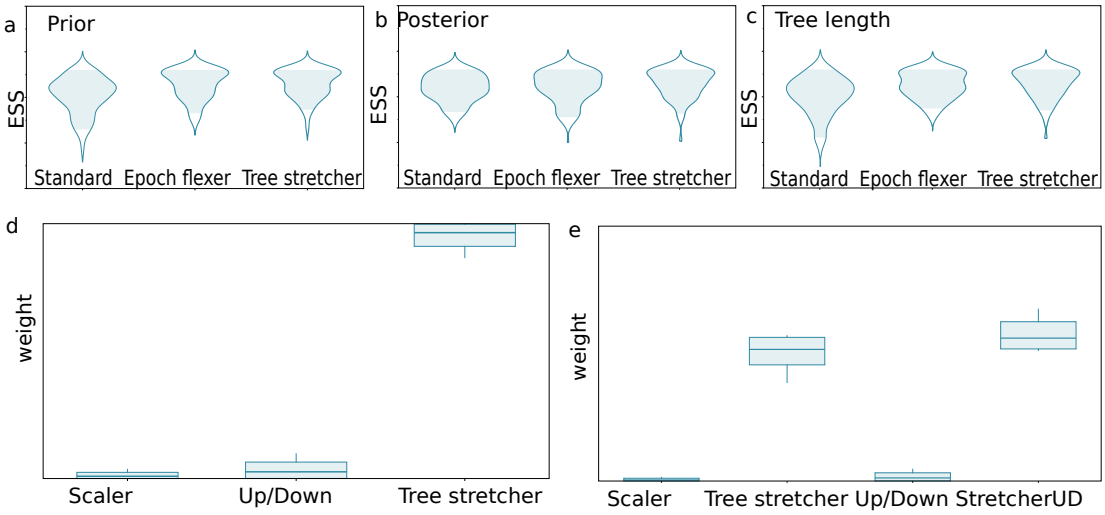


Figure 3: Performance of epoch flexer and tree stretch operators. Operator are weighted such that run times of various combinations are similar, so ESSs are comparable. ESSs improve a little, and weights favour the new operators over standard ones. a) ESSs on a scale of 0 to 2500 of the prior for different operator combinations – classic, with epoch operator and with epoch operator and scaler. b) ESSs for posterior, c) ESSs for tree length. d) Weights on a scale of 0 to 1 assigned to operators by the adaptable operator sampler for standard tree scaler, up/down operator and tree stretcher. e) Weights for standard tree scaler, tree stretcher, up/down operator and tree stretcher with up/down combination.

4.2 COVID-19 analysis

For the 10 runs of the 257 taxa COVID-19 analysis, MCMC convergences (all parameters having ESSs larger than 200) around 20 million samples for the BICEPS analysis while the BSP analysis still struggles to achieve mixing. In particular the tree length only achieves single digit or less than 20 ESSs when taking favourable burn-in values for the 10 runs in Tracer. Fig. 4 shows a typical trace of the tree length for one of the BSP and one of the BICEPS analyses, highlighting how BICEPS achieves convergence much faster.

The COVID-19 analysis from Douglas et al. (2021a) required 8 chains running 1 billion samples each and were combined to obtain satisfactory ESSs over 200. In contrast, for the same analysis with BICEPS prior and new operators a single run converged in 1

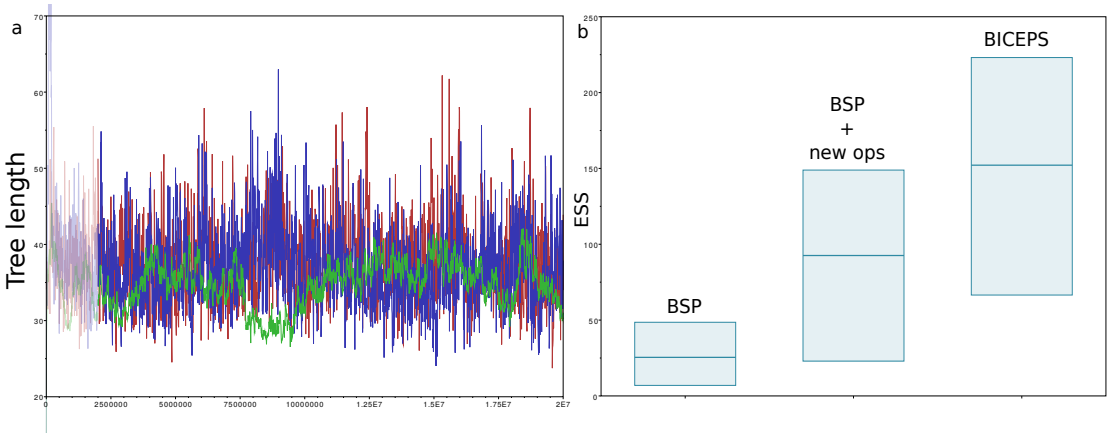


Figure 4: MCMC efficiency a) Trace of tree length for BSP (green line with very low period), BSP with new operators (blue line with higher period) and BICEPS analysis (red line forming a satisfying hairy caterpillar pattern). The BSP analysis typically does not reach an ESS of 10 when the BICEPS analysis already has ESSs around 200. b) Posterior ESS for 10 runs of BSP, BSP + new operators, and BICEPS. Both new operators and the BICEPS prior contribute to improving ESSs.

billion samples to ESSs over 200, a factor 8 speed up. Since these analyses use a different though related tree prior, we compare the tree posteriors (see Fig. 5a) and conclude these tree priors lead to very similar results in posterior tree distributions. Clade support is very similar (red dots in Fig. 5a) except for a handful of clades, which may be due to imperfect mixing of the trees, something not unexpected with this many taxa and sequences with relatively little variation (some sequences are even identical).

The estimate of most clade ages and in particular the root ages are consistent with each other. However, the BSP analysis puts the root age a fraction lower (at 1.24 year) than the BICEPS analysis (at 1.25). This can be explained when considering the demographic reconstruction, shown for BSP in Fig. 5b, and for BICEPS in Fig. 5c. Over all, the reconstructions are quite similar, but note that the population size estimates near the root (right hand side of plots) are lower and with higher uncertainty for the BSP reconstruction. The BICEPS reconstruction assumes a constant population size for each epoch and number of coalescent intervals are fixed to 29 or 30 (making 30 groups for 887 taxa). Therefore, the last 29 coalescent events to the root are assumed to be under a constant population. The BSP analysis on the other hand estimates group sizes and it is 22 on average with 95% credible range of 11 to 35, resulting in a smaller population size estimate, hence a slightly reduced root age estimate. When running the BICEPS with 10 epochs instead of 30, the effect is enlarged (giving a root age estimate of 1.29 year).

A general rule of thumb in statistics is that 30 observations are sufficient to estimate the mean of a parameter. Given that epochs can be linked through posterior mean

population size estimates in BICEPS using epochs that cover more than 30 observations does not seem necessary. By default, the model uses 10 groups unless group sizes are larger than 30, then the group count is set to thenumber of taxa divided by 30. However, if group sizes are less than 6 then group count is set to the number of taxa divided by 6.

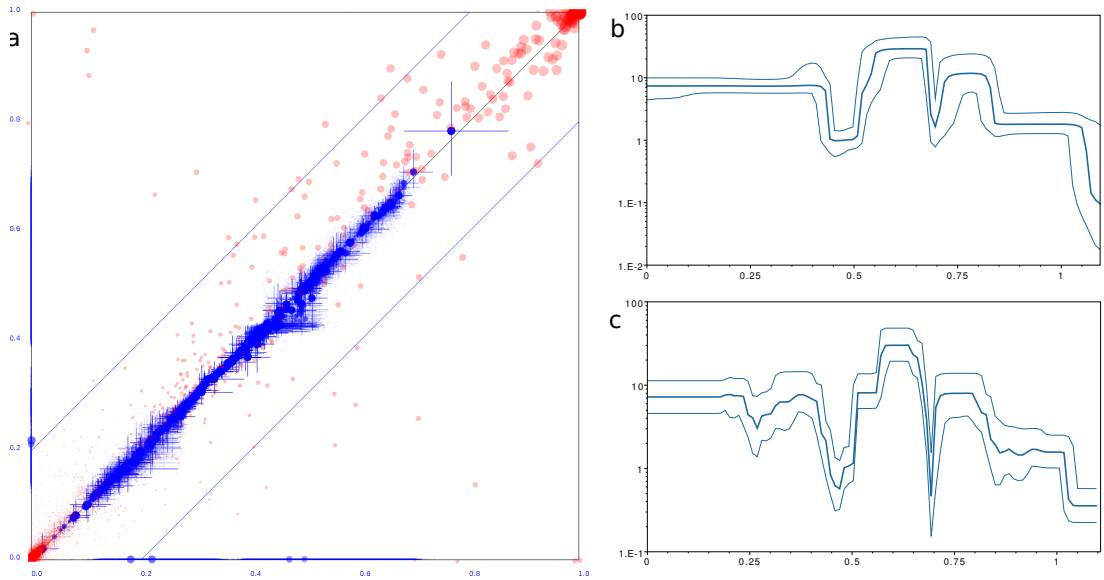


Figure 5: BSP and BICEPS compared. a) Difference in clade support and clade heights for the Bayesian skyline analysis from Douglas et al. (2021a) and the same analysis with a BICEPS prior. Red dots indicate clade support between 0 and 1 on both axis, blue dots indicate mean clade heights with cross hairs showing the 95% HPD intervals of height estimates. The axis are scaled between zero and the highest tree height found in either tree set. b) Population history for COVID-19 inferred with the BSP model and c) BICEPS model. Dark middle line indicates the median, lighter outer lines cover the 95%HPD intervals. The x-axis shows time in years going backward from left to right, the y-axis shows population size on a log scale. BSP and BICEPS analyses largely agree.

4.3 HCV analysis

To demonstrate BICEPS does not only perform well with serially sampled data, we analysed a dataset of 63 hepatitis-C virus sequences sampled in Egypt in 1993, which was earlier analysed in Drummond et al. (2005) and Stadler et al. (2013). We analysed with a GTR substitution model with gamma rate heterogeneity with 4 categories and fixed the clock rate at 7.9×10^{-4} substitutions per site per year.

Where BSP requires 30 million samples for MCMC to converge, BICEPS requires only 5 million samples, demonstrating that BICEPS can be considerably faster. A comparison similar to shown in Fig. 5 for COVID-19 can be found in Appendix A.

5 Conclusions

We introduced a two headed approach for improving the efficiency of Bayesian inference under epoch models: a flexible tree prior based coalescent epoch model that integrates out population size parameters and a set of new MCMC proposals directly targeting tree lengths. Both these elements contribute to more efficient inference, in particular with COVID-19 data and with serially sampled sequence data. The behaviour of BICEPS tree prior is very similar to that of the popular Bayesian skyline plot, and allows for reconstruction of demographic histories through time. The model is implemented in BEAST 2 Bouckaert et al. (2019) and can be used in combination with a large range of different data types, substitution and site models as well, a number of clock models and in combination with various types of data, including geographical locations, morphological characters, micro satellite, etc.

Coalescent models assume that the samples represent a small number of individuals from a much larger population. When this assumption does not hold, birth death models may be more appropriate. However, it is more challenging to extend the idea of integrating out parameters to birth death sampling models. For the Yule model (Yule, 1924; Aldous et al., 2001), a pure birth model this is straightforward (Appendix B). An epoch version of the Yule model assuming death and sample rates of zero and sampling all extant taxa at the same time (that is, rho-sampling with rho=1) can be found in Appendix C. This provides a flexible prior for the case where tips are not sampled through time, but are all taken at the same time. The model is implemented in BEAST 2 and a well calibrated simulation study (Appendix C) passed.

For more general cases this approach is hampered by the large number of parameters (birth, death, sampling rate, etc.) and because the tree likelihood is of a form that does not appear to lend itself for integrating out parameters.

Availability

The open source `BICEPS` package for BEAST 2 (Bouckaert et al., 2019) is available under GPL at `https://github.com/rbouckaert/biceps`. An analysis can be set up through BEAUti, the user friendly GUI for BEAST, both for the BICEPS and Yule Skyline models.

Acknowledgements

I thank Alexei Drummond and Jordan Douglas for stimulating discussions and Jordan Douglas and Cinthy Jimenez-Silva for proofreading the manuscript. The study was funded by Marsden grant 18-UOA-096 from the Royal Society of New Zealand and a contract from the Health Research Council of New Zealand (20/1018).

References

Aldous, D. J. et al. (2001). Stochastic models and descriptive statistics for phylogenetic trees, from yule to today. *Statistical Science*, 16(1):23–34.

Bouckaert, R., Vaughan, T. G., Barido-Sottani, J., Duchêne, S., Fourment, M., Gavryushkina, A., Heled, J., Jones, G., Kühnert, D., De Maio, N., et al. (2019). BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. *PLoS computational biology*, 15(4):e1006650.

Bouckaert, R. R. (2020). OBAMA: OBAMA for Bayesian amino-acid model averaging. *PeerJ*, 8:e9460.

Douglas, J., Geoghegan, J. L., Hadfield, J., Bouckaert, R., Storey, M., Ren, X., de Ligt, J., French, N., and Welch, D. (2021a). Real-time genomics to track COVID-19 post-elimination border incursions in Aotearoa New Zealand. *medRxiv*.

Douglas, J., Mendes, F. K., Bouckaert, R., Xie, D., Jimenez-Silva, C. L., Swanepoel, C., de Ligt, J., Ren, X., Storey, M., Hadfield, J., et al. (2020). Phylodynamics reveals the role of human travel and contact tracing in controlling COVID-19 in four island nations. *medRxiv*.

Douglas, J., Zhang, R., and Bouckaert, R. (2021b). Adaptive dating and fast proposals: Revisiting the phylogenetic relaxed clock model. *PLoS computational biology*, 17(2):e1008322.

Drummond, A. J. and Bouckaert, R. R. (2015). *Bayesian evolutionary analysis with BEAST*. Cambridge University Press.

Drummond, A. J., Rambaut, A., Shapiro, B., and Pybus, O. G. (2005). Bayesian coalescent inference of past population dynamics from molecular sequences. *Molecular biology and evolution*, 22(5):1185–1192.

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of molecular evolution*, 17(6):368–376.

Gill, M. S., Lemey, P., Faria, N. R., Rambaut, A., Shapiro, B., and Suchard, M. A. (2013). Improving bayesian population dynamics inference: a coalescent-based model for multiple loci. *Molecular biology and evolution*, 30(3):713–724.

Hill, V. and Baele, G. (2019). Bayesian estimation of past population dynamics in BEAST 1.10 using the skygrid coalescent model. *Molecular biology and evolution*, 36(11):2620–2628.

Jones, G. (2017). Algorithmic improvements to species delimitation and phylogeny estimation under the multispecies coalescent. *Journal of mathematical biology*, 74(1-2):447–467.

Kingman, J. F. C. (1982). The coalescent. *Stochastic processes and their applications*, 13(3):235–248.

Kuhner, M. K., Yamato, J., and Felsenstein, J. (1998). Maximum likelihood estimation of population growth rates based on the coalescent. *Genetics*, 149(1):429–434.

Liu, L., Pearl, D. K., Brumfield, R. T., and Edwards, S. V. (2008). Estimating species trees using multiple-allele DNA sequence data. *Evolution: International Journal of Organic Evolution*, 62(8):2080–2091.

Minin, V. N., Bloomquist, E. W., and Suchard, M. A. (2008). Smooth skyride through a rough skyline: Bayesian coalescent-based inference of population dynamics. *Molecular biology and evolution*, 25(7):1459–1471.

Ogilvie, H. A., Bouckaert, R. R., and Drummond, A. J. (2017). StarBEAST2 brings faster species tree inference and accurate estimates of substitution rates. *Molecular biology and evolution*, 34(8):2101–2114.

Parag, K. V., du Plessis, L., and Pybus, O. G. (2020). Jointly inferring the dynamics of population size and sampling intensity from molecular sequences. *Molecular biology and evolution*, 37(8):2414–2429.

Parag, K. V. and Pybus, O. G. (2019). Robust design for coalescent model inference. *Systematic biology*, 68(5):730–743.

Pybus, O. G., Rambaut, A., and Harvey, P. H. (2000). An integrated framework for the inference of viral population history from reconstructed genealogies. *Genetics*, 155(3):1429–1437.

Stadler, T., Kühnert, D., Bonhoeffer, S., and Drummond, A. J. (2013). Birth–death skyline plot reveals temporal changes of epidemic spread in HIV and hepatitis C virus (HCV). *Proceedings of the National Academy of Sciences*, 110(1):228–233.

Strimmer, K. and Pybus, O. G. (2001). Exploring the demographic history of DNA sequences using the generalized skyline plot. *Molecular Biology and Evolution*, 18(12):2298–2305.

Thawornwattana, Y., Dalquen, D., Yang, Z., et al. (2018). Designing simple and efficient Markov chain Monte Carlo proposal kernels. *Bayesian Analysis*, 13(4):1037–1063.

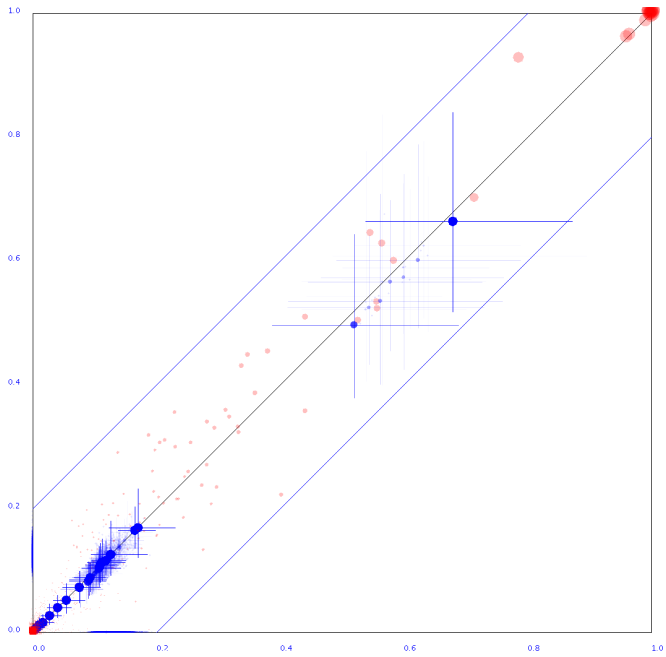
Yang, Z. and Rodríguez, C. E. (2013). Searching for efficient Markov chain Monte Carlo proposal kernels. *Proceedings of the National Academy of Sciences*, 110(48):19307–19312.

Yule, G. U. (1924). A mathematical theory of evolution, based on the conclusions of dr. JC Willis. *Philosophical transactions of the Royal Society of London. Series B, containing papers of a biological character*, 213(402-410):21–87.

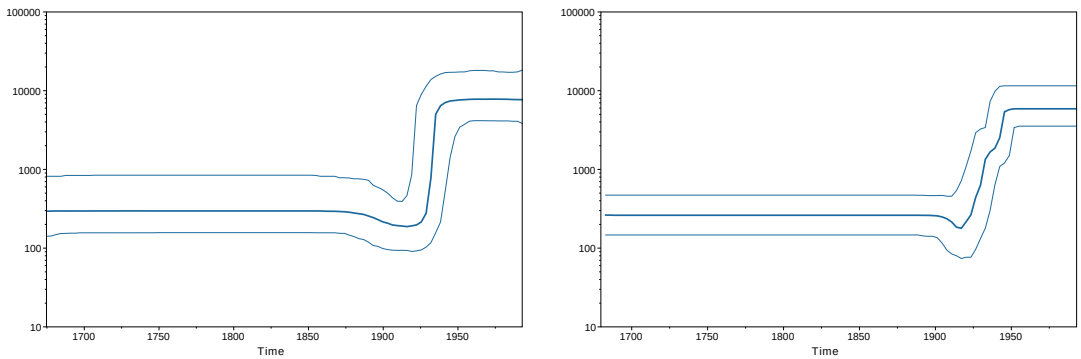
Appendices

A BSP vs BICEPS comparison for HCV

Clade set comparison as in Fig.5. BSP on y-axis, BICEPS on x-axis.



BSP and BICEPS demographic plots



B Yule prior

Consider a binary tree T with n taxa, leaf nodes numbered $0, \ldots, n-1$, internal node numbered $n, \ldots, 2n-2$ and root numbered $2n-1$. Let h_i denote the height of the ith node.

The Yule prior is defined as $p_{Yule}(T|\lambda) = \lambda^{n-1}e^{-\lambda L}$ where $L = h_{2n-1} + \sum_{i=n}^{2n-1} h_i$ is the sum of lengths of all branches in T, or the length of the tree and λ the birth rate.

Uniform prior

Assuming uniform prior on λ on $[0, u)$, integrating out λ means solving $\int_0^u p_{Yule}(T|\lambda)d\lambda = \int_0^u \lambda^{n-1}e^{-\lambda L}d\lambda$, which equals $u^{n-1}(Lu)^{(-(n-1)}(\Gamma(n) - \Gamma(n, Lu))/L$ where $\Gamma[n, Lu]$ the incomplete gamma function. This simplifies to

$$\frac{\Gamma(n) - \Gamma(n, Lu)}{L^n} \tag{7}$$

Gamma prior

Assuming gamma prior on λ with shape parameter α and rate parameter β, $\Gamma(x; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)}x^{\alpha-1}e^{-\beta x}$. Integrating out λ means solving

$$\int_{\lambda=0}^{\infty} \lambda^n e^{-L\lambda}\Gamma(\lambda|\alpha, \beta)d\lambda = \int_{\lambda=0}^{\infty} \lambda^n e^{-L\lambda}\frac{\beta^\alpha}{\Gamma(\alpha)}x^{\alpha-1}e^{-\beta x}d\lambda$$

Putting constants outside the integral, and combining terms gives

$$\frac{\beta^\alpha}{\Gamma(\alpha)}\int_{\lambda=0}^{\infty} \lambda^{n+\alpha-1}e^{-(L-\beta)\lambda}d\lambda$$

Now note that the part inside the integral is of the form $\Gamma(x; \alpha + n, \beta + L)$ but divided by the constant $\frac{(\beta+L)^{\alpha+n}}{\Gamma(\alpha+n)}$ so we can write

$$\frac{\beta^\alpha}{\Gamma(\alpha)}\frac{\Gamma(\alpha + n)}{(\beta + L)^{\alpha+n}}\int_{\lambda=0}^{\infty} \Gamma(\lambda; \alpha + n, \beta + L)d\lambda$$

and since the gamma density integrates to 1, this leaves

$$\frac{\beta^\alpha}{\Gamma(\alpha)}\frac{\Gamma(\alpha + n)}{(\beta + L)^{\alpha+n}} \tag{8}$$

In summary, we can integrate out a uniform (Eq(7)) or gamma prior (Eq(8)) on the birth rate for a Yule prior. Note that the Yule prior is conditioned on all leafs being sampled at the same time, so this should not be used when sampling tips through time.

C Yule skyline prior

Consider m epochs with assignment A defined as for the coalescent prior before and let $\Lambda = \{\lambda_1, \ldots, \lambda_m\}$ be a vector of m birth rates. We are interested in the density of a tree T given A and Λ $f(T|\Lambda, A)$.

Starting with the general birth-death skyline model defined in Theorem 1 of Stadler et al. (2013) and assuming no sampled tips through time, setting death, sampling rates

to zero for all epochs ($\forall_i \mu_i = \psi_i = 0$) and only allow sampling of all lineages ($\rho = 1$) at the same time we obtain for all $0 < i < m$, $A_i = \lambda_i$, $B_i = 1$ and $q_i = 1/exp(-\lambda(t - t_i))$. Filling this into the tree likelihood in Theorem 1 of Stadler et al. (2013) gives us

$$f(T|\Lambda, A) = \prod_{i=1}^{m} \lambda_i^{a_i} e^{-\lambda_i L_i}$$

where L_i the sum of the lengths of branches inside epoch i (remember a_i is the number of coalescent events for epoch i). Note that each term in the product ($\lambda_i^{a_i} e^{-\lambda_i L_i}$) is of the same form as for the Yule prior, so we can integrate out λ_i under a uniform prior bounded by 0 and finite upper bound u. We can also assume a gamma distribution $\Gamma(x; \alpha, \beta)$ for each of the epochs and obtain posterior

$$f(T|\Lambda, A) = \prod_{i=1}^{m} \frac{\beta^\alpha}{\Gamma(\alpha)} \frac{\Gamma(\alpha + a_i)}{(\beta + L_i)^{\alpha + a_i}}$$

As for BICEPS, epochs can be linked obtaining a smoothing prior; we fix the shape parameter α, but instead of having a fixed scale β for each epoch, the rate β_{i-1} for epoch $i - 1$ can be chosen such that the posterior mean of the previous epoch (which is $\frac{\alpha + a_i}{\beta_i + L_i}$ for epoch i) is used apriori. So, $\frac{\alpha}{\beta_{i-1}} = \frac{\alpha + a_i}{\beta_i + L_i}$ giving $\beta_{i-1} = \alpha \frac{\beta_i + L_i}{\alpha + a_i}$

Note 1: The birth process moves forward in time, unlike the coalescent process, which moves backward in time. Therefore, when putting the smoothing prior on birth rates, we put a $\Gamma(\alpha, \beta)$ prior on λ_m, and proceed in reverse order from epoch m to epoch 1 from the way the coalescent smoothing prior is calculated.

Note 2: Like for the Yule prior, this assumes all tips are sampled at the same time. The Yule skyline is not appropriate for data that is sampled through time.

Validation

A well calibrated simulation study was performed with 50 contemporary taxa, and a root age prior normally distributed with mean 100 and standard deviation 0.5. The site and clock priors are as for the study for the BICEPS prior. The gamma prior used for the birth rate is with shape of 2 and rate estimated with log-normal($\mu = 1, \sigma = 1.25$), and the 8 epochs are unlinked in one set and linked in the other. Alignments of 250 sites were randomly generated. Coverage for the various parameters is listed in the table below, and it passes the test.

parameter	coverage unlinked	coverage linked
prior	94	96
Tree.height	93	95
Tree.treeLength	94	95
YuleSkylineBirthRateRate	93	97
YuleSkyline	95	97
BirthRates.1	97	93
BirthRates.2	97	96
BirthRates.3	94	92
BirthRates.4	96	95
BirthRates.5	93	97
BirthRates.6	97	100
BirthRates.7	95	95
BirthRates.8	99	98
gammaShape	96	92
kappa	96	91
freqParameter.1	92	95
freqParameter.2	98	96
freqParameter.3	97	94
freqParameter.4	93	95
clockRate	93	99