Supplementary Material for "How trustworthy is your tree? Bayesian phylogenetic effective sample size through the lens of Monte Carlo error"

Visualizing convergence of a single chain

To explore the uncertainty of estimates from a single MCMC chain through time, we employ a block-bootstrap approach in which we resample from the MCMC sample [Politis, 2003, Suchard et al., 2003]. This approach requires a vector of subsample sizes, $n_1, \ldots, n_s$. For a given subsample length n_i, we define the batch size to be $b = \lfloor \sqrt{n_i} \rfloor$ and the number of batches $a = \lfloor n_i/b \rfloor$. The summary tree is computed for the first ab samples of the real chain, and then for some number of bootstrap replicates r we re-estimate the summary tree. We use block-bootstrapping to preserve autocorrelation in the samples. Thus, for each of the r bootstrap replicates (at a given n_i), we draw a starting indices uniformly on $1, \ldots, n - b + 1$, and concatenate the resulting a blocks of length b into a bootstrap replicate chain. Then we compute the median RF distance from the real-chain-subsample summary tree to the r bootstrap replicates, as well as the 5th and 95th percentiles. As the longer subsamples of the real chain include the shorter subsamples (all real-chain subsamples start at the first sample), this procedure allows us to track how the summary tree converges over the course of the MCMC run. We can similarly track split or topology probabilities over the course of the run, in which case we use the average standard deviation of split frequencies (ASDSF) [Lemey et al., 2009] and the Euclidean distance, respectively, to compare the real chain estimates to the bootstrap estimates.

In Figure S1, we explore the convergence behavior of chain 1 of the Paroedura dataset using three summary measures. These measures are the ASDSF between bootstrap and real-chain split probabilities, the Euclidean distance between bootstrap and real-chain estimates of the vector of split probabilities, and the RF distance between bootstrap and real-chain summary trees. While both the split probabilities and tree probabilities appear to converge relatively well (the ASDSF quickly declines below the usual field-standard for good convergence of 0.01), there is still considerable Monte Carlo variability evident in summary trees. This pattern holds across all datasets and almost all chains, indicating that classical ASDSF cutoffs for convergence of chains are not guarantees of the convergence of summary trees from those chains. We note, however, that this can only ever help determine whether estimates from a single run have stabilized. To diagnose issues such as convergence to a local mode, practitioners must run multiple chains. We note that this is suggested as standard practice [Lemey et al., 2009] and is a widely available option, including in BEAST [Suchard et al., 2018] and RevBayes [Höhna et al., 2016], and is the default in MrBayes [Ronquist et al., 2012].

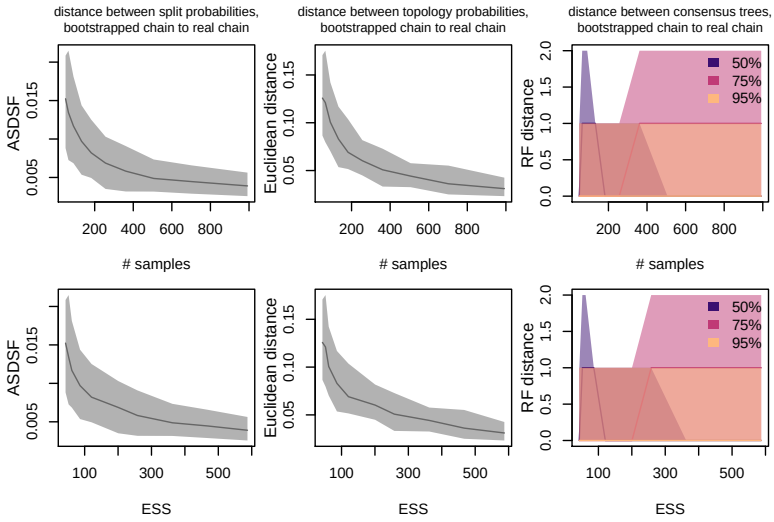

Figure S1: Monte Carlo error visualized over the length of one chain of the Paroedura dataset from Scantlebury [2013]. The top and bottom rows are equivalent except that the x-axis is scaled to the absolute number of MCMC samples (top), and the split-frequency ESS (bottom). The left column plots the ASDSF between bootstrap replicate estimates of the split probabilities and the split probabilities estimated from the first n_i samples of the chain. The central column plots the Euclidean distance between bootstrap replicate estimates of the (vector of) tree probabilities and the (vector of) tree probabilities estimated from the first n_i samples of the chain. The right column plots the RF distance between bootstrap replicate estimates of consensus trees and the consensus trees estimated from the first n_i samples of the chain. The different colors show consensus trees constructed with different minimum inclusion probabilities of splits, such that the purple curve shows the classical MRC tree, and the yellow curve shows a consensus tree containing only splits with 95% probability. In all cases, the dark lines are the median and the shaded region is the central 90% range.

More efficient simulated phylogenetic MCMC

Recall that our simulated phylogenetic MCMC is based on real-data phylogenetic posterior distributions, potentially truncated. This consists of a vector of trees, $\boldsymbol{\tau}$, and an associated probability mass function, $\widehat{\Pr}(\boldsymbol{\tau})$ (we use the hat as a reminder that this target is based, indirectly, on real data). We use NNIs to move between tree topologies, by uniformly drawing a tree from the set of neighbors, $N(\Psi)$. Then we accept or reject according to the estimated topology probability $\widehat{\Pr}(\Psi)$ (any tree not in the real-data posterior has probability 0). If we redefine $N(\Psi)$ to instead be the NNI neighbors of Ψ with positive probability (*e.g.* $\Psi \in \boldsymbol{\tau}$, which also requires far less storage), we can instead simulate the proposal in two steps. First, draw $u \sim \text{Uniform}(0, 1)$, and if $u < |N(\Psi)|/|N|$ (where $|N|$ is the number of NNI neighbors of any tree in the posterior), we draw our proposed tree Ψ^* uniformly at random from $N(\Psi)$ and set $A = \min(1, \Pr(\Psi^*)/\Pr(\Psi))$. Otherwise, we have drawn a tree outside the set of supported neighbors of Ψ ($\Psi^* \notin \boldsymbol{\tau}$) and we do not need to specify which tree, as in this case it has probability 0 and so $A = 0$ and we will always reject the proposal. Then we accept or reject the move with probability A and proceed normally. This approach requires us only to know what

trees in the support of the posterior are neighbors, which for real phylogenetic posterior distributions is a much smaller set than the set of all NNI neighbors.

Explicit definitions and derivations of tree ESS measures

In the following sections, we present more thorough derivations of the frechetCorrelationESS and approximateESS use in the main text, and derivations for 6 other potential tree ESS measures. We have not presented these additional ESS measures in the main text as their performance is at best no better than the performance of the methods presented above, and in some cases is markedly worse (Figures S5-S7). The ten total methods fall into the same three categories as the main text and are (using $*$ to denote those appearing in the main text):

- ESS measures based on Fréchet generalizations of Equation 5 to trees
  - $*$The Fréchet Correlation ESS (frechetCorrelationESS)
  - The split frequency ESS (splitFrequencyESS)
- ESS measures based on projecting the tree to a single dimension and computing the ESS of that using standard univariate approaches
  - The folded rank-medoid ESS (foldedRankmedoidESS)
  - $*$The median pseudo-ESS (medianPseudoESS)
  - $*$The minimum pseudo-ESS (minPseudoESS)
  - The total distance ESS (totalDistanceESS)
  - The classical multidimensional scaling ESS (CMDSESS)
- Ad-hoc ESS measures
  - $*$The approximate ESS (splitFrequencyESS)
  - The unsmoothed bootstrap jump-distance ESS (jumpDistanceBootstrapUnsmoothedESS)
  - The (smoothed) bootstrap jump-distance ESS (jumpDistanceBootstrapESS)

Calculating the ESS by generalizing previous definitions

In this section, we provide more in-depth derivations of our two ESS approaches that generalize Equation 5 using concepts borrowed from the notions of Fréchet mean and Fréchet variance. For a continuous random variable X, the sample mean minimizes the sum of squared deviations to all sampled points. The Fréchet mean generalizes this concept to other metric spaces and higher dimensions by keeping the idea of minimizing the sum of squared distances. The Fréchet mean of a set of samples is,

$$\bar{x} = \text{argmin}_y \sum_{i=1}^{n} d(x_i, y)^2,$$

where $d(\cdot, \cdot)$ is a distance metric. Note that for the rest of this subsection on Fréchet generalizations of univariate ESS approaches, we will use $\bar{x}$ to refer to the Fréchet mean. The Fréchet mean may not be unique, in which case the collection of values that minimize the sum of squared distances are known as Karcher means. Where the variance is the average squared deviation from the mean, the Fréchet variance is the average squared distance from the Fréchet mean. In the case where X is continuous and one-dimensional and $d(\cdot, \cdot)$ is the Euclidean distance, the Fréchet mean is the mean and the Fréchet variance is the variance.

These definitions take some adaptation to the setting considered here. When using RF distances between trees, one can think of an "RF space" where topologies are encoded as a binary vector. For a tree with n_{taxa} tips, there are $2^{n_{\text{taxa}}} - n_{\text{taxa}}$ possible non-trivial splits. Thus we can represent a tree as a vector of length $2^{n_{\text{taxa}}} - n_{\text{taxa}}$ which has a one entry exactly when the corresponding split is present in the tree. There are $n_{\text{taxa}} - 3$ non-trivial splits in a fully resolved tree, thus the sum of entries in such a vector representation is $n_{\text{taxa}} - 3$. The Hamming distance (or equivalently the Manhattan distance) between two trees represented as coordinate vectors in RF space is the classical RF distance. This also means that we only need to consider coordinates in RF space which are non-zero in at least one tree in the set. As we use RF distances in this paper, all this work can be seen to live in RF space.

The frechetCorrelationESS

In this section, we will explore how to generalize the sum-of-correlations ESS of Equation 8 to trees. To do so, we first review several key identities, including relationships between pairwise distances and both covariance and variance. For two real-valued variables, X and Y, we can express the expected squared Euclidean distance as a function of the variances, the difference in means, and the covariance. For convenience, we will write $\Delta^2 = (X-Y)^2$. Then, taking advantage of the fact that $\text{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$, we get,

$$\begin{aligned}
\mathbb{E}[\Delta^2] &= \mathbb{E}[(X-Y)^2] \\
&= \mathbb{E}[X^2] - 2\mathbb{E}[XY] + E[Y^2] \\
&= \text{Var}(X) + E[X]^2 + \text{Var}(Y) + E[Y]^2 - 2(\text{Cov}(X, Y) + E[X]E[Y]) \\
&= \text{Var}(X) + \text{Var}(Y) - 2\text{Cov}(X, Y) + (\mathbb{E}[X] - \mathbb{E}[Y])^2 \\
&\geq \text{Var}(X) + \text{Var}(Y) - 2\text{Cov}(X, Y)
\end{aligned}$$

Where the last line follows because $(\mathbb{E}[X] - \mathbb{E}[Y])^2 > 0$. The last two lines of this equation block rearrange to:

$$\text{Cov}(X, Y) = \frac{1}{2}(\text{Var}(X) + \text{Var}(Y) - \mathbb{E}[\Delta^2] + (\mathbb{E}[X] - \mathbb{E}[Y])^2). \tag{10}$$

If $\mathbb{E}[X] \approx \mathbb{E}[Y]$, then we have the approximate equality,

$$\text{Cov}(X, Y) \approx \frac{1}{2}(\text{Var}(X) + \text{Var}(Y) - \mathbb{E}[\Delta^2]) \tag{11}$$

It is worth noting that the sum of pairwise distances for a sample of a random variable can be used to estimate its variance.

$$\widehat{\mathrm{Var}}(X) = \frac{1}{n-1}\sum_i (x_i - \bar{x})^2 = \frac{1}{n(n-1)}\sum_{j>i}(x_i - x_j)^2. \tag{12}$$

To show this, first we need that,

$$\sum_i \sum_j (x_i - x_j)^2 = 2n \sum_i (x_i - \bar{x})^2. \tag{13}$$

This can be shown as follows,

$$\begin{aligned}
\sum_i \sum_j (x_i - x_j)^2 &= \sum_i \sum_j ((x_i - \bar{x}) - (x_j - \bar{x}))^2 \\
&= \sum_i \sum_j (x_i - \bar{x})^2 - 2(x_i - \bar{x})(x_j - \bar{x}) + (x_j - \bar{x})^2 \\
&= n\sum_i (x_i - \bar{x})^2 + n\sum_j (x_j - \bar{x})^2 - 2\sum_i \sum_j (x_i - \bar{x})(x_j - \bar{x}) \\
&= 2n\sum_i (x_i - \bar{x})^2 - 2\sum_i \sum_j (x_i - \bar{x})(x_j - \bar{x}) \\
&= 2n\sum_i (x_i - \bar{x})^2 - 2\sum_i \sum_j (x_i x_j - x_i\bar{x} - x_j\bar{x} + \bar{x}^2) \\
&= 2n\sum_i (x_i - \bar{x})^2 - 2(n^2\bar{x}^2 - n^2\bar{x}^2 - n^2\bar{x}^2 + n^2\bar{x}^2) \\
&= 2n\sum_i (x_i - \bar{x})^2.
\end{aligned}$$

Having shown that Equation 13 is true, from it we can get,

$$\frac{1}{2n(n-1)}\sum_i \sum_j (x_i - x_j)^2 = \frac{1}{n-1}\sum_i (x_i - \bar{x})^2. \tag{14}$$

And Equation 12 results by noting that,

$$\sum_i \sum_j (x_i - x_j)^2 = 2\sum_{j>i}(x_i - x_j)^2.$$

(We note this relationship can also be derived analogously to Equation 10 by letting X and Y be IID.) Letting $d(\cdot,\cdot)$ be a distance measure, we can write a Fréchet generalization of Equation 12 as,

$$\widehat{\mathrm{Var}}(X) = \frac{1}{n(n-1)}\sum_{j>i} d(x_i, x_j)^2 \tag{15}$$

In the same way that mean and variance can be generalized using the Fréchet mean and variance, Equation 10 allows us to generalize covariance to a Fréchet covariance.

This is accomplished by defining $\mathrm{Var}(X)$ and $\mathrm{Var}(Y)$ to be the Fréchet variances, $\mathbb{E}[X]$ and $\mathbb{E}[Y]$ to be the Fréchet means, and redefining $\Delta^2 = d(X,Y)^2$. Note that $\mathbb{E}[\Delta^2]$ is simply the average distance between X and Y,

$$\mathbb{E}[\Delta^2] = \frac{1}{n}\sum_{i=1}^{n} d(x_i, y_i)^2. \tag{16}$$

Thus, we can get a single-dimensional summary of the dependency of two random variables, and compute a single ESS measure for a high-dimensional object. Equation 11 is particularly useful in this circumstance because it avoids the need to compute the topological mean of a set of trees, $\mathbb{E}[X]$ and $\mathbb{E}[Y]$, which may not be unique. Equations 15, and 16 are also useful, and they allow us to compute everything we need for Equation 11 from the sample distance matrix.

To compute the ESS for trees using Equations 8 and 11, we specifically need to be able to compute the Fréchet autocorrelation ρ_s of the chain at time lag t, and thus X and Y are actually X_t and X_{t+s}. If the chain is stationary, then the mean does not change over time, and we should expect that $\mathbb{E}[X_t] \approx \mathbb{E}[X_{t+s}]$, and the use of Equation 11 rather than Equation 10 is justified. If instead of trees we had a time series of a Euclidean variable $\boldsymbol{X}$, the estimated autocorrelation is the sample Pearson correlation coefficient between samples at the given time lag,

$$\hat{\rho}_s = \frac{\widehat{\mathrm{Cov}}(\boldsymbol{X}_t, \boldsymbol{X}_{t+s})}{\sqrt{\widehat{\mathrm{Var}}(\boldsymbol{X}_t)\widehat{\mathrm{Var}}(\boldsymbol{X}_{t+s})}}. \tag{17}$$

For trees, instead we use Fréchet variances and Equation 11 to get an approximation to the Fréchet covariance, and plug this into Equation 17,

$$\hat{\rho}_s = \frac{\frac{1}{2}(\widehat{\mathrm{Var}}(\boldsymbol{\tau}_t) + \widehat{\mathrm{Var}}(\boldsymbol{\tau}_{t+s}) - \hat{\mathbb{E}}[\Delta^2])}{\sqrt{\widehat{\mathrm{Var}}(\boldsymbol{\tau}_t)\widehat{\mathrm{Var}}(\boldsymbol{\tau}_{t+s})}} \tag{18}$$

Once we obtain our estimates $\hat{\rho}_s$, we use Equation 8 to estimate the ESS. We call this approach the Fréchet correlation ESS, or frechetCorrelationESS. In this set up, we must estimate $\mathrm{Var}(\boldsymbol{\tau}_t)$, $\mathrm{Var}(\boldsymbol{\tau}_{t+s})$, and $\mathbb{E}[\Delta^2]$, which we compute using Equations 15 and 16. Let n be the total number of tree samples, $\boldsymbol{\tau}$ be the vector of tree samples, and let us define $\boldsymbol{\tau}_t, \boldsymbol{\tau}_{t+s}$ to be the pair of vectors of trees separated by time lag s. Then, we can compute the terms as,

$$\widehat{\mathrm{Var}}(\boldsymbol{\tau}_t) = \frac{1}{(n-s)(n-s-1)}\sum_{i=1}^{n-s-1}\sum_{j=i+1}^{n-s} d(\tau_i, \tau_j)^2,$$

$$\widehat{\mathrm{Var}}(\boldsymbol{\tau}_{t+s}) = \frac{1}{(n-s)(n-s-1)}\sum_{i=s+1}^{n-1}\sum_{j=i+1}^{n} d(\tau_i, \tau_j)^2,$$

$$\hat{\mathbb{E}}[\Delta^2] = \frac{1}{n-s}\sum_{i=1}^{n-s} d(\tau_i, \tau_{i+t})^2.$$

Given that most distances will appear in several calculations, it is most efficient to pre-compute the sample distance matrix $\boldsymbol{D}$ with $D_{ij} = d(\tau_i, \tau_j)$.

In practice, while the above definition could theoretically permit ESS $> n$, we enforce frechetCorrelationESS $\leq n$.

The splitFrequencyESS

Our next generalization approach we term the split frequency ESS, or splitFrequencyESS. This is a generalization of the univariate Vats and Knudson [2021] estimator of the effective sample size, which we will call the batch means ESS, which we will now review. The batch means ESS is based on the relationship,

$$\widehat{\text{ESS}} = n\frac{\hat{\sigma}^2_\pi}{\hat{\lambda}^2_L}, \tag{19}$$

where $\hat{\lambda}^2_L$ is an estimate of the limiting variance (σ^2_{lim}) and $\hat{\sigma}^2_\pi$ is the estimate of the posterior variance computed from the samples. Let $\boldsymbol{X}$ be the vector of MCMC samples, B be a batch size (define a to be the according number of batches), and $\boldsymbol{Y}$ the vector of batch means, with Y_i the mean in the ith batch (subset of the chain). Then, Vats and Knudson [2021] define,

$$\hat{\lambda}^2_B = \frac{B}{a-1}\sum_{i=1}^{a}(Y_i - \bar{X})^2.$$

To use the batch means approach in practice, the batch size must scale with n. Following Vats and Knudson [2021], we use a batch size $b = \lfloor n^{1/2} \rfloor$. Then, the estimate of the limiting variance from the batch-means approach is given by,

$$\hat{\lambda}^2_L = 2\hat{\lambda}^2_b - \hat{\lambda}^2_{b/3},$$

where $\hat{\lambda}^2_{b/3}$ is computed using a batch size $\lfloor b/3 \rfloor$ [Equation 5, Vats and Knudson, 2021].

To apply the batch means ESS to trees, we represent trees as vectors of splits. We now walk through this generalization. If the posterior distribution contains S non-trivial splits, $s_1, \ldots, s_S$, then we transform the vector of trees into a matrix, where in each row we represent that tree as its vector of coordinates in RF-space $\boldsymbol{X}$. Namely,

$$X_{ij} = \begin{cases} 1 & \text{if } s_j \in \tau_i, \\ 0 & \text{otherwise,} \end{cases}$$

As our distance metric we take the Euclidean distance, so the Fréchet mean is the arithmetic mean, $\bar{\boldsymbol{X}}$. We choose a batch size b that scales with n (we use $\lfloor n^{1/2} \rfloor$). Again, $a = \lfloor n/b \rfloor$ is the number of batches, and $\boldsymbol{Y}$ the matrix of batch means, with $\boldsymbol{Y}_i$ the vector of means in the ith batch (subset of the chain). For a fixed split j,

$$Y_{ij} = \frac{1}{b}\sum_{k=(i-1)b+1}^{ib} X_{kj}.$$

We use a Fréchet-based generalization for $\hat{\lambda}_b^2$, namely,

$$\hat{\lambda}_b^2 = \frac{b}{a-1}\sum_{i=1}^{a} d(\boldsymbol{Y}_i, \bar{\boldsymbol{X}})^2.$$

Similarly, We use a Fréchet-based generalization for $\hat{\sigma}_\pi^2$,

$$\hat{\sigma}_\pi^2 = \frac{1}{n-1}\sum_{i=1}^{n} d(\boldsymbol{X}_i, \bar{\boldsymbol{X}})^2.$$

Given these modified $\hat{\lambda}_b^2$ and $\hat{\sigma}_\pi^2$, we use Equation 19 to compute the ESS. We note that all the batch means,$\boldsymbol{Y}_i$, are in fact the split frequencies (or estimates split probabilities) in those batches, and the global mean, $\bar{\boldsymbol{X}}$ are the marginal split frequencies across the entire posterior distribution. We thus call this the split frequency ESS, or splitFrequencyESS.

Approaches to calculating the ESS by projecting the tree to a single dimension

All dimension-reduction approaches entail first transforming the trees into a 1-D representation, then taking the ESS of that. We use the The R package `coda` [Plummer et al., 2006] implementation of the ESS, and before we discuss our approaches we first outline how it works.

The ESS computation in `coda`

The R package `coda` [Plummer et al., 2006], commonly used for MCMC diagnostics, fits an autoregressive model to the MCMC samples to estimate the ESS. Specifically, the `coda` estimate of the ESS, which we will call the power spectrum ESS, is,

$$\text{ESS} = n\frac{\hat{\sigma}_\pi^2}{\widehat{\Gamma(0)}}, \tag{20}$$

where $\widehat{\Gamma(0)}$ is an estimate of the power spectrum at frequency 0 [see Heidelberger and Welch, 1981, for details], and $\hat{\sigma}_\pi^2$ is the estimate of the posterior variance computed from the samples. This follows from Equation 5 and the fact that the standard error of the mean of a covariance-stationary process is $\Gamma(0)/n$ [Heidelberger and Welch, 1981]. The power spectrum at 0, $\Gamma(0)$, can be linked to the autoregressive parameters by,

$$\Gamma(0) = \frac{\sigma_e^2}{\left(1-\sum_{i=1}^{p}\phi_i\right)^2},$$

where σ_e^2 is the error variance (the variance unexplained by the autoregressive model, also called the noise variance), also called the noise variance [Von Storch and Zwiers, 2001]. In practice, `coda` estimates $\widehat{\Gamma(0)}$ using an autoregressive process of unknown order

p. With an estimated order, $\hat{p}$, a fitted set of autoregression coefficients $\phi_1, \ldots, \phi_{\hat{p}}$, and an estimated error variance $\hat{\sigma}_e^2$, the estimate is,

$$\widehat{\Gamma(0)} = \frac{\hat{\sigma}_e^2}{\left(1 - \sum_{i=1}^{\hat{p}} \hat{\phi}_i\right)^2}.$$

The foldedRankmedoidESS

Vehtari *et al.* introduce two new approaches for computing ESS measures, one of which, the folded rank-transformed ESS, can be co-opted for phylogenies relatively painlessly [Vehtari et al., 2021]. For a real-valued parameter x, this ESS is computed for the transformed variable z, where there are a few layers of transformations:

$$\begin{aligned} \zeta &= |x - \text{median}(x)|, \\ r &= \text{rank}(\zeta), \\ z &= \Phi^{-1}\left(\frac{r - 3/8}{n - 1/4}\right). \end{aligned}$$

The first step is to "fold" the variable, and track the absolute deviations from the median. Then a rank transformation is applied, which stabilizes for any extreme deviations. Lastly, a Normal inverse-CDF is applied (with an offset). Vehtari et al. [2021] then take the folded rank-transformed ESS to be the ESS of z using Equation 8. In the case there is not a unique medoid tree, we compute the ESS using all possible reference trees and take the minimum.

To use this approach for trees, we make a few generalizations, and we call the resulting ESS the foldedRankmedoidESS. First, we replace the sample median with the medoid, which is a generalization of the median to higher dimensions. Specifically, the medoid tree is the (sampled) tree with the minimum sum of distances to all other sampled trees, $\text{medoid}(\boldsymbol{\tau}) = \underset{\Psi \in \boldsymbol{\tau}}{\text{argmin}} \sum_i d(\Psi, \tau_i)$. Then, we replace the absolute divergence with the distance (in one dimension, these are equivalent). The foldedRankmedoidESS is computed for the transformed variable z, where we obtain z through the following transformations:

$$\begin{aligned} \zeta &= d(\tau, \text{medoid}(\tau)), \\ r &= \text{rank}(\zeta), \\ z &= \Phi^{-1}\left(\frac{r - 3/8}{n - 1/4}\right). \end{aligned}$$

The totalDistanceESS

As an alternative to picking a specific reference tree, as in the foldedRankmedoidESS, medianPseudoESS, or minPseudoESS, we also consider an ESS based on the sum of distances between each tree and all the other trees. In this setup, we compute the ESS of the transformed variable y, defined by $y_i = \sum_{j=1}^{n} d(\tau_i, \tau_j)$. We call this the total distance ESS, or totalDistanceESS.

The CMDSESS

We also consider multidimensional scaling of the (squared) distance matrix $\boldsymbol{D}^2$ to compute an ESS. Specifically, we use classical multidimensional scaling. This approach seeks to find a matrix $\boldsymbol{Y}$ which minimizes a loss function called the strain between $\boldsymbol{Y}$ and the $\boldsymbol{B}$, a doubly centered version of $\boldsymbol{D}$. As our new variable we take the first column of the new matrix, $\boldsymbol{Y}_{\cdot 1}$ We call this the classical multidimensional scaling ESS, or CMDSESS.

Ad-hoc approaches to computing the ESS

If we define s_0 to be the time lag at which samples from our MCMC become independent of each other, then we could somewhat conservatively estimate the ESS as n/s_0. This approach can be seen as a naive implementation of the idea that the effective sample size is the hypothetical number of independent samples contained within the n MCMC samples. This is not tied to any mathematical definition of the ESS, and is not without problems. For one, the approach is expected to be overly conservative, as it effectively discards all samples in between an estimated autocorrelation time, whereas classical ESS approaches keep fractions of all samples. Additionally, in this approach ESS can take on only n distinct values because it is guaranteed that s_0 is an integer between 1 and n (inclusive). The approximate ESS of Lanfear et al. [2016] can be seen as one approach to overcoming these limitations, as it requires estimating s_0 then uses identities about expected distances. In the following sections, we consider methods for estimating s_0 and simply using this to estimate the ESS directly, $\widehat{\text{ESS}} = n/\hat{s}_0$.

The jumpDistanceBootstrapUnsmoothedESS and the jumpDistanceBootstrapESS

We now define two new approaches to computing the ESS based on estimating s_0. In both approaches, we start with a similarity or dissimilarity measure for trees at time lag s, $g(s)$, which we then smooth into a monotonically increasing function $G(s)$. We do this by defining $G(s) = \max(g(s), g(t-1))$ for dissimilarity measures and $G(s) = \max(-g(s), -g(t-1))$ for similarity measures. In essence, regardless of $g(s)$, $G(s)$ is a distance or dissimilarity measure. We also consider a smoother version of $G(s)$, which we call $G^*(s)$, which we define below. We do not search for an asymptote of either curve directly, as Lanfear et al. [2016] do for the approximate ESS. Rather, we seek the point at which the dissimilarity of trees at time lag s is indistinguishable from the dissimilarity of a pair of trees drawn independently from the posterior distribution.

Let $\Pr(G(1) \mid \text{iid})$ be the distribution of $G(1)$ given a set of iid samples from the posterior. Given a probability α, we define a threshold ϵ to be the $(1-\alpha)$th percentile of $\Pr(G(1) \mid \text{iid})$. We estimate $\hat{\epsilon}$ using bootstrap resampling of the posterior samples, which breaks the autocorrelation but preserves the fact that the samples are from the posterior distribution. Given a choice of α and an estimate $\hat{\epsilon}$, s_0 is the first time lag s for which $G(s) > \epsilon$. In this paper, we define $G(s)$ to be the median RF distance between trees at time lag s, and call the resulting estimate the unsmoothed jump-distance bootstrap ESS, jumpDistanceBootstrapUnsmoothedESS, though we note that any choice of $G(s)$

could rightfully be called a bootstrap ESS. In practice, we set $\alpha = 0.05$, such that s_0 is the time lag at which the tree-to-tree dissimilarity is at least as big as the 5th percentile of the tree-to-tree dissimilarity for trees drawn identically and independently from the posterior distribution.

To circumvent the fact that the jumpDistanceBootstrapUnsmoothedESS can only take on values in $n/1, n/2, n/3, \ldots, n/(n-1), 1$, we also consider using smoothing. Specifically, we use linear interpolation to smooth $G(s)$ into $G^*(s)$. If $\boldsymbol{s}^{\text{step}}$ is the vector of times at which $G(s)$ changes, we can define a piecewise linear function $G^*(s)$ as,

$$G^*(s) = G(s_i^{\text{step}}) + \frac{s - s_i^{\text{step}}}{s_{i+1}^{\text{step}} - s_i^{\text{step}}}(G(s_{i+1}^{\text{step}}) - G(s_i^{\text{step}})), \tag{21}$$

where s is in the ith interval ($s_i^{\text{step}} \leq s < s_{i+1}^{\text{step}}$). Defining s_0 to be the time s such that $G^*(s) >= \epsilon$ allows us to assign fractional s_0, and have a continuous estimator. We call the resulting estimator the (smoothed) jump distance bootstrap ESS, jumpDistanceBootstrapESS.

There are a few constraints that must be imposed to complete this approach. In the pathological case where all trees are the same tree, we set $\widehat{\text{ESS}} = 1$, as clearly if we have only sampled one topology we have an effective sample size of 1. The unsmoothed approach would not have a defined answer, as there is s for which $G(s) > \epsilon = 0$, and the smoothed approach would yield $\widehat{\text{ESS}} = n$ as $G(1) = \epsilon = 0$. It is also possible that there is no observed s for which $G^*(s) = \hat{\epsilon}$ and that $G^*(s) < \hat{\epsilon}$ for all s, in which case we enforce a minimum ESS of 1. Further, while $g(0)$ is defined, and we could infer $s_0 < 1$, we enforce a maximum ESS of n.

There is evidence that the curve-fitting approach of Lanfear et al. [2016] for underestimates s_0. It is not completely clear how well these jump-distance bootstrap approaches work to estimate s_0, but it is possible that they may be useful in combination with the approximate ESS, which requires estimating s_0. We leave this to future work.

Performance of the ESS measures below ESS = 500

As an alternative to binning ESS performance using a cutoff of 500, we also consider a laxer cutoff of 250. From Figures S2 and S4 it is evident a cutoff of 250 is not sufficient. In the $250 \leq \text{ESS} < 500$ regime, it would appear most methods are generally conservative and underestimate the ESS. However, there are splits and tree topologies where the error in the estimated probability can be quite large compared to the ESS ≥ 500 regime.

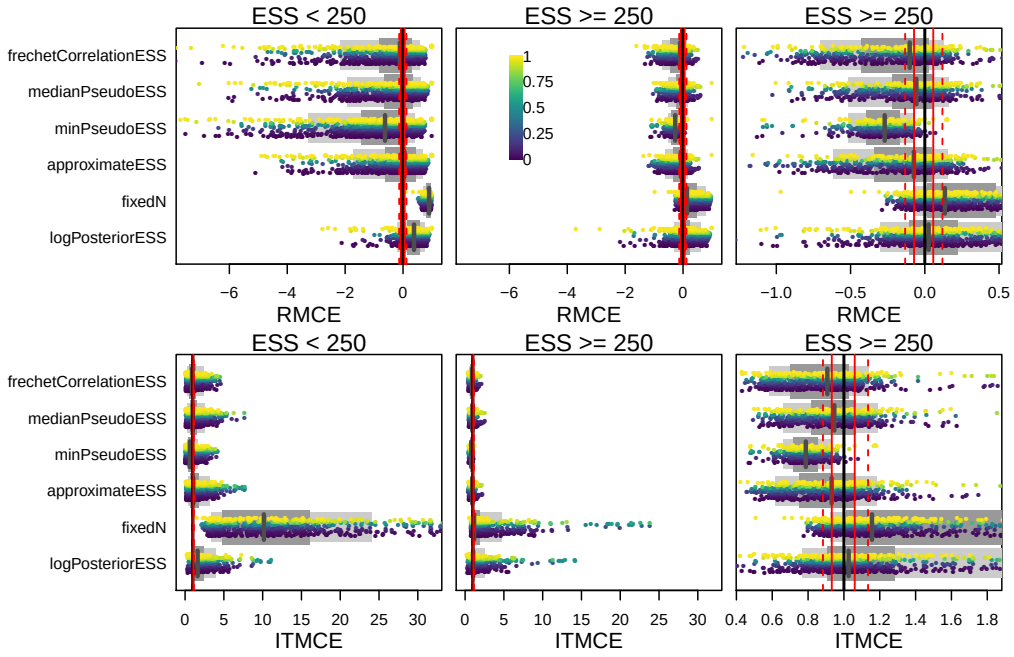

Figure S2: The RMCE $((\widehat{\mathrm{SE}}_{\mathrm{MCMC}} - \widehat{\mathrm{SE}}_{\mathrm{MCESS}})/\widehat{\mathrm{SE}}_{\mathrm{MCMC}})$ and ITMCE $(\widehat{\mathrm{SE}}_{\mathrm{MCMC}}/\widehat{\mathrm{SE}}_{\mathrm{MCESS}})$ for split probabilities for all topological ESS measures and all 45 combinations of 9 datasets and 5 run lengths. This figure uses an ESS cutoff of 250 instead of 500, but is otherwise the same as Figure 2.

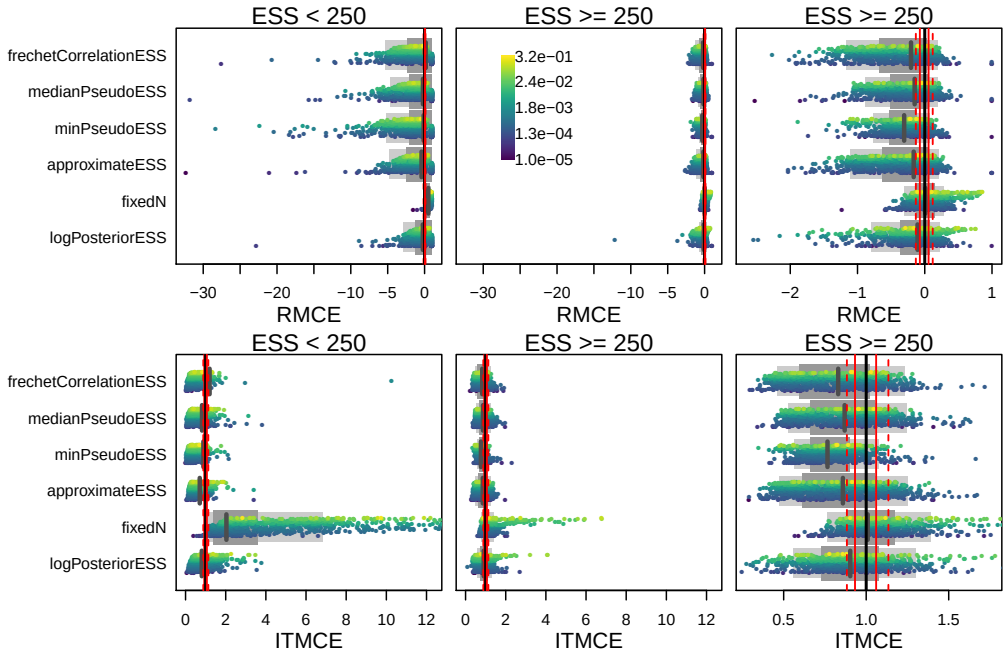

Figure S3: The RMCE $((\widehat{\mathrm{SE}}_{\mathrm{MCMC}} - \widehat{\mathrm{SE}}_{\mathrm{MCESS}})/\widehat{\mathrm{SE}}_{\mathrm{MCMC}})$ and ITMCE $(\widehat{\mathrm{SE}}_{\mathrm{MCMC}}/\widehat{\mathrm{SE}}_{\mathrm{MCESS}})$ for topology probabilities for all topological ESS measures and all 45 dataset by run length combinations. This figure uses an ESS cutoff of 250 instead of 500, but is otherwise the same as Figure 3.

Comparing the estimates of the effective sample size

In the main text, we focused on the performance of each ESS measure separately. Here we examine how similar their estimates are, using the 4500 simulated analyses (9 datasets, 5 chain lengths, 100 replicates each). There is large-scale agreement, but there are also clearly effects of different datasets and run lengths.

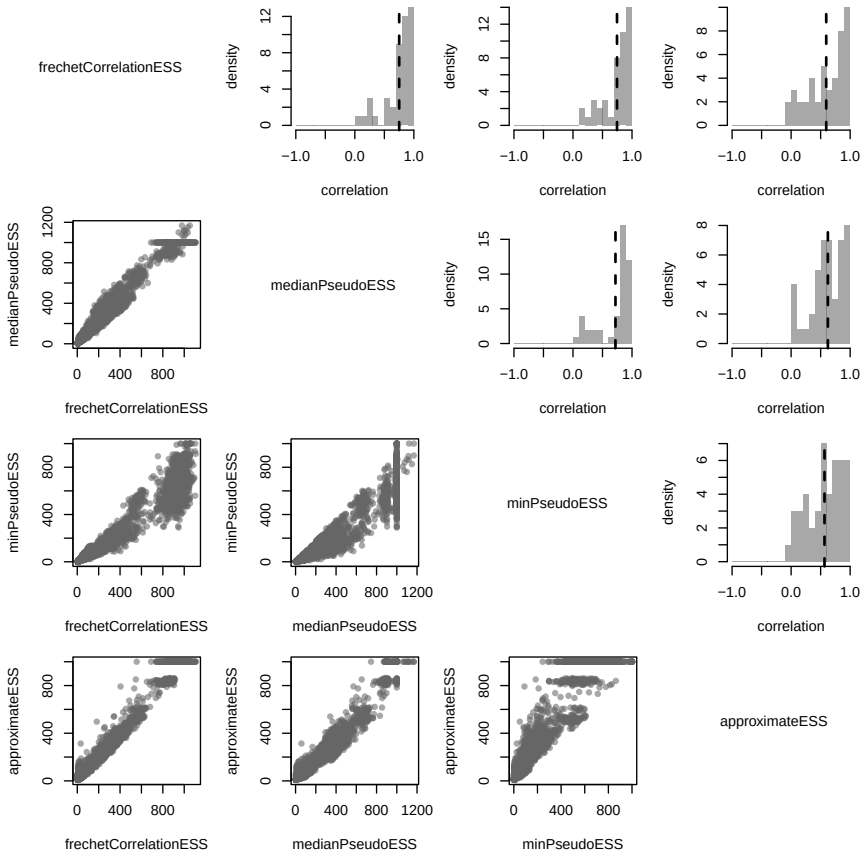

Figure S4: Comparison of the estimated ESS for the main-text tree ESS measures on all 4500 simulated analyses (9 datasets, 5 chain lengths, 100 replicates each). Below the diagonal, we simply plot the estimated ESS for each pair of methods. However, there is clearly variability across the different datasets and run lengths in concordance. We summarize this variability in the histograms above the diagonal. These summarize the 45 correlation coefficients computed for all 100 replicates for each dataset and run length combination. The dashed vertical line is the mean.

Performance of all additional tree ESS measures

Across all 10 ESS methods (4 main text ESS methods and the 6 introduced in the supplement), performance is mostly similar to the main-text results. The two *ad-hoc* "jump distance" approaches that only use s_0 to estimate the ESS drastically underestimate the ESS. The splitFrequencyESS performs about as well as the frechetCorrelationESS, with slightly worse performance in the ESS < 500 regime and slightly better performance in the ESS ≥ 500 regime. The dimension-reduction approaches all perform relatively similarly. The foldedRankmedoidESS generally performs equivalently to medianPseudoESS, the CMDSESS is slightly more conservative, and the performance of the totalDistanceESS is a bit more variable. In Figures S5-S7, we plot all 10 methods for all 3 MCMCSE measures. For simplicity, and since the results are similar, we present only the RMCE. Overall, a combination of the minPseudoESS and either the frechetCorrelationESS, splitFrequencyESS, foldedRankmedoidESS, or medianPseudoESS should cover both the ESS < 500 and ESS ≥ 500 regimes in practice.

Scalability

Computing most of the tree ESS measures described requires computing the entire $n \times n$ distance matrix, which is computationally costly and scales with the square of the number of trees. Many of the described methods can be altered to accommodate subsampling, and the RWTY implementation implements this for both the approximate ESS and medianPseudoESS [Warren et al., 2017]. Future work will be needed to determine whether any methods perform adequately with subsampling, and which methods provide an adequate runtime for either very large samples of trees or samples of very large trees.

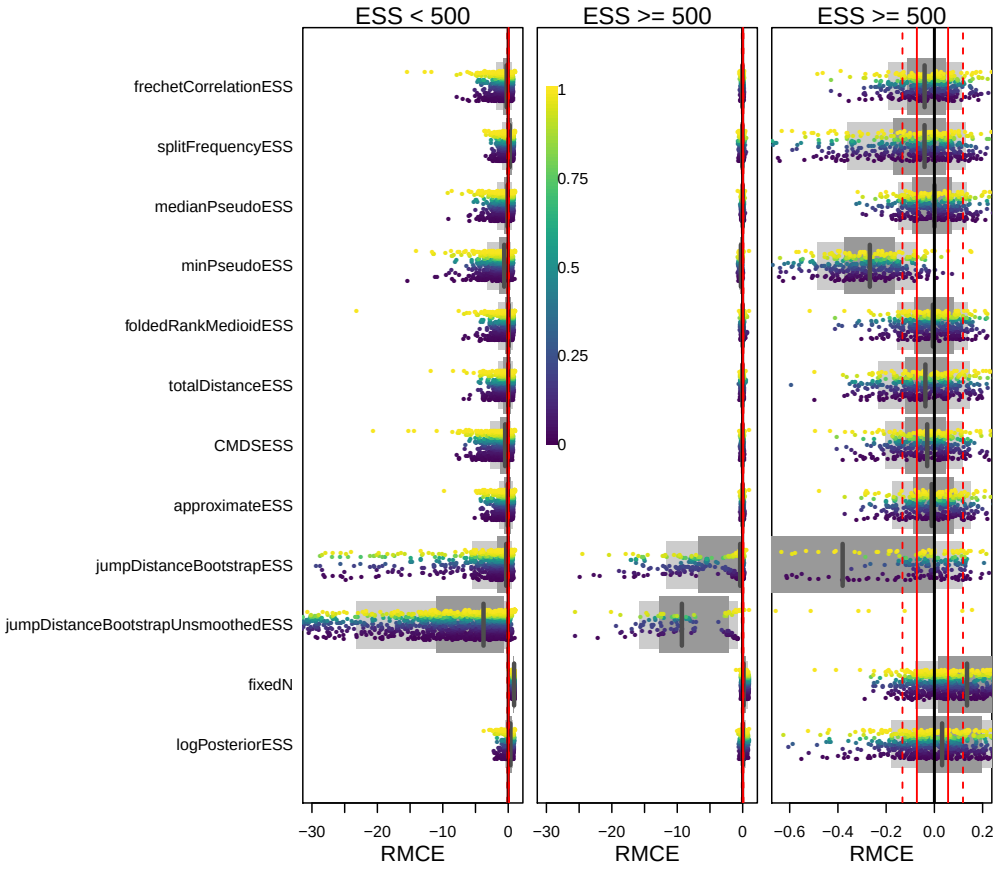

Figure S5: The RMCE $((\widehat{\mathrm{SE}}_{\mathrm{MCMC}} - \widehat{\mathrm{SE}}_{\mathrm{MCESS}})/\widehat{\mathrm{SE}}_{\mathrm{MCMC}})$ for split probabilities for all topological ESS measures and all 45 combinations of 9 datasets and 5 run lengths. This figure is a more comprehensive version of Figure 2 including all ESS measures considered in the paper.

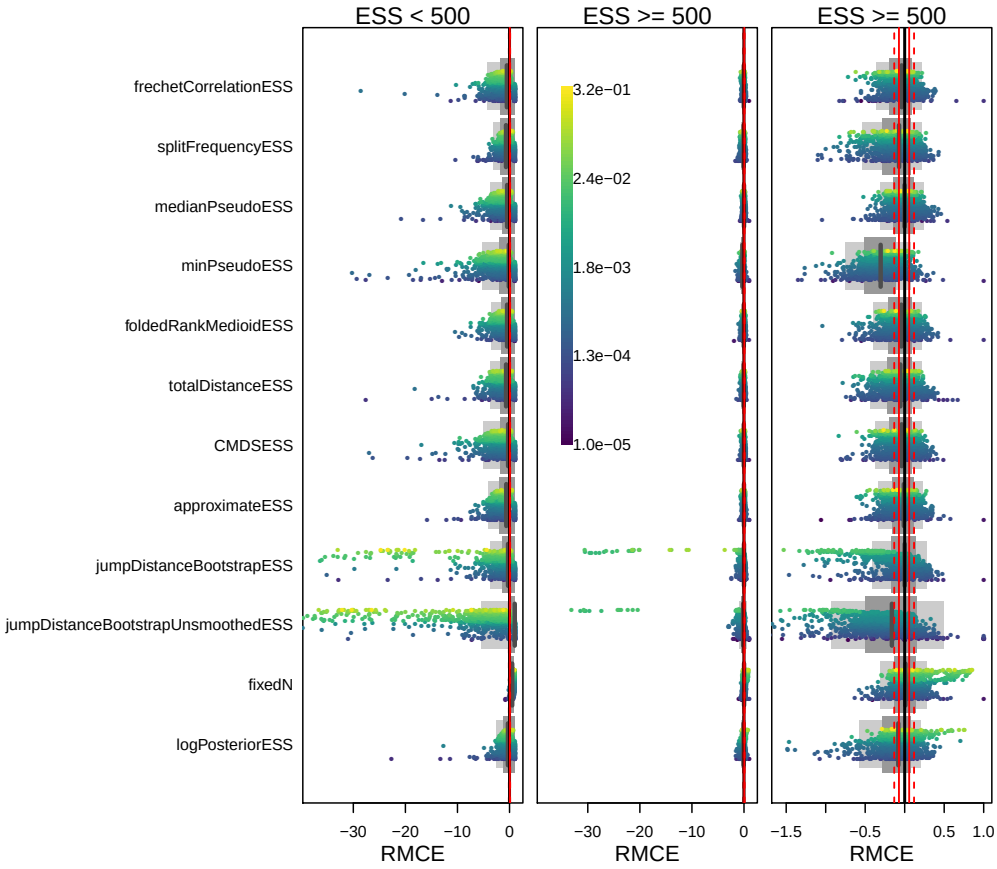

Figure S6: The RMCE $((\widehat{\mathrm{SE}}_{\mathrm{MCMC}} - \widehat{\mathrm{SE}}_{\mathrm{MCESS}})/\widehat{\mathrm{SE}}_{\mathrm{MCMC}})$ for topology probabilities for all topological ESS measures and all 45 dataset by run length combinations. This figure is a more comprehensive version of Figure 3 including all ESS measures considered in the paper.

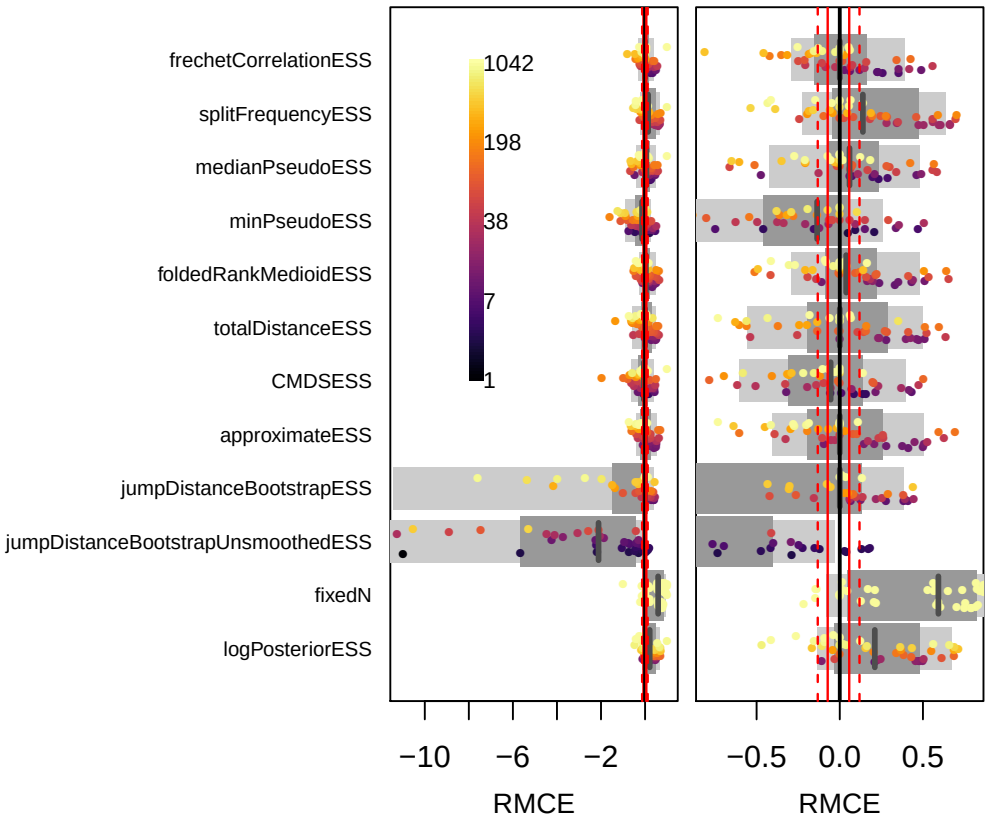

Figure S7: The RMCE $((\widehat{\mathrm{SE}}_{\mathrm{MCMC}} - \widehat{\mathrm{SE}}_{\mathrm{MCESS}})/\widehat{\mathrm{SE}}_{\mathrm{MCMC}})$ for the majority-rule consensus (MRC) tree for all topological ESS measures and all 45 dataset by run length combinations. This figure is a more comprehensive version of Figure 4 including all ESS measures considered in the paper.

Additional empirical results

For completeness, we now present split-split plots with confidence intervals for the other 5 datasets of Scantlebury [2013]. The confidence intervals for split probabilities here are computed using the Jeffreys interval [Brown et al., 2001], while the confidence intervals for the difference in split probabilities are computed here using the approach of Agresti and Caffo [Agresti and Caffo, 2000]. These approaches appear to work well in practice, though in `treess` we implement alternatives to both. We also present an aggregation across all these plots comparing the confidence interval approach to more standard approaches based on the average and maximum standard deviations of split frequencies (ASDSF and MSDSF). This aggregation highlights the fact that similar ASDSF or MSDSF can correspond to a range of numbers of splits whose probabilities disagree across runs (and vice-versa).

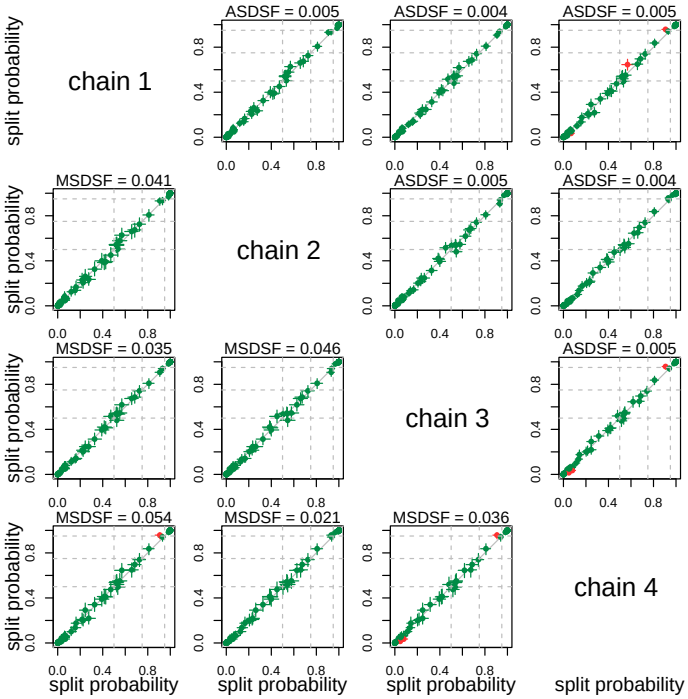

Figure S8: Split probabilities computed for all chains of the *Cophyline* dataset of Scantlebury [2013], plotted against the probabilities computed for all other chains, with confidence intervals. Comparisons above the diagonal use the frechetCorrelationESS to compute confidence intervals, while comparisons below the diagonal use the minPseudoESS, which is generally smaller and thus leads to larger confidence intervals. Each confidence interval is colored by whether or not the 95% CI for the difference in split probability between chains i and j includes 0 (green for including 0, red for excluding 0). CIs for differences in probability that exclude 0 (or non-overlapping confidence intervals) are more likely to be indicative of convergence issues between chains. Narrower confidence intervals from larger tree ESS estimates will flag more splits as problematic (as in chains 1 and 4). Dashed grey lines indicate posterior probabilities of 0.5 (threshold for inclusion in the MRC tree), 0.75 (moderate support for a split), and 0.95 (strong support for a split). For comparison, we include the average standard deviation of split frequencies (ASDSF, above the diagonal) and maximum standard deviation of split frequencies (MSDSF, below the diagonal).

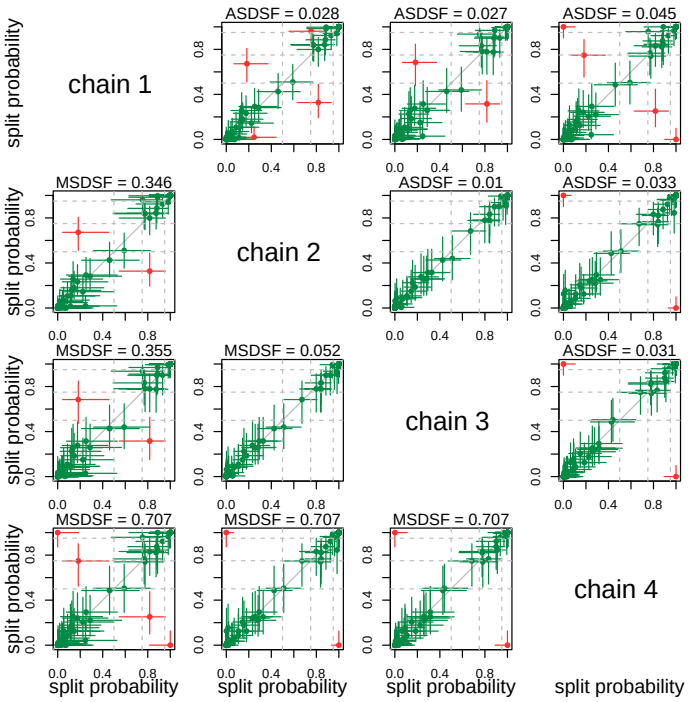

Figure S9: Split probabilities computed for all chains of the *Gephyromantis* dataset of Scantlebury [2013], plotted against the probabilities computed for all other chains, with confidence intervals. For more explanation, see Figure S8 caption.

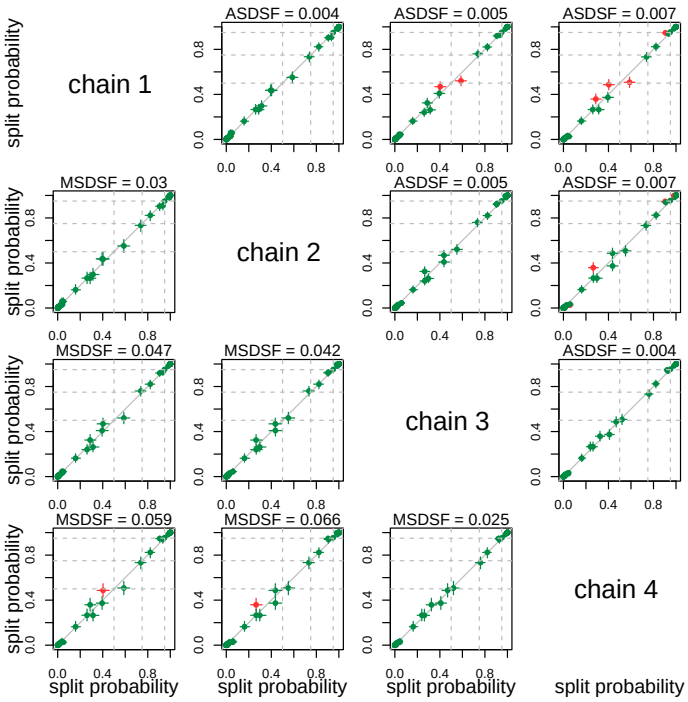

Figure S10: Split probabilities computed for all chains of the *Heterixalus* dataset of Scantlebury [2013], plotted against the probabilities computed for all other chains, with confidence intervals. For more explanation, see Figure S8 caption.

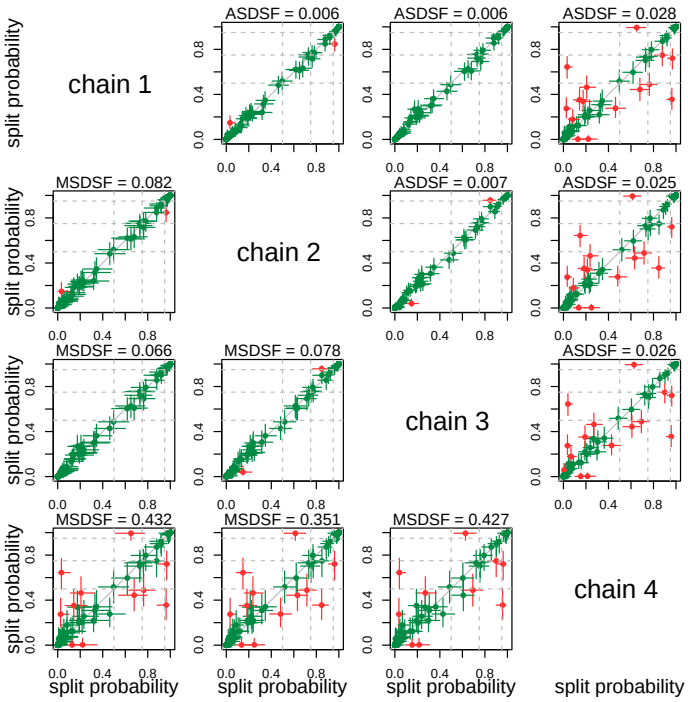

Figure S11: Split probabilities computed for all chains of the *Phelsuma* dataset of Scantlebury [2013], plotted against the probabilities computed for all other chains, with confidence intervals. For more explanation, see Figure S8 caption.

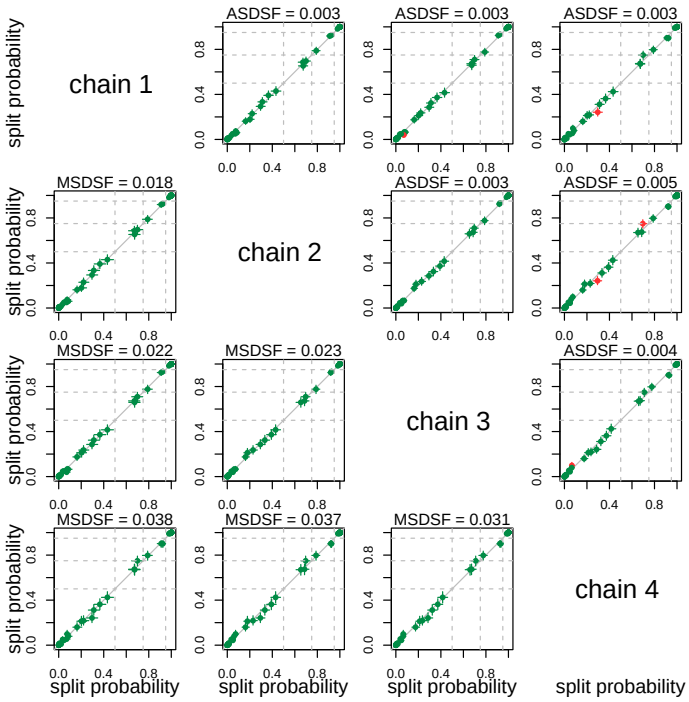

Figure S12: Split probabilities computed for all chains of the *Uroplatus* dataset of Scantlebury [2013], plotted against the probabilities computed for all other chains, with confidence intervals. For more explanation, see Figure S8 caption.

Comparing split confidence intervals and the ASDSF

Here we consider how the ASDSF (and MSDSF) compare to our confidence-interval-based approach to comparing splits.

Let us first take the *Gephyromantis* and *Phelsuma* datasets together as a case study. The ESS is low for both (Figure 6), with an average frechetCorrelationESS of 26 for the *Gephyromantis* dataset and 129 for the *Phelsuma* dataset. The MSDSF is large for many pairwise chain comparisons, and visual inspection of split-split probability plots shows notable discrepancies. Clearly the analyses of these datasets encountered MCMC difficulties. What do we learn from the various comparisons available to us? The MSDSF clearly indicates that there are between-chain convergence problems in both datasets, at which point we might plot the split probabilities to see what is going on. We would clearly see that chain 4 is distinct in both datasets, and that chain 1 also appears discordant in the *Gephyromantis* dataset. The confidence intervals provide additional information and suggest different failure modes between the chains. That many splits (roughly a dozen) disagree for the *Phelsuma* dataset suggests the possibility that the fourth chain has converged to a different local mode than the others. In this case, running the chains longer should solve the problem. Without accounting for the effective sample size, we might think that the *Gephyromantis* dataset experienced similar problems. But accounting for the low ESS, we see that the pattern is being driven by only 2-4 splits. When considered with the fact that there appear to be 3 clusters of chains, 1, 2+3, and 4, we may begin to suspect that a peculiarity of the treespace which is causing difficulty mixing. In this case, longer runs alone may not easily solve the problem and we may wish to consider alternatives like Metropolis-coupling.

We also consider a coarser comparison of the confidence interval approach with the ASDSF and MSDSF. In Figure S13, we plot the number of splits which appear to be distinctly different using the confidence interval approach (those colored red) against the ASDSF and MSDSF. While the measures are correlated, we can see that for a given ASDSF or MSDSF, there is a notable range of numbers of splits which differ, and vice-versa.

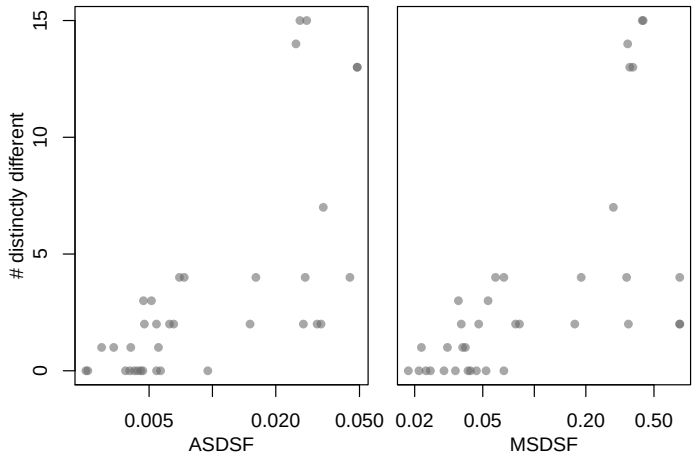

Figure S13: An aggregated comparison of our confidence interval approach to comparing split probabilities and previously-existing approaches. For all 24 pairwise comparisons of chains in the Malagasy analyses, we compute the average standard deviation of split frequencies (ASDSF) and maximum standard deviation of split frequencies (MSDSF). We also count the number of splits for which the 95% CI for the difference in split probability between chains excludes 0, which we term the number of distinctly different splits. We us the frechetCorrelationESS to compute the CIs. While this number is correlated with the ASDSF and MSDSF, there is notable variation. Similar ASDSF and MSDSF can correspond to a range of numbers of failing splits, and vice-versa.

Comparing ESS-based measures of Monte Carlo error to multiple-chain-based measures

For splits in the consensus tree, MrBayes reports the standard deviation (across runs) of the split freqencies (the SDSF). This is a direct approach to quantifying the Monte Carlo error in split probabilities. To determine how well this approach can capture Monte Carlo error, we perform an additional experiment. For each of our 45 dataset x run length combinations, we take a small number of the independent MCMC runs (2, 4, 10, and 20) and use them to estimate the Monte Carlo error. These runs are a subset of the 100 runs used to compute the $\widehat{\text{SE}}_{\text{MCMC}}$. We compare the performance to the main-text ESS measures for split probabilities (Figure S14), tree probabilities (Figure S15), and the MRC tree (Figure S16). We find that using 2 or 4 chains fails to capture the Monte Carlo error well for any of these three quantities, performing notably worse than the logPosteriorESS and fixedN approaches. Using 10 chains can capture the Monte Carlo error for the MRC tree adequately, but not for split or tree probabilities. Using 20 chains can capture the Monte Carlo error in the MRC tree about as well as the tree ESS measures. For split and tree probabilities, the performance using 20 chains falls between the logPosteriorESS and fixedN approaches and using the tree ESS measures considered in this paper.

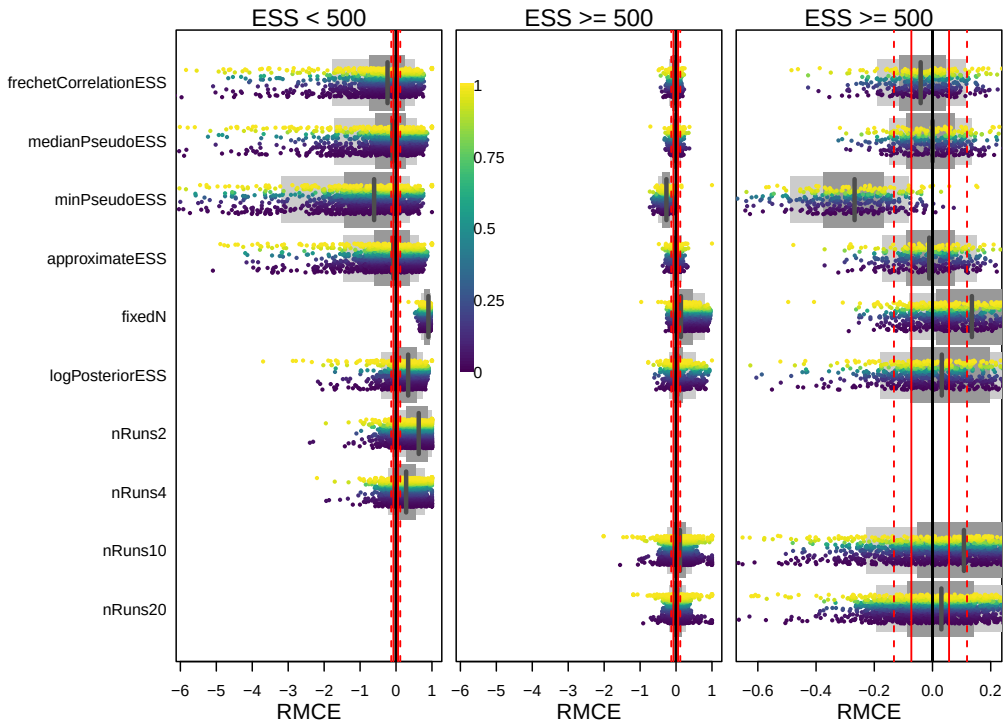

Figure S14: The RMCE $((\widehat{\text{SE}}_{\text{MCMC}} - \widehat{\text{SE}}_{\text{MCESS}})/\widehat{\text{SE}}_{\text{MCMC}})$ for split probabilities for all topological ESS measures and all 45 combinations of 9 datasets and 5 run lengths. This figure reproduces Figure 3 and adds four approaches to directly estimating the Monte Carlo error. We use 2, 4, 10, or 20 independent MCMC runs (nRuns2 to nRuns20) and the same brute-force approach employed to obtain $\widehat{\text{SE}}_{\text{MCMC}}$. These brute-force approaches do not use an ESS to estimate the Monte Carlo error. Given the performance differential between < 10 and ≥ 10 runs, we place the results from 2 and 4 chains in the ESS < 500 column and the results from 10 and 20 chains in the ESS ≥ 500 columns.

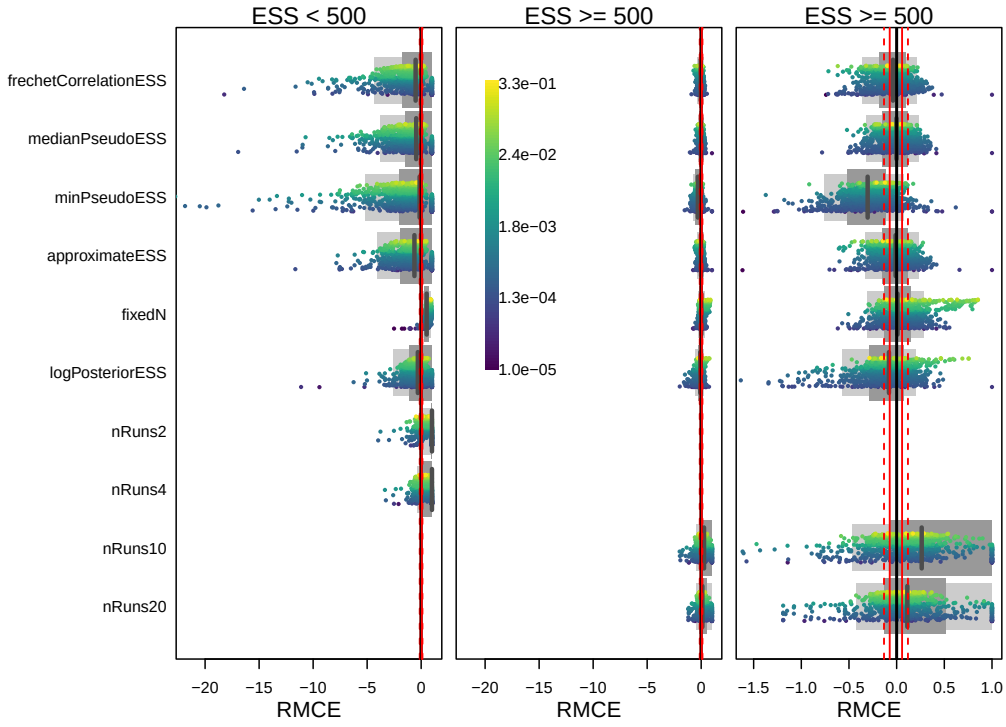

Figure S15: The RMCE $((\widehat{\text{SE}}_{\text{MCMC}} - \widehat{\text{SE}}_{\text{MCESS}})/\widehat{\text{SE}}_{\text{MCMC}})$ for topology probabilities for all topological ESS measures and all 45 dataset by run length combinations. This figure reproduces Figure 4 and adds four approaches to directly estimating the Monte Carlo error as in Figure S14.

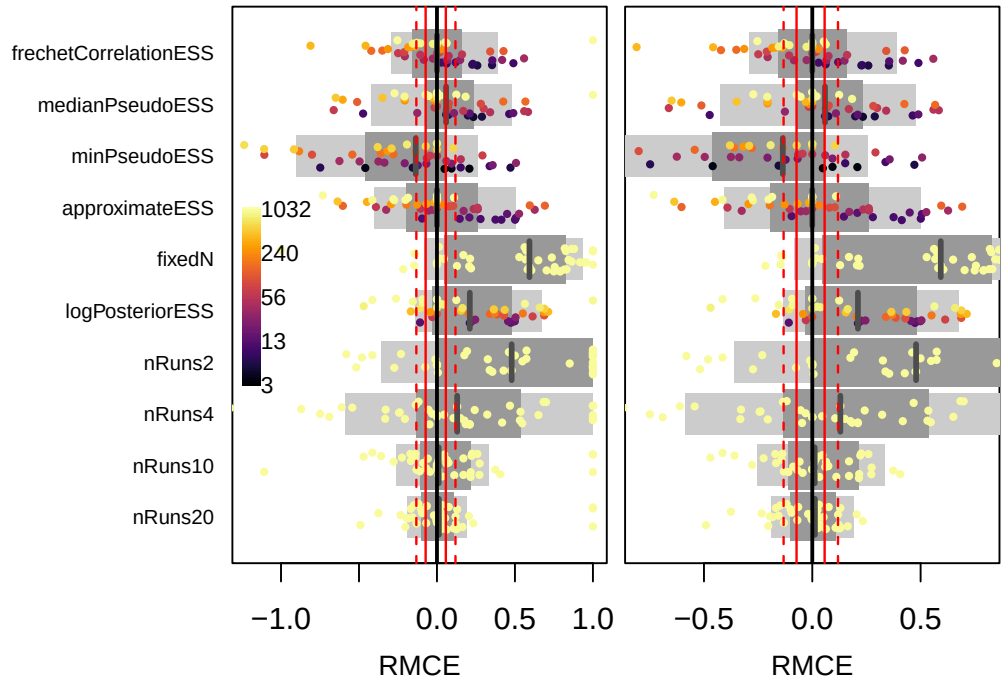

Figure S16: The RMCE $((\widehat{\mathrm{SE}}_{\mathrm{MCMC}} - \widehat{\mathrm{SE}}_{\mathrm{MCESS}})/\widehat{\mathrm{SE}}_{\mathrm{MCMC}})$ for the majority-rule consensus (MRC) tree for all topological ESS measures and all 45 dataset by run length combinations. This figure reproduces Figure 5 and adds four approaches to directly estimating the Monte Carlo error as in Figure S14. As these brute-force approaches do not use an ESS to estimate the Monte Carlo error, we arbitrarily color the points for the nRuns approaches as if they had ESS 1000.

References

Alan Agresti and Brian Caffo. Simple and effective confidence intervals for proportions and differences of proportions result from adding two successes and two failures. *The American Statistician*, 54(4):280–288, 2000.

Lawrence D Brown, T Tony Cai, and Anirban DasGupta. Interval estimation for a binomial proportion. *Statistical Science*, pages 101–117, 2001.

Philip Heidelberger and Peter D Welch. A spectral method for confidence interval generation and run length control in simulations. *Communications of the ACM*, 24 (4):233–245, 1981.

Sebastian Höhna, Michael J Landis, Tracy A Heath, Bastien Boussau, Nicolas Lartillot, Brian R Moore, John P Huelsenbeck, and Fredrik Ronquist. RevBayes: Bayesian phylogenetic inference using graphical models and an interactive model-specification language. *Systematic Biology*, 65(4):726–736, 2016.

Robert Lanfear, Xia Hua, and Dan L Warren. Estimating the effective sample size of tree topologies from bayesian phylogenetic analyses. *Genome Biology and Evolution*, 8(8):2319–2332, 2016.

Philippe Lemey, Marco Salemi, and Anne-Mieke Vandamme. *The Phylogenetic Handbook: A Practical Approach to Phylogenetic Analysis and Hypothesis Testing*. Cambridge University Press, 2009.

Martyn Plummer, Nicky Best, Kate Cowles, and Karen Vines. Coda: Convergence diagnosis and output analysis for mcmc. *R News*, 6(1):7–11, 2006. URL `http://CRAN.R-project.org/doc/Rnews/`.

Dimitris N Politis. The impact of bootstrap methods on time series analysis. *Statistical Science*, 18:219–230, 2003.

Fredrik Ronquist, Maxim Teslenko, Paul Van Der Mark, Daniel L Ayres, Aaron Darling, Sebastian Höhna, Bret Larget, Liang Liu, Marc A Suchard, and John P Huelsenbeck. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Systematic Biology*, 61(3):539–542, 2012.

Daniel P Scantlebury. Diversification rates have declined in the Malagasy herpetofauna. *Proceedings of the Royal Society B: Biological Sciences*, 280(1766):20131109, 2013.

Marc A Suchard, Robert E Weiss, Janet S Sinsheimer, Karin S Dorman, Megha Patel, and Edward R B McCabe. Evolutionary similarity among genes. *Journal of the American Statistical Association*, 98(463):653–662, 2003.

Marc A Suchard, Philippe Lemey, Guy Baele, Daniel L Ayres, Alexei J Drummond, and Andrew Rambaut. Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evolution*, 4(1):vey016, 2018.

Dootika Vats and Christina Knudson. Revisiting the gelman–rubin diagnostic. *Statistical Science*, 36(4):518–529, 2021.

Aki Vehtari, Andrew Gelman, Daniel Simpson, Bob Carpenter, and Paul-Christian

Bürkner. Rank-normalization, folding, and localization: An improved $\widehat{R}$ for assessing convergence of MCMC. *Bayesian Analysis*, 2021.

Hans Von Storch and Francis W Zwiers. *Statistical Analysis in Climate Research*. Cambridge University Press, 2001.

Dan L Warren, Anthony J Geneva, and Robert Lanfear. RWTY (R We There Yet): an R package for examining convergence of Bayesian phylogenetic analyses. *Molecular Biology and Evolution*, 34(4):1016–1020, 2017.