Supplementary Information for Seo et al.'s manuscript

[A] Justification of translocation factor in bootstrap procedure.

In Equation (12) and (13), we employed the translocation factors, $\left\{\widehat{\tilde{I}}^{*}_{gfii} - \widehat{\tilde{I}}_{gfii}\right\}$ and $\left\{\widehat{\tilde{I}}^{*}_{ffii}\widehat{I}_{ffii}/\widehat{\tilde{I}}_{ffii} - \widehat{I}_{ffii}\right\}$ In this subsection, we explain why they are required.

Although the parameter-wise Fisher information quantities are all univariate, we represent them in the 2-dimensional plane of Figure S1 for the convenience of explanation. The colored crosses and points on the left side of Figure S1 represent random quantities obtained from the original data and the original simulated data. Likewise, the colored crosses and points on the right side represent random quantities obtained from bootstrap resampling. In the center of Figure S1, the black cross denotes $\tilde{I}_{gfii}$. This is unknown but can be estimated via $\widehat{\tilde{I}}_{gfii}$. The random position of $\widehat{\tilde{I}}_{gfii}$ is represented by the blue cross on the left. The blue points on the left are sitewise Fisher information values from the original incomplete data. The position of $\widehat{\tilde{I}}_{ffii}$ is represented by the red cross on the left with the red dots being sitewise Fisher information values from extremely long simulated incomplete data. If the model is correct, the red cross will approach the blue cross as the number of sites n goes to infinity. Also, the blue cross on the left will approach the black cross in the center as n goes to infinity.

The reason why translocation is required in Equation (12) is that the positions of the blue and red crosses on the left are correlated. Because $\widehat{\tilde{I}}_{gfii}$ was calculated at $\widehat{\boldsymbol{\theta}}$ and the red dots on the left were generated by $\widehat{\boldsymbol{\theta}}$, the separation between the red and blue crosses should be mimicked during bootstrap resampling. We do this by translocating the position of the red cross by the separation of the corresponding blue crosses. The amount of translocation, $\left\{\widehat{\tilde{I}}^{*}_{gfii} - \widehat{\tilde{I}}_{gfii}\right\}$, is represented by brown-colored arrow in Figure S1. After translocating the y_r to their new positions, we re-use the bootstrap indices $p(r)$'s as in Equation (12).

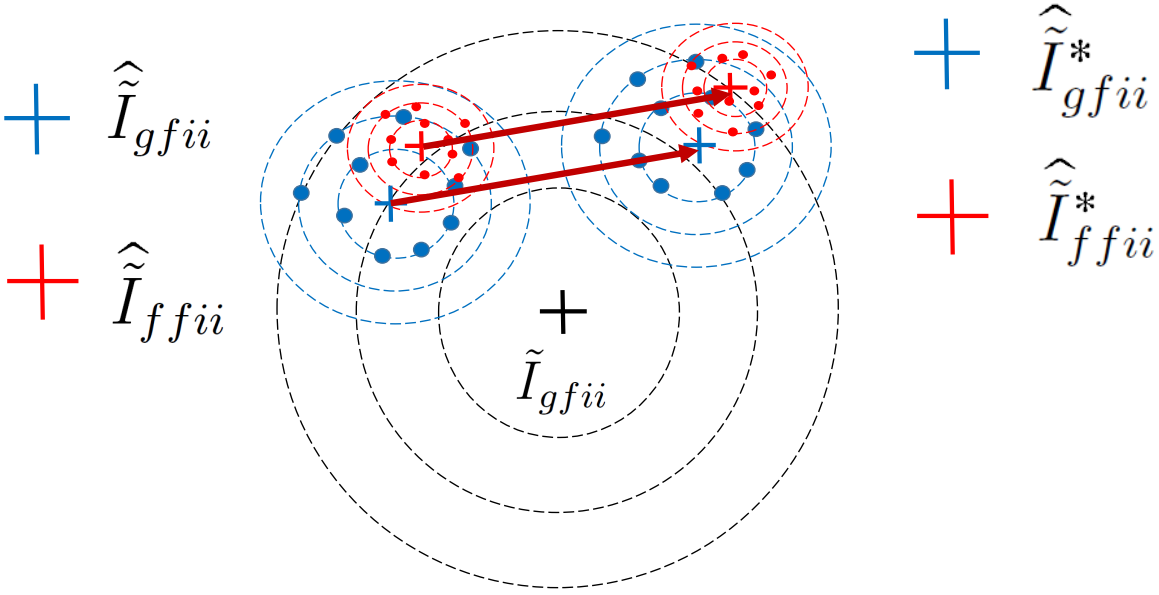

Figure S1: Summary of bootstrap scheme. The unknown Fisher information $\tilde{I}_{gfii}$ is in the center. The estimate of $\widehat{\tilde{I}}_{gfii}$ is represented by the blue cross on the left and is a realization of $\tilde{I}_{gfii}$. The blue dots around the blue cross correspond to the sitewise Fisher information $\widehat{\tilde{I}}_{gfii}$. The red cross and red dots on the left correspond to $\widehat{\tilde{I}}_{ffii}$ and its sitewise Fisher information. The boostrap resampled quantities are shown as blue and red crosses and dots on the right. The relative difference between the red cross and the blue cross on the left should be preserved on average during the bootstrap resampling shown on the right.

Conditioning on the $y_{(\cdot)}$, $\widehat{\tilde{I}}_{gfii}$, and $\widehat{\tilde{I}}_{ffii}$ in Equation (12) when taking the expectation of $y^*_{(\cdot)}$ and $\widehat{\tilde{I}}^*_{gfii}$ with respect to the random variable $p(r)$ leads to $\mathrm{E}_{p(r)}\left[\widehat{\tilde{I}}^*_{gfii} - \widehat{\tilde{I}}^*_{ffii}\right] = \widehat{\tilde{I}}_{gfii} - \widehat{\tilde{I}}_{ffii}$. That is, the difference between $\widehat{\tilde{I}}_{gfii}$ and $\widehat{\tilde{I}}_{ffii}$ – the distance between red and blue crosses on the left side of Figure S1 – is preserved 'on average' during the bootstrap procedure. This assumption of distance preservation appears reasonable when n is moderately large and is supported by our simulation results (see RESULTS).

Furthermore, conditioning on the $z_{(\cdot)}$, $\widehat{I}_{ffii}$, $\widehat{\tilde{I}}_{ffii}$ and $\widehat{\tilde{I}}^*_{ffii}$ in Equation (13) when taking the expectation of $z^*_{(\cdot)}$ with respect to the random variable $p(r)$ leads to $\mathrm{E}_{p(r)}\left[\widehat{I}^*_{ffii}\right] \Big/ \widehat{\tilde{I}}^*_{ffii} = \widehat{I}_{ffii}/\widehat{\tilde{I}}_{ffii}$. That is, the ratio of $\widehat{I}_{ffii}/\widehat{\tilde{I}}_{ffii}$ (i.e., the inverse of $\widehat{G}_i$) is preserved 'on average' during the bootstrap procedure.

[B] Approximate equality of variance and bootstrap variance

Our bootstrap procedure can estimate the variance of G-Factors and M-Factors. In this subsection, we focus on the example of G-Factors.

Using the following distribution approximation,

$$\frac{\widehat{\tilde{I}}_{ffii}}{\widehat{I}_{ffii}} \doteqdot \frac{\widehat{\tilde{I}}^*_{ffii}}{\widehat{I}^*_{ffii}}, \tag{S1}$$

we estimate the variance of left-hand side with bootstrap samples of the right-hand side. Let us represent each term of Equation (S1) as

$$\begin{aligned}
\widehat{\tilde{I}}_{ffii} &= \mu_Y + \frac{1}{\sqrt{nm}}U && \text{(S2)}\\
\widehat{I}_{ffii} &= \mu_Z + \frac{1}{\sqrt{nm}}V && \text{(S3)}\\
\widehat{\tilde{I}}^*_{ffii} &= \widehat{\tilde{I}}_{ffii} + \frac{1}{\sqrt{nm}}U^* \\
\widehat{I}^*_{ffii} &= \widehat{I}_{ffii} + \frac{1}{\sqrt{nm}}V^*,
\end{aligned}$$

where

$$\begin{aligned}
\mathrm{E}[U] &= 0, \quad \mathrm{Var}[U] = \sigma_Y^2 \\
\mathrm{E}[V] &= 0, \quad \mathrm{Var}[V] = \sigma_Z^2 \\
\mathrm{E}[U^*] &= 0, \quad \mathrm{Var}[U^*] = \sigma_{Y^*}^2 \approx \sigma_Y^2 \qquad \text{(S4)} \\
\mathrm{E}[V^*] &= 0, \quad \mathrm{Var}[V^*] = \sigma_{Z^*}^2 \approx \sigma_Z^2, \qquad \text{(S5)}
\end{aligned}$$

and where the approximations of Equation (S4, S5) hold because the pool of y^*'s and z^*'s will be similar to the pool of y's and z's and similar to the pool from model $f(\cdot)$ when n is large. Replacing the left side of Equation (S1) with Equations (S2–S3) leads to

$$\begin{aligned}
\frac{\hat{\tilde{I}}_{ffii}}{\hat{I}_{ffii}} &= \frac{\mu_Y + \frac{1}{\sqrt{nm}}U}{\mu_Z + \frac{1}{\sqrt{nm}}V} \\
&= \frac{\mu_Y}{\mu_Z} + \frac{-\frac{\mu_Y}{\mu_Z\sqrt{nm}}V + \frac{1}{\sqrt{nm}}U}{\mu_Z + \frac{1}{\sqrt{nm}}V}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
\mathrm{Var}\left[\frac{\hat{\tilde{I}}_{ffii}}{\hat{I}_{ffii}}\right] &= \mathrm{Var}\left[\frac{-\frac{\mu_Y}{\mu_Z\sqrt{nm}}V + \frac{1}{\sqrt{nm}}U}{\mu_Z + \frac{1}{\sqrt{nm}}V}\right] \\
&= \mathrm{E}\left[\left\{\frac{\frac{1}{\sqrt{nm}}U - \frac{\mu_Y}{\mu_Z\sqrt{nm}}V}{\mu_Z + \frac{1}{\sqrt{nm}}V}\right\}^2\right] - \mathrm{E}\left[\frac{\frac{1}{\sqrt{nm}}U - \frac{\mu_Y}{\mu_Z\sqrt{nm}}V}{\mu_Z + \frac{1}{\sqrt{nm}}V}\right]^2 \\
&= \mathrm{E}\left[\left\{\frac{\mu_Z}{\mu_Z + \frac{1}{\sqrt{nm}}V}\right\}^2 \cdot \left\{\frac{\frac{1}{\sqrt{nm}}U - \frac{\mu_Y}{\mu_Z\sqrt{nm}}V}{\mu_Z}\right\}^2\right] \\
&\quad -\mathrm{E}\left[\frac{\mu_Z}{\mu_Z + \frac{1}{\sqrt{nm}}V} \cdot \frac{\frac{1}{\sqrt{nm}}U - \frac{\mu_Y}{\mu_Z\sqrt{nm}}V}{\mu_Z}\right]^2. \qquad \text{(S6)}
\end{aligned}$$

By Chebyshev's inequality,

$$P\left\{|V| < k\sigma_Z\right\} > 1 - \frac{1}{k^2},$$

Then, with probability greater than $1 - 1/k^2$,

$$\frac{\mu_Z}{\mu_Z + k\frac{\sigma_Z}{\sqrt{nm}}} < \frac{\mu_Z}{\mu_Z + \frac{1}{\sqrt{nm}}V} < \frac{\mu_Z}{\mu_Z - k\frac{\sigma_Z}{\sqrt{nm}}},$$

when n is sufficiently large that the denominator of the rightmost term above will be positive. Then, with probability greater than $1 - 1/k^2$, the following inequalities hold,

$$\begin{aligned}
\mathrm{Var}\left[\frac{\widehat{\tilde{I}}_{ffii}}{\widehat{I}_{ffii}}\right] &< \left\{\frac{\mu_Z}{\mu_Z - k\frac{\sigma_Z}{\sqrt{nm}}}\right\}^2 \cdot \mathrm{E}\left[\left\{\frac{\frac{1}{\sqrt{nm}}U - \frac{\mu_Y}{\mu_Z\sqrt{nm}}V}{\mu_Z}\right\}^2\right] \\
&\quad - \left\{\frac{\mu_Z}{\mu_Z + k\frac{\sigma_Z}{\sqrt{nm}}}\right\}^2 \cdot \mathrm{E}\left[\frac{\frac{1}{\sqrt{nm}}U - \frac{\mu_Y}{\mu_Z\sqrt{nm}}V}{\mu_Z}\right]^2 \\
&\text{and} \\
\mathrm{Var}\left[\frac{\widehat{\tilde{I}}_{ffii}}{\widehat{I}_{ffii}}\right] &> \left\{\frac{\mu_Z}{\mu_Z + k\frac{\sigma_Z}{\sqrt{nm}}}\right\}^2 \cdot \mathrm{E}\left[\left\{\frac{\frac{1}{\sqrt{nm}}U - \frac{\mu_Y}{\mu_Z\sqrt{nm}}V}{\mu_Z}\right\}^2\right] \\
&\quad - \left\{\frac{\mu_Z}{\mu_Z - k\frac{\sigma_Z}{\sqrt{nm}}}\right\}^2 \cdot \mathrm{E}\left[\frac{\frac{1}{\sqrt{nm}}U - \frac{\mu_Y}{\mu_Z\sqrt{nm}}V}{\mu_Z}\right]^2.
\end{aligned}$$

Therefore, for large n, we obtain the following approximation,

$$\mathrm{Var}\left[\frac{\widehat{\tilde{I}}_{ffii}}{\widehat{I}_{ffii}}\right] \approx \mathrm{E}\left[\left\{\frac{\frac{1}{\sqrt{nm}}U - \frac{\mu_Y}{\mu_Z\sqrt{nm}}V}{\mu_Z}\right\}^2\right] - \mathrm{E}\left[\frac{\frac{1}{\sqrt{nm}}U - \frac{\mu_Y}{\mu_Z\sqrt{nm}}V}{\mu_Z}\right]^2, \tag{S7}$$

because

$$\frac{\mu_Z}{\mu_Z - k\frac{\sigma_Z}{\sqrt{nm}}} \approx \frac{\mu_Z}{\mu_Z + k\frac{\sigma_Z}{\sqrt{nm}}} \approx 1.$$

In a similar way,

$$\begin{aligned}
\frac{\widehat{\tilde{I}}^*_{ffii}}{\widehat{I}^*_{ffii}} &= \frac{\widehat{\tilde{I}}_{ffii} + \frac{1}{\sqrt{nm}}U^*}{\widehat{I}_{ffii} + \frac{1}{\sqrt{nm}}V^*} \\
&= \frac{\tilde{I}_{ffii}}{\widehat{I}_{ffii}} + \frac{-\frac{\tilde{I}_{ffii}}{\widehat{I}_{ffii}\sqrt{nm}}V^* + \frac{1}{\sqrt{nm}}U^*}{\widehat{I}_{ffii} + \frac{1}{\sqrt{nm}}V^*}
\end{aligned}$$

and

$$\mathrm{Var}\left[\frac{\widehat{\tilde{I}}^*_{ffii}}{\widehat{I}^*_{ffii}}\right] \approx \mathrm{E}\left[\left\{\frac{\frac{1}{\sqrt{nm}}U^* - \frac{\widehat{\tilde{I}}_{ffii}}{\widehat{I}_{ffii}\sqrt{nm}}V^*}{\widehat{I}_{ffii}}\right\}^2\right] - \mathrm{E}\left[\frac{\frac{1}{\sqrt{nm}}U^* - \frac{\widehat{\tilde{I}}_{ffii}}{\widehat{I}_{ffii}\sqrt{nm}}V^*}{\widehat{I}_{ffii}}\right]^2. \tag{S8}$$

As $n \to \infty$, $\widehat{\tilde{I}}_{ffii} \to \mu_Y$, $\widehat{I}_{ffii} \to \mu_Z$, $U \doteq U^*$ and $V \doteq V^*$. Therefore, we can expect the variances of Equation (S7) and Equation (S8) to be similar to each other for large n.

[C] Assessing normality of branch length estimates

The overall M-Factor and G-Factor are weighted averages that are motivated by the asymptotic normal distributions of maximum likelihood estimators around the true parameter values. This asymptotic behavior is subject to regularity conditions that do not apply when the true parameter values are on the boundary of the parameter space. When a maximum likelihood estimate is at or sufficiently near the boundary of parameter space, this is a strong indication that the distribution of the maximum likelihood estimator is not approximately normally distributed. The deviation from normality could be due to the true parameter value being on the boundary of the parameter space and/or could be due to the amount of data being insufficient for the asymptotic behavior (e.g., see Susko and Roger 2019). In phylogenetics, branch length estimates that are 0 or close to 0 can disrupt the interpretation of overall M-Factor and G-Factor values. To avoid this disruption, our practice is to exclude branches with length estimates that are close to 0 from the G-Factor and M-Factor calculations.

To explain the details of how we decide which branches (parameters) to exclude, let us ignore off-diagonal elements of the second derivatives of the log-likelihood function and consider a univariate case for simplicity. When the PSL is large, the MLE of ith branch length, $\tilde{\theta}_i$, asymptotically follows a normal distribution subject to regularity conditions (White 1982; see APPENDIX [B] of the main text)

$$\tilde{\theta}_i - \theta_i \quad \dot{\sim} \quad N\left(0, \tilde{I}_{gfii}^{-2}\tilde{J}_{gfii}/n\right), \tag{S9}$$

where $\tilde{J}_{gfii}$ is obtained with the outer product of sitewise first derivatives of the log-likelihood function (see APPENDIX [B] of the main text). This leads to the following $100(1-\alpha)\%$ Confidence Interval (CI),

$$\begin{aligned} \theta_i \quad &\in \quad \left(\tilde{\theta}_i - z_\alpha\sqrt{\tilde{I}_{gfii}^{-2}\tilde{J}_{gfii}/n}, \quad \tilde{\theta}_i + z_\alpha\sqrt{\tilde{I}_{gfii}^{-2}\tilde{J}_{gfii}/n}\right) \\ &=: \quad \left(A(\tilde{\theta}_i, \alpha), \quad B(\tilde{\theta}_i, \alpha)\right) \end{aligned} \tag{S10}$$

where z_α is the $100(1-\alpha/2)\%$ percentile of the standard normal distribution. $A(\tilde{\theta}_i, \alpha)$ and $B(\tilde{\theta}_i, \alpha)$ are respectively defined as the lower and the upper bound of the above CI.

Another way to obtain a CI is via the difference of log-likelihood scores (Pawitan 2001). By using Wilk's likelihood ratio statistic, the $100(1-\alpha)\%$ CI can be obtained as

$$\theta_i \in \left\{ \theta \;\middle|\; 2\log\frac{L(\tilde{\theta}_i)}{L(\theta)} < \chi^2_{1,(1-\alpha)} \right\} =: \Big(C(\tilde{\theta}_i, \alpha), \quad D(\tilde{\theta}_i, \alpha) \Big), \tag{S11}$$

where $\chi^2_{1,(1-\alpha)}$ is $100(1-\alpha)$ percentile of χ^2_1 distribution. $C(\tilde{\theta}_i, \alpha)$ and $D(\tilde{\theta}_i, \alpha)$ are respectively defined as the lower and the upper bound of the above CI. When $\tilde{\theta}_i$ is normally distributed, the CI's of Equation (S10) and Equation (S11) are identical. In cases such as ours where $\tilde{\theta}_i$ is not normally distributed, the two CI's approaches each other as sample size (PSL) increases.

Our empirical observation is that the phylogenetic likelihood function decreases faster (slower) than the normal density on the left (right) side of the MLE. Therefore, the CI of Equation (S11) is asymmetric whereas that of Equation (S10) is symmetric around the MLE. Although there is some stochastic variation, the general tendency is $A(\tilde{\theta}_i, \alpha) < C(\tilde{\theta}_i, \alpha)$ and $B(\tilde{\theta}_i, \alpha) < D(\tilde{\theta}_i, \alpha)$.

To select branches with MLE's that are far enough from the boundary, we can consider constraints that we term 'Weak', 'Moderate', and 'Stringent':

$$\mathrm{R}_0 \text{ [Weak]:} \qquad \tilde{\theta}_i > 0, \tag{S12}$$

$$\mathrm{R}_1(\tilde{\theta}_i, \alpha) \text{ [Moderate]:} \qquad A(\tilde{\theta}_i, \alpha) > 0, \tag{S13}$$

$$\mathrm{R}_2(\tilde{\theta}_i, \alpha, \beta) \text{ [Stringent]:} \qquad \frac{|C(\tilde{\theta}_i,\alpha) - A(\tilde{\theta}_i,\alpha)|}{B(\tilde{\theta}_i,\alpha) - A(\tilde{\theta}_i,\alpha)} < \beta, \tag{S14}$$

where β is a pre-defined tolerance.

Constraint R_0 is the minimal constraint. This is required to remove branches whose length estimate is zero and should be satisfied so that Fisher information can be defined. Because $\tilde{\theta}_i$ is the branch length estimate in our setting, $A(\tilde{\theta}_i, \alpha)$ – the lower bound of the CI of Equation (S10) – needs to be greater than zero. This requirement is implemented as R_1. Because the two CI's of

Equations (S10, S11) approach each other as the PSL increases, we can quantify their difference measured in terms of the relative proportion of the left bound difference. This is implemented with R_2. In our empirical experience, R_2 seems too harsh even for a value of $\beta = 0.1$ because it excludes too many branches. Therefore, in most cases, we suggest applying the only the $R_1(\cdot, 0.05)$ and R_0 constraints. to select meaningful branches.

Applications of constraints to mouse lemur mt-DNA: Because its PSL is smaller than for the n-DNA, the mt-DNA inferences are more subject to the effect of short branches. Figure S2 uses the log-likelihood surface around the maximum likelihood branch length estimate to illustrate how the three criteria assess normality. For the mt-DNA data, only 46 branches out of 107 satisfy constraint R_0. Using just this constraint, the mean and standard deviation of the G-Factor and M-Factor are 0.588 (0.0567) and 0.945 (0.0521). Applying the additional constraint $R_1(\cdot, 0.05)$ excludes 27 branches so that only 19 branches remain for the mt-DNA data. With this additional constraint, the G-Factor and M-Factor become 0.596 (0.0421) and 0.955 (0.0288). Adding the constraint $R_2(\cdot, 0.05, 0.1)$ further excludes 8 mt-DNA branches so that only 11 branches are used for estimating the G-Factor and the M-Factor with the mt-DNA. With this third constraint added, the G-Factor and M-Factor become 0.541 (0.0428) and 0.858 (0.0277). In contrast, none of the 111 branches are excluded from the n-DNA tree even when all three constraints are applied.

REFERENCES

Pawitan Y. 2001. In All Likelihood:Statistical Modelling and Inference Using Likelihood. Oxford University Press. pp 35–41.

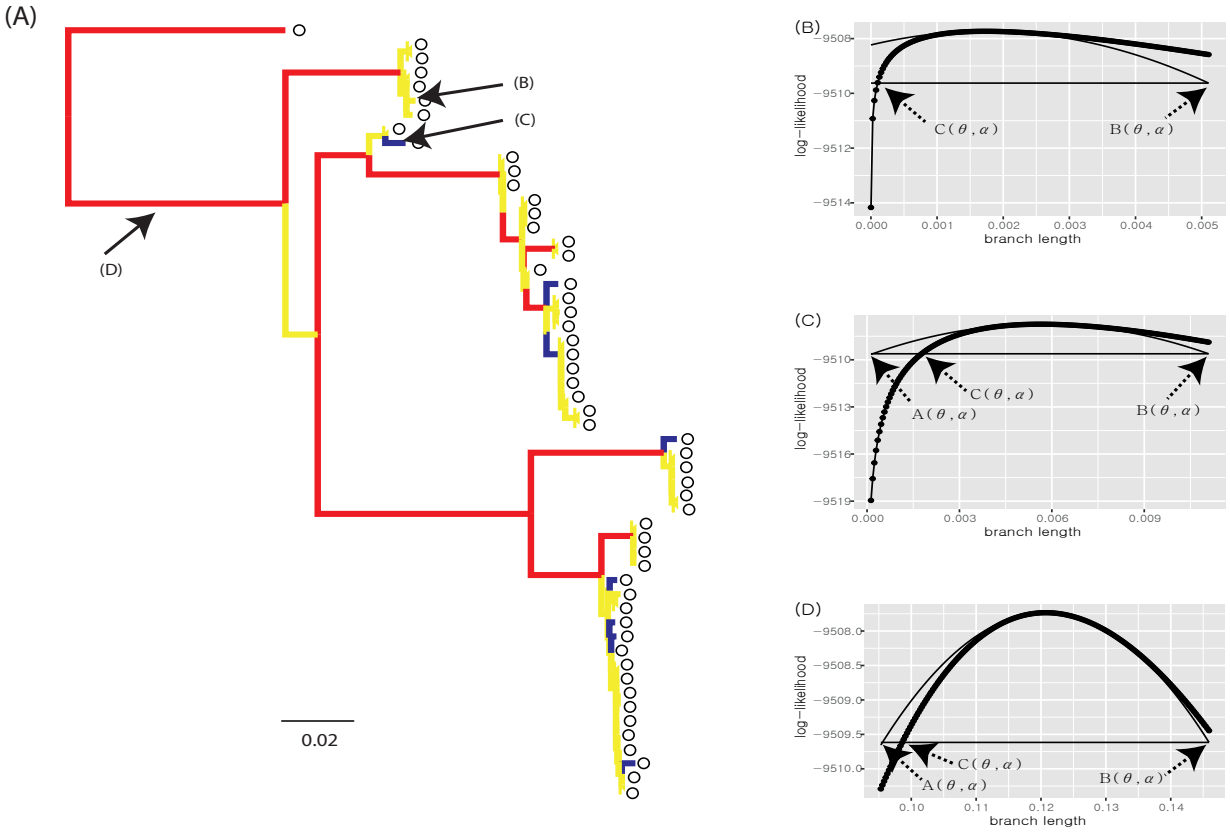

Figure S2: Applying the normality constraints to branches of the mouse lemur mt-DNA phylogeny. (A) Branches shaded blue satisfy R_0 and $R_1(\cdot, 0.05)$, but not $R_2(\cdot, 0.05, 0.1)$. Branches shaded red satisfy all three constraints (R_0, $R_1(\cdot, 0.05)$, and $R_2(\cdot, 0.05, 0.1)$). The remaining branches are shaded yellow to indicate that they either have length zero or satisfy only R_0. The labels B, C, and D respectively refer to the branches featured in the plots shown in (B), (C), and (D). For these three plots, the thick (dotted) lines represent the log-likelihood as a branch length varies around its MLE when all other parameters are held constant. The solid thin lines show the approximated normal distributions and the horizontal solid lines show the y-axis position of $\{\log L(\tilde{\theta}_i) - \chi^2_{1,(1-\alpha)}/2\}$. (B) R_0 is satisfied for this branch, but $R_1(\cdot, 0.05)$ is not satisfied. Approximated normal distribution is truncated. (C) Both R_0 and $R_1(\cdot, 0.05)$ are satisfied for this branch, but $R_2(\cdot, 0.05, 0.1)$ is not satisfied. (D) All of R_0, $R_1(\cdot, 0.05)$ and $R_2(\cdot, 0.05, 0.1)$ are satisfied for this branch.