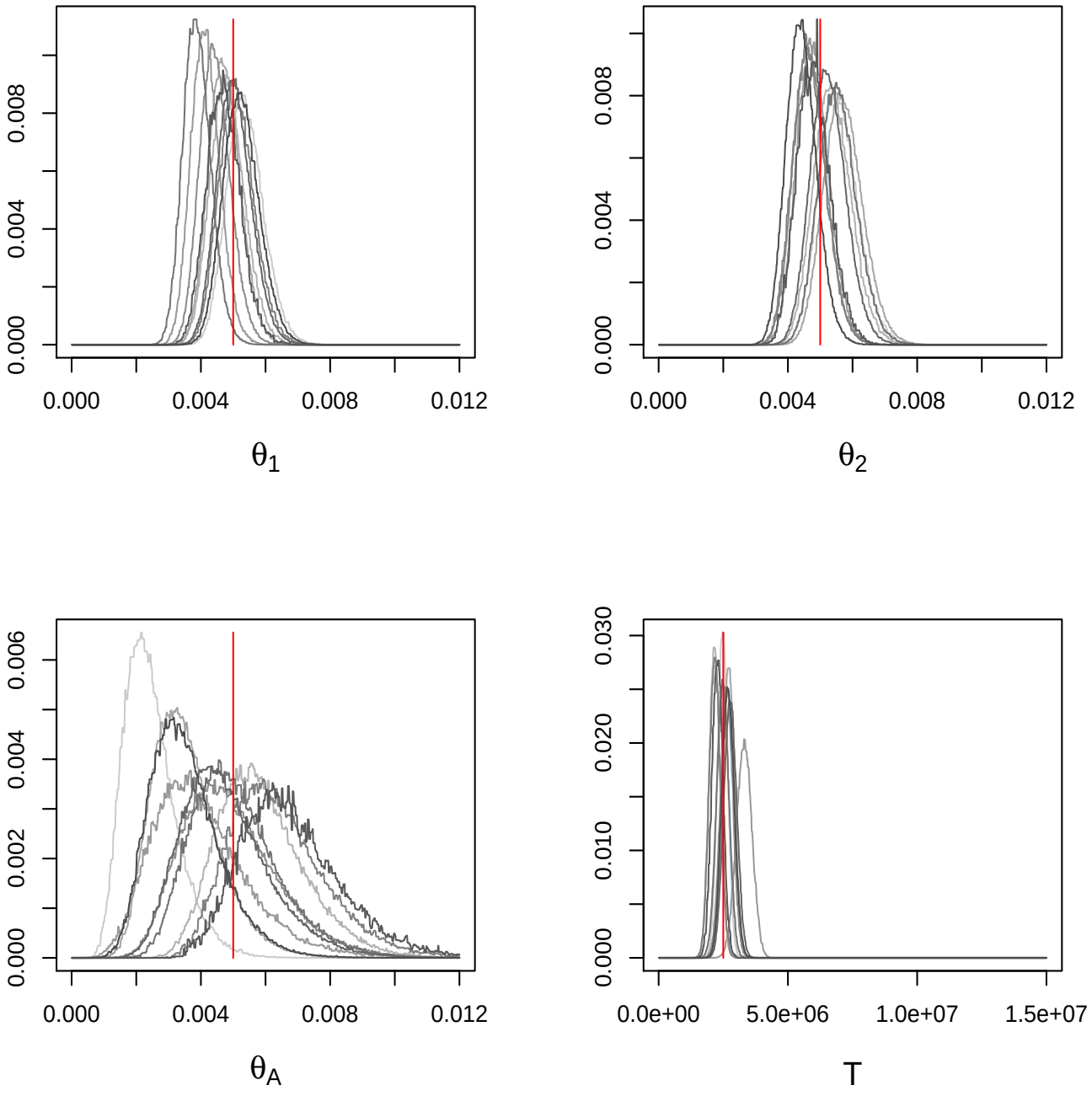
Supplementary materials

Supplementary equations

Recall that our goal is to sample from the posterior distribution: $\pi(\mathbf{\Theta}|\mathbf{D}) \propto p(\mathbf{D}|\mathbf{\Theta})p(\mathbf{\Theta})$. To do so, we perform MCMC on the set of all genealogies, $\mathbf{G} = (G_{11}, \cdots, G_{1X}, \cdots, G_{Y1}, \cdots, G_{YX})$, and the parameters, $\mathbf{\Theta} = (\theta_1, \theta_2, \theta_A, T, M, \mathbf{P})$, by sampling sets $(\mathbf{G}, \mathbf{\Theta})$ from the target distribution, $\pi(\mathbf{G}, \mathbf{\Theta})$ (see eq. 3 in Methods). By integrating equation 3 over genealogies $\mathbf{G}$, we show that the marginal distribution of sampled values of $\mathbf{\Theta}$ from the target distribution is the posterior distribution of interest, $\pi(\mathbf{\Theta}|\mathbf{D})$:

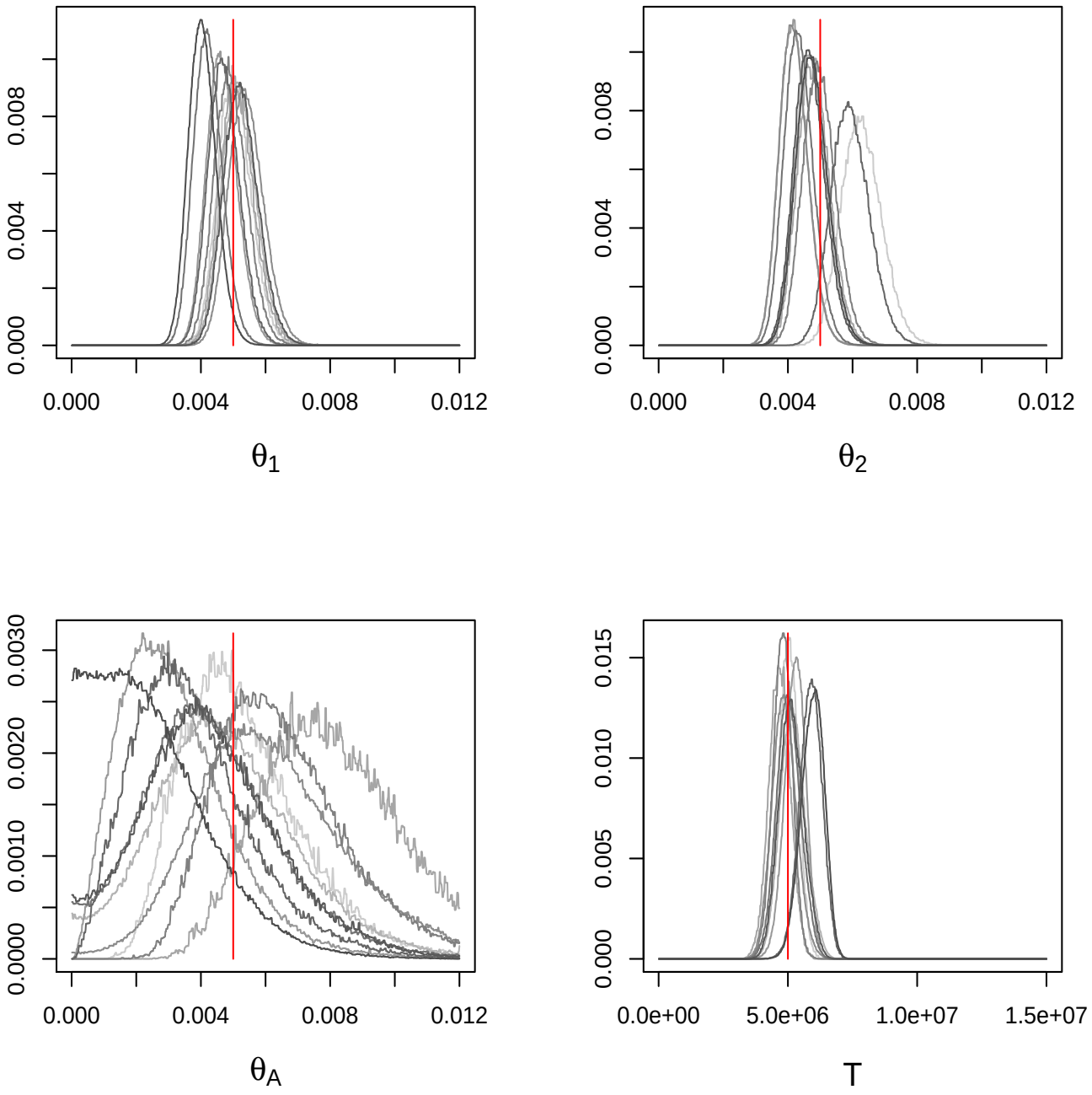
$$
\begin{aligned}
\int \pi(\mathbf{G}, \mathbf{\Theta}) d_{\mathbf{G}} &\propto \int \prod_{y=1}^{Y} \left(\frac{1}{X} \sum_{x=1}^{X} p(\mathbf{D_y}|G_{yx}, \mathbf{\Theta}) \right) p(\mathbf{\Theta}) p(\mathbf{G}|\mathbf{\Theta}) d_{\mathbf{G}} \\
&\quad \text{(realizing that the integration is over independent } G_{yx}\text{)} \\
&= \prod_{y=1}^{Y} \left(\frac{1}{X} \sum_{x=1}^{X} \int p(\mathbf{D_y}|G_{yx}, \mathbf{\Theta}) p(G_{yx}|\mathbf{\Theta}) d_{G_{yx}} \right) p(\mathbf{\Theta}) \\
&= \prod_{y=1}^{Y} \left(\frac{1}{X} \sum_{x=1}^{X} p(\mathbf{D_y}|\mathbf{\Theta}) \right) p(\mathbf{\Theta}) \\
&= p(\mathbf{D}|\mathbf{\Theta}) p(\mathbf{\Theta}).
\end{aligned}
$$

Supplementary Figures



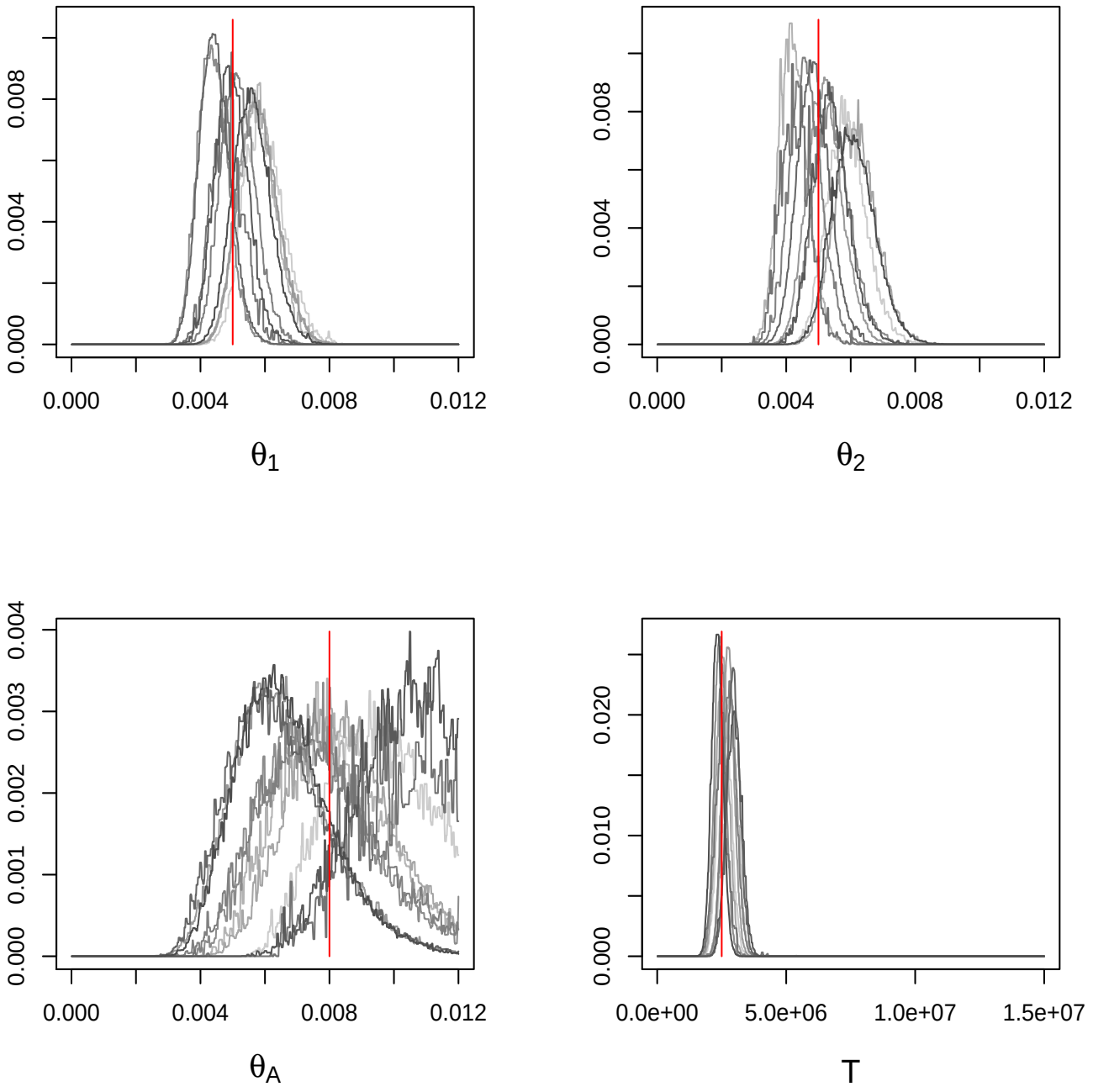
(a)

Figure S1: **Smoothed posterior distributions estimated by `MIMAR` from simulated data sets** (averaging over two independent runs). Each panel (a−c) shows results for 10 simulated data sets consisting of 20 loci surveyed at 20 chromosomes from each population, with no intra-locus recombination or migration and $\theta_1 = \theta_2 = .005$ per base pair, a) with $\theta_A = .005$ per base pair and $T = 2.5 \times 10^6$ generations, b) with $\theta_A = .005$ and $T = 5 \times 10^6$ and c) with $\theta_A = .008$ and $T = 2.5 \times 10^6$. The range of the x-axis corresponds to the support of the prior and the red vertical line denotes the true value of the parameter (see Methods for details).



(b)

Figure S1 – continued: Results for 10 data sets simulated with no intra-locus recombination or migration, $\theta_1 = \theta_2 = \theta_A = .005$ per base pair and $T = 5 \times 10^6$ generations.



(c)

Figure S1 − continued: Results for 10 data sets simulated with no intra-locus recombination or migration, $\theta_1 = \theta_2 = .005$ and $\theta_A = .008$ per base pair and $T = 2.5 \times 10^6$ generations.

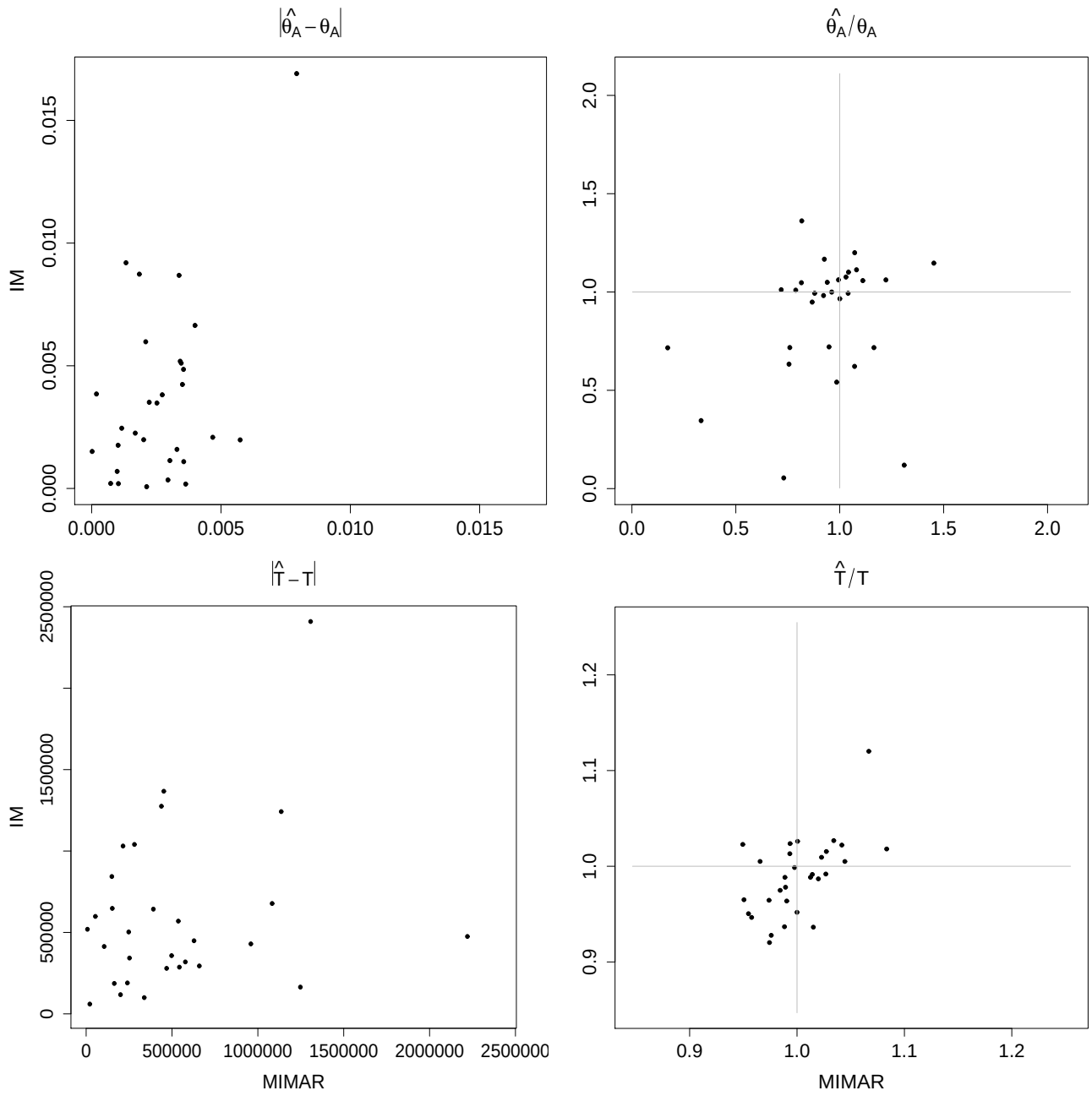
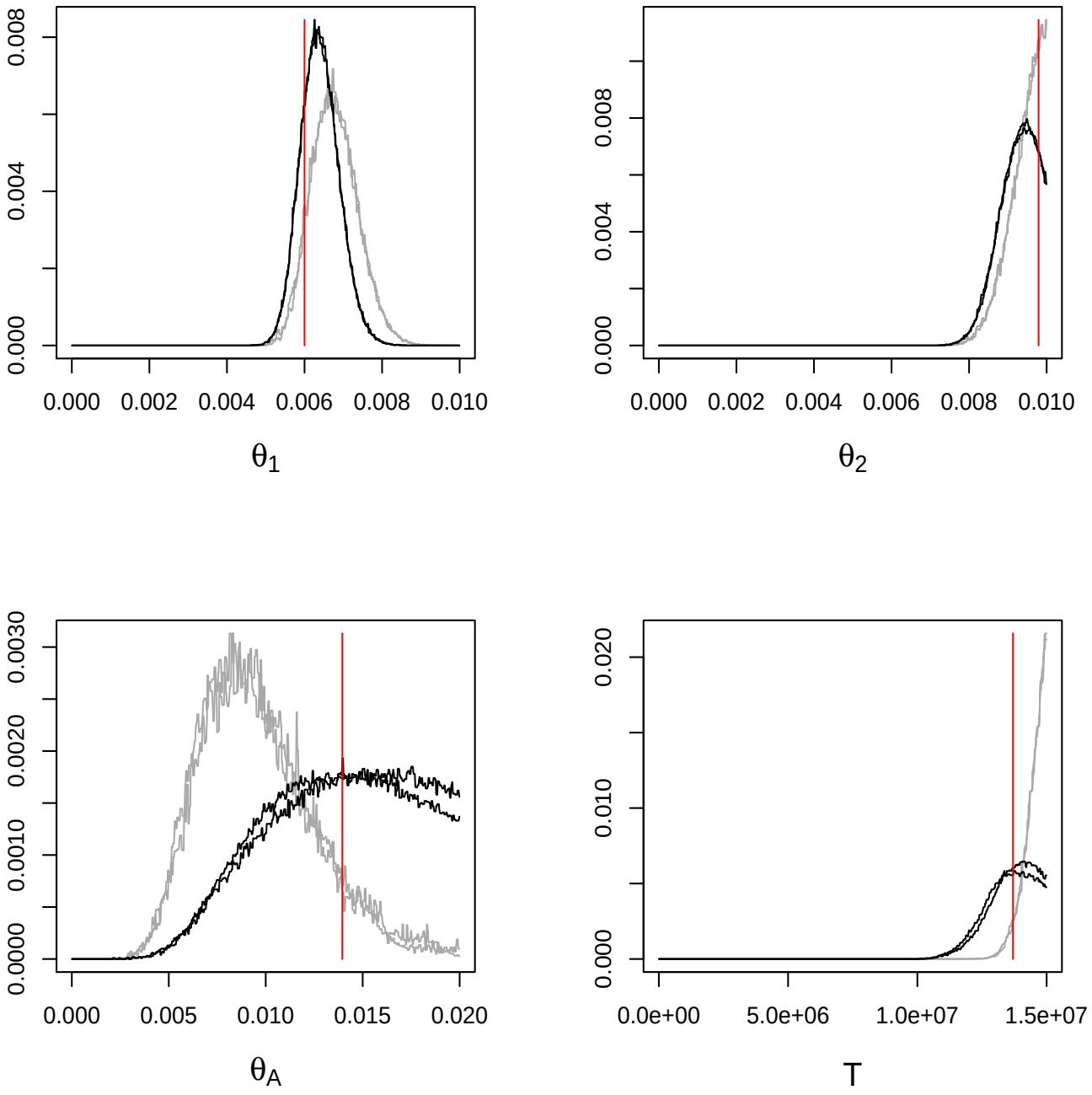
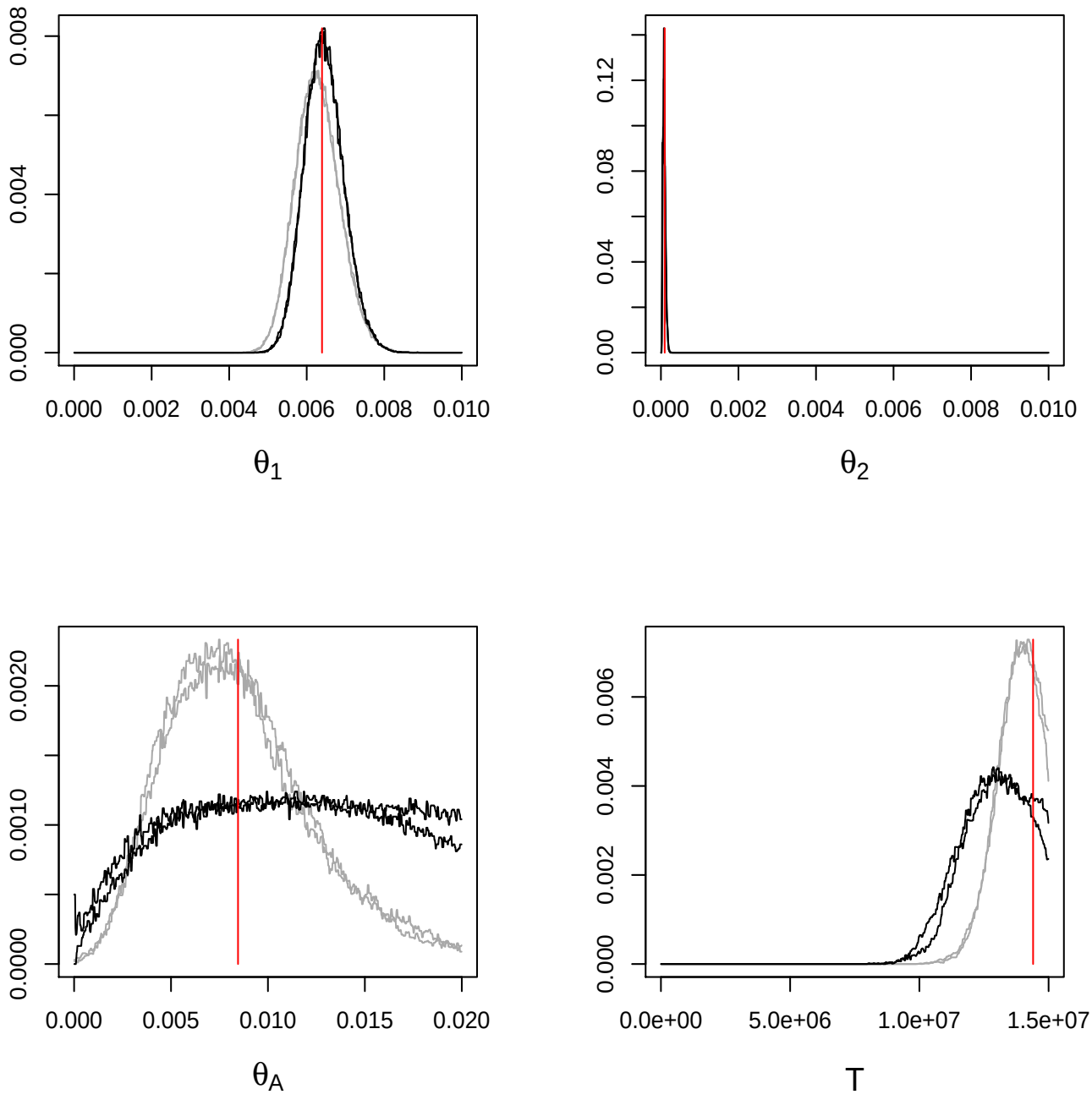


Figure S2: **Performance of `MIMAR` (x-axis) and IM (y-axis)**. Precision and accuracy of the estimates of θ_A (upper panel) and T (lower panel) obtained from the mode of the marginal posterior distribution samples. Each data point is the result from one of 30 data sets simulated under the allopatric model, with parameter values sampled from prior distributions (see Methods for details). If both methods provided estimates with perfect precision, all the data points in the left panels would be located at (0,0). If one method provided systematically more precise estimates, the data points would align along its axis: x-axis for `MIMAR` or y-axis for IM. Similarly, if both methods provided estimates with perfect accuracy, all the data points in the right panels would be at the intersection of the two grey lines (1,1). If one method provided more precise estimates, the data points would align along the vertical grey line for `MIMAR` or the horizontal grey line for IM. As can be seen, both methods perform similarly well (see Table 1; p-value from Wilcoxon signed rank test $> 5\%$, after Bonferroni correction for multiple tests).



(a)

Figure S3: **Smoothed posterior distributions estimated by IM (black) and `MIMAR` (grey) from simulated data sets** (results from two independent runs are shown). Both methods ran for the same number of steps and were smoothed similarly. a) and b) are the results for two data sets chosen from the 30 simulated to show a case in which IM (a) or `MIMAR` (b) performed better (see Fig. S2). The range of the x-axis corresponds to the support of the prior and the red vertical line denotes the true value of the parameter (see Methods for details).



(b)

Figure S3 – continued. A case in which `MIMAR` performed better.

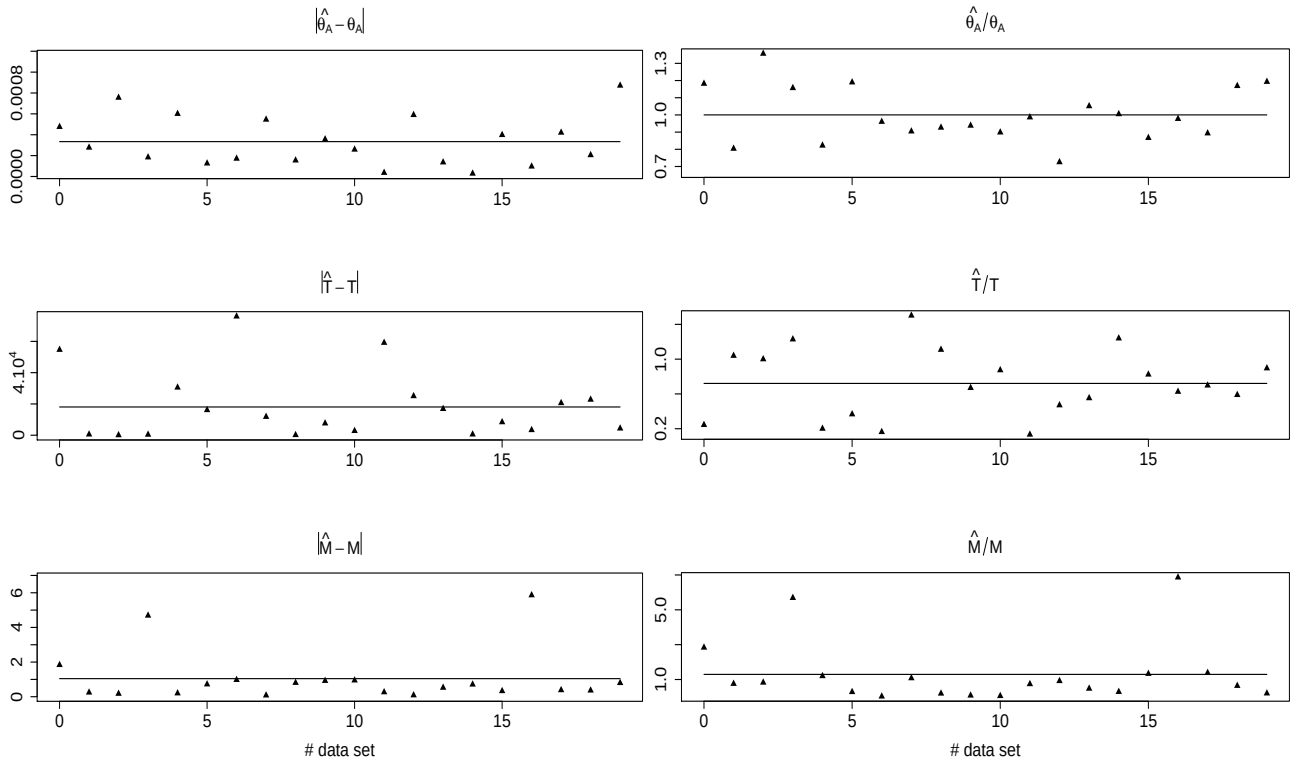


Figure S4: **Performance of MIMAR in the presence of gene flow**. Precision and accuracy of the estimates of θ_A (upper panel), T (middle panel) and M (lower panel) obtained from the mode of the marginal posterior distribution samples. Each data point is the result from one of 20 data sets (randomly numbered along the x-axis) simulated under the parapatric model with intra-locus recombination, with parameter values sampled from prior distributions. The black lines are the means over the 20 simulated data sets (see Table 2 and Methods for details).

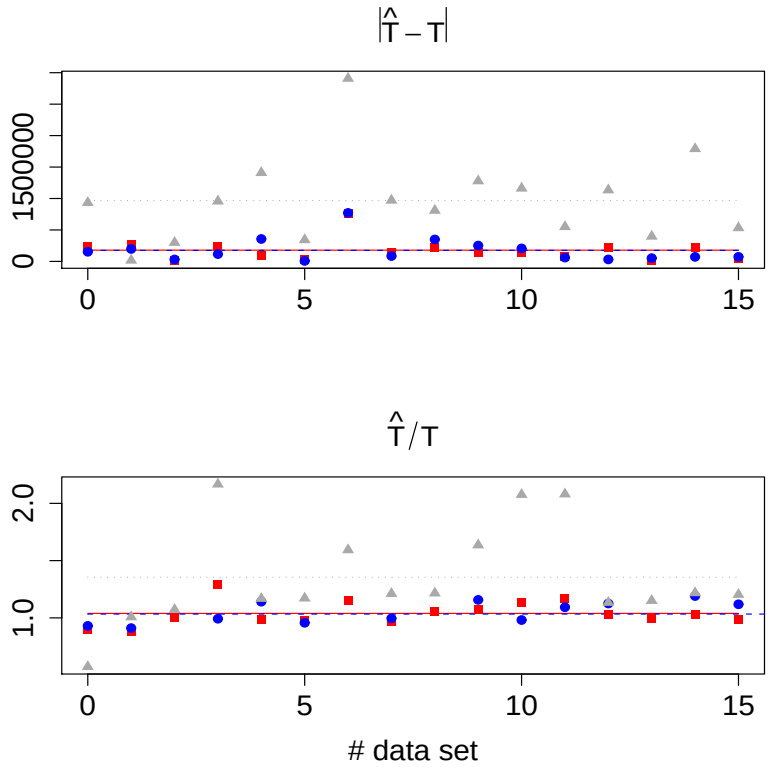
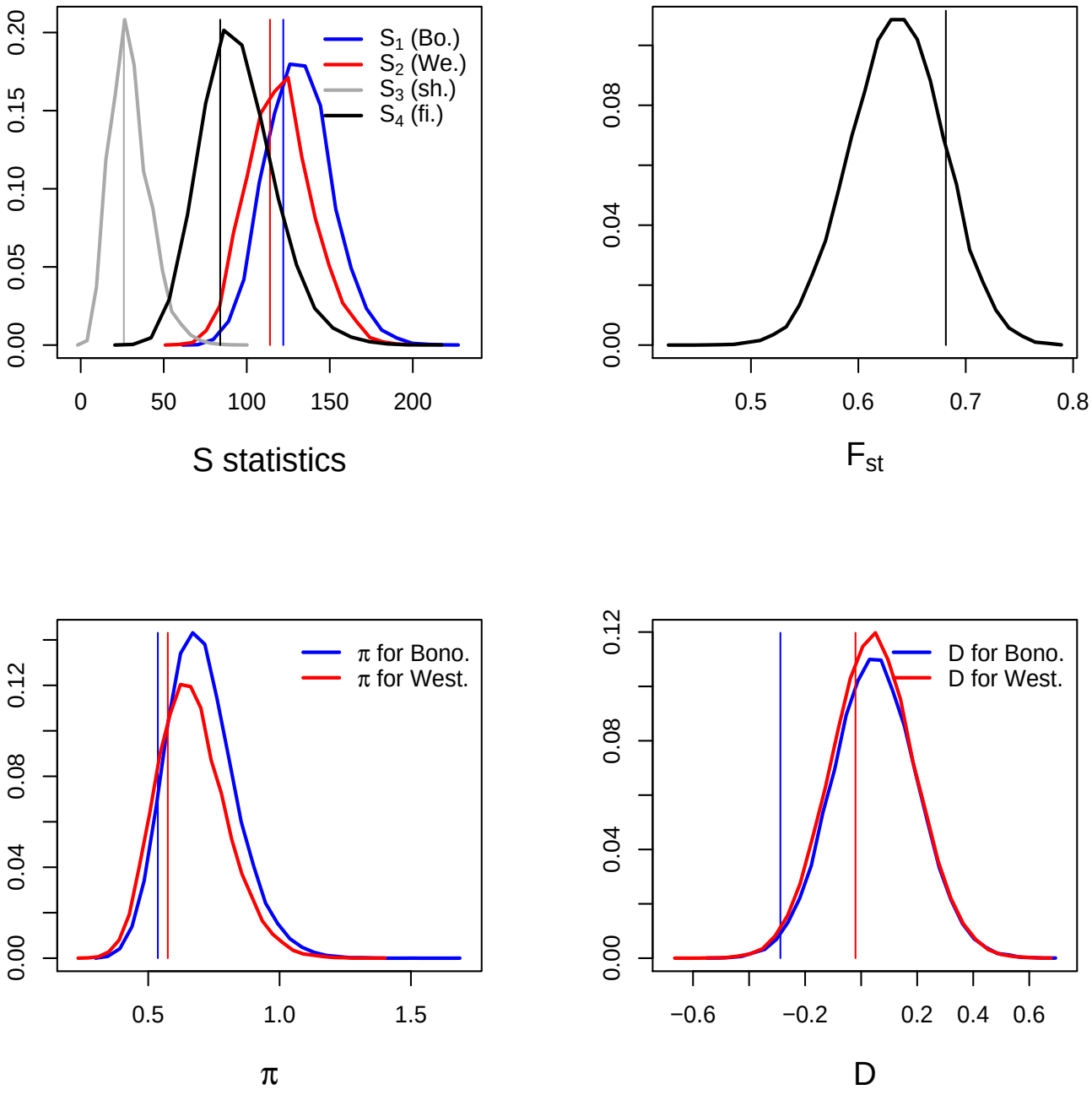
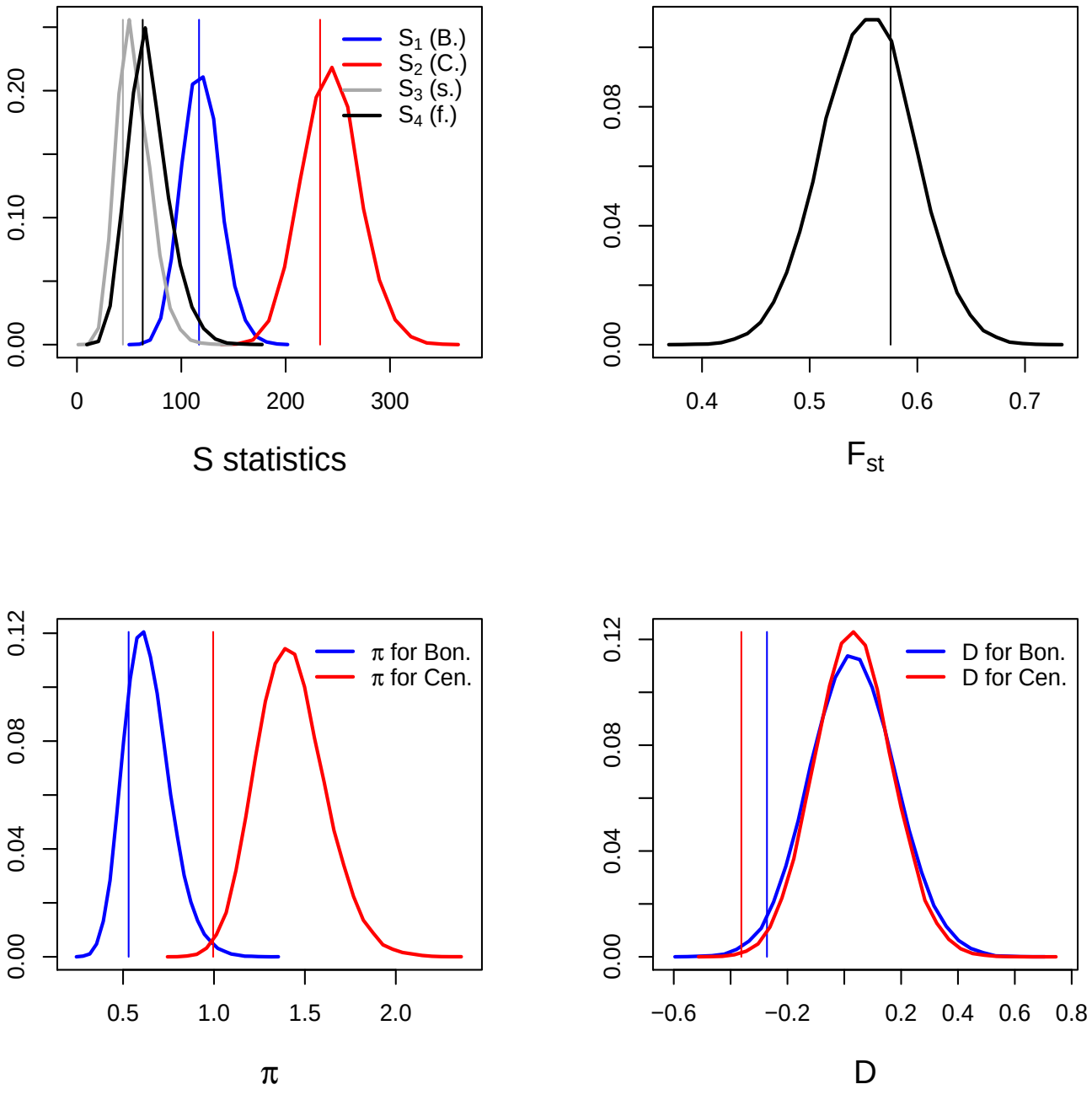


Figure S5: **Sensitivity of `MIMAR` to intra-locus recombination**. Precision (upper panel) and accuracy (lower panel) of the estimate of T obtained from the mode of the marginal posterior distribution samples. Each data point is the result from one of 16 data sets (randomly numbered along the x-axis) simulated under the allopatric model with intra-locus recombination and parameter values sampled from prior distributions. These data sets were analyzed by fixing all the parameters but T to their true values, and i) fixing the locus-specific recombination rates to their true value (red squares), ii) sampling the recombination rates from the same prior as used when generating the simulated data (blue discs) and iii) setting the intralocus recombination rates to 0 (grey triangles). The lines are the means over the 16 simulated data sets (see Methods for details). The precision and bias in the estimates of T from iii) and i) (or ii)) are significantly different at the 5% level (p=6×10^{-5} and p=0.002, respectively, using paired Wilcoxon signed rank tests).



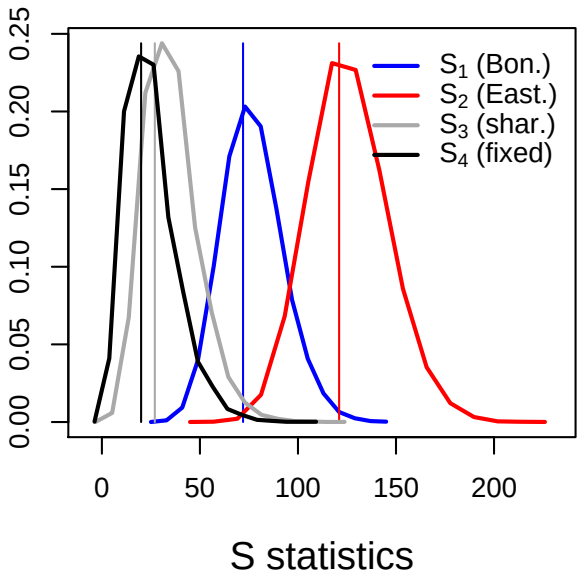
(a)

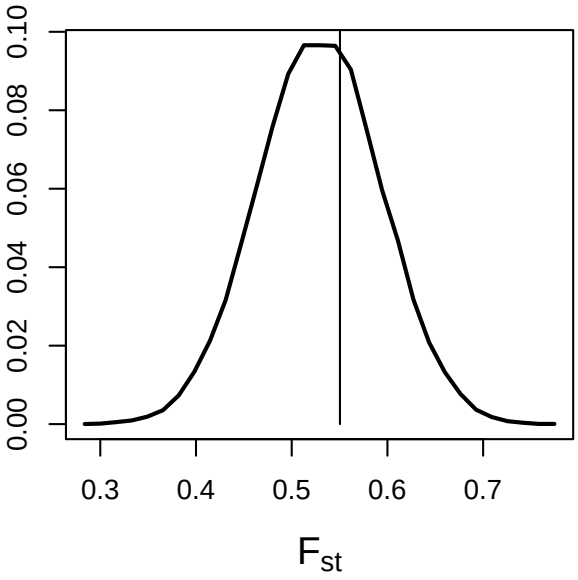
Figure S6: **Goodness of fit of the isolation-migration model for the ape species and subspecies data**. Distributions of the four statistics used by `MIMAR` (the polymorphisms specific to the first and second sample and shared and fixed between the two samples, S_1, S_2, S_3 and S_4, respectively), as well as F_{st}, π and Tajima's D (see Methods for details). The vertical lines correspond to the observed values. Shown are results for a) bonobos (blue) and Western chimpanzees (red), b) bonobos and Central chimpanzees, c) bonobos and Eastern chimpanzees, d) Western and Central chimpanzees, e) Western and Eastern chimpanzees, f) Central and Eastern chimpanzees, g) Western and Eastern gorillas and h) Sumatran and Bornean orangutans. In this case, the estimated model does not seem to provide a good fit to D for bonobos.

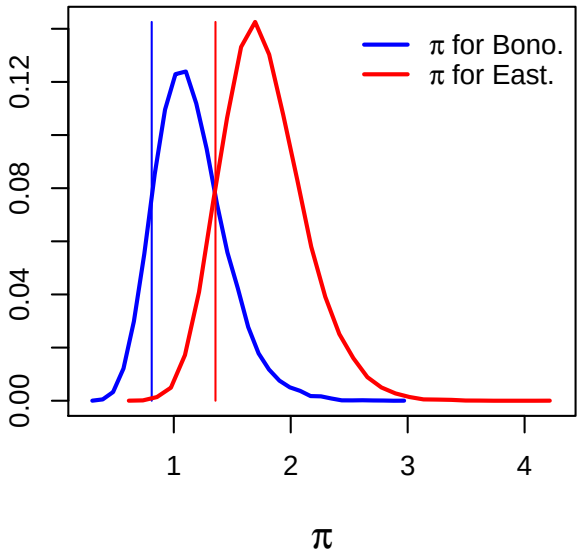


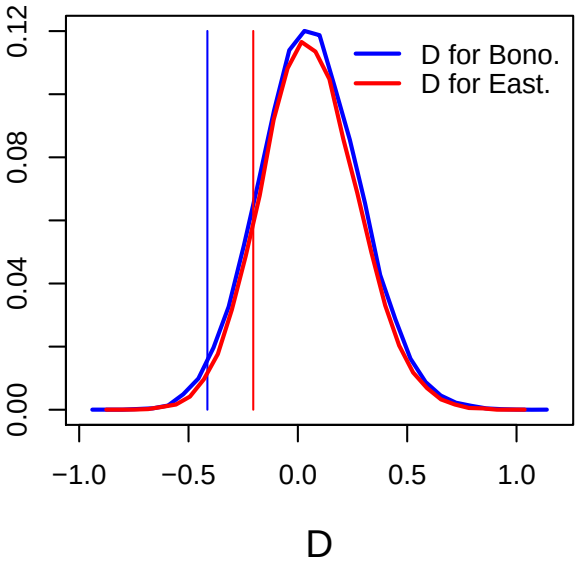
(b)

Figure S6 – continued: Results for bonobos (blue) and Central chimpanzees (red). In this case, the estimated model does not seem to provide a good fit to D for both bonobos and Central chimpanzees and as well as to π for Central chimpanzees.



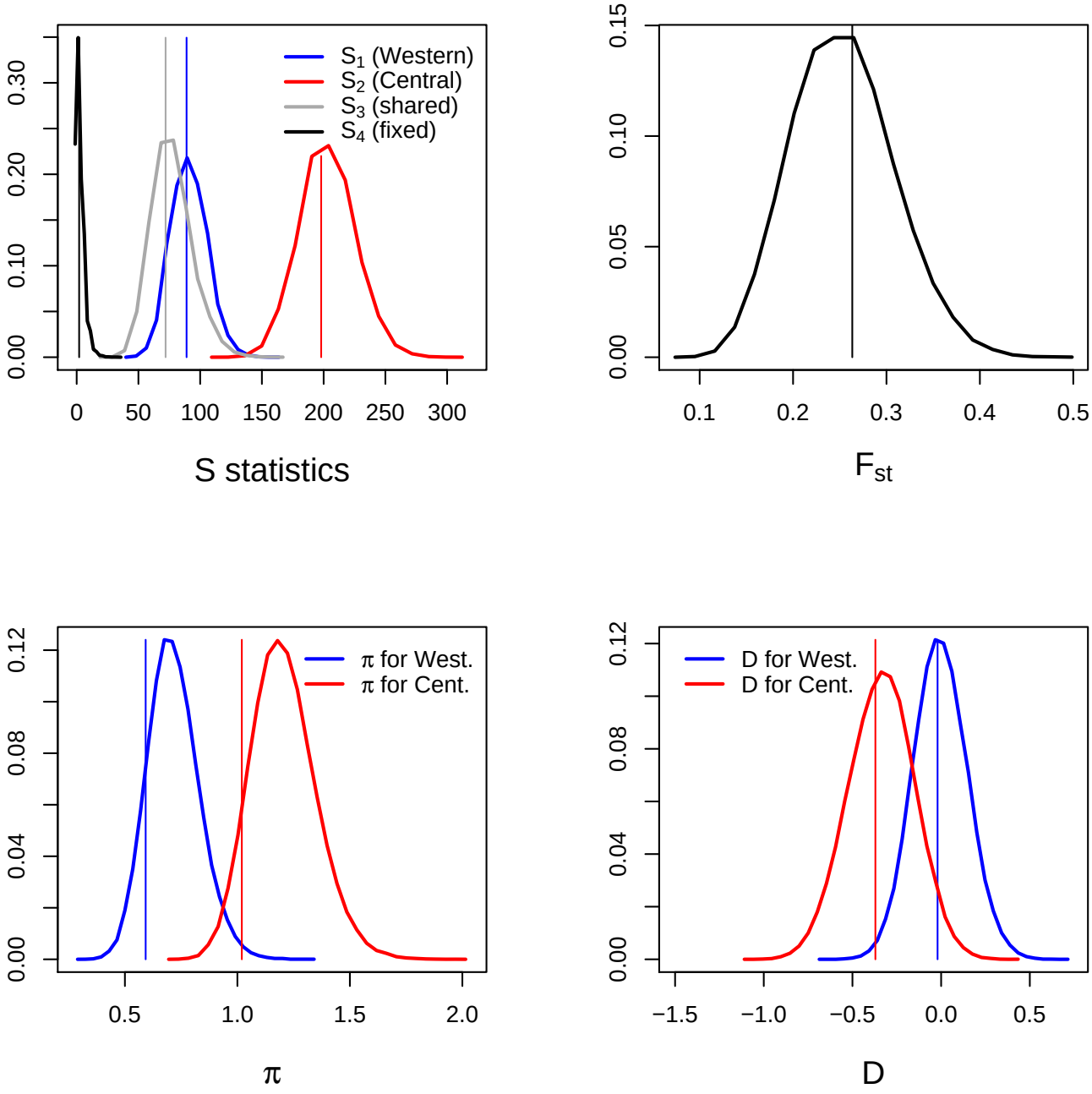






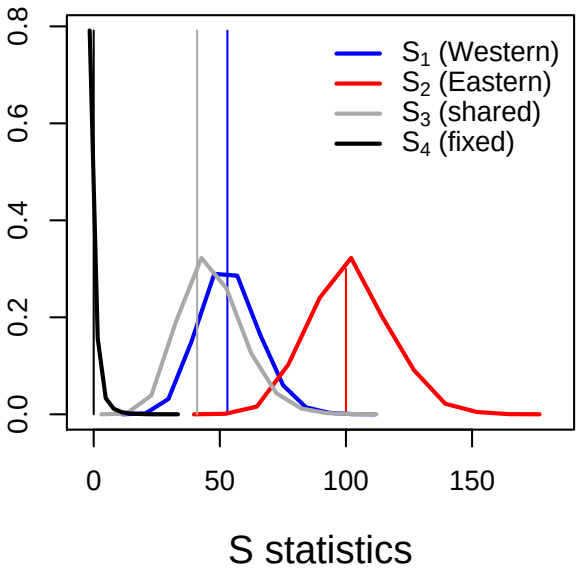
(c)

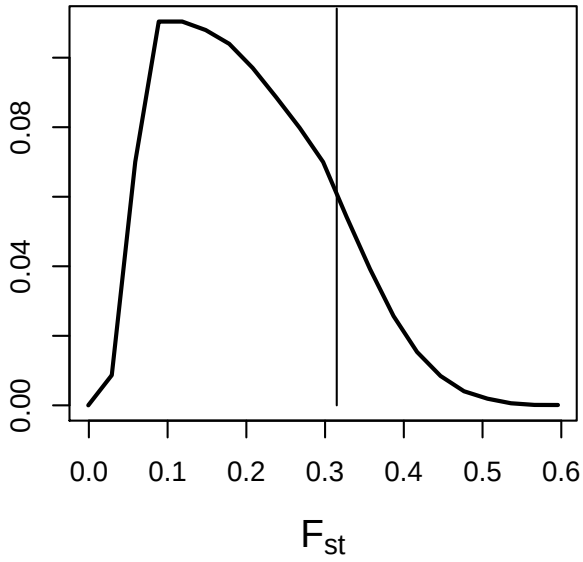
Figure S6 – continued: Results for bonobos (blue) and Eastern chimpanzees (red). In this case, the estimated model does not seem to provide a good fit to D for bonobos.

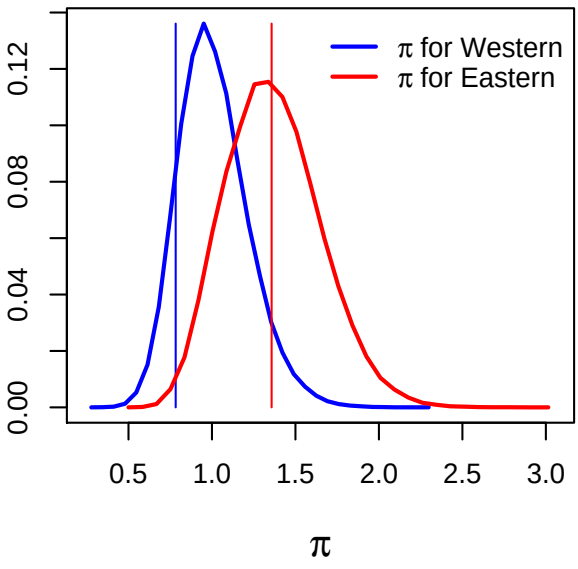


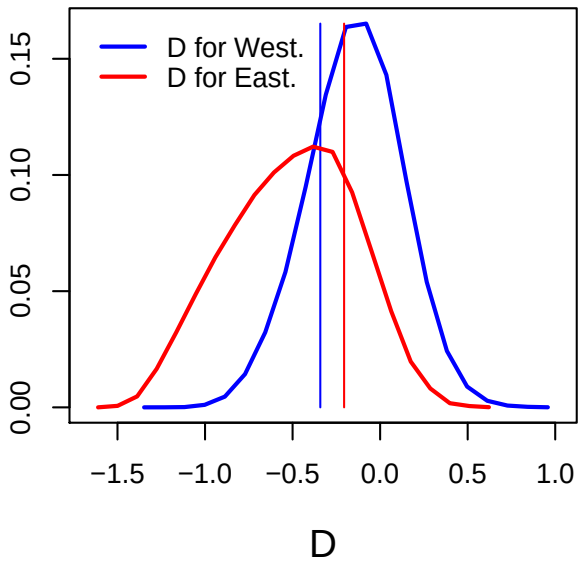
(d)

Figure S6 – continued: Results for Western (blue) and Central chimpanzees (red).



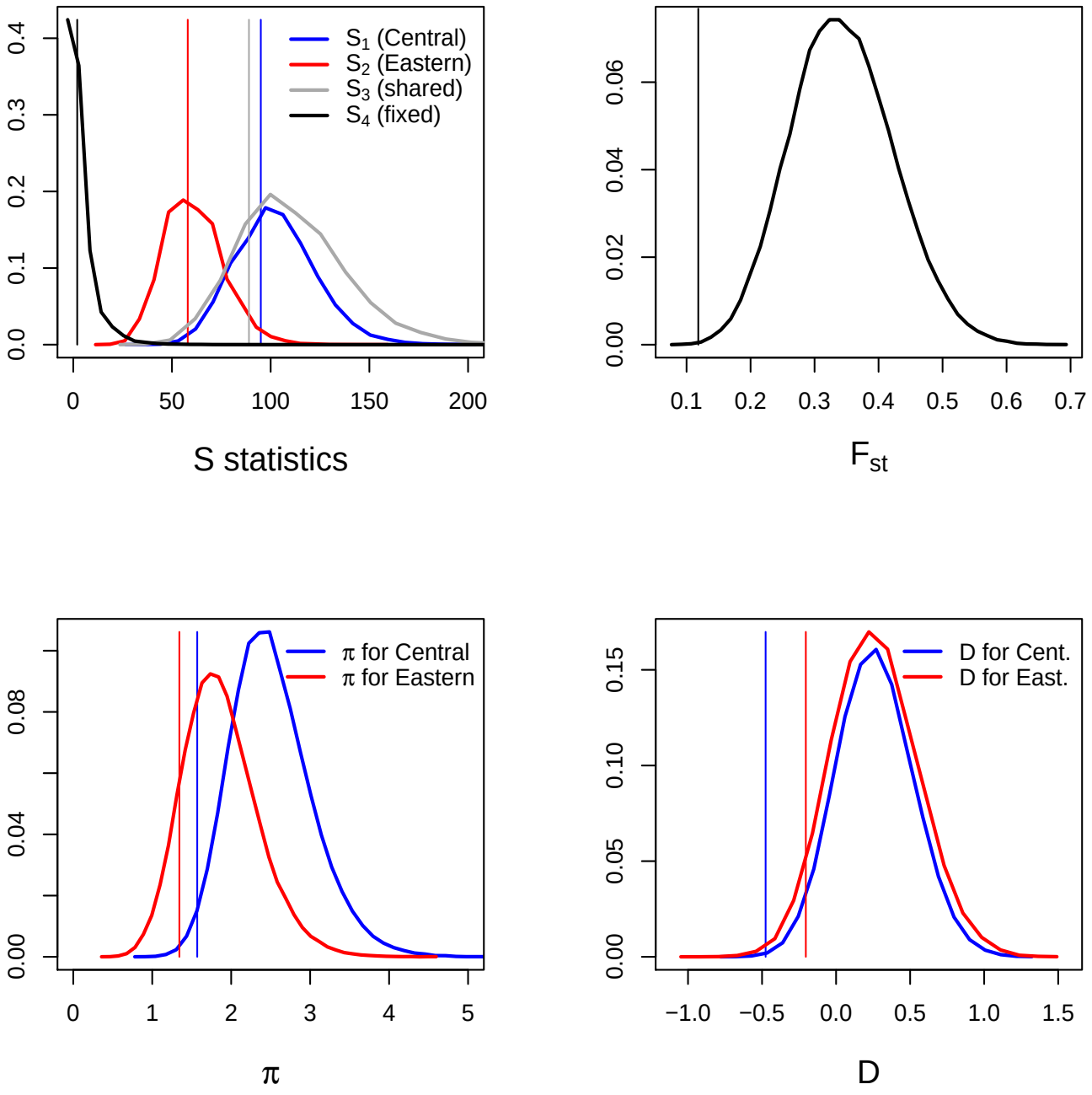






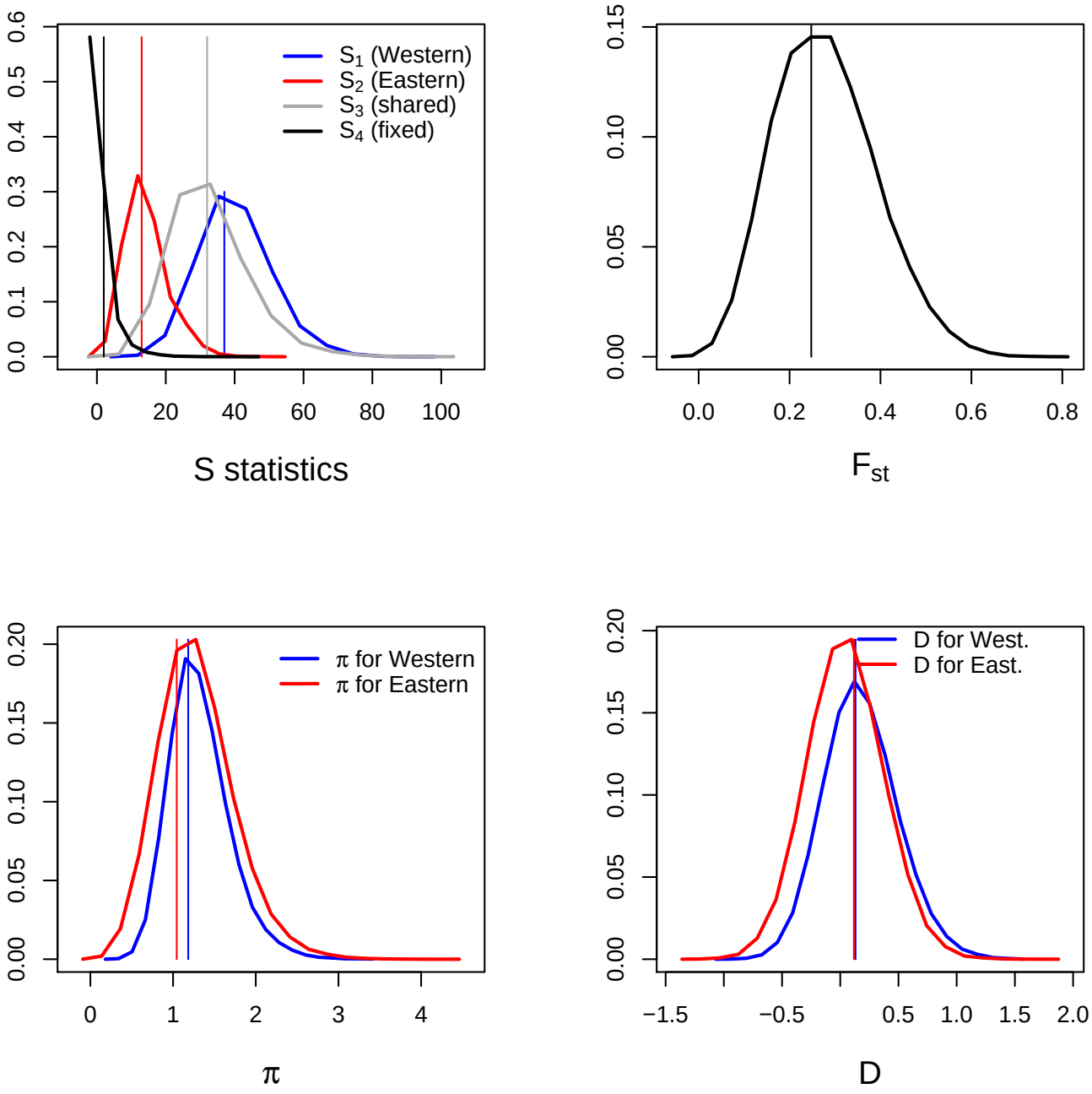
(e)

Figure S6 – continued: Results for Western (blue) and Eastern chimpanzees (red).



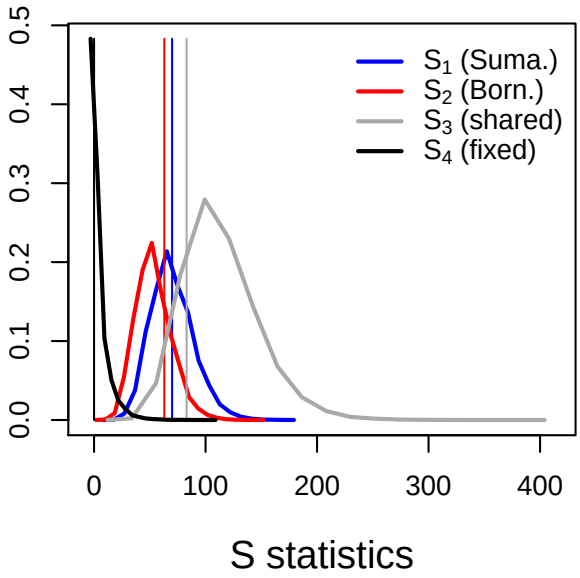
(f)

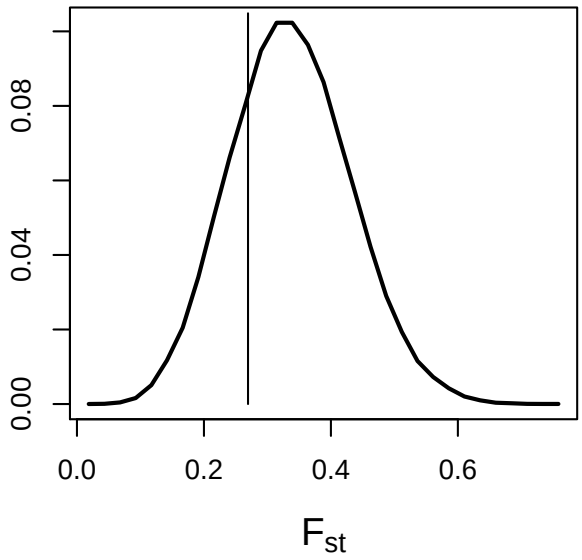
Figure S6 – continued: Results for Central (blue) and Eastern chimpanzees (red). In this case, the estimated model appears to provide a poor fit to the observed F_{ST}, and D for Central chimpanzees.

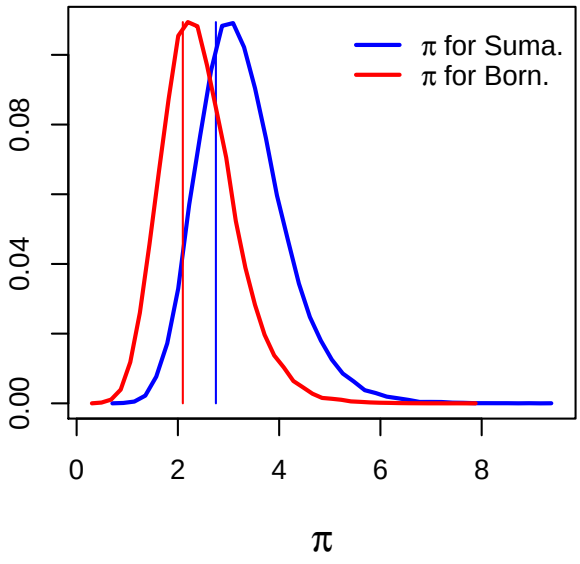


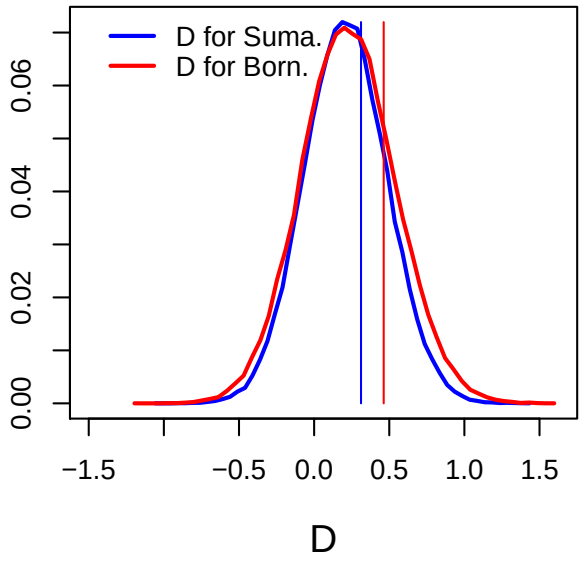
(g)

Figure S6 – continued: Results for Western (blue) and Eastern (red) gorillas.









(h)

Figure S6 – continued: Results for Sumatran (blue) and Bornean (red) orangutans.