# Supplementary material for A RESPONSE TO MAYROSE ET AL. (2014) AND THE FATE OF POLYPLOID LINEAGES

Graham Jones

February 4, 2015

Figures S1 and S2 are based on 200 simulated trees with similar characteristics to those analyzed in Mayrose et al (2011). Table 1 shows summaries of various estimates from these trees.

| | Min. | Mean | Std. dev. | Median | Max. |
|---|---|---|---|---|---|
| posterior mean of $\lambda_D$ | 0.0698 | 0.2016 | 0.0447 | 0.2009 | 0.3468 |
| MLE of $\lambda_D$ | 0 | 0.214 | 0.103 | 0.1893 | 0.8542 |
| posterior mean of $\lambda_P$ | 0.0526 | 0.1899 | 0.0638 | 0.1892 | 0.4135 |
| MLE of $\lambda_P$ | 0 | 0.1876 | 0.1427 | 0.1579 | 0.7812 |
| posterior mean of $\mu_D$ | 0.032 | 0.0968 | 0.0397 | 0.0905 | 0.251 |
| MLE of $\mu_D$ | 0 | 0.1182 | 0.1469 | 0.072 | 0.8741 |
| posterior mean of $\mu_P$ | 0.0277 | 0.1779 | 0.0936 | 0.1601 | 0.5755 |
| MLE of $\mu_P$ | 0 | 0.2108 | 0.3619 | 0.094 | 2.6906 |
| posterior mean of $q_{DP}$ | 0.0098 | 0.0406 | 0.019 | 0.0368 | 0.1216 |
| MLE of $q_{DP}$ | 0 | 0.0336 | 0.0335 | 0.0236 | 0.2158 |
| posterior mean of $r_D - r_P$ | -0.1849 | 0.0929 | 0.1312 | 0.062 | 0.5754 |
| MLE of $r_D - r_P$ | -0.77 | 0.119 | 0.3801 | 0.0414 | 2.9324 |

Table 1: Summaries of posterior means and maximum likelihood estimates of the five parameters and the difference between divergence rates. The values are rounded to 4 digits.

Figure S3 is based on a set of 280 larger trees. The maximum time for the trees to evolve was doubled from 26 to 52, and the transition rate $q_{DP}$ was halved to 0.01, to keep the the proportion of polyploid tips roughly 1/3.

## Two simpler examples of bias

These examples may give some insight into why BiSSE is biased. They both both concern estimates of a single parameter, from samples of different sizes. It seems likely that in BiSSE that there is also some interaction between $\mu_P$ and $q_{DP}$ which is not captured by these examples.

### Example 1

Here we make some greatly simplifying assumptions about the diversification process and the nature of the data, so that an approximate result can be found without the need for a program. The mathematical results used here are very old and appear in many textbooks; Gernhard (2008) is a recent reference.

Suppose all extinction rates are zero, in other words the Yule model applies. Suppose there are two types of species, 'red' and 'blue'. For each, we have a collection of observations $(n_i, t_i)$, where $n_i$ is a clade size, and $t_i$ is the time of origin of the $i$th clade. (The origin time is when the clade split off from the rest of the tree of life, and so consisted of just one species.) The birth rates $\lambda_i$ are all

unknown. Finally, assume that all the $t_i$ for red species are 1My, and all the $t_i$ for blue species are very big.

We will now see what happens if the true value of all the $\lambda_i$ is in fact $0.1\text{My}^{-1}$ and we estimate the values using maximum likelihood estimation (MLE). Under the Yule model, the probability that $n$ species are present after time $t$ given birth rate $\lambda$ is $e^{-\lambda t}(1-e^{-\lambda t})^{n-1}$. So for the red species, the probability of observing a clade of size one is $e^{-0.1} = 0.905$, for a size two clade it is $e^{-0.1}(1-e^{0.1}) = 0.086$, for a size three clade it is $e^{-0.1}(1-e^{0.1})^2 = 0.008$ and larger clade sizes are even less likely. The MLE of $\lambda_i$ is $\log(n_i)/t_i$. So, for a large collection of red clades, the mean of our estimates will about $0.905 \times 0 + 0.086 \times \log(2) + 0.008 \times \log(3) \approx 0.07$. Since the blue clades are all very old, they will almost all be large, the MLE will work well, and our estimate for them will be close to 0.1. If we didn't understand what was going on, we might conclude that blue species diversify faster than red species.

The following R code calculates the mean MLE over a large collection of clades for several origin times.

```
prob.clade <- function(n, t) {
  exp(-lambda*t) * (1 - exp(-lambda*t)) ^ (n-1)
}

mean.MLE <- function(t) {
  ns <- 1:100000
  stopifnot(sum(prob.clade(ns,t)) > .999999999) # check we are summing far enough
  sum(prob.clade(ns,t) * log(ns) ) / t
}

lambda <- 0.1
for (t in c(1,5,10,20,40,80)) {
  cat(paste0("Origin time ", sprintf("%2d", t), "Mya, mean MLE of birth rate ",
             round(mean.MLE(t), digits=5), "\n"))
}
```

Here is the output. The bias diminishes very slowly with tree size. The average tree size after 80My is about $e^8 \approx 3000$.

```
Origin time  1Mya, mean MLE of birth rate 0.0699
Origin time  5Mya, mean MLE of birth rate 0.07219
Origin time 10Mya, mean MLE of birth rate 0.07491
Origin time 20Mya, mean MLE of birth rate 0.07971
Origin time 40Mya, mean MLE of birth rate 0.08656
Origin time 80Mya, mean MLE of birth rate 0.0928
```

## Example 2

Consider sampling from the density $f(x;u) = (1-u)e^{-(1-u)x}$, where the parameter $u$ is known to be in $[0,1]$. This is just an exponential density with a non-standard parameterization, and the rate restricted to be below 1. If the true value of $u$ is near zero, it will be difficult to get a good estimate of it: for example, $f(x;0.1) = 0.9e^{-0.9x}$ is similar to $f(x;0.2) = 0.8e^{-0.8x}$ and it will be hard to distinguish samples from them.

Now suppose we have two random samples, namely $X = X_1, \ldots, X_m$ from the density $f(x;v)$, and $Y = Y_1, \ldots, Y_n$ from the density $f(x;w)$. We would like to test whether $v > w$. The likelihood functions are

$$(1-v)^m \exp(-m(1-v)\bar{X}) \quad \text{and} \quad (1-w)^n \exp(-n(1-w)\bar{Y})$$

where $\bar{X}$ and $\bar{Y}$ are the sample means. If we assume uniform priors for $v$ and $w$ over $[0,1]$, we can estimate $v$ and $w$ as the posterior mean in the usual way:

$$\hat{v} = \frac{\int_0^1 v(1-v)^m \exp(-m(1-v)\bar{X})dv}{\int_0^1 (1-v)^m \exp(-m(1-v)\bar{X})dv}$$

with a similar expression for $\hat{w}$.

Suppose $m = 10$ and $n = 20$. These are quite small sample sizes, so estimates may be poor, but suppose we can repeat the experiment many times. The true values of $v$ and $w$ may be different for each experiment, but the distribution of $(\hat{v} - \hat{w})$ over many experiments may give us what we want. Or maybe not, as the R script `simple-analogy.r` shows. In the code, `X.ssize` is $m$, `Y.ssize` is $n$, and the true value of both $v$ and $w$ is 0.2 for all experiments. The only asymmetry is that $n$ and $m$ are different. There are `N`=63 experiments. Using a t-test on the set of 63 values $(\hat{v} - \hat{w})$, a p-value is found. The whole thing is then repeated `M`=100 times. With these settings, the p-value is less than 0.05 about 2/3 of the time.

There is a loose analogy in which $m$ is a bit like the number of polyploids, $n$ is a bit like the number of diploids, and $v$ and $w$ are a bit like extinction rates.

# References

T Gernhard (2008) The conditioned reconstructed process. J. Theo. Biol. **253**, 769–778.

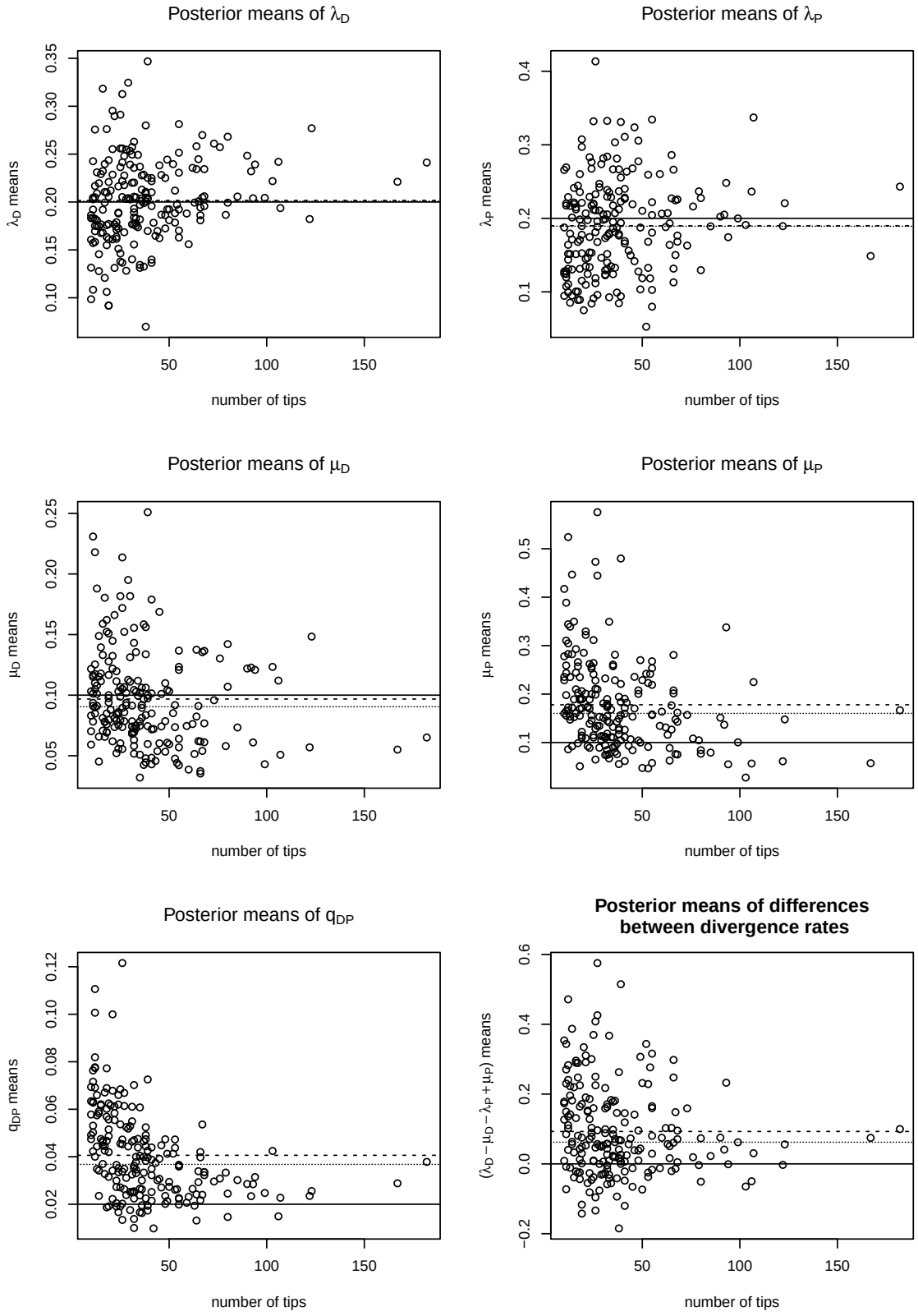


Figure S1: Posterior means of the parameters and the differences between divergence rates. The 6 graphs show results for the same 200 simulated trees. These trees are similar to those of Mayrose et al. (2011). The solid lines show true values, the dashed lines show the means, and the dotted lines show medians.

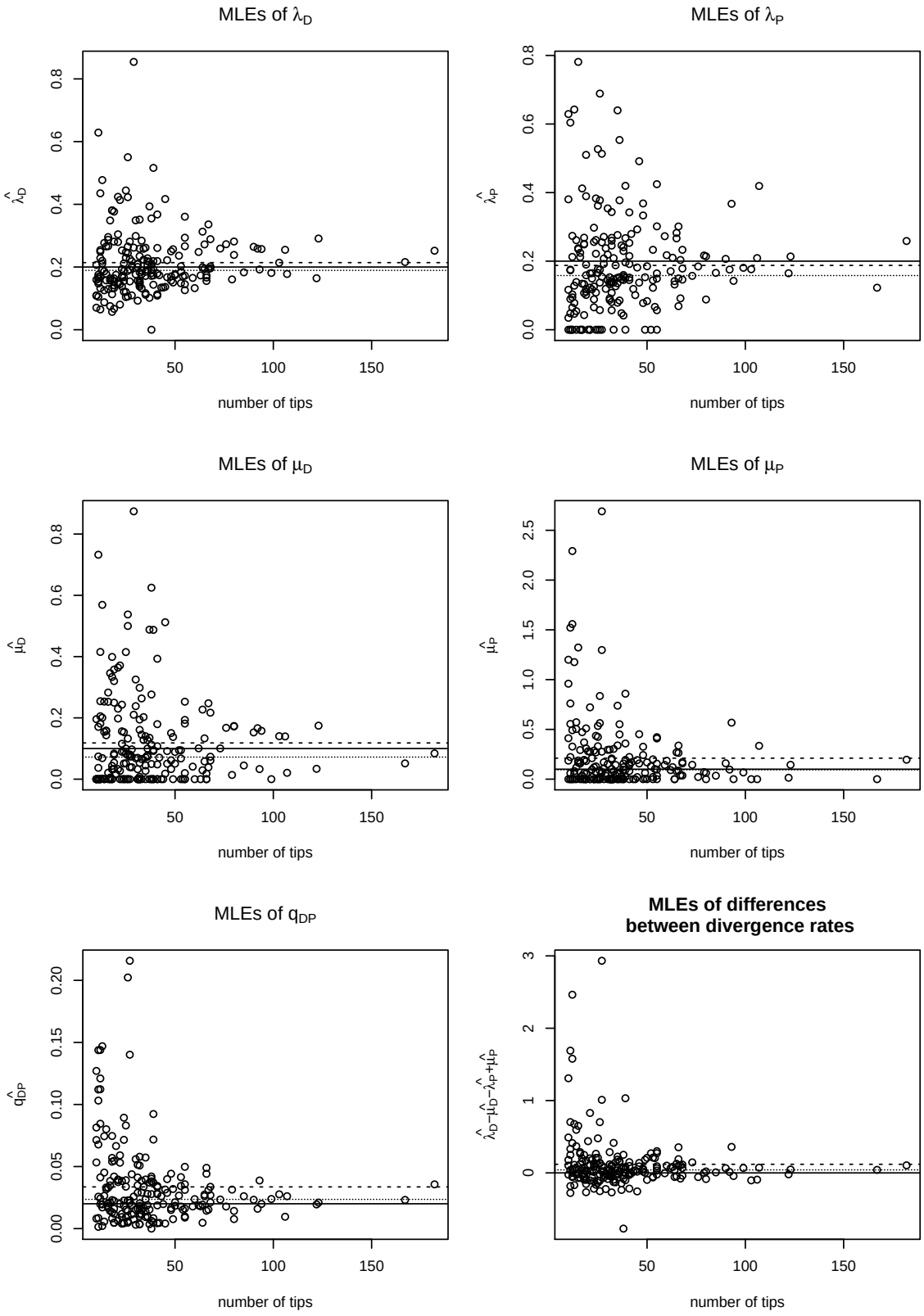


Figure S2: Maximum likelihood estimates of the parameters and the differences between divergence rates. Other details as figure S1. Note the large scale of some y-axes.

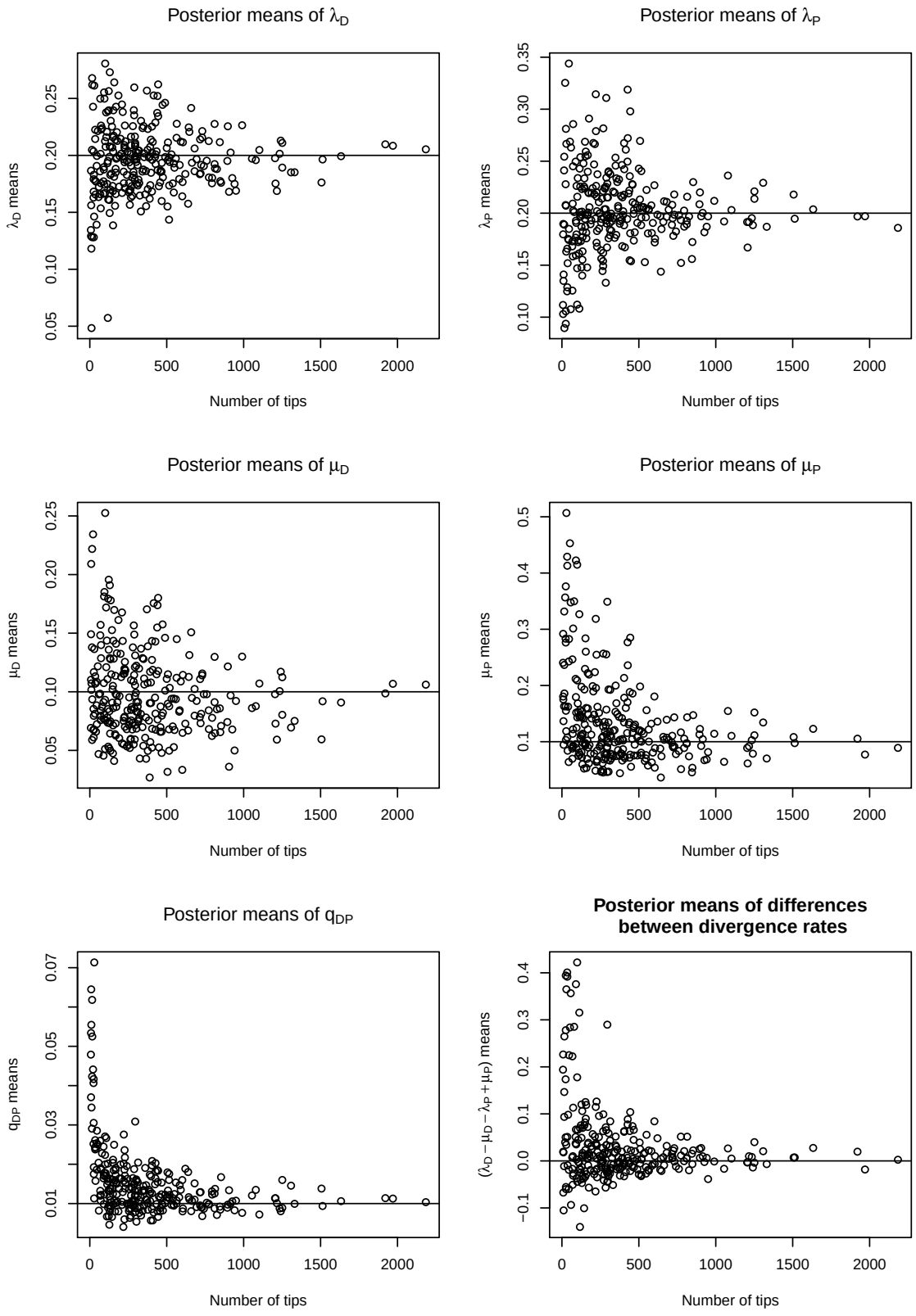


Figure S3: Posterior means of the parameters and the differences between divergence rates, for a set of 280 larger trees. The solid lines show true values.