The fate of polyploid lineages:

A response to Mayrose et al. (2014)

Douglas E. Soltis[2,3,4], María Claudia Segovia-Salcedo[2], Ingrid Jordon-Thaden[2,5], Lucas C. Majure[2,6], Nicolas M. Miles[3], Evgeny V. Mavrodiev[3], Wenbin Mei[2], Mark E. Mort[7], Pamela S. Soltis[3,4], Graham R. Jones[8], Thomas Marcussen[9], Bengt Oxelman[8], and Matthew A. Gitzendanner[2,4]

[2]Department of Biology, University of Florida, Gainesville, Florida, 32611 USA

[3]Florida Museum of Natural History, University of Florida, Gainesville, Florida, 32611 USA

[4]Genetics Institute, University of Florida, Gainesville, Florida, 32611 USA

[5]Department of Biology, Bucknell University, Lewisburg, PA 17837 USA

[6] Desert Botanical Garden, Phoenix, AZ 85008 USA

[7]Department of Ecology and Evolutionary Biology & Biodiversity Institute, University of Kansas, Lawrence, KS 66045 USA

[8]Department of Biological and Environmental Sciences. University of Gothenburg, Box 461, SE 405 30 Göteborg, Sweden

[9]Monrads Gate 21a, NO-0564 Oslo, Norway

[1] We thank Mike Moore for helpful discussion. This work was supported by DEB-1146065 and DEB-0922003.

"The report of my death was an exaggeration."

Mark Twain, *New York Journal*, June 1897

INTRODUCTION

Mayrose et al. (2011) and Arrigo and Barker (2012) concluded that young polyploid lineages have higher extinction rates than diploid lineages. Arrigo & Barker (2012, p. 140) refer to 'rarely successful' polyploid lineages and state that "…most [polyploids] are evolutionary dead-ends". We have already written at length regarding our concerns on the general approach and conclusions of Mayrose et al. (2011; see Soltis et al. 2014a), and we will minimize repeating them here. Our goal is to respond as succinctly as possible to the recent reply by Mayrose et al. (2014), and to address some additional, fundamental problems with the approach used by Mayrose et al. (2011, 2014). We stress again that the conclusions of Mayrose et al. (2011, 2014) and Arrigo and Barker (2012) regarding the fate of polyploid lineages were premature (or perhaps an "exaggeration" in the words of Mark Twain). In the spirit of cooperation typical of this research area, we also propose a research path forward.

Polyploids may well have higher extinction rates than diploids, as Mayrose et al. (2011, 2014) and Arrigo and Barker (2012) concluded. Our point is simple—these papers do not convincingly demonstrate this fate in established polyploid lineages; certainly

strong conclusions regarding the fate of polyploids (as in Arrigo and Barker, 2012, in particular) should not be made. In fact, a recent paper by this team using the same methods finds the reverse—in fish, polyploids diversify at higher rates than diploids (see below). We stress again that there are methodological as well as sampling issues with Mayrose et al. (2011), which render their conclusions unjustified. In particular, we show that their main method is biased in several ways that have not been previously emphasized.

MISSING LINKS FROM MAYROSE ET AL. (2011)

We acknowledge that Soltis et al. (2014a) misinterpreted the error in Table S2 of Mayrose et al. (2011), and we are glad that the error has now been corrected. Similarly, a bug in Mesquite (Maddison and Maddison, 2015) has now been fixed. Lastly, the details provided in Mayrose et al. (2014) make the methods of data cleansing much more transparent and allay many of the concerns about inconsistencies highlighted in Soltis et al. 2014. We thank the authors for their clarifications and corrections, and it is gratifying that both Soltis et al. (2014a) and Mayrose et al. (2014) have clarified many points. However, many of our concerns remain, and we present some further details on these and additional cautions.

SYSTEMATIC BIAS IN BiSSE

To test the approach of Mayrose et al. (2011), we simulated trees with similar characteristics to the empirical trees and then followed their method for assessing differences in divergence rates. The analysis was carried out in R (R Core Team, 2013)

using the diversitree package which implements the BiSSE model (Maddison, 2007). All of the R scripts used for this paper are in the supplementary material. In the simulations, both the diploid speciation rate λ_D and the polyploid speciation rate λ_P were set to 0.2, and the corresponding extinction rates μ_D and μ_P were both set to 0.1. The transition rate q_{DP} was set to 0.02. The trees were allowed to grow for a maximum of 26 time units. Trees with fewer than ten tips, and those with diploids only or polyploids only at the tips, were rejected. These settings were found to produce trees with a similar range of sizes to those in the empirical data set of Mayrose et al. (2011), with about one third of the tips being polyploid. The simulated trees were then analyzed using the methods of Mayrose et al. (2011), using a Bayesian approach and the Markov chain Monte Carlo (MCMC) method described in FitzJohn et al. (2009).

The main result is shown in Figure 1, which is directly comparable to Figure 1A of Mayrose et al. (2011). In our case, however, the bias towards one (high posterior probabilities for diploids having higher diversification rates) is a pure artifact. The reason for the bias is not yet fully understood. However, the situation is clearly asymmetric between diploids and polyploids: diploids can become polyploids, but not vice versa, and there are more diploids than polyploids in the set of trees. The supplementary material contains two examples of how bias can occur in similar, but simpler situations.

Mayrose et al. (2014) argue that "approaches like…[theirs] …that consider the preponderance of evidence across multiple clades is the best way to assess whether a trait, like polyploidy, affects diversification in a repeatable way." A meta-analysis of the type conducted by Mayrose et al. (2011) is indeed attractive. It can reduce the variance of individual analyses so that a useful signal can be revealed. However, it is vulnerable to a

systematic bias affecting all the individual analyses, as our results show. Used on a single tree, BiSSE is unlikely to produce misleading results. The bias will usually be overwhelmed by uncertainty. Figures S1 and S2 show that the bias diminishes with tree size, and other results with larger trees (such as Figure S3, supplementary material) confirm this. The results of meta-analyses using BiSSE are likely to be affected by bias if the true transition rates are significantly different and the tree sizes are fairly small, as in the case of the empirical data analyzed by Mayrose et al. (2011).

In support of their argument, Mayrose et al. (2014) present data from Goldberg and Igic (2008), indicating in Figure 1 that speciation rate estimates are centered near the true values, and report, but do not show, 'similar' results for extinction rates. These results appear at first sight to be in conflict with our results, so some explanation seems warranted.

In Figure 1 of Mayrose et al. (2014), there is no information about the number of terminals (species) in the trees of the Goldberg and Igic (2008) simulations. However, in all but the center and left panels of the bottom row of the figure, which have 55 and 56 terminals, respectively, the number of terminals is over 100. The genera included in Mayrose et al. (2011) have a mean of only 38.7 terminals, and over half have fewer than 30 terminals. The estimates in Figure 1 of Mayrose et al (2014) show a large uncertainty in these two panels, and it is not clear whether this hides a systematic bias. Furthermore, Mayrose et al. (2014) do not specify what measure of centrality they use. Because some of the distributions are quite skewed, a median and mean (for example) could give different results.

In the supplementary material we show results that shed more light on the issue. In Figure S1 we show posterior means of the parameters λ_D, λ_P, μ_D, μ_P, q_{DP}, and of the difference $r_D - r_P$ between the diploid divergence rate $r_D = \lambda_D - \mu_D$ and the polyploid divergence rate $r_P = \lambda_P - \mu_P$, plotted against tree size. The estimates for λ_D, λ_P, and μ_D are not noticeably biased, but μ_P, q_{DP}, and $r_D - r_P$ are clearly overestimated. In Figure S2 we show maximum likelihood estimates (MLEs) of the same quantities. These estimates have much wider spread than the posterior means. The results for speciation rates and for μ_D are similar in general form to those in Figure 1 of Mayrose et al. (2014), but others are not. The distributions of the MLEs of μ_P and q_{DP} are extremely diffuse and highly skewed. For example, estimates of μ_P can exceed 2.0, more than twenty times its true value of 0.1, while around a quarter of the estimates are less than 0.0001. The median of the MLEs of μ_P is 0.094, a little low, but the mean is 0.211, over double the true value. The only safe conclusion we can draw is that MLEs of μ_P and q_{DP} are almost useless for trees of this size.

Further supporting our call for caution in interpreting the Mayrose et al. (2011, 2014) results, Robosky and Goldberg (2015) recently explored the use of BiSSE and showed the "ease with which neutral traits are inferred to have statistically significant associations with speciation rate." They conclude that the "surprising severity of this phenomenon suggests that many trait diversification relationships reported in the literature may not be real."

RETICULATION ISSUES

A large fraction of polyploidization events are thought to have arisen by hybridization (allopolyploidy), thus making the tree model inadequate for species phylogenies. Mayrose et al. (2014) acknowledge this as a potential problem for their approach, but argue that their conclusions are unaffected by this, because the results between analyses made on cpDNA trees do not differ significantly from those from other trees that are based on cpDNA and concatenated nuclear data or nuclear data only. Although non-recombined plastid genomes will produce tree-like phylogenies, this disregards the fact that the actual diploid parents of an allopolyploid may be extinct, something that could greatly overestimate the age of an allopolyploid (Doyle and Egan 2010; Fig. 2).

Reticulation presents another problem for the method of Mayrose et al. (2011). There is an increased awareness that gene and species trees are fundamentally different (e.g., Pamilo and Nei, 1988; Doyle, 1992; Edwards, 2009), and that concatentation of unlinked genes may result in trees that are poor estimates of species trees (e.g., Degnan and Rosenberg, 2009). In the case of allopolyploid species networks, the situation is further complicated by the fact that the homeologues of an allopolyploid usually cannot be assigned to its parental lineages a priori (e.g., Huber et al., 2006, Jones et al., 2013). Thus, uncritical concatenation of sequences from the plastid genome and unlinked nuclear sequences may result in chimeras (Bertrand et al., 2015). Such “hybrid taxa” are known to tend to branch off close to the root if analyzed in a phylogenetic tree context (e.g., McDade, 1992). Similar, but perhaps less serious, concerns can be made on the use of nuclear ribosomal DNA cistron repeats (i.e., ITS, Álvarez and Wendel, 2003). Again,

the consequence is that the method of Mayrose et al. (2011) may result in an underestimate of the diversification rate in allopolyploids.

Mayrose et al. (2014) state that polyploids "arise but fail to persist." A fundamental problem with this statement is that it does not take into account the long-term evolutionary history of most groups of angiosperms, in which all contemporary taxa are derived from polyploidization (e.g., Jiao et al., 2011; *Amborella* Genome Project, 2014), be it neo- or paleopolyploidization. In fact, most of the oldest angiosperm clades demonstrate chromosome numbers (and gene content; e.g., Cui et al., 2006) consistent with polyploidy, which would suggest that polyploids are truly the taxa that persist on long-term evolutionary scales (see also mature polyploid complexes; Stebbins, 1971). We realize that Mayrose et al. (2014) were framing that statement in the context of neopolyploids (although they have no real temporal component in their analyses), but we caution that the statement is an overgeneralization and diminishes what we know regarding the repeated evolutionary success of polyploids throughout angiosperms. Mayrose et al. (2011) did address the long-term issue in their supplemental material. It is possible for polyploids to have a lower divergence rate but nonetheless dominate diploids in the long term, if the rate at which diploids produce new polyploids is high enough. This can be proved mathematically and shown in easy simulations.

The temporal component remains a key element of our critique of their work. As they point out, diversification rates are calculated per unit time, such that estimates for younger polyploid clades are not biased relative to those of older diploid clades. However, the amount of diversification in two clades of different ages but diversifying at equal rates will not be the same, and it is these patterns of diversity that are used to

estimate rates. Thus, clade age becomes important. Mayrose et al. (2011) compared the relative diversification rates of polyploids and diploids within a clade and assumed therefore that absolute age is not an issue. However, as they later point out (and as we noted in Soltis *et al.* 2014a), a polyploid clade is necessarily younger than its diploid parents (see below for further discussion).

DIPLOIDS HAVE A HEAD START

Soltis et al. (2014a) stressed that diploids had a head start over any subsequently formed polyploids in terms of diversification—this is a logical point (Marcussen et al., 2015), and Mayrose et al. (2014) ironically use this to argue against sister-clade comparisons (suggested by Soltis et al., 2014; see below), stating: "In addition, such sister-clade comparisons have been shown recently to be inherently biased in cases where one character state is more often the derived one. In these cases, there must be a transition from the ancestral to the derived state on the branch subtending the derived-state clade. The ancestral-state clade, however, gets a head start by already being in that state, so the time available for diversification of the derived state is less, causing it to artificially appear in clades with lower species richness (Käfer and Mousset, 2014)." This is exactly our original criticism—the Mayrose et al. (2011) approach is biased in that diploids have more time to diversify than do the polyploid lineages they spawn.

CLADE-SPECIFIC ERRORS

Mayrose et al. (2014) report that 21 of 63 clades show significantly higher diversification rates for diploids than polyploids, meaning that 2/3 of the clades do NOT

show higher diversification rates in diploids. Furthermore, at least a few of those clades that are reported to exhibit higher diploid rates continue to be plagued by misinterpretation of evolutionary history and ploidy: the GAMA clade and *Tiquilia* are presented below.

GAMA

Mayrose et al. (2014) continue to argue that the GAMA clade of *Greenovia/Aeonium/Monanthes/Aichryson* (Crassulaceae) is an example of higher diversification of diploids vs. tetraploids, stating: "Nevertheless, even if we use the three datasets as reanalyzed by S2014 (= Soltis et al., 2014), the preponderance of evidence … continues to support higher diversification rates for diploids…" We find this example particularly egregious, as our lab generated the original paper on the GAMA clade (Mort et al., 2001). We know these results well, and in fact, this example is what first caused us to question Mayrose et al. (2011) more broadly. Mort et al. (2001) inferred $n = 15$ or $n = 18$ as the ancestral haploid chromosome number for the GAMA clade. This clade comprises exclusively polyploid taxa and thus is clearly an example of diversification at the polyploid level.

As stressed in Soltis et al. (2014), Mayrose et al. (2011) violated their practice of focusing at the genus level in that the GAMA clade comprises four genera. If Mayrose et al. (2011) had taken just one step out to include the well-supported sisters to the GAMA clade, *Sedum modestum* or *S. jaccardianum*, the picture would change dramatically in that members of the GAMA clade have $2n = 30$, 34, 36, and 72; the sister taxa are $2n = 18$ (Mort et al. 2001). Soltis et al. (2014) used the Mayrose et al. (2011) data and did not

add counts for species of *Aichryson*, so the Mayrose et al. (2011) methods erroneously inferred *Aichryson* to be diploid. If, as Mayrose et al. (2014) have now done, counts for *Aichryson* are added, the whole GAMA clade is polyploid (as inferred previously by Mort *et al*., 2001), and there is not even an appropriate data set to analyze with BiSSE. A clade with no known diploids cannot have higher rates of diversification for diploids as asserted by Mayrose et al. (2011, 2014)! The GAMA clade can demonstrate nothing other than that diversification is at the polyploid level, with additional polyploidy (*Aeonium steussyi*, $2n = 72$) having occurred—exactly the conclusion of Mort et al. (2001). This example also points to the larger problem stressed in Soltis et al. (2014)—genera are artificial constructs, and moving out a few nodes can change the results.

TIQUILIA

Mayrose et al. (2014) concede that the *Tiquilia* data set they used was an example of a general bias in their data because polyploids are often underrepresented in gene sampling (one of the points of Soltis et al., 2014). However, their new attempt to discount our conclusion of a high polyploid speciation rate in the clade is unfounded. They state: "Thus, *Tiquilia* represents a potential example where a bias against genotyping polyploids led to their underrepresentation in our dataset. However, because the genes used to generate the S2014 phylogeny (*rps16* and ITS) were said by Moore et al. (2006) to be incongruent and, for ITS, difficult to align, reliable conclusions about *Tiquilia* must await a more complete and robust genetic dataset." However, the genes used to generate the original phylogeny for *Tiquilia* (Moore et al., 2006) or our re-estimated phylogeny (Soltis et al., 2014) do not show incongruences in the backbone phylogeny or in the

placement of the polyploid taxa, which are stably placed. A more complete and robust genetic dataset is not needed to know the placement of the polyploid taxa, nor to see that the polyploid clade is a sterling example of rapid speciation at the polyploid level that has facilitated morphological and geographical evolution (see also Moore et al., 2006).

These errors in the analyses of the GAMA clade and *Tiquilia* were obvious to us because we are familiar with these data sets; we cannot comment on the veracity of the remaining individual analyses, but we caution that meta-analyses are only as powerful as the underlying data. No matter the statistical rigor or precision of an analysis, fundamental misinterpretations, as occurred here for GAMA and *Tiquilia*, cannot be overcome.

FISH—REVERSAL OF FORTUNE

If polyploids diversify less than do diploids, then why was the reverse result obtained for fish in a paper produced by the Mayrose et al. (2011) team (Zhan et al., 2014)? Based on the Mayrose et al. (2011) conclusion that polyploid plants have lower rates of diversification than diploids, it seems counterintuitive to propose that fish are doing something fundamentally different with polyploidy than are angiosperms or ferns. However, Zhan et al. (2014) assert: "our results suggest that polyploidy is associated with different diversification patterns in these two major branches of the eukaryote tree of life." This statement points to different rates of diversification following polyploidy in different groups of organisms and therefore contradicts the strong conclusions leveled against the fate of polyploids in Arrigo and Barker (2012). The differing results might reflect differences in the depth of where polyploidy occurred in the phylogeny—Zhan et

al. (2014) and Mayrose et al. (2014) suggest that their analysis of fish (at the family level) is deeper phylogenetically than the Mayrose et al. (2011) work on plants (at the genus level—usually). Mayrose et al. (2014) admit that the fish example is problematic: "conducting analyses at deeper phylogenetic levels might reveal greater evolutionary success for polyploidization events earlier in plant evolution (just as deeper phylogenetic analyses carried out in fishes did not find lower polyploid diversification; Zhan et al., 2014)."

This statement from Mayrose et al. (2014) highlights additional problems with the logic of Mayrose et al. (2011, 2014). First, they equate taxonomy with age--angiosperm and fern genera are "young", but fish families are "old". But there is no age component to the analyses of Mayrose et al. (2011, 2014) or Zhan et al. (2014). Importantly, Soltis et al. (2014) showed a huge range in ages across the genera sampled by Mayrose et al. (2011)—not all genera are young in age. Relative age is an important component to this discussion (as we repeatedly have noted); it is not considered by the methods of Mayrose et al. (2011) but can be added using the approach we advocate below.

In addition, the opposing results for fish may simply reflect problems of small sample size in the studies of both plants and fish. That is, the small number of taxa used and the actual choice of target genera may have unduly impacted the results in Mayrose et al. (2011). Perhaps if we picked a suite of 50 or more different angiosperm lineages, we would get the results obtained for fish. This was one of the points of Soltis et al. (2014), that sample sizes in Mayrose et al. (2011) were very small and not representative of angiosperms in general, as no truly complex polyploid species groups were included. Totals of 63 of ~12,000 -angiosperm genera (a lineage of 300,000-400,000 species) and 5

of ~550 families of fish, including ~450 of the roughly 27,000 recognized fish species, are not enough to make strong generalizations regarding the fate of polyploid lineages. Whether due to sampling issues or phylogenetic depth, or both, the contrasting results of Zhan et al. (2014) and Mayrose et al. (2011) indicate that broad generalizations regarding the fate of polyploids (as in Arrigo and Barker, 2012) were premature.

ONE PATH FORWARD

Although clarifications by Mayrose et al. (2014) enable further use of their methods, we do not advocate their approach for studies of recent polyploids because reticulation violates the paradigm of bifurcating, tree-like evolution (Soltis et al., 2014 and above). However, the fate of polyploid and diploid lineages through evolutionary time remains an important and unanswered problem, and we suggest investigation of more ancient polyploidy events across phylogenies, at timescales that (generally) are consistent with the tree paradigm. We endorse the use of large, densely sampled, ultrametric trees to plot whole-genome duplications and examine diversification rates after those events for significant shifts in diversification associated with the whole-genome duplications. Mayrose et al. (2014) criticize the use of sister-clade comparisons, but sister-clade comparisons might be just one approach that could be employed across phylogenies. Note that our proposed approach does not rely explicitly on sister clades but evaluates shifts in diversification across the tree. Furthermore, the fates of polyploid lineages can be evaluated regardless of arbitrary taxonomic rank (as in the use of 'genus' by Mayrose et al., 2011 and Zhan et al., 2014), across densely sampled phylogenies and along a common, absolute time scale. In angiosperms, large dated trees are available, making this approach possible (e.g., Smith et al., 2011; Zanne et al., 2014), and the Open

Tree of Life (opentreeoflife.org) provides trees and methods of tree synthesis at truly grand scales. This large phylogeny approach was recently taken by Tank et al. (accepted pending revision) and gives a fresh perspective on polyploidy and diversification; this is a general methodology that we hope others will embrace and develop further.

Lastly, we do agree with our close colleagues (Mayrose et al., 2014) that the best thing that can come from this friendly back and forth discussion is an increased research interest in polyploids.

LITERATURE CITED

ÁLVAREZ, I. AND J. F. WENDEL. 2003. Ribosomal ITS sequences and plant phylogenetic inference. Molecular Phylogenetics and Evolution 29: 417–434.

AMBORELLA GENOME PROJECT. 2013. The complete nuclear genome of *Amborella trichopoda*: an evolutionary reference genome for the angiosperms. *Science* 342: no 6165.

ARRIGO, N., AND M. S. BARKER. 2012. Rarely successful polyploids and their legacy in plant genomes. *Current Opinion in Plant Biology* 15: 140–146.

BERTRAND, Y. J. K., A. C. SCHEEN, T. MARCUSSEN, B. E. PFEIL, F. DE SOUSA, and B. OXELMAN. 2015. Assignment of homoeologues to parental genomes in allopolyploids for species tree inference, with an example from *Fumaria* (Papaveraceae). *Systematic Biology* (in press.). doi:10.1093/sysbio/syv004

CUI, L., P. K. WALL, J. LEEBENS-MACK, B.G. LINDSAY, D.E. SOLTIS, J. J. DOYLE, P. S. SOLTIS, J. CARLSON, A. ARUMUGANATHAN, A. BARAKAT, V. ALBERT, H. MA, AND C.W. DEPAMPHILIS. 2006. Widespread genome duplications throughout the history of flowering plants. *Genome Research* 16: 738-749.

DOYLE, J. J. 1992. Gene trees and species trees - molecular systematics as one-character taxonomy. *Systematic Botany* 17: 144-163.

DOYLE, J. J., AND A. N. EGAN. 2010. Dating the origins of polyploidy events. *New Phytologist*. 186:73–85.

EDWARDS, S. V. 2009. Is a new and general theory of molecular systematics emerging? *Evolution* 63: 1-19.

ESTEP, M. C., M. R. MCKAIN, D. V. DIAZ, J. ZHONG, J.G. HODGE, T. R. HODKINSON, D. J. LAYTON, S. T. MALCOMBER, R. PASQUET, AND E. A. KELLOGG. 2014. Allopolyploidy, diversification, and the Miocene grassland expansion. *Proceedings of the National Academy of Sciences (USA)* 111; 15149-15154.

FITZJOHN, R.G.,, W. P. MADDISON, AND S. P. OTTO. 2009. Estimating trait-dependent speciation and extinction rates from incompletely resolved phylogenies. *Syst. Biol.* 58: 595. doi:10.1093/sysbio/syp067 Medline

JIAO, Y, N. J. WICKETT, S. AYYAMPALAYAM, A. S. CHANDERBALI, L. LANDHERR, P. E. RALPH ET AL. 2011. Ancestral polyploidy in seed plants and angiosperms. *Nature* 473: 97–100.

JONES, G., S. SAGITOV, and B. OXELMAN. 2013. Statistical inference of allopolyploid species networks in the presence of incomplete lineage sorting. *Systematic Biology* 62: 467-478.

HUBER, K. T., B. OXELMAN, M. LOTT, and V. MOULTON. 2006. Reconstructing the evolutionary history of polyploids from multilabeled trees. *Molecular Biology and Evolution* 23: 1784-1791.

KAFER, J, AND S. MOUSSET. 2014. Standard sister-clade comparison fails when testing derived character states. *Systematic Biology* 63: 601–609.

MADDISON, W. P. AND D. R. MADDISON. 2014. Mesquite: a modular system for evolutionary analysis. Version 3.01 http://mesquiteproject.org

MADDISON, W. P., P. E. MIDFORD, AND S. P. OTTO. 2007. Estimating a binary character's effect on speciation and extinction. *Systematic Biology* 56: 701-710.

MARCUSSEN, T., L. HEIER, A. K. BRYSTING, B. OXELMAN, AND K. S. JAKOBSEN. 2015. From gene trees to a dated allopolyploid network: insights from the angiosperm genus *Viola* (Violaceae). *Systematic Biology* 64: 84–101.

MAYROSE, I, S. H. ZHAN, C.J. ROTHFELS, K. MAGNUSON-FORD, M. S.BARKER, L. H. RIESEBERG, S. P. AND OTTO. 2011. Recently formed polyploid plants diversify at lower rates. *Science* 333: 1257.

MAYROSE, I, S. H. ZHAN, C.J. ROTHFELS, N. ARRIGO, M. S. BARKER, L. H. RIESEBERG, AND OTTO SP. 2014. Methods for studying polyploid diversification and the dead end hypothesis: a reply to Soltis et al. (2014). *New Phytologist* in press.

MCDADE, L. A. 1992. Hybrids and phylogenetic systematics II. The impact of hybrids on cladistic analysis. *Evolution* 46: 1329–1346.

MORT, M. E., D. E. SOLTIS, P. S. SOLTIS, J. FRANCISCO-ORTEGA, AND A. SANTOS-GUERRA. 2001. Phylogenetic relationships and evolution of Crassulaceae inferred

from *matK* sequence data. *American Journal of Botany* 88: 76–91.

PAMILO, P., and M. NEI. 1988. Relationships between gene trees and species trees. *Molecular Biology and Evolution* 5: 568-583.

R: A LANGUAGE AND ENVIRONMENT FOR STATISTICAL COMPUTING. R FOUNDATION FOR STATISTICAL COMPUTING, R CORE TEAM. 2013. Vienna, Austria. URL http://www.R-project.org/.

RABOSKY, D. L. 2010. Extinction rates should not be estimated from molecular phylogenies. *Evolution* 64: 1816–1824.

RABOSKY, D.L. AND E.E. GOLDBERG. 2015. Model inadequacy and mistaken inferences of trait-dependent speciation. *Systematic Biology* in press.

SMITH, S. A., J. M. BEAULIEU, A. STAMATAKIS, AND M. J. DONOGHUE. 2011. Understanding Angiosperm diversification using small and large phylogenetic trees. *American Journal of Botany* 98: 404-414.

SOLTIS, D. E., M. C. SEGOVIA-SALCEDO, I. JORDON-THADEN, L. MAJURE, N. M. MILES, E. V. MAVRODIEV, W. MEI, M. B. CORTEZ, P. S. SOLTIS, AND M. A. GITZENDANNER. 2014. Are polyploids really evolutionary dead-ends (again)? A critical reappraisal of Mayrose et al. (2011). *New Phytologist* 202: 1105–17.

STEBBINS, G. L. 1971. Chromosomal evolution in higher plants. Edward Arnold, London, UK.

TANK, D. C., J. M. EASTMAN, M.W. PENNELL, P. S. SOLTIS, D. E. SOLTIS, C. E. HINCHLIFF, J. W. BROWN, AND L. J. HARMON. Progressive Radiations and the Pulse of Angiosperm Diversification. *New Phytologist*, in press.

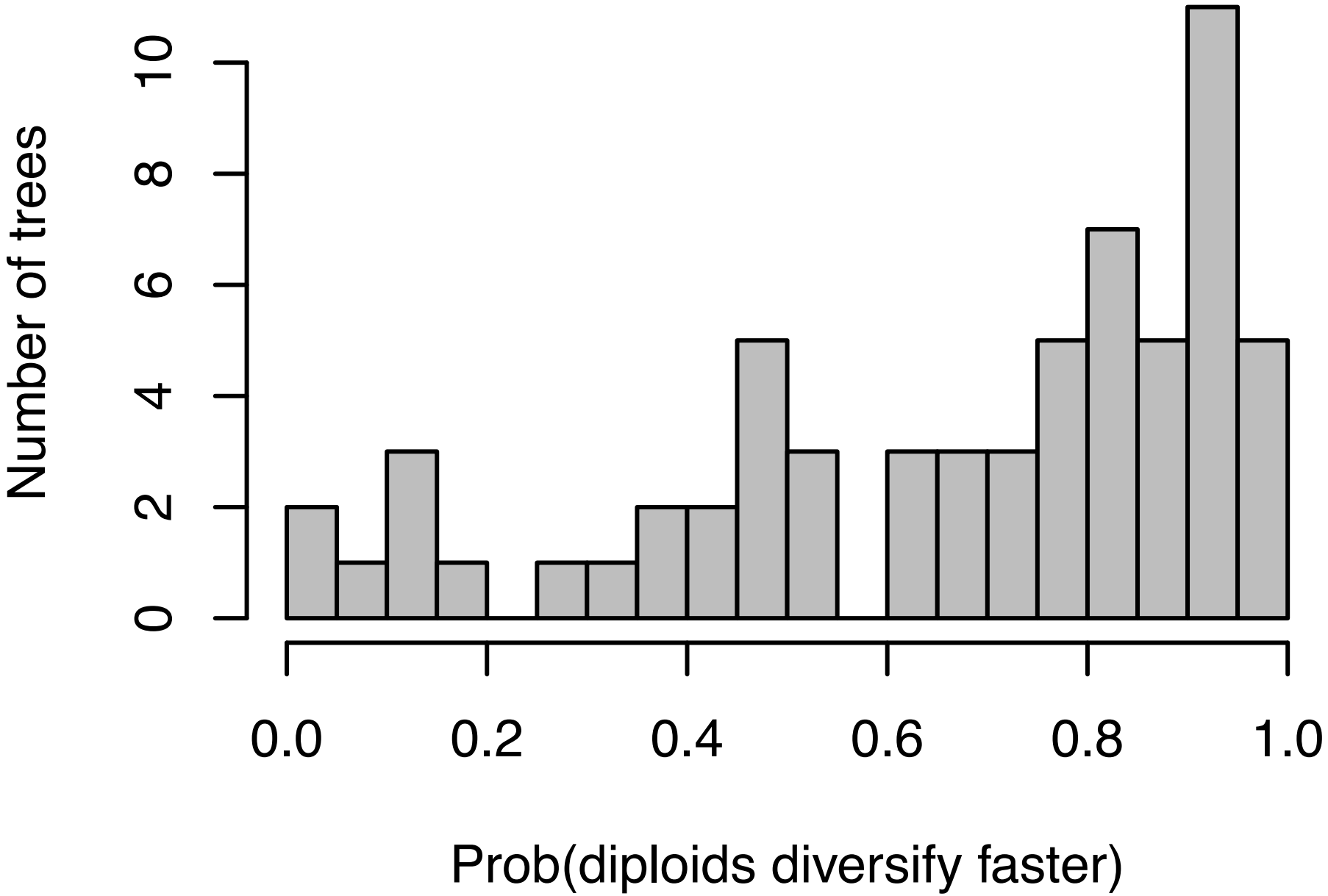
ZANNE, A. E. , D. C. TANK, W. K. CORNWELL, J. M. EASTMAN, S. A. SMITH, R. G. FITZJOHN, D. J. MCGLINN, B. C. O'MEARA, A. T. MOLES, P. B. REICH, D. L. ROYER, D. E. SOLTIS, P. F. STEVENS, M. WESTOBY, I. J. WRIGHT, L. AARSSEN, R. I. BERTIN, A. CALAMINUS, R. GOVAERTS, F. HEMMINGS, M. R. LEISHMAN, J. OLEKSYN, P. S. SOLTIS, N. G. SWENSON, L. WARMAN, AND J. M. BEAULIEU. 2014. Into the cold - three keys to radiation of angiosperms into freezing environments. *Nature* doi:10.1038/nature12872.

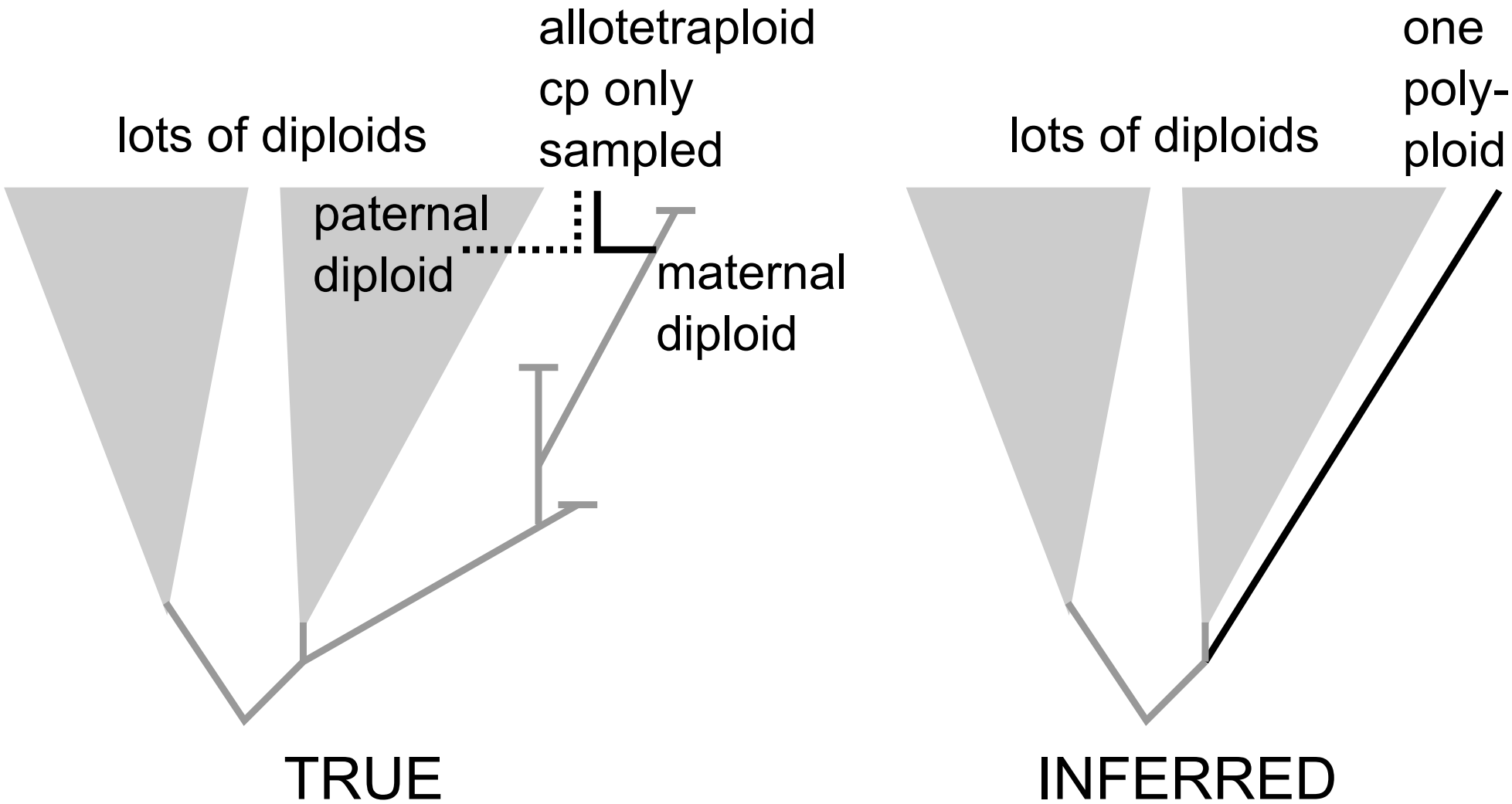
ZHAN, S. H., L. GLICK, C. S. TSIGENOPOULOS, S. P. OTTO, AND I. MAYROSE. 2014. Comparative analysis reveals that polyploidy does not decelerate diversification in fish. *Journal of Evolutionary Biology* 27: 391–403.

Figure legends

Figure 1. Number of trees vs. the probability that diploids diversify faster than polyploids in diversitree analysis implementing BiSSE. This analysis is directly comparable to Figure 1A of Mayrose et al. (2011) (see text). In our analyses, the bias towards 1 (high posterior probabilities for diploids having higher diversification rates) is an artifact.

Figure 2. Hypothetical true vs. inferred trees illustrating one of the difficulties in using a phylogenetic approach to examine polyploidy. In this example, non-recombined plastid genomes will produce tree-like phylogenies, disregarding the fact that the actual diploid parents of an allopolyploid may be extinct. This approach could greatly overestimate the age of an allopolyploid.


Number of trees
0
2
4
6
8
10
0.0
0.2
0.4
0.6
0.8
1.0
Prob(diploids diversify faster)


allotetraploid
cp only
sampled
lots of diploids
paternal
diploid
maternal
diploid
TRUE
one
poly-
ploid
lots of diploids
INFERRED

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

Graham Jones

February 2, 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 λ_D	0.0698	0.2016	0.0447	0.2009	0.3468
MLE of λ_D	0	0.214	0.103	0.1893	0.8542
posterior mean of λ_P	0.0526	0.1899	0.0638	0.1892	0.4135
MLE of λ_P	0	0.1876	0.1427	0.1579	0.7812
posterior mean of μ_D	0.032	0.0968	0.0397	0.0905	0.251
MLE of μ_D	0	0.1182	0.1469	0.072	0.8741
posterior mean of μ_P	0.0277	0.1779	0.0936	0.1601	0.5755
MLE of μ_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.

A simpler example of bias

We describe a much simpler but analogous statistical analysis which demonstrates a similar type of bias.

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.

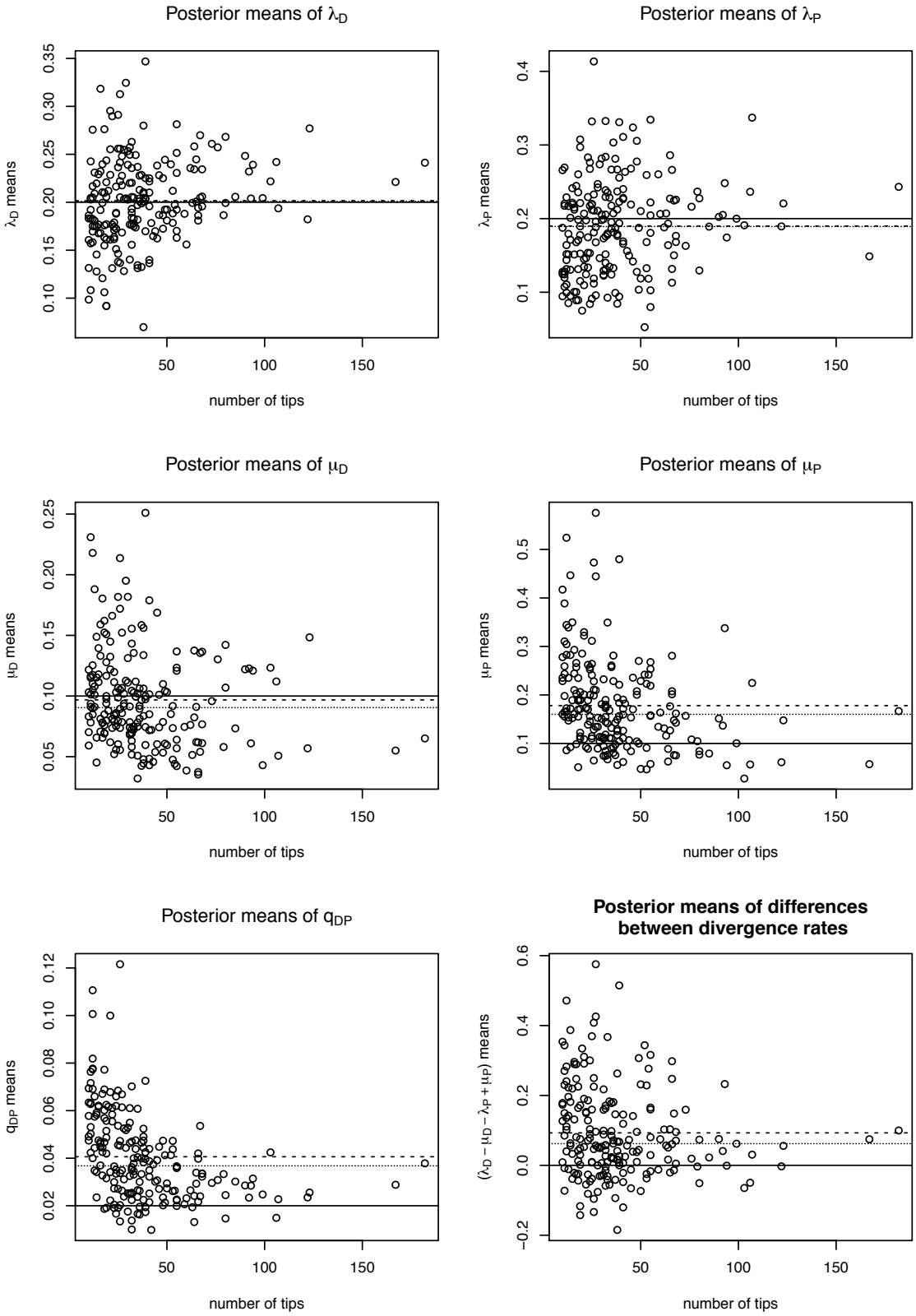


Figure S1: Posterior means of the parameters and the differences between divergence rates. The 6 graphs show results for the same 200 simulated trees. These trees are similar to those of Mayrose et al. (2011). The solid lines show true values, the dashed lines show the means, and the dotted lines show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

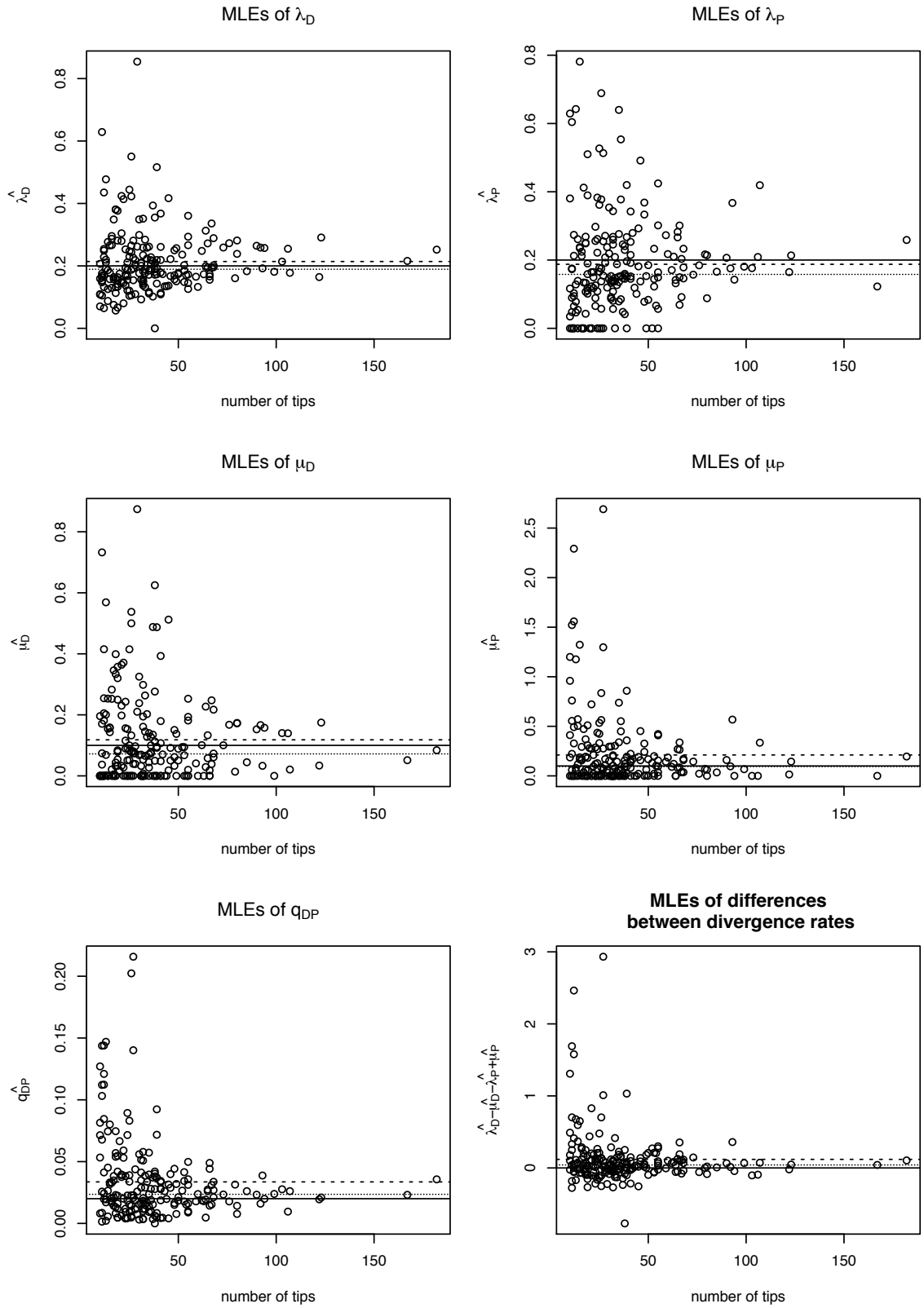


Figure S2: Maximum likelihood estimates of the parameters and the differences between divergence rates. Other details as figure S1. Note the large scale of some y-axes.

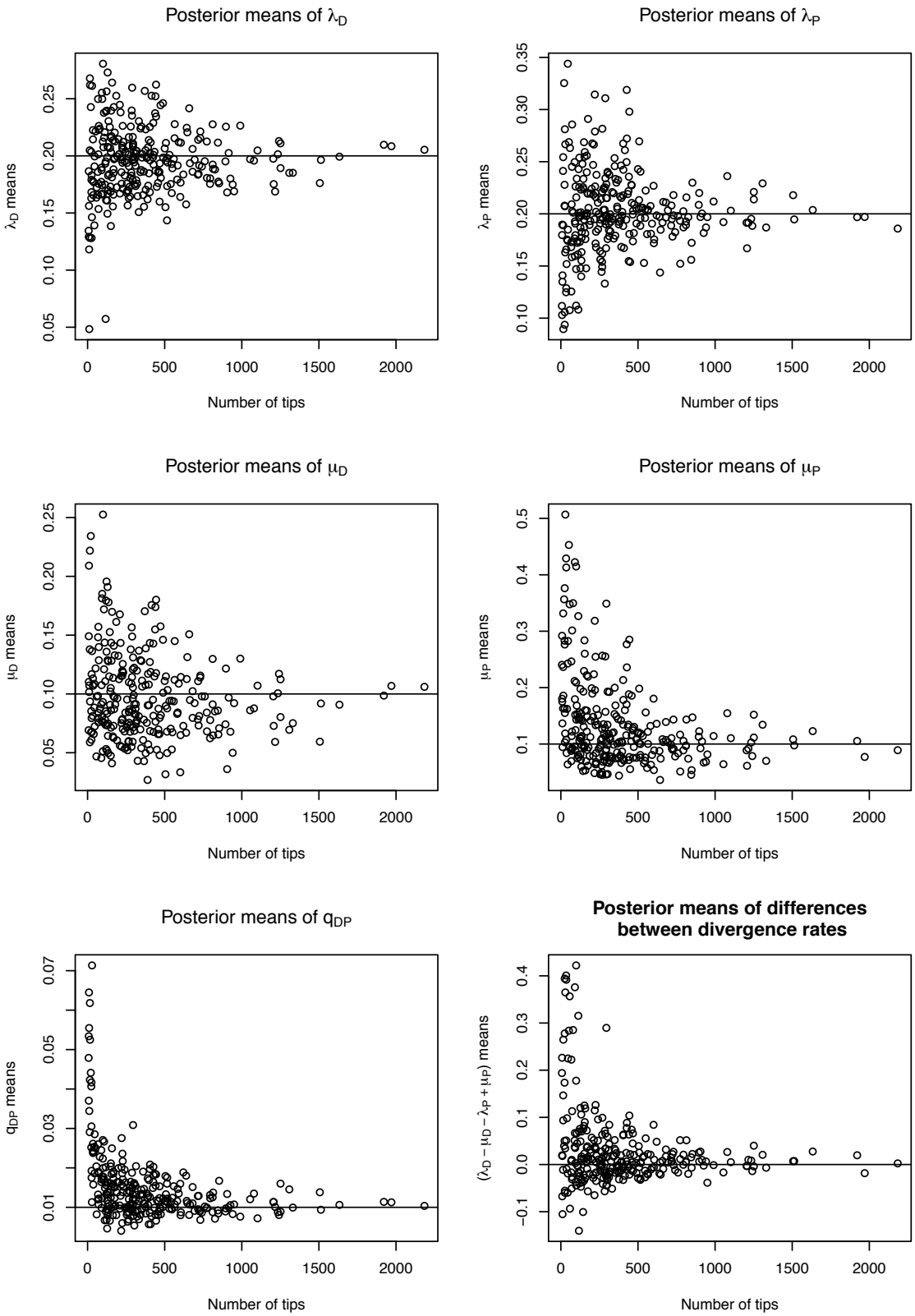


Figure S3: Posterior means of the parameters and the differences between divergence rates, for a set of 280 larger trees. The solid lines show true values.