Supplementary material: Model-based detection of whole-genome duplications in a phylogeny

Arthur Zwaenepoel[1,2,3,*], Yves Van de Peer[1,2,3,4,*]

1. Department of Plant Biotechnology and Bioinformatics, Ghent University, 9052 Ghent, Belgium
2. Center for Plant Systems Biology, VIB, 9052 Ghent, Belgium
3. Bioinformatics Institute Ghent, 9052 Ghent, Belgium
4. Department of Biochemistry, Genetics and Microbiology, University of Pretoria, Pretoria 0028, South Africa

[*] Corresponding author: arzwa@psb.vib-ugent.be, yvpee@psb.vib-ugent.be

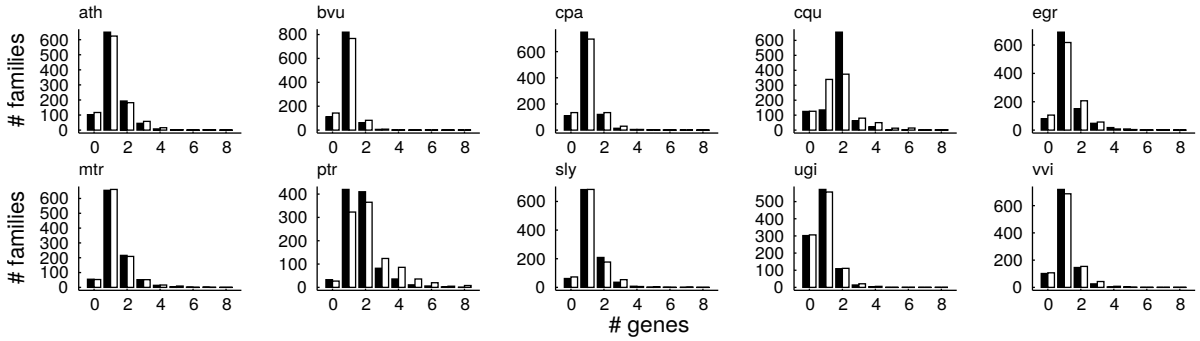

Figure S1: Comparison of the distribution of the number of genes of each species in a gene family for the observed data (1000 gene families across 10 plant species) (in black) with simulations from the posterior predictive distribution (in white) under the BM model of duplication and loss rate variation. Considering for example *Chenopodium quinoa* (cqu) or *P. trichocarpa* (ptr) (two species with a well-described highly preserved ancient WGD), and comparing for instance with *Vitis vinifera* (vvi) or *Carica papaya* (cpa) (two species with no known WGDs in their recent evolutionary past), these simulations clearly illustrate that WGDs are a major source of model violation when modeling gene family evolution by a flexible linear birth-death process.

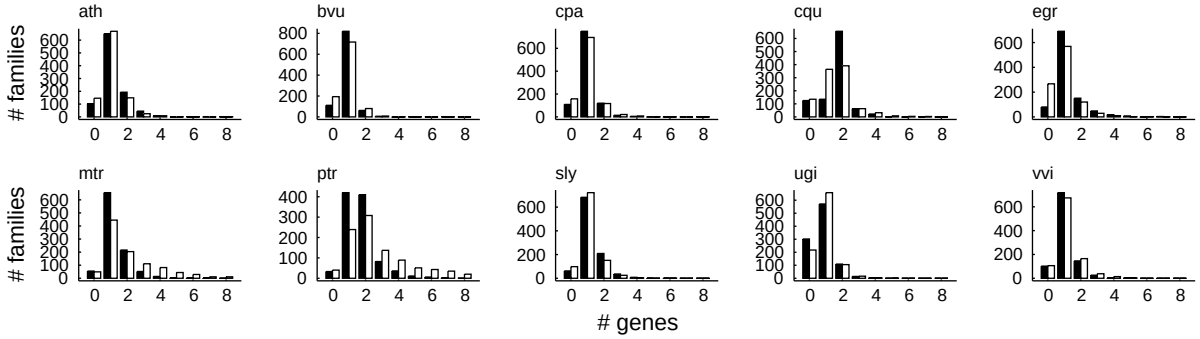

Figure S2: As in fig. S1 but under the IR model for rate variation across the species tree

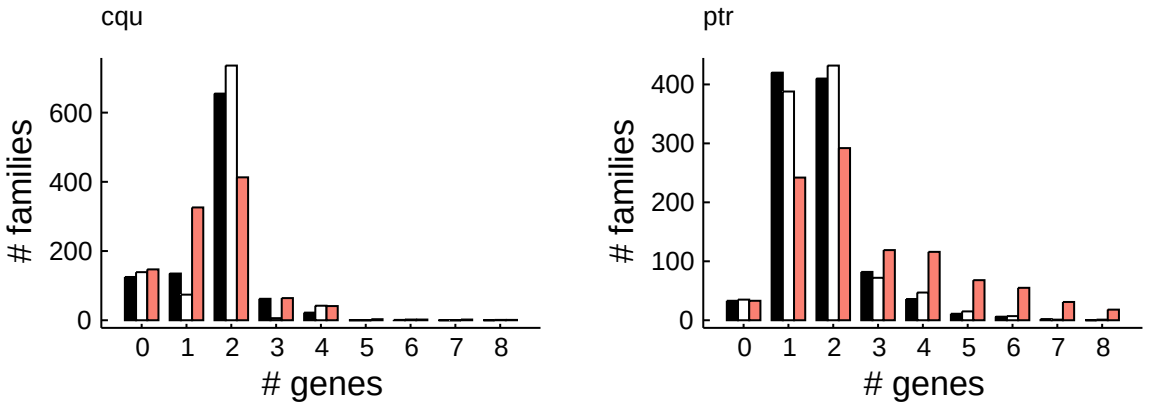

Figure S3: As in fig. S2 but now the SSDL-only model is shown in red and we include in white posterior predictive simulations under a model including a WGD on the branches leading to *C.quinoa* (cqu) and *P. trichocarpa* (ptr).

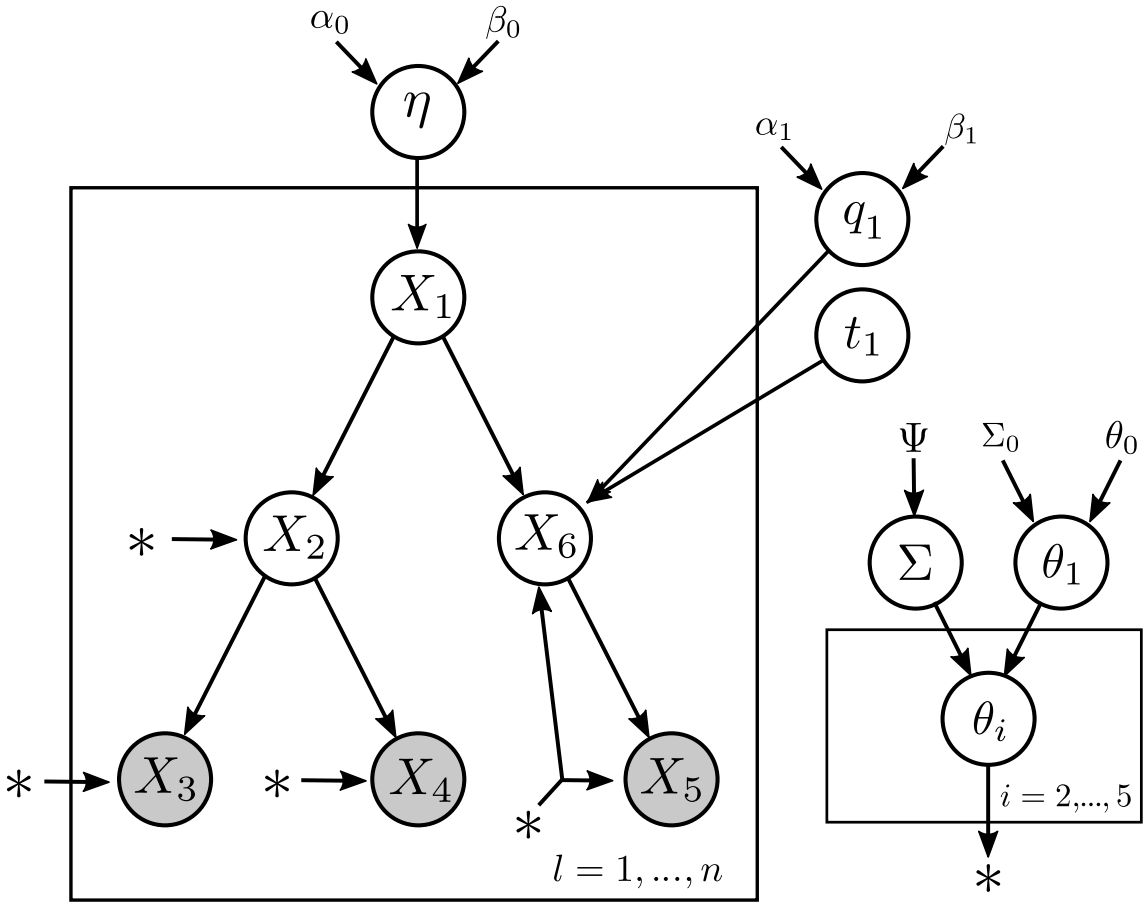

Figure S4: Probabilistic graphical model (PGM) illustrating the hierarchical structure of the IR-model for a hypothetical three-taxon species tree with one WGD node (X_6).

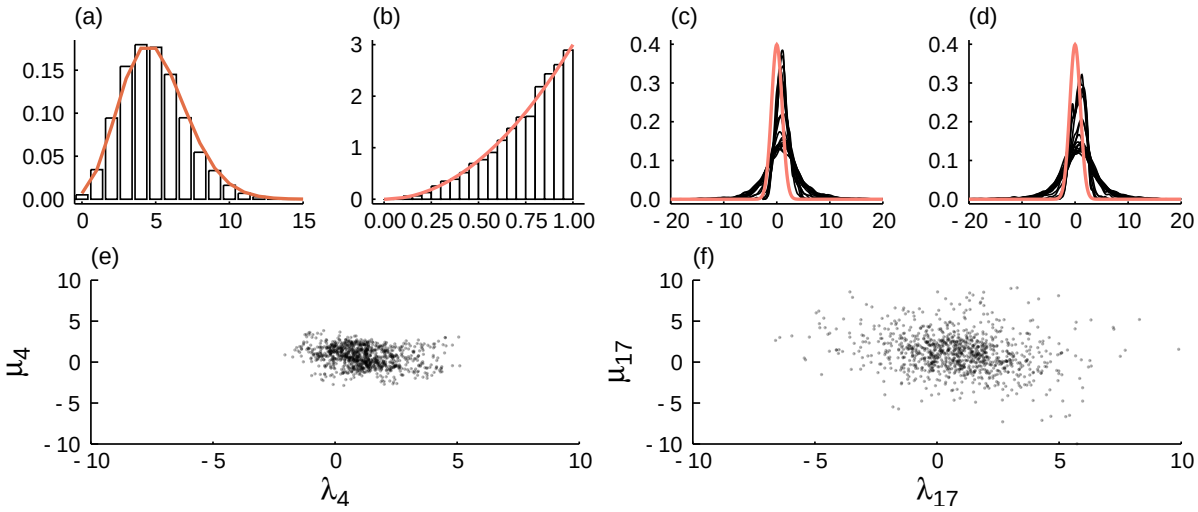

Figure S5: MCMC samples from the BM prior. (a) The number of WGDs k, with a Poisson prior (in orange), (b) η with a Beta prior (in orange), (c) duplication rates, with in orange the prior on the root duplication rate $\log \lambda_1$, (d) as in (c) but for loss rates (e) scatter plot of $\log \mu_4$ vs. $\log \lambda_4$ and (f) as in (e) but for node 17.

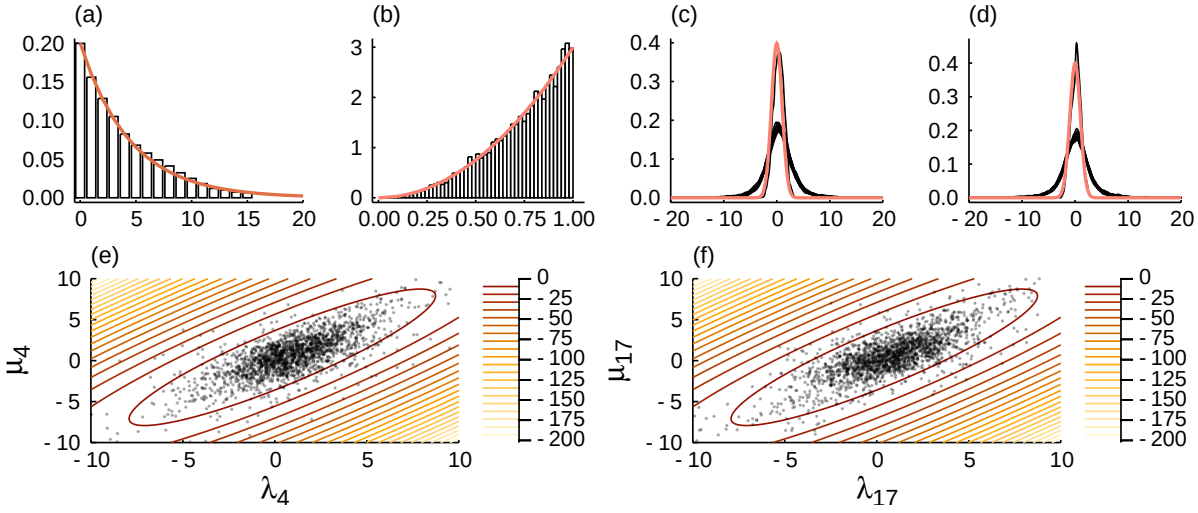

Figure S6: MCMC samples from the IR prior with a Geometric(0.25) prior with upper bound of 15 on the number of WGDs, and a prior covariance matrix with non-zero non-diagonal entries. See fig. S5 for explanation of the panels. The overlayed contour plot in panels (e) and (f) shows the log density of the ultivariate Normal distribution with mean θ_0 and covariance matrix $\Sigma_0 = \Psi$.

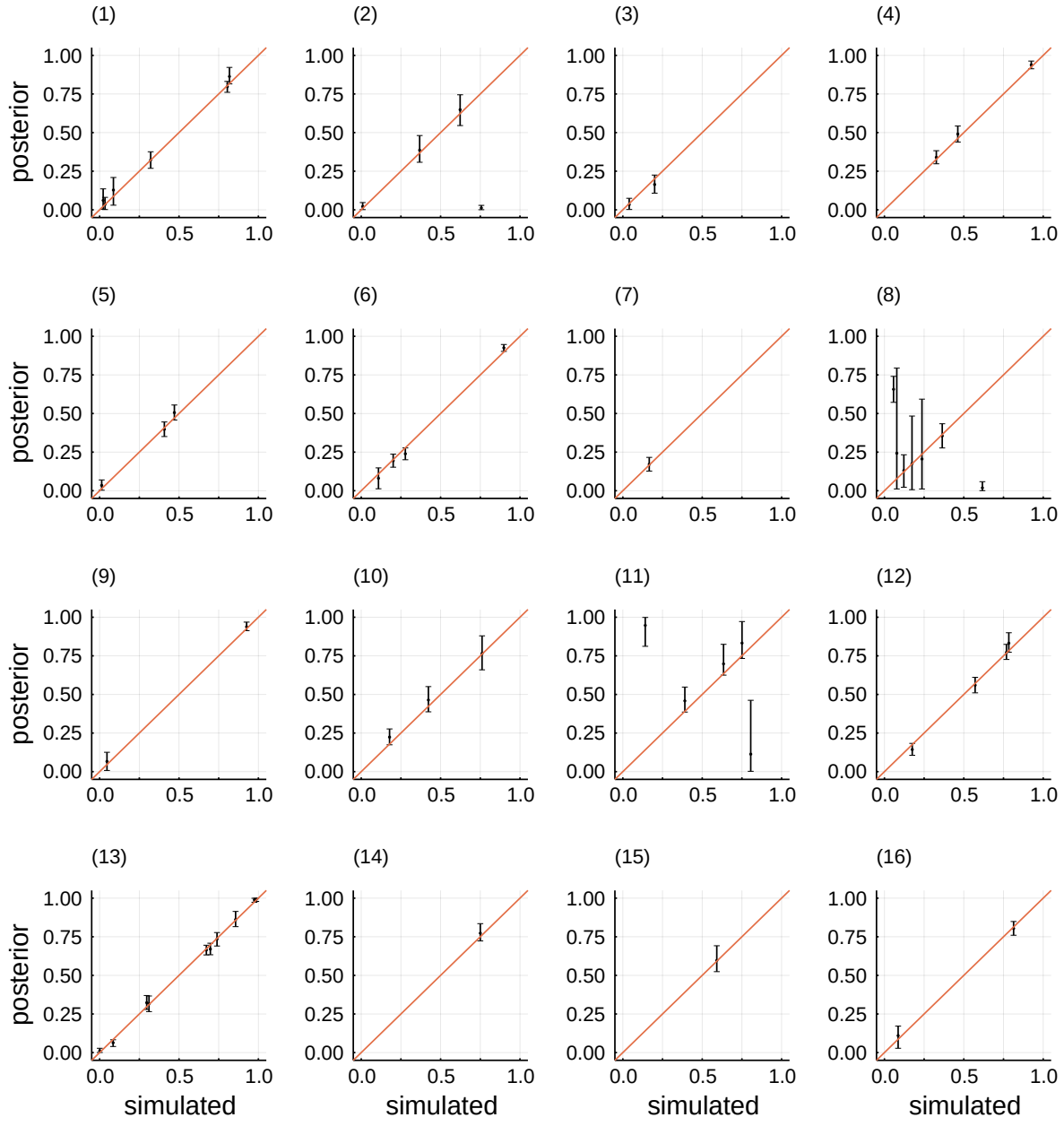

Figure S7: Posterior mean estimates and associated 95% credibility intervals for retention rates, as in fig. 1 but here we show results for each of the 16 replicates (that included WGDs) separately. We note that replicate 8 and 11 show the issue where retention rates for WGDs on the same branch are swapped (see main text).

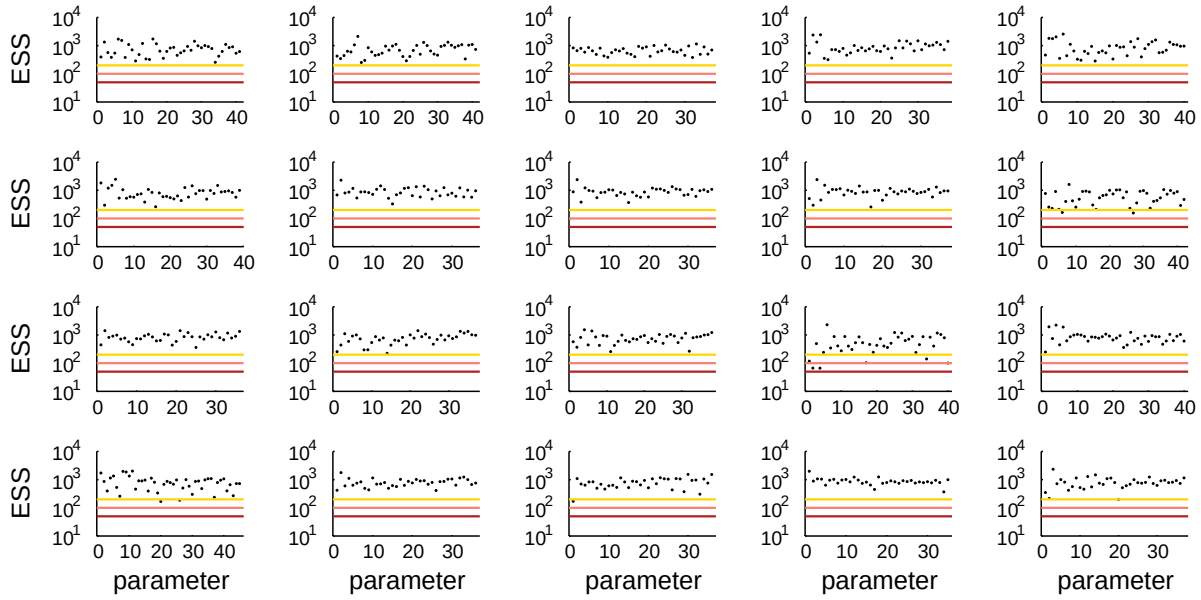

Figure S8: Effective sample size (ESS) associated with each parameter for each replicate for the first set of simulations employing the fixed-dimensional MCMC sampler (see fig. 1).

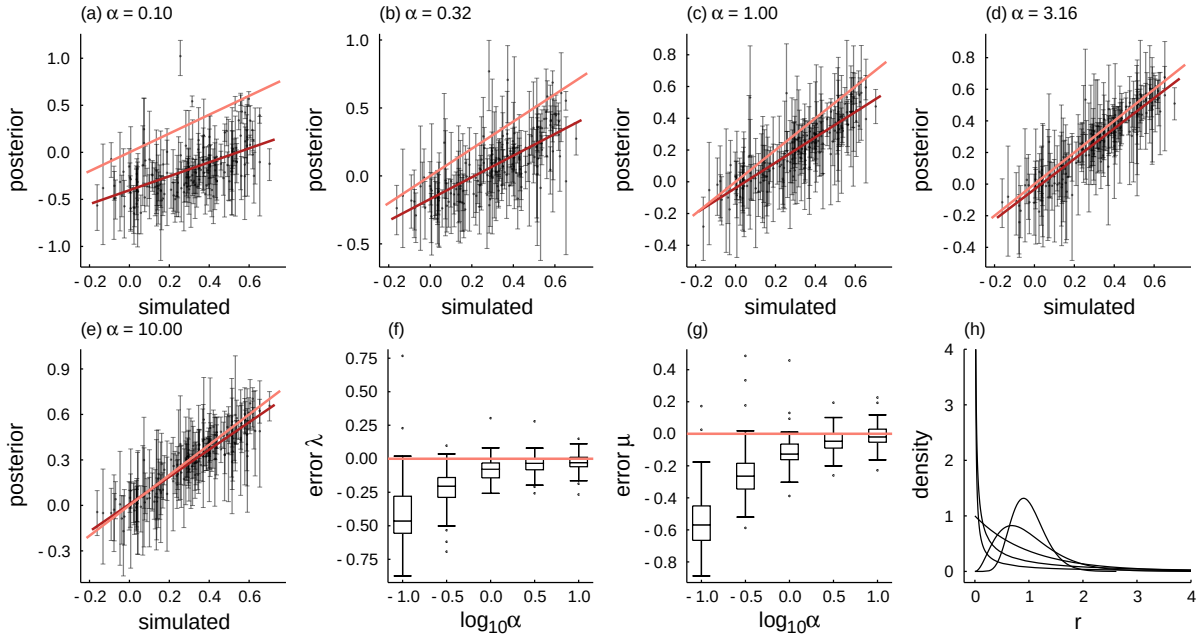

Figure S9: The effects of family × lineage variation on the posterior inference of duplication and loss rates. Simulations with Gamma distributed family-specific relative rates and without WGDs were performed for different shapes of the Gamma distribution, i.e. Gamma(α, $1/\alpha$) (such that the mean is 1) for different α. Posterior inference was performed using the fixed-dimensional MCMC sampler, assuming constant rates of duplication and loss across families (but not across lineages). Duplication and loss rate estimates (on a $\log_{10}$ scale) for five replicate simulations of $N = 500$ families are shown pooled together in panels (a-e). Panels (f) and (g) show the distributions of the difference between the posterior geometric mean and true value. Panel (h) illustrates the different shapes of the Gamma distribution considered in these simulations

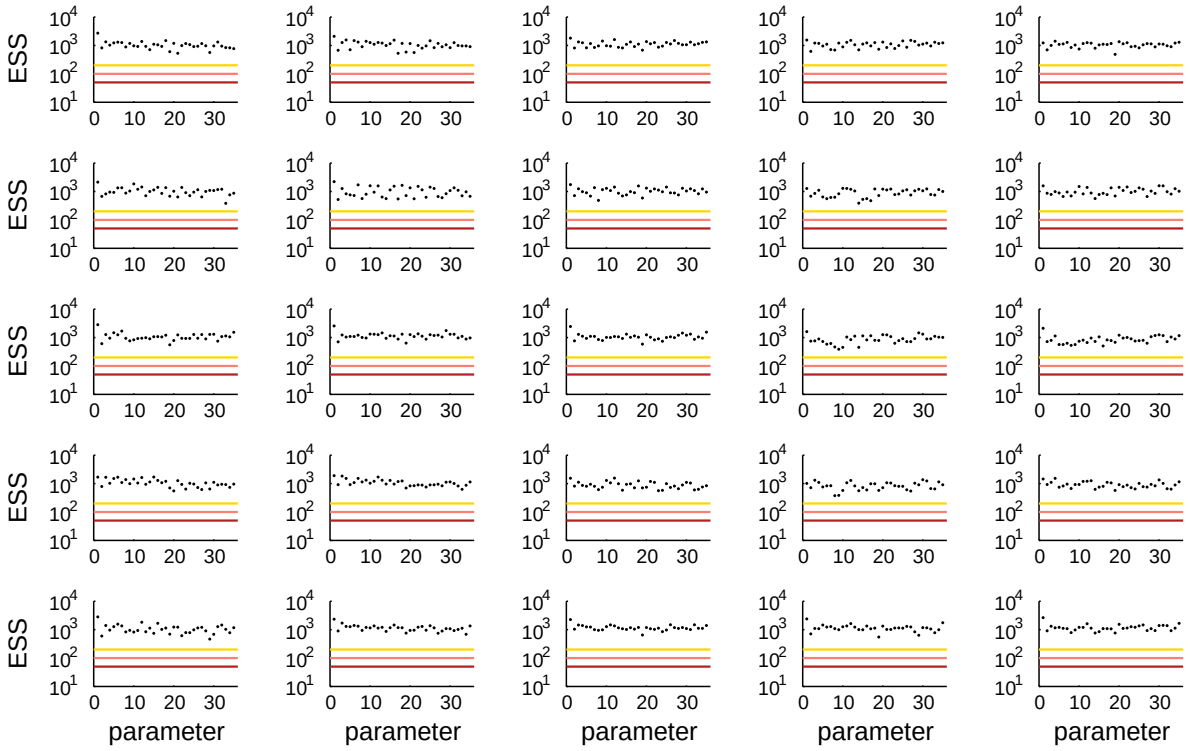

Figure S10: Effective sample size (ESS) associated with each parameter for each replicate for the second set of simulations (with rate variation across gene families) employing the fixed-dimensional MCMC sampler (see fig. S9).

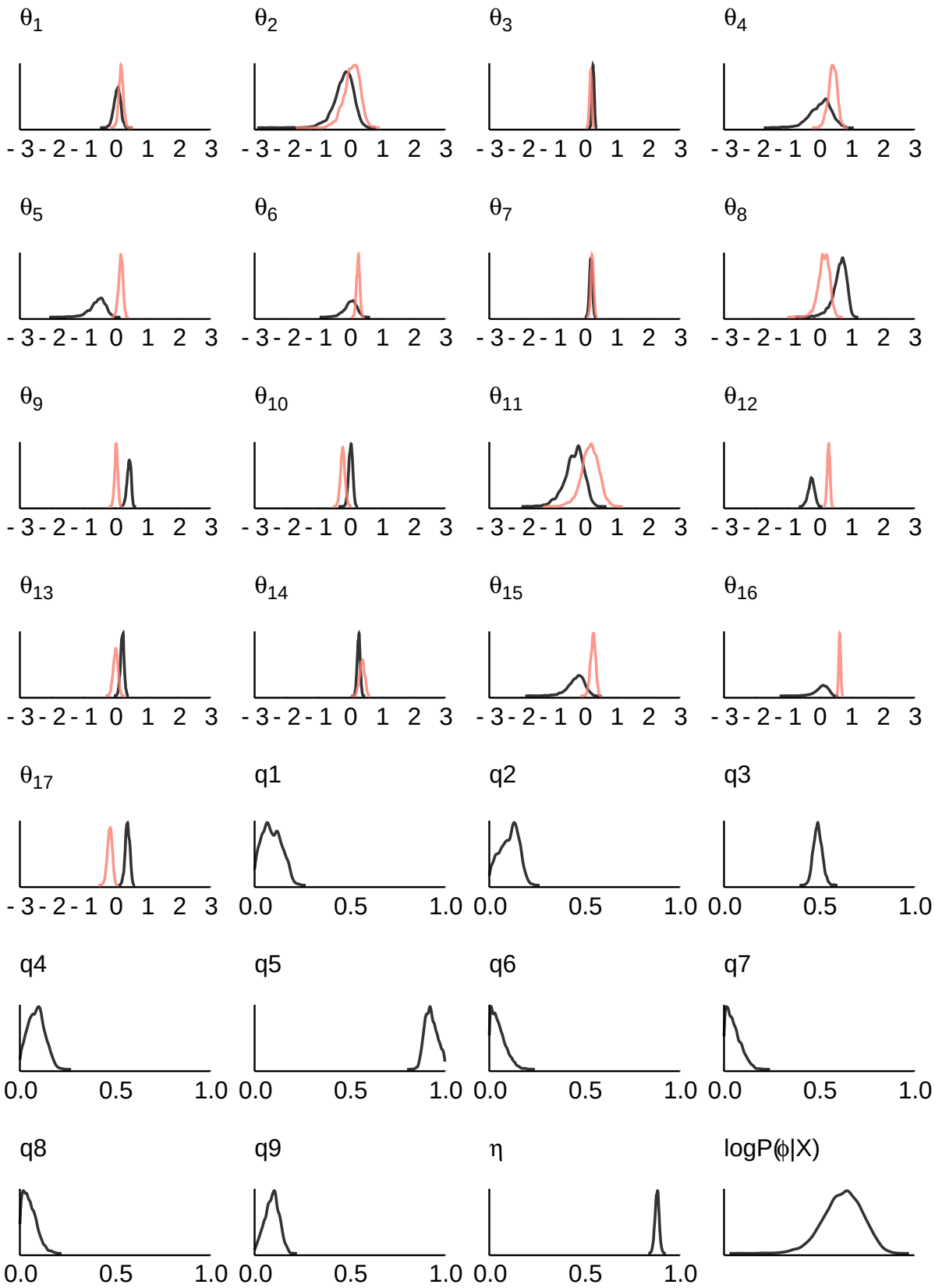

Figure S11: Approximate marginal posterior densities for branch-wise duplication and loss rates (θ_i, black and red respectively, on a $\log_{10}$ scale), retention rates q_j and η for the dicot data set obtained with a fixed-dimensional MCMC sampler for a nine-WGD model. Duplication and loss rates for the second set of simulations shown in fig. 2 were drawn from the joint posterior distribution associated with these results.

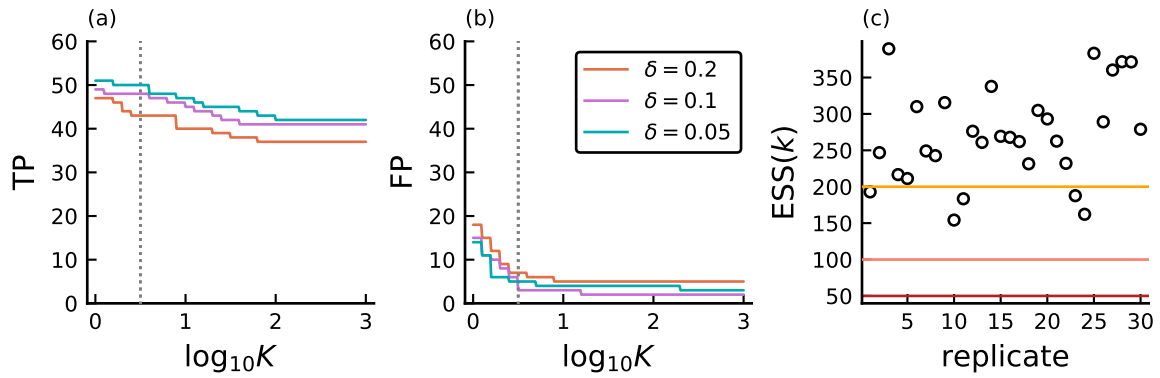

Figure S12: Performance of the rjMCMC algorithm for WGD inference for simulated data sets of $N = 500$ gene families with branch-wise duplication and loss rates, where for each branch the duplication and loss rate are equal. Other simulation settings were as in fig. 2 (upper two rows). Note that for each set of WGDs and duplication and loss rates inference was done using three δ values for te constraint prior. (a) The number of true positive WGD inferences (TP, in a total of 60) as a function of the Bayes factor (K) in favor of the relevant WGD. (b) The number of falsely positive WGD inferences as a function of the Bayes factor. (c) ESS estimates for all replicates.

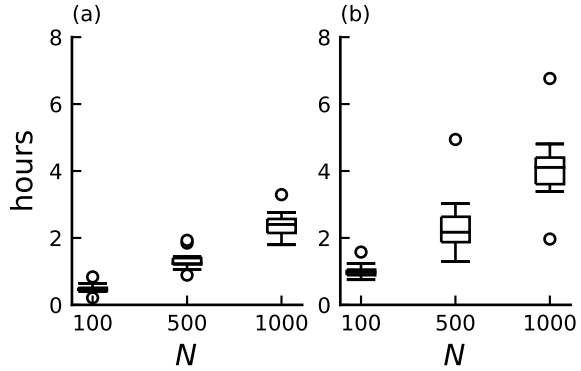

Figure S13: Computing times per 10.000 iterations as a function of data set size N for the simulations in fig. 2. (a) Times for inferences under the constant-rates model for simulations from the IR prior. (b) Times for inferences under the branch-wise rates model for simulations from the IR prior. All results were obtained on the same machine using 5 CPUs to compute the likelihood in parallel.

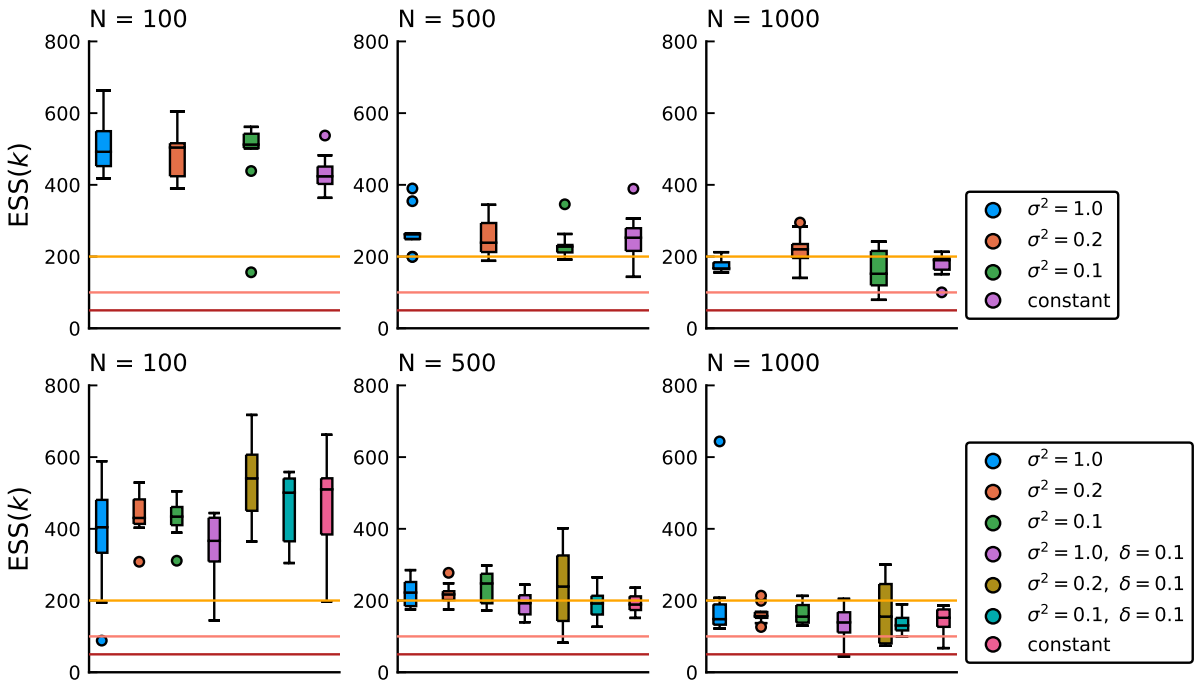

Figure S14: ESS estimates for the model indicator variable k (i.e. the number of WGDs) for the rjMCMC samples associated with the simulations in fig. 2.

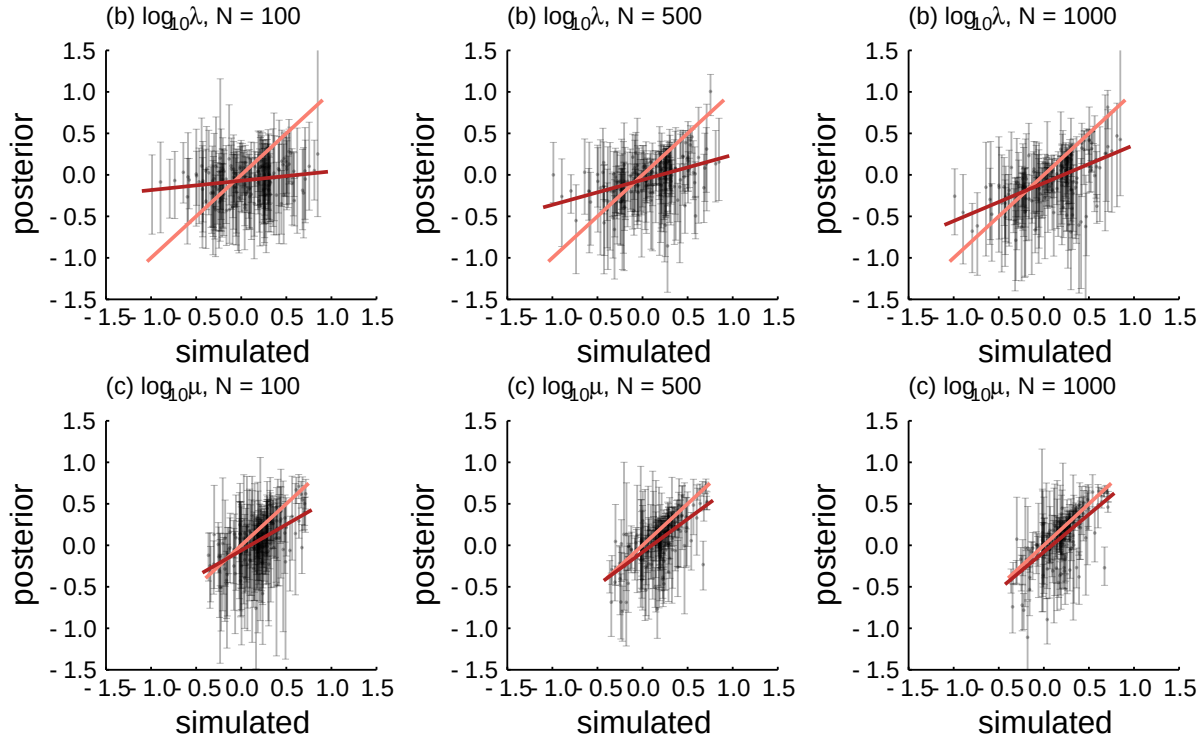

Figure S15: Posterior mean duplication and loss rates with 95% credibility intervals for the second set of simulations (i.e. the data sets simulated from the approximate joint posterior for the nine dicots data set acquired with the fixed-dimensional sampler), obtained using the rjMCMC sampler with prior covariance matrix $\Psi = 0.2I_2$. In orange the true relationship is shown whereas in red the least squares regression of posterior means on the true simulated values is shown. Results for all branch-wise rates for 10 simulated replicates are shown pooled together. Note how the variation in the duplication rates is much higher than than the variation in the loss rates.

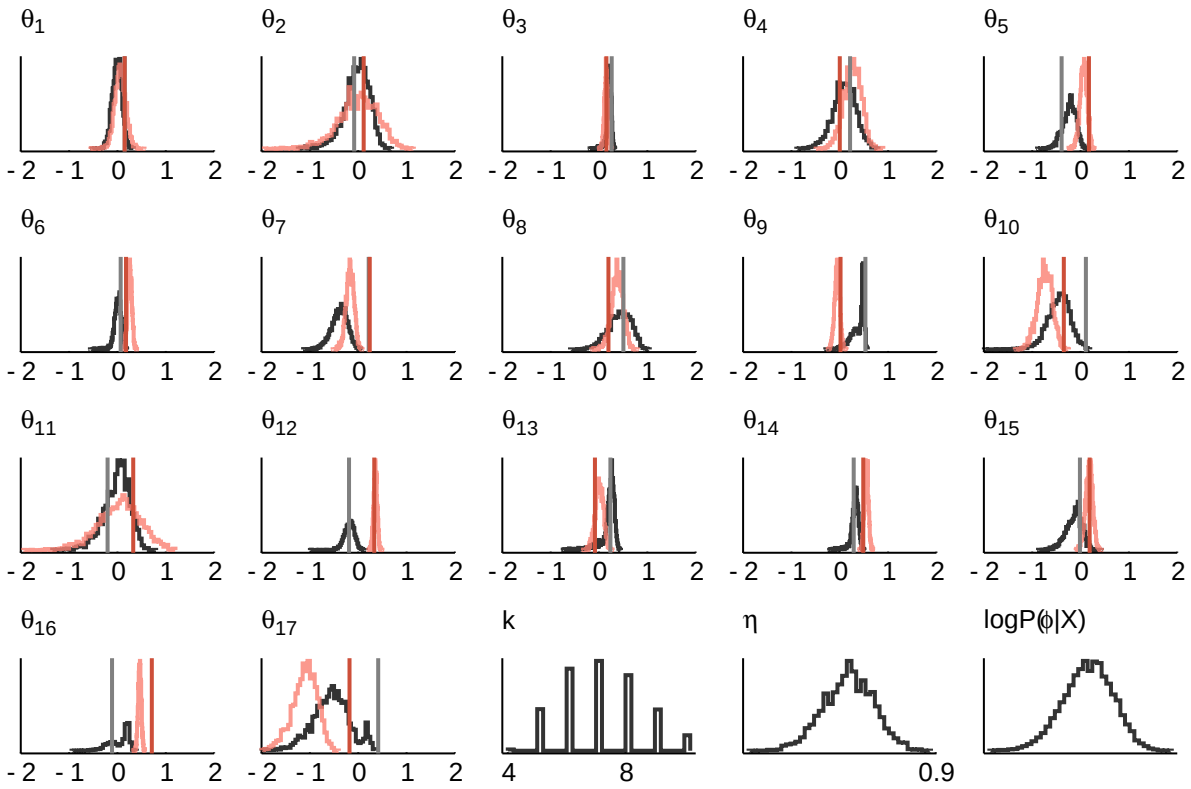

Figure S16: Approximate marginal posterior densities for one of the simulated replicates ($N = 1000$, $\Psi = 0.2I_2$, where simulated duplication and loss rates were drawn from the joint posterior for the dicots data set under a fixed model, see main text for more details) obtained with the rjMCMC algorithm. The branch-wise duplication and loss rates (θ_i, on a $\log_{10}$ scale) are shown in black and red respectively. See also fig. S17.

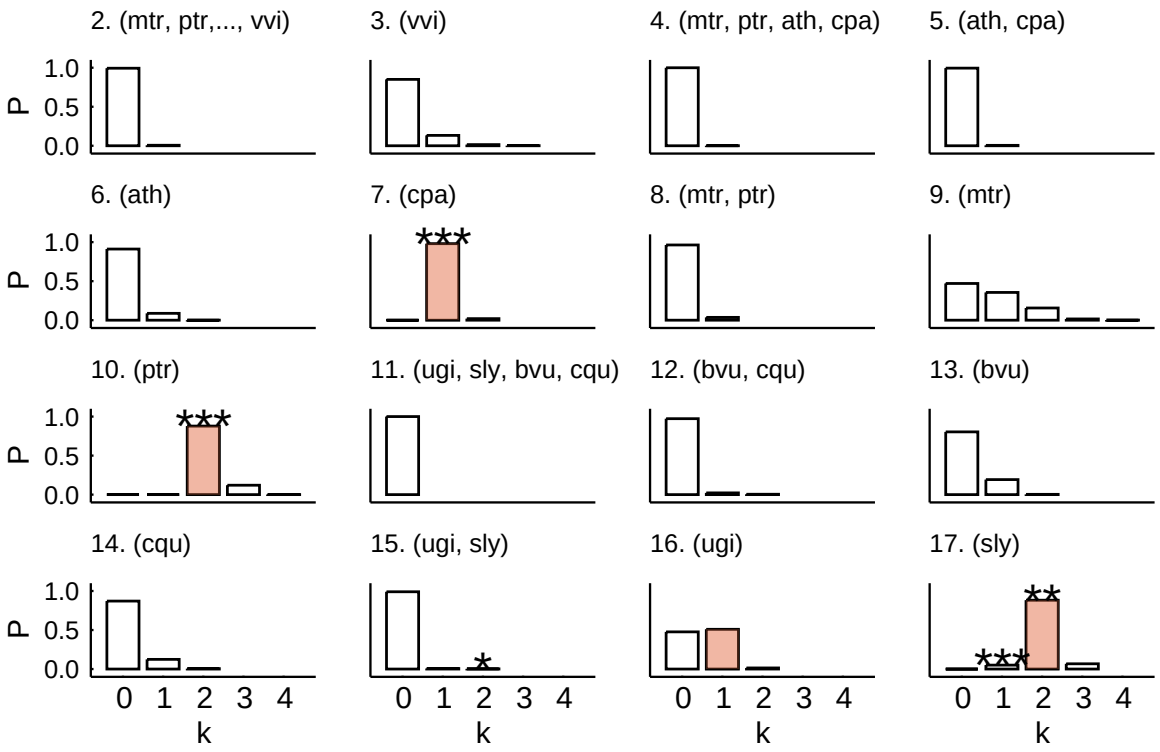

Figure S17: Posterior probabilities for the number of WGDs on each branch for the same simulated data as in fig. S16. Asterisks indicate the magnitude of the Bayes factor (*) $0.5 < \log_{10} K < 1$; (**) $1 < \log 10 K < 2$; (***) $\log 10 K > 2$. The true number of WGDs are marked in red.

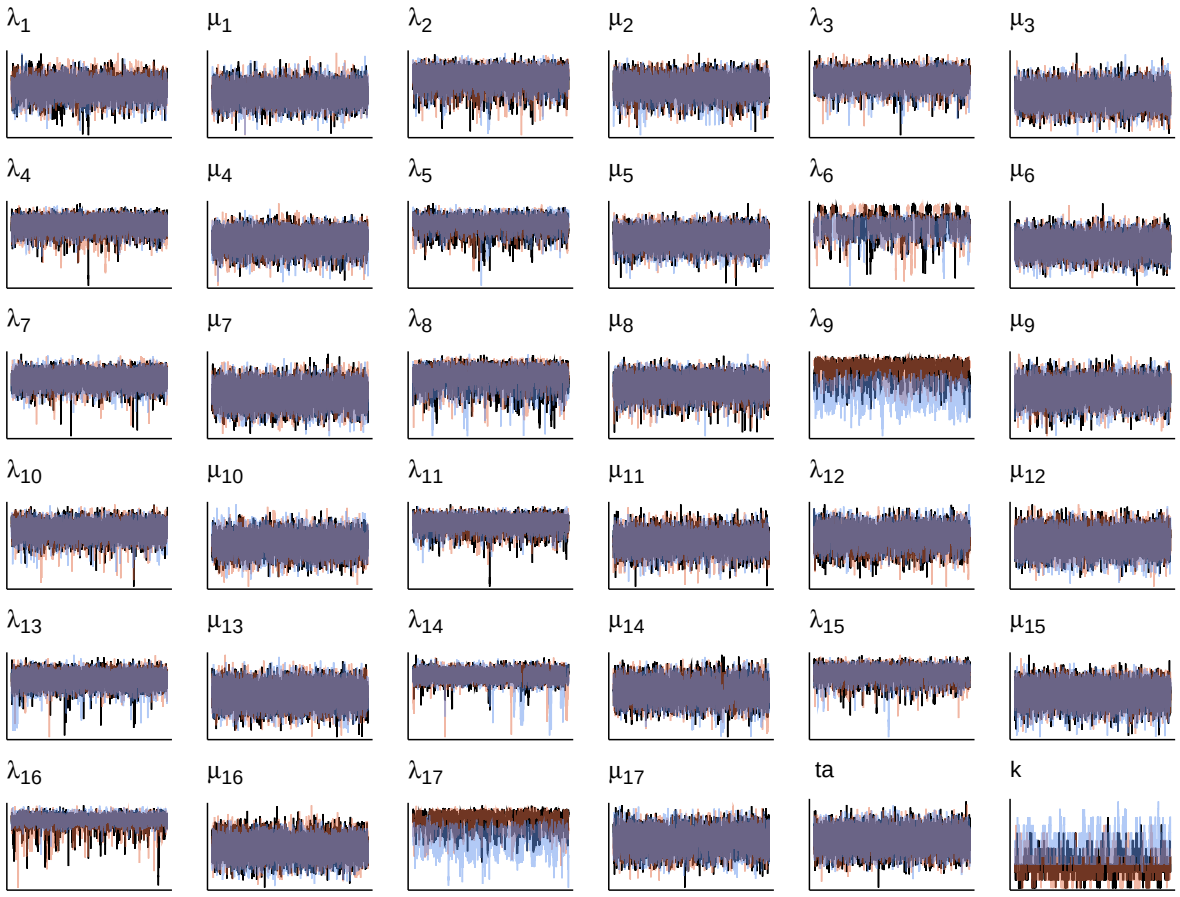

Figure S18: Trace plots for two realizations of the rjMCMC chain with $\delta = 0.1$ (black and red) and one realization with $\delta = 0.05$ (blue) based on 1000 gene families for the nine dicot species tree. Log-transformed duplication (λ) and loss (μ) rates are shown for each branch of the species tree. k denotes the number of WGDs in the model. Traces are based on 50.000 iterates from the rjMCMC algorithm.

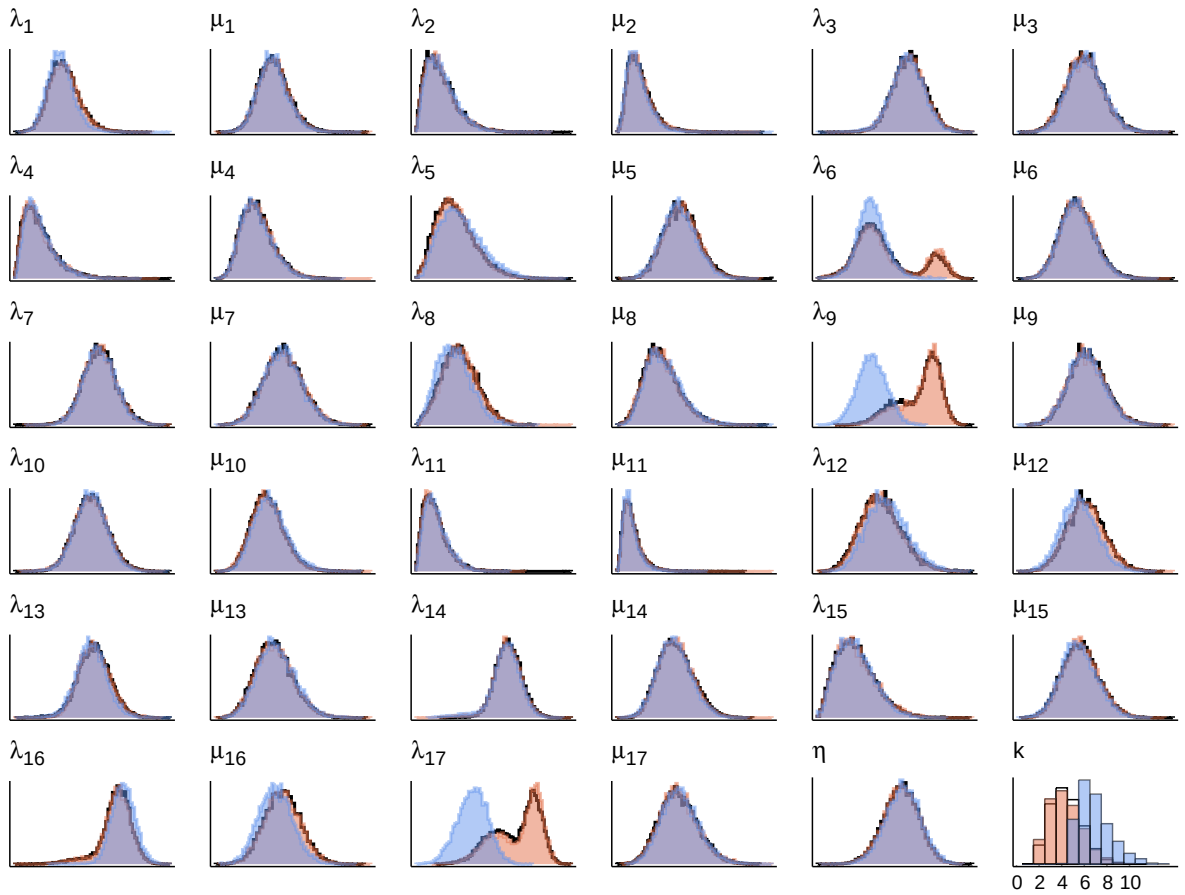

Figure S19: Approximate marginal posterior distributions corresponding to the trace plots shown in fig. S18.

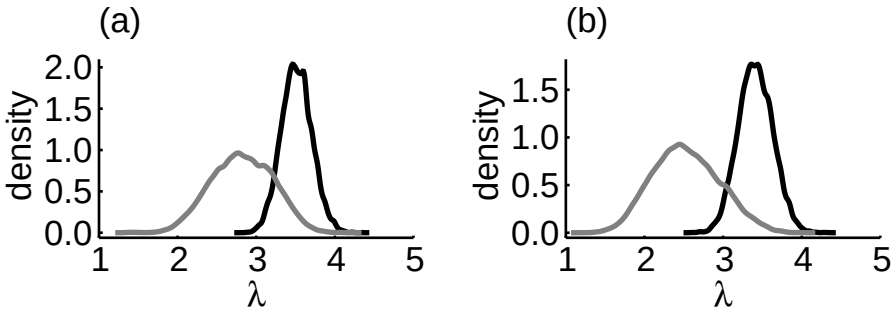

Figure S20: Conditional posterior duplication rates for (a) *Medicago* and (b) *Solanum* , conditioning on the no-WGD model (black) and one-WGD model (gray) for the results using the prior with $\delta = 0.1$. The x-axis is on a scale of number of events per gene lineage per billion years.

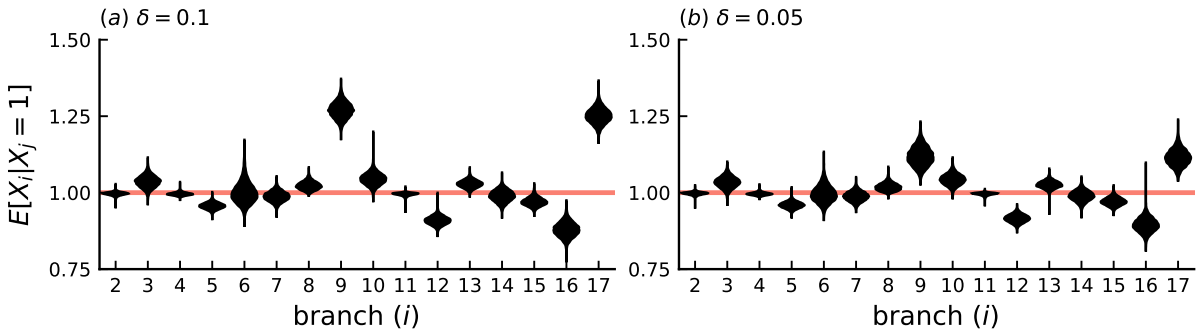

Figure S21: Posterior distributions for the expected number of genes per ancestral gene at the end of each branch under the SSDL process alone for the dicot dataset, showing results for two different prior settings, (a) $\delta = 0.1$ and (b) $\delta = 0.05$. For correspondence of branch numbers to clades, refer to fig. 3. The marginal posterior is shown over all different WGD models in the rjMCMC sample.

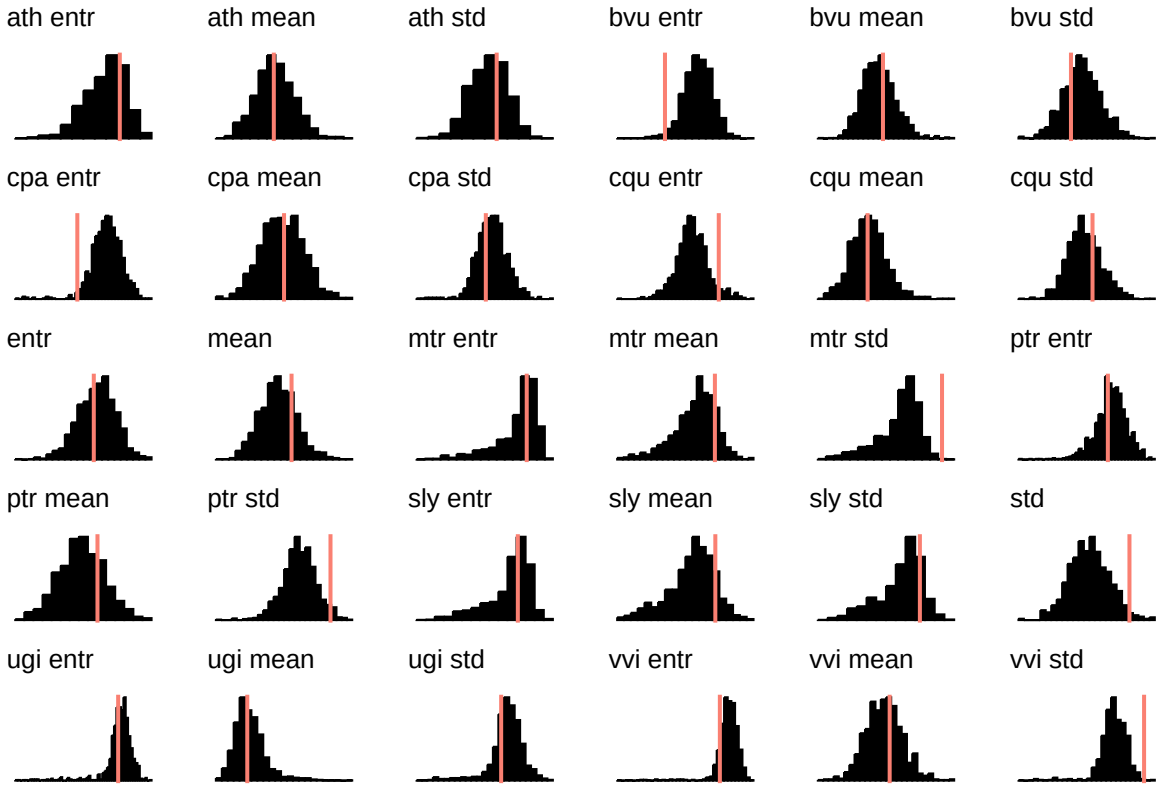

Figure S22: Posterior predictive distributions (histograms) and observed values (vertical lines) for 30 summary statistics of the phylogenetic profile matrix for the dicots data set with the $\delta = 0.1$ prior. Summary statistics are the mean, standard deviation (std) and entropy (entr) of the number of genes across families in the leaves of the species tree, as well as the total number of genes in a family.

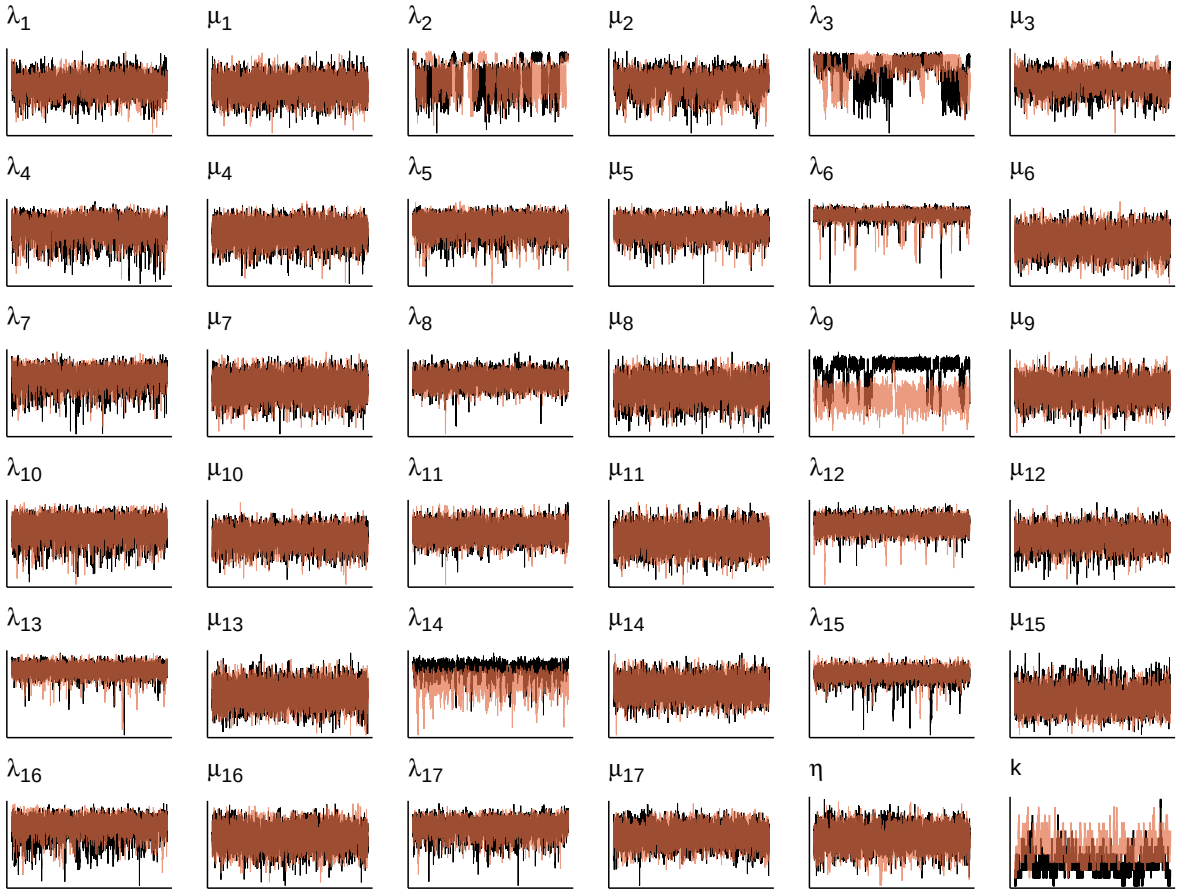

Figure S23: Trace plots for two realizations of the rjMCMC chain with 1000 gene families for the monocot species tree. In black a chain with $\delta = 0.1$ is shown whereas in orange a chain with $\delta = 0.05$ is shown. Log-transformed duplication (λ) and loss (μ) rates are shown for each branch of the species tree. k denotes the number of WGDs in the model. Traces are based on 50.000 iterates from the rjMCMC algorithm.

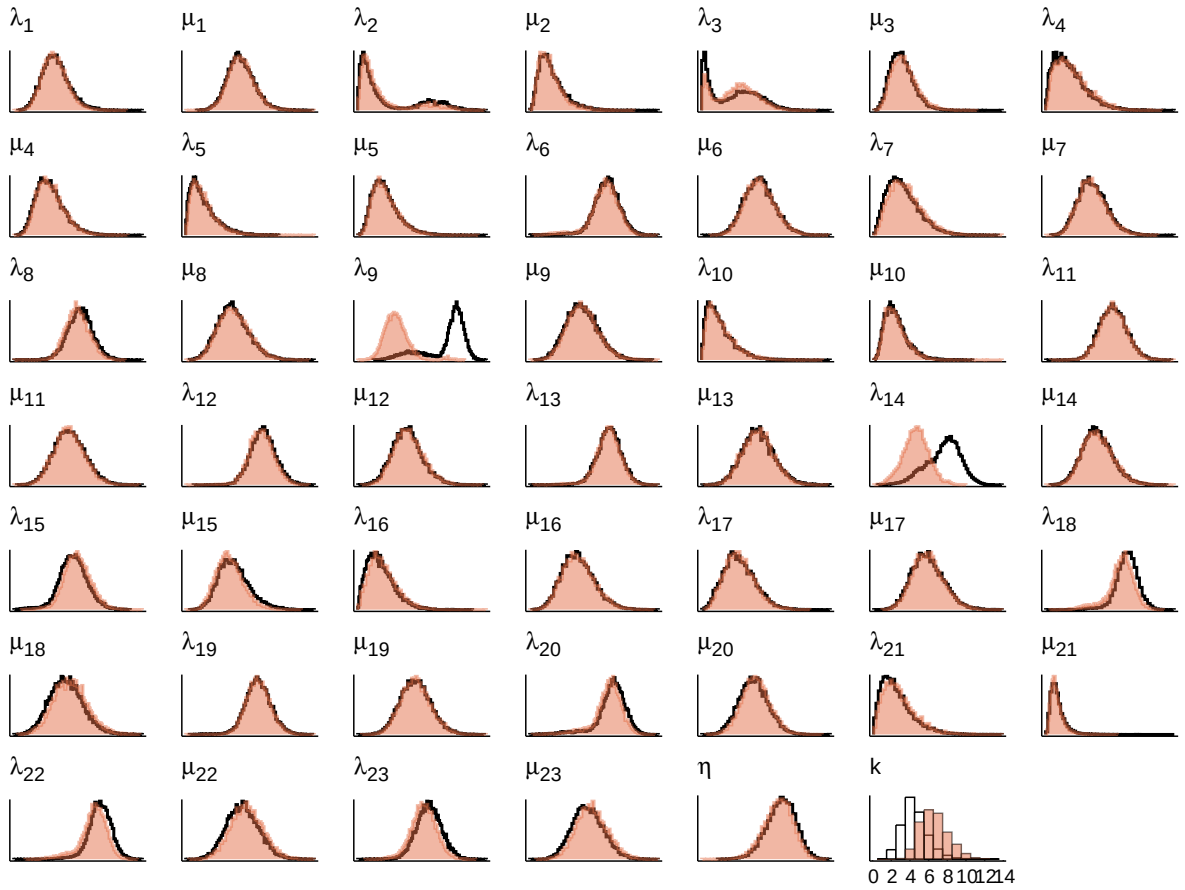

Figure S24: Approximate marginal posterior distributions for two realizations (black and red histograms) of the rjMCMC chain with 1000 gene families for the monocot species tree. Interpretation is as in fig. S23.

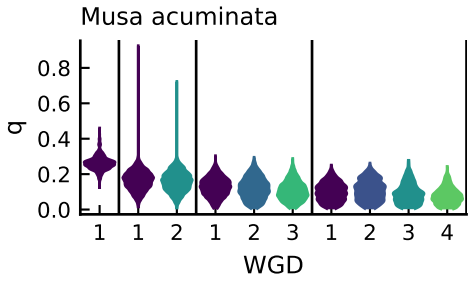

Figure S25: Marginal posterior distributions of retention rates for the *Musa acuminata* WGDs across different models. Vertical lines separate different models with different numbers of WGDs.

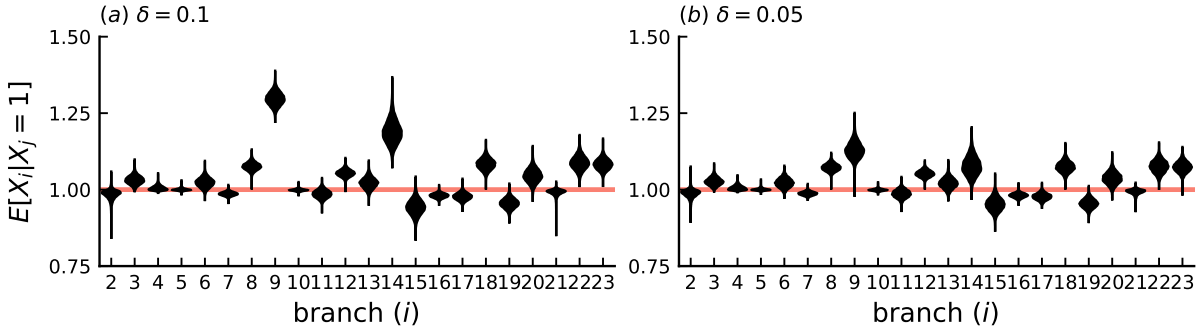

Figure S26: Posterior distributions for the expected number of genes per ancestral gene at the end of each branch under the SSDL process alone for the monocot dataset, showing results for two different prior settings, (a) $\delta = 0.1$ and (b) $\delta = 0.05$. For correspondence of branch numbers to clades, refer to figs. 4, 5. The marginal posterior is shown over all different WGD models in the rjMCMC sample.

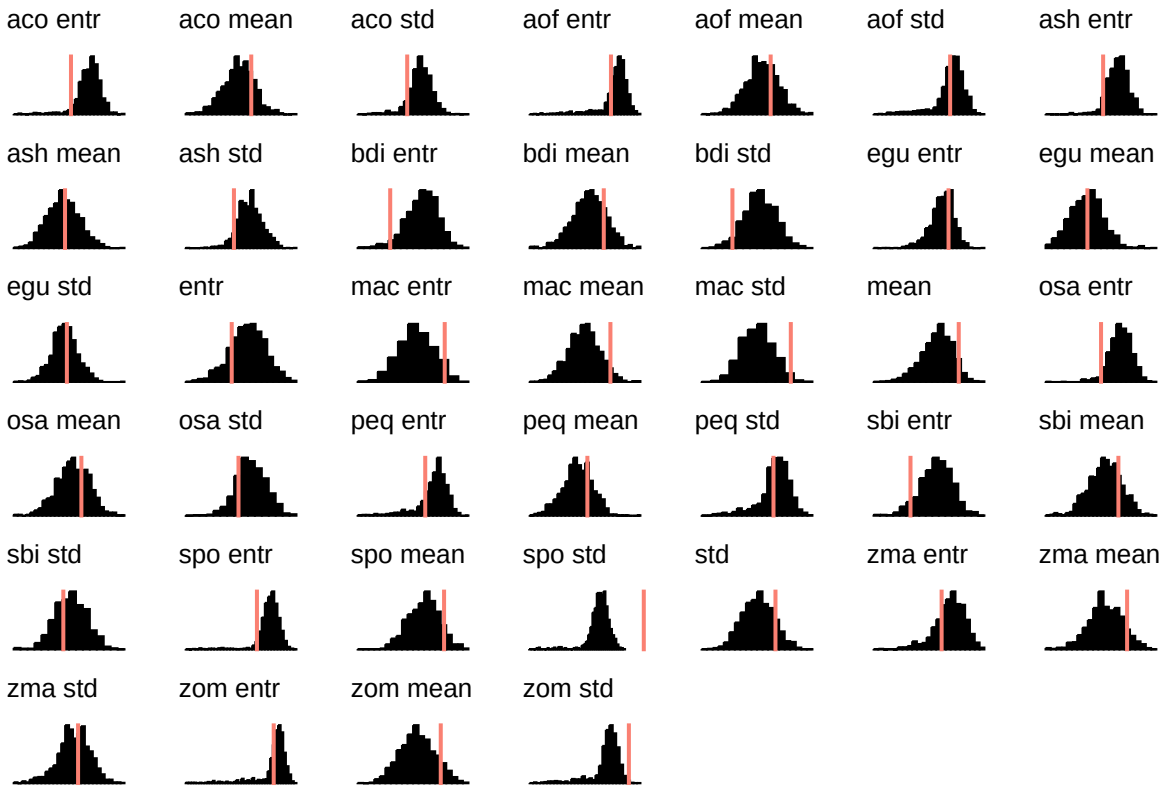

Figure S27: Posterior predictive distributions (histograms) and observed values (vertical lines) for 39 summary statistics of the phylogenetic profile matrix for the monocots data set. Summary statistics are the mean, standard deviation (std) and entropy (entr) of the number of genes across families in the leaves of the species tree, as well as the total number of genes in a family. These results are based on the chain with $\delta = 0.05$.