Interacting associations between ploidy, breeding system, and lineage diversification

Rosana Zenil-Ferguson[1,†], J. Gordon Burleigh[2], William A. Freyman[3], Boris Igić[4], Itay Mayrose[5], and Emma E. Goldberg[6]

[1]Department of Ecology, Evolution, and Behavior, University of Minnesota, Saint Paul, MN 55108, U.S.A.

[2]Department of Biology, University of Florida, Gainesville, FL 32611, U.S.A.

[3]23andMe, Inc., 899 W Evelyn Ave, Mountain View, CA 94041, U.S.A.

[4]Department of Biological Sciences, University of Illinois at Chicago, Chicago, IL 60607, U.S.A.

[5]Department of Molecular Biology and Ecology of Plants, Tel Aviv University, Tel Aviv, Israel

[6]Department of Ecology, Evolution, and Behavior, University of Minnesota, Saint Paul, MN 55108, U.S.A.

†Author for correspondence.

Running head: Ploidy and breeding systems in Solanaceae

Keywords: Polyploidy, Breeding System, Diversification, SSE models

Abstract

If particular traits consistently affect rates of speciation and extinction, broad macroevolutionary patterns can be understood as consequences of selection at high levels of the biological hierarchy. Identifying traits associated with diversification rate differences is tricky, though, because of the many traits available to consider and the statistical challenge of testing for associations from phylogenetic data. Two traits that have been repeatedly suggested as drivers of differential diversification are whether a lineage is diploid or polyploid, and whether it is self-incompatible or self-compatible. We investigate the role of each of these traits, and their interaction, on speciation and extinction rates in Solanaceae. We find that the effect of ploidy can largely be explained by its correlation with breeding system, and that additional unknown factors work with breeding system to determine diversification rates. These results are largely robust to assumptions about whether diploidization occurs. Finally, we show that allowing for state-dependent diversification affects conclusions about the relative contribution of different evolutionary pathways to self-compatible polyploids.

Introduction

Species accumulate around the tree of life at different rates. The search for traits that explain these differences has been accelerated by dramatic increases in phylogenetic data and, despite some setbacks (**???**), advances in analytical methods (Maddison et al. 2007; **?**; Goldberg and Igić 2012; Beaulieu and O'Meara 2016; **?**) are finding interesting lineages where potential interesting biological processes are changing the diversification patterns.

In these studies, it is common to identify a single focal trait and investigate its association with rates of speciation and extinction. This is problematic because the context in which traits occur can lead to complex interactions. Scientists are well aware of the contribution of complex interactions to the diversification process, thus, investigating multivariate traits linked to diversification is becoming the next best step in the development of new models (see Caetano et al. (2018); Herrera-Alsina et al. (2018)).

In this paper, we focus on two of the best studied traits in flowering plants, polyploidy and breeding system and their contributions to diversification with the goal of investigating how the complex interactions between these two traits can lead to problematic diversification patterns when considering them as independent.

Polyploidy events multiply the genomic content of cells, and they consequently have the potential to affect many or all traits and a great variety of evolutionary (**?**) and ecological processes(**?**). It is also a mutation that occurs commonly in plants, at it is widespread at both population and lineage scales(**??**). The prevalence of variation in chromosome number, and especially ploidy change, has been broadly considered a salient feature of flowering plants for nearly a century (**?**). The prevalence of polyploidy, its variation across clades, and its large effects on genotypes and phenotypes raise the hypothesis that polyploidy plays an important role in shaping rates of speciation and extinction. However, previous stochastic diversification models found a lower and negative rate of lineage diversification for polyploids Mayrose et al. (2011, 2015) when applied to 49 clades from the angiosperm phylogeny. This key finding lead to a the renaissance of the "polyploidy is an evolutionary dead-end" idea and initiated a discussion about the location and the diversification effect of whole genome duplications in flowering plants. Genomic evidence has shown that at the base of highly diverse angiosperm clades (including the root of all angiosperms (**?**)) it is possible to find at least a polyploidy event Soltis et al. (2014). Most recently, the genomic evidence from the 1KP make it possible for scientist to locate 106 whole genome duplications in the angiosperm phylogeny, and using a state depended diversification model that accounted for the frequency of whole genome duplications in a lineage they found that approximately 60-% of these whole genome duplications are enhancing diversification(Landis et al. 2018). This recent dramatic increases in the scale of available genome sequences uncovering ancient rounds of whole-genome duplications likeLandis et al. (2018) work bring into question the role of genome downsizing and diploidization in the diversification process(**??**). The presence of a significant and positive net diversification rate for diploids can be the result of a polyploid lineage that has been diploidized. Under this scenario, polyploidy could not longer be considered a macroevolutionary dead-end but instead, polyploidy would be the initiator of the diversification process. In the present work, we explore this polyploidy-diploidization scenario by adding diploidization rates in our proposed stochastic models.

Another approach useful to understand the differences between the net diversification processes of diploids and polyploids is the work of Beaulieu and O'Meara (2016). Beaulieu et al. proposed to allow for a 'hidden' or unspecified state linked to the trait of interest for models of trait linked to diversification. The presence of a hidden state can point out at sudden changes in the diversification rate under a value of a trait that are due to some unobserved source of heterogeneity.

The hidden-state models have effectively shown that traits that were believed to be drivers of diversification, in reality are not, and that it is an unobserved but associated trait responsible for differences that are initially found. When the hidden state and not the focal trait, is responsible for changes in the diversification process, it leaves behind the question of whether the hidden state corresponds to a real state of a trait that should be sought, or whether it is an approximation of some unknowable heterogeneity, as well as whether interactions between the known and hidden trait were modeled appropriately. Alternatively, one can simultaneously consider the effects of more than one known trait, especially in systems where multiple traits are suspected to influence diversification, and where interactions between those traits are well understood.

In the case of polyploidy, one of the most prominent complex interactions known is the association between ploidy and propensity for self-fertilization (**?**). In some cases, the evidence for a correlated shift in mating system along with polyploidization appears limited and sometimes contradictory (**???**). In other cases, however, polyploidy is not only a suspected correlate of breeding systems but indeed a causal link (**??**). Doubled number of alleles in pollen is thought to effect disruption of the genetic mechanisms in gametophytic self-incompatibility systems, which prevent self-fertilization (**???**). This creates a correlation between polyploidy and self-compatibility by precluding the existence of self-incompatible polyploids. In clades with these systems, it is thus natural to consider the simultaneous macroevolution of polyploidy and breeding system.

Breeding system shifts—changes in the collection of physiological and morphological traits that determine the likelihood that any two gametes unite—are remarkably common and affect the distribution and amount of genetic variation in populations (**??**). In particular, self-incompatibility (SI) systems cause a plant to reject its own pollen, and their loss, yielding self-compatibility (SC), is one of the most replicated transitions in flowering plant evolution. Previous phylogenetic analyses have reported higher rates of diversification for SI than for SC lineages (Goldberg et al. 2010; **?**), but they have not considered the possibility of other correlated traits driving this pattern. Given that changes in ploidy and breeding systems may be causally related and have profound affects on the fate of lineages, it seems particularly profitable to examine possible interactions in their macroevolutionary effects. This includes their joint influence on lineage diversification, and also potential patterns in the order of their transitions. For example, do losses of SI more commonly occur tied to polyploidization, or without a ploidy shift? Do polyploids arise more commonly from SI or SC diploids? Robertson et al. (2011) found that the pathway from SI diploids to SC polyploids is dominated by loss of SI followed later by polyploidization over long timescales, but proceeds in one step via polyploidization of SI species over short timescales. We revisit this question with a greatly improved phylogeny and methods that allow for diversification rate differences.

Here, we use extensive data on ploidy and breeding system in 595 Solanaceae taxa to investigate the associations of these two traits with lineage diversification. Considering each trait separately indicates that each is connected to diversification differences. Considering them jointly, however, reveals that the ploidy connection is removed by incorporating breeding system. We further show that the general results are robust to allowing for diploidization and a hidden trait, and something about pathways. Our results emphasize the importance of considering traits not only in isolation, especially when there are strong correlations between them.

Methods

Data

Chromosome number data were obtained for all Solanaceae taxa in the Chromosome Counts Database (CCDB; Rice et al. 2015), and the ca. 14,000 records were cleaned semi-automatically using the CCDBcurator R package (Rivero et al. 2019). This large dataset includes the compilation of Solanaceae ploidy states from Robertson et al. (2011). Species were coded as either diploid (D) or polyploid (P). For the majority of species, ploidy was assigned according to information from the original publications and the Kew Royal Botanic Gardens C-value DNA resource (Bennett and Leitch 2005). For taxa without ploidy information but with information about chromosome number, we assigned ploidy based on the multiplicity of chromosomes within the genus. For example, *Solanum betaceum* did not include information about ploidy level but it has 24 chromosomes, so because $x = 12$ is the base chromosome number of the *Solanum* genus (Olmstead and Bohs 2007), we assigned *S. betaceum* as diploid. Species with more than one ploidy level were assigned the smallest and most frequent ploidy level recorded. Breeding system was scored as self-incompatible (I) or self-compatible (C) based on results curated from the literature and original experimental crosses (as compiled in Igić et al. 2006; Goldberg et al. 2010; Robertson et al. 2011; Goldberg and Igić 2012). Most species could unambiguously be coded as either I or C (Raduski et al. 2012). Following previous work, we coded as I any species with functional I systems, even if C or dioecy was also reported. Dioecious species without functional I were coded as C.

To those existing data sets, we added some additional records for chromosome number and breeding system. The Supplementary Information contains citations for the numerous sources for the added data. Resolution of taxonomic synonymy followed the conventions provided in Solanaceae Source (PBI *Solanum* Project 2012). Hybrids and cultivars were excluded because ploidy and breeding system can be affected by artificial selection during domestication. Following the reasoning outlined in Robertson et al. (2011), we examined closely the few species for which the merged ploidy and breeding system data indicated the presence of self-incompatible polyploids. Although SI populations frequently contain some SC individuals, and diploid populations frequently contain some polyploid individuals, in no case did we find a convincing case of a naturally occurring SI and polyploid population. The single instance of an SI and polyploid individual appears to be an allopentaploid hybrid of *Solanum oplocense* Hawkes x *Solanum gourlayii* Hawkes, reported by **?**. Under exceedingly rare circumstances, it is possible for polyploids containing multiple copies of S-loci to remain SI, so long as they express a single allele at the S-locus (discussed in Robertson et al. 2011). Because of the resulting absence of SI and polyploid populations, as well as the linked functional explanation for disabling of gametophytic self-incompatibility systems with non-self recognition, following whole genome duplication (reviewed in Ramsey and Schemske 1998; Stone 2002), we consider only three observed character states: self-incompatible diploids (ID), self-compatible diploids (CD), and polyploids which are always self-compatible (CP).

Matching our character state data to the largest time-calibrated phylogeny of Solanaceae (Särkinen et al. 2013) yielded 595 species with ploidy and/or breeding system information on the tree. Binary or three-state classification of ploidy and breeding system for the 595 taxa is summarized in Fig. 1. We retained all of these species in each of the analyses below because pruning away tips lacking breeding system in the ploidy-only analyses (and vice versa) would discard data that could inform the diversification models. A total of 405 taxa without any information about breeding system or polyploidy were excluded. Tips without trait data are much less informative for diversification parameters linked to trait values. Including this many more species would have prohibitively slowed our analyses, especially those implementing the most complex models.

Models for ploidy and diversification

To investigate the association between ploidy level and diversification, we first defined a binary state speciation and extinction model (BiSSE, Maddison et al. 2007) in which taxa were classified as diploid (D) or polyploid (P) (Fig. 1). We call this the *D/P ploidy* model. In a Bayesian framework, we obtained posterior probability distributions of speciation rates (λ_D, λ_P), extinction rates (μ_D, μ_P), net diversification rates ($r_D = \lambda_D - \mu_D$, $r_P = \lambda_P - \mu_P$), and relative extinction rates ($\nu_D = \mu_D/\lambda_D$, $\nu_D = \mu_D/\lambda_D$) associated with each state. This analysis explores the same question as Mayrose et al. (2011, 2015), but our analyses differ because we include not only polyploidization (parameter ρ, the transition rate from D to P), but also diploidization (parameter δ, the transition rate from P to D).

Our second model assesses the signal of diversification due to ploidy differences while also parsing out the heterogeneity of diversification rates due a possible unobserved trait. BiSSE-like models can suffer from a large false discovery rate because they fail to account for diversification rate changes that do not directly depend on the trait of interest (**?**Beaulieu and O'Meara 2016). Diversification rate differences explained by something (trait) other than ploidy, are accommodated by adding a hidden state (HiSSE model; Beaulieu and O'Meara 2016). In this model, each of the observed diploid and polyploid states is subdivided by a binary hidden trait with states *A* and *B*. We call this the *D/P+A/B ploidy and hidden state* model. We estimated the posterior probability distributions of speciation rates (λ_{D_A}, λ_{D_B}, λ_{P_A}, λ_{P_B}), extinction rates (μ_{D_A}, μ_{D_B}, μ_{P_A}, μ_{P_B}), net diversification rates (r_{D_A}, r_{D_B}, r_{P_A}, r_{P_B}), and relative extinction rates (ν_{D_A}, ν_{D_B}, ν_{P_A}, ν_{P_B}). In this model polyploidization rate ρ and diploidization rate δ are also included, and changes between hidden states are symmetrical with rate α.

Models for breeding system and diversification

To assess the effects of breeding system in the diversification process, we first fit model in which the states are self-incompatible (I) or self-compatible (C). This is the same as the analysis of Goldberg et al. (2010), save for an updated phylogeny (Särkinen et al. 2013). We call this BiSSE model the *I/C breeding system* model. To parse out the effect of breeding system on diversification, while allowing for the possibility of heterogeneous diversification rates unrelated to breeding system, we subdivided each of those states into hidden states *A* and *B*. We call this HiSSE model the *I/C+A/B breeding system and hidden state model.*

For all breeding system models, we allow transitions from *I* to *C* (at rate q_{IC}) but not the reverse. Within Solanaceae, self-incompatibility is homologous in all species in which S-alleles were cloned, and controlled crosses performed. All species sampled to date, possess a non-self recognition, RNase-based, gametophytic self-incompatibility (shared even with other euasterid families; **?**). Furthermore, species that are distantly related within this family carry closely-related alleles, with deep trans-specific polymorphism, at the S-locus, which controls the SI response (?Igić et al. 2006). This represents very strong evidence that the SI mechanism, and our *I* state is ancestral to the Solanaceae, and did not arise independently within the family ($q_{CI} = 0$).

Models for ploidy, breeding system, and diversification

If ploidy and breeding system each influence lineage diversification individually, it is logical to examine their possible joint effects. We thus fit a multi-state model that includes both traits (MuSSE, FitzJohn 2012). The three states in this model are self-incompatible diploids (ID), self-compatible diploids (CD), and polyploids, which are always self-compatible (CP). As explained above, we did not include a state for self-incompatible polyploids because they are not observed in the data, and that trait combination state is mechanistically predicted not to occur. We call this the *ID/CD/CP ploidy and breeding system* model. The model has 10 parameters, six for diversification in each state (λ_{ID}, λ_{CD}, λ_{CP} for speciation, μ_{ID}, μ_{CD}, μ_{CP} for extinction) and four for transitions between states (ρ_I, ρ_C for polyploidization transitions from *ID* and *CD* to *CP*, respectively; δ for diploidization from *CP* to *CD*; q_{IC} for loss of self-incompatibility without polyploidization, from *ID* to *CD*). The total rate of loss of self-incompatibility, i.e., transitions out of *ID*, is $q_{IC} + \rho_I$. Diploidization from *CP* to *ID* is not allowed because it would represent a simultaneous regain of SI.

The ID/CD/CP model could potentially capture similar dynamics as earlier models, if the effects of the hidden state in D/P+A/B were effectively caused by breeding system (or its correlates), and the hidden state in I/C+A/B was effectively caused by ploidy. There is also the potential, however, for a hidden factor to be influencing diversification beyond both of our focal traits, and this could again mislead inferences. We therefore added a hidden trait layer on top of our three-state model (analogous to Caetano et al. 2018; Herrera-Alsina et al. 2018; **?**). We refer to this as the *ID/CD/CP+A/B* model. A fully parameterized version of this model would have 26 rate parameters (Herrera-Alsina et al. 2018). Because our goal was to look for diversification rate differences associated with ploidy and breeding system rather than the specific effects of the hidden states, we fitted a simplified version with 16 parameters. The reduction in parameter space is achieved by fixing the rates for transitions among hidden states to be equal with rate α, and fixing the transition rates between observed states to be independent of the hidden state (rates ρ_I, ρ_C, δ, q_{IC} as defined for the ID/CD/CP model). There are additionally twelve diversification rate parameters (λ_{ID_A}, λ_{ID_B}, λ_{CD_A}, λ_{CD_B}, λ_{CP_A}, λ_{CP_B}, μ_{ID_A}, μ_{ID_B}, μ_{CD_A}, μ_{CD_B}, μ_{CP_A}, μ_{CP_B}).

Pathways to polyploidy

Considering ploidy and breeding system together, there are two evolutionary pathways from SI diploid to SC polyploid (**?**Robertson et al. 2011). In the one-step pathway, the CP state is produced directly from the ID state when whole genome duplication disables SI. In the two-step pathway, the CD state is an intermediate: SI is first lost, and later the SC diploid undergoes polyploidization. We quantify the relative contribution of these pathways to polyploidy in two ways, each using the MAP rate estimates from the IC/CD/CP model.

Both of our methods are based on a propogation matrix that describes flow from ID to CP, as in Robertson et al. (2011). We insert an artificial division in the CP state, so that one substate contains the CP species that arrived via the one-step pathway and the other substate contains the CP species that arrived via the two-step pathway. We consider only unidirectional change along each step of the pathway in order to separate them into clear alternatives, and because in this family there is no support for regain of SI, and not strong support for diploidization. First, we consider only the rates of transitions between these states, placing them in the propogation matrix Q. The matrix $\mathsf{P} = \exp(\mathsf{Q}t)$ then provides the probabilities of changing from one state to any other state after time t. Closed-form solutions for the two pathway probabilities are provided in Robertson et al. (2011). Our results will differ from those of Robertson et al. (2011) because our transition rate estimates come from a dated phylogeny and a model that allows for state-dependent diversification. Second, we consider not only transitions between states but also diversification within each state. State-dependent diversification can change the relative contributions of the two pathways. In particular, if the net diversification rate is small for CD, the two-step pathway will contribute relatively less. We therefore include the difference between speciation and extinction along the diagonal elements of the propogation matrix. As before matrix exponentiation provides the relative chance of changing from one state to any other state after time t. In this case, however, these are not probabilities because diversification changes the number of lineages as time passes. We can still use their ratios to consider the relative contribution of each pathway, though, analogous to the normalized age structure in a growing population.

Diploidization as an exploratory hypothesis

For all four models that consider ploidy changes, we allowed diploidization. Previous modeling approaches (Mayrose et al. 2011) have argued against inferring diploidization rates when using ploidy data that comes from classifications based on chromosome number multiplicity or chromosome number change models like chromEvol (**?**). These types of classifications do not allow for a ploidy reversion. Where indicated, the classification of ploidy for the data used in our models was based on chromosome multiplicity at the genus level. However, the majority of the ploidy classifications were adopted from original studies with alternative sources of information (e.g., geographic distribution, genus ploidy distribution) where ploidy was defined by authors that found evidence for it. Since it is not clear whether diploidization can be detected under alternative ploidy classifications or even classifications based on chromosome number multiplicity at the genus level, we also fit the models without diploidization in order to test whether the conclusions about diversification are sensitive to including diploidization. As discussed by Servedio et al. (2014), the presence or absence of a hypothesis can have an exploratory goal. In our case the diploidization parameter (or its absence, $\delta = 0$) in our models is an opportunity to explore an assumption that might be important but that is not the single definitive process to understand the interactions among polyploidy, breeding system, and diversification.

Statistical inference under the models

Parameters for each of the 10 diversification models were estimated using custom code in the RevBayes (Höhna et al. 2016) environment. Code for analyses and key results is available at `https://github.com/roszenil/solploidy`. We included a correction for incomplete sampling in all analyses, based on assuming that the Solanaceae family has approximately 3,000 species ($s = 595/3000$) as estimated by the Solanaceae Source project (PBI *Solanum* Project 2012). For all 10 models, we assumed that speciation and extinction parameters had log-normal prior distributions with means equal to the expected net diversification rate (number of taxa$/[2 \times$ root age$]$) and standard deviation 0.5. Priors for parameters defining trait changes were assumed to be gamma distributed with parameters $k = 0.5$ and $\theta = 1$. For each model, an MCMC chain was run for 96 hours in the high-performance computational cluster at the Minnesota Supercomputing Institute, which allowed for 5,000 generations of burn-in and a minimum of 200,000 generations of MCMC for each of the 6 models. For each model, convergence and mixing of the MCMC was tested using the R library `coda` and the software package Tracer (see supplementary information for convergence plots).

Model selection

We calculated the marginal likelihood for each of the 10 models in RevBayes (Höhna et al. 2016). Marginal likelihoods were calculated using 50 stepping stone steps under the methodology of Xie et al. (2010). Each stepping stone step was found by calculating 500 generations of burn-in followed by a total of 1,000 MCMC steps (Table 1). The calculation of each marginal likelihood ran for 24 hours on a high-performance computational cluster.

Using the marginal likelihood values, we calculated thirteen different Bayes factors. Six compared the models of ploidy against one other (D/P and D/P+A/B, each with or without diploidization), one compared the breeding system models (I/C and I/C+A/B), and six compared the models with both traits

(ID/CD/CP and ID/CD/CP+A/B, each with or without diploidization) (Table 2). Other comparisons between these models are not valid because the input data are different under the different state space codings (Fig. 1). In mathematical terms, the D/P, I/C, and ID/CD/CP state spaces are not 'lumpable' with respect to one another (Tarasov 2018). Each model comparison is reported with a Bayes factor on the natural log scale: the comparison between models M_0 and M_1 is $BF(M_0,\ M_1) = \ln[P(\mathbf{X}|M_0) - P(\mathbf{X}|M_1]$. There is 'positive' support for M_0 when this value is more than 2, 'strong' support when it is more than 6, and 'very strong' support when it is more than 10 (**?**).

Results

Traits and diversification

Individual trait-dependent diversification inference, of ploidy and breeding system considered separately, each showed significant associations with net diversification differences. When considered together, however, the effect of breeding system dominated the effect of ploidy, although hidden factors played an important role as well.

When considering ploidy alone (D/P model), we found a greater net diversification rate for diploids than for polyploids, in agreement with (Mayrose et al. 2011, 2015). This result holds with (Fig. 2A) or without (Fig. 3A) the diploidization parameter, though including diploidization shifts the net diversification rate of polyploids to be non-negative. Incorporating a hidden state, however, removes the clear separation in diversification between diploids and polyploids (D/P+A/B model; Fig. 2B, Fig. 3B). Thus, differences in net diversification are better explained by an unknown factor than by ploidy. Statistical model comparisons show very strong support for including the hidden state and strong support for including diploidization (Table 2).

When considering breeding system alone (I/C model, Fig. 2C), we found a larger net diversification rate for SI than for SC species, in agreement with Goldberg et al. (2010). When a hidden state is included (I/C+A/B model), the large net diversification difference persists for one of the hidden states but is removed for the other (Fig. 2D). Thus, differences in net diversification are best explained by both breeding system and an unknown factor. The statistical model comparison shows very strong support for including the hidden state (Table 2).

When considering ploidy and breeding system together (ID/CD/CP model), the net diversification rate for SI diploids was greater than for either SC diploids or SC polyploids, with or without diploidization (Fig. 2E, Fig. 3C). It thus appears that the difference in net diversification with breeding system persists when ploidy is included in the model, but not the reverse. The association of ploidy with net diversification in the

D/P model (Fig. 2A, Fig. 3A) appears to be driven by the subset of diploids that are SI; among SC species, net diversification rates for diploids and polyploids are similar. When a hidden state is included (IC/CD/CP+A/B model), the same general pattern remains when diploidization is prevented (Fig. 3D), although the higher net diversification rate of ID is less clear within one of the hidden states. With diploidization, the net diversification rate of ID is still greater than CD within each hidden state, but diversification for P is highly uncertain and perhaps bimodal. Statistical model comparisons show very strong support for including the hidden state and at most positive support for including diploidization (Table 2).

Pathways to polyploidy

Evolution from SI diploids to SC polyploids can proceed through two different pathways. Determining the relative contribution of these pathways based on our estimated transition rates from the ID/CD/CP model, we find that the one-step pathway is more likely on short timescales and the two-step pathway is more likely on long timescales (Fig. 4, left panels). When not much time has elapsed, the one-step pathway is more likely because only one event need happen. When more time has elapsed, the two-step pathway is more likely because the rate of loss of SI within diploids, q_{IC}, is greater than the rate of polyploidization for SI species, ρ_I (Fig. S8). These conclusions are the same as those of Robertson et al. (2011) under one of their branch length approximations.

When rates of lineage diversification are considered, however, the conclusions change. Even over long timescales, the two-step pathway contributes less (Fig. 4, right panels). The lower rate of net diversification in the CD state, relative to ID, means that relatively fewer lineages are available to complete the second step of the two-step pathway.

Diploidization as an exploratory hypothesis

We considered models both with or without diploidization in order to explore the effects of this process on the estimates of state-dependent diversification. For the two models that only include diploid and polyploid states (D/P and D/P+A/B), the net diversification rate of polyploid lineages is likely negative when diploidization (parameter δ) is excluded but positive when δ is included (Fig. 2A versus Fig. 3A). For the two models that also included breeding system (ID/CD/CP and ID/CD/CP+A/B), the main effect of including diploidization is increasing the uncertainty of the estimate of polyploid net diversification rate (Fig. 2EF versus Fig. 3CD). For all models, there is greater uncertainty in the estimate of diploidization rate than polyploidization rate, as judged by the width of the credibility intervals (see supplementary information figures).

Discussion

The present work shows the importance of considering the trait linked diversification patterns under a multivariate approach. Species are created and go extinct based un multiple and often highly correlated phenotypes, understanding the speciation and extinction processes requires understanding of the evolutionary consequences that those trait correlations produce in organisms. In the present work we show how considering both polyploidy and breeding system can disentangle the importance (or the lack of) of polyploidy when confronted with the evidence brought by breeding system.

Using the most complete dataset for polyploidy in a phylogenetic tree in Solanaceae, we were able to replicate the results found by Mayrose et al. (2011), polyploids have a slower net diversification compare to diploids. Furthermore, we also found polyploids had a high probability of having negative diversification which implies that polyploids can become a macroevolutionary dead-end, a result that was also found in the two large angiosperm diversification studies Mayrose et al. (2011) and Mayrose et al. (2015). However, we expanded this study to accommodate background heterogeneity in the diversification process. When adding heterogeneity we found that it was more likely that an unobserved trait linked to diploid state was the one leading the net diversification patterns, and that there were some “second-class" diploids that were not different from polyploids in diversification terms (Figure 2A). This result lead us to our central question: *what is that other trait linked diploids that makes them different in the diversification process?* .

In Solanaceae, our immediate intuition was to look into breeding system. Previous studies shown that self-incompatible Solanaceae species have also higher rates of diversification compared to their self-compatible counterparts (Goldberg and Igić 2012). Self-incompatible species are diploid in our sample and also expected to be diploid due to ... (citation?). By considering both polyploidy and breeding system simultaneously for every species in our sample, it was possible to disentangle why some diploids were quantitatively different than polyploids. In the three-state diversification model ID/CD/CP, we found that self-incompatible diploids have faster and positive rates than self-compatible diploids and polyploids, and that the difference between the rates of net diversification of self-compatible diploids is not as large (Figures 2C) as first found by binary trait diversification models (Figures 2A). This result is important, since it aligns with the net diversification results of the D/P+A/B model, where a ‘ ‘hidden-trait" seem to be dictating the diversification pattern. By adding breeding system, we were able to hint at which that hidden-trait possibly is. Therefore, we consider that finding a heterogenous result in the hidden trait approaches should be be treated as evidence of a second trait that is necessary to consider. Pursuing knowledge of such trait can result on a clearer picture of the importance of trait linked diversification patterns, but also on a better reconstruction on past events in phylogenies.

Diploidization needs to be considered in diversification models because higher rates of net diversification (speciation minus extinction) in diploids can be obtained from models that ignore the possibility of a polyploid lineage that has diploidized being the diversification enhancer. Hence, in a diploidization lacking model, diploids will show higher rates of diversification when in fact it is a polyploidy event and its subsequent diploidization that generated higher speciation and less extinction. By adding diploidization to models of polyploidy linked to diversification it is possible to recover this complicated scenario and to reconcile genomic evidence with stochastic models.

Acknowledgements

NSF DEB-1655478. The Minnesota Supercomputing Institute (MSI) at the University of Minnesota provided computing resources for this project.

Literature Cited

Beaulieu, J. M. and B. C. O'Meara, 2016. Detecting hidden diversification shifts in models of trait-dependent speciation and extinction. Syst Biol 65:583–601.

Bennett, M. D. and I. J. Leitch, 2005. Plant DNA C-values database.

Caetano, D. S., B. C. O'Meara, and J. M. Beaulieu, 2018. Hidden state models improve state-dependent diversification approaches, including biogeographical models. Evolution 72:2308–2324.

FitzJohn, R. G., 2012. Diversitree : comparative phylogenetic analyses of diversification in r. Methods Ecol Evol 3:1084–1092.

Goldberg, E. E. and B. Igić, 2012. Tempo and mode in plant breeding system evolution. Evolution 66:3701–3709.

Goldberg, E. E., J. R. Kohn, R. Lande, K. A. Robertson, S. A. Smith, and B. Igić, 2010. Species selection maintains self-incompatibility. Science 330:493–495.

Herrera-Alsina, L., P. van Els, and R. S. Etienne, 2018. Detecting the dependence of diversification on multiple traits from phylogenetic trees and trait data. Systematic biology .

Höhna, S., M. J. Landis, T. A. Heath, B. Boussau, N. Lartillot, B. R. Moore, J. P. Huelsenbeck, and F. Ronquist, 2016. RevBayes: Bayesian phylogenetic inference using graphical models and an interactive model-specification language. Syst Biol 65:726–736.

Igić, B., L. Bohs, and J. R. Kohn, 2006. Ancient polymorphism reveals unidirectional breeding system shifts. Proc Natl Acad Sci USA 103:1359–1363.

Igić, B. and J. W. Busch, 2013. Is self-fertilization an evolutionary dead end? New Phytol 198:386–397.

Landis, J. B., D. E. Soltis, Z. Li, H. E. Marx, M. S. Barker, D. C. Tank, and P. S. Soltis, 2018. Impact of whole-genome duplication events on diversification rates in angiosperms. Am J Bot 105:348–363.

Maddison, W. P., P. E. Midford, and S. P. Otto, 2007. Estimating a binary character's effect on speciation and extinction. Syst Biol 56:701–710.

Mayrose, I., S. H. Zhan, C. J. Rothfels, N. Arrigo, M. S. Barker, L. H. Rieseberg, and S. P. Otto, 2015. Methods for studying polyploid diversification and the dead end hypothesis: a reply to soltis et al. (2014). New Phytol 206:27–35.

Mayrose, I., S. H. Zhan, C. J. Rothfels, K. Magnuson-Ford, M. S. Barker, L. H. Rieseberg, and S. P. Otto, 2011. Recently formed polyploid plants diversify at lower rates. Science 333:1257.

Olmstead, R. G. and L. Bohs, 2007. A summary of molecular systematic research in Solanaceae: 1982–2006. Acta Horticulturae Pp. 255–268.

PBI *Solanum* Project, 2012. Solanaceae Source: a global taxonomic resource for the nightshade family.

Raduski, A. R., E. B. Haney, and B. Igić, 2012. The expression of self-incompatibility in angiosperms is bimodal. Evolution 66:1275–1283.

Ramsey, J. and D. W. Schemske, 1998. Pathways, mechanisms, and rates of polyploid formation in flowering plants. Annual Review of Ecology and Systematics 29:467–501.

Rice, A., L. Glick, S. Abadi, M. Einhorn, N. M. Kopelman, A. Salman-Minkov, J. Mayzel, O. Chay, and I. Mayrose, 2015. The chromosome counts database (CCDB) - a community resource of plant chromosome numbers. New Phytol 206:19–26.

Rivero, R., E. B. Sessa, and R. Zenil-Ferguson, 2019. Eyechrom and ccdb curator: Visualizing chromosome count data from plants. Applications in Plant Sciences P. e01207.

Robertson, K., E. E. Goldberg, and B. Igić, 2011. Comparative evidence for the correlated evolution of polyploidy and self-compatibility in solanaceae. Evolution 65:139–155.

Särkinen, T., L. Bohs, R. G. Olmstead, and S. Knapp, 2013. A phylogenetic framework for evolutionary study of the nightshades (Solanaceae): a dated 1000-tip tree. BMC Evol Biol 13:214.

Servedio, M. R., Y. Brandvain, S. Dhole, C. L. Fitzpatrick, E. E. Goldberg, C. A. Stern, J. Van Cleve, and

D. J. Yeh, 2014. Not just a theory?the utility of mathematical models in evolutionary biology. PLoS biology 12:e1002017.

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 mayroseetăal . (2011). New Phytol 202:1105–1117.

Stone, J. L., 2002. Molecular mechanisms underlying the breakdown of gametophytic self-incompatibility. The Quarterly Review of Biology 77:17–30.

Tarasov, S., 2018. Integration of anatomy ontologies and evo-devo using structured markov models suggests a new framework for modeling discrete phenotypic traits. BioRxiv P. 188672.

Xie, W., P. O. Lewis, Y. Fan, L. Kuo, and M.-H. Chen, 2010. Improving marginal likelihood estimation for bayesian phylogenetic model selection. Systematic biology 60:150–160.

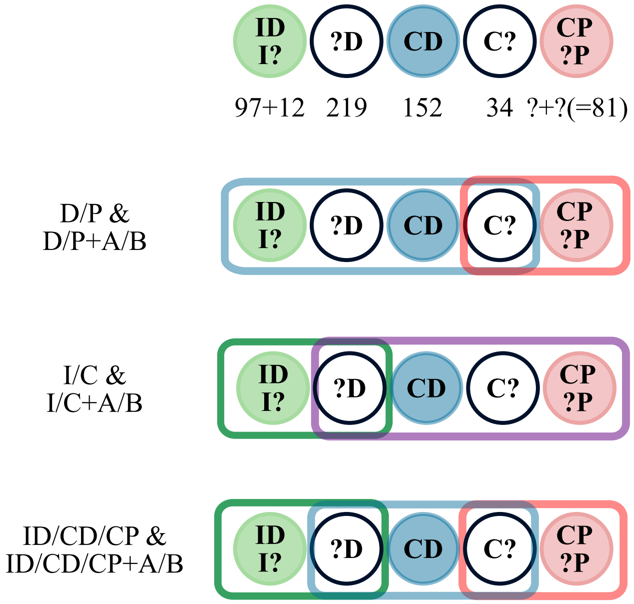


Figure 1: Character states used for each of the models. Each species retained on the tree belonged to one of five possible categories, depending on whether ploidy and/or breeding system were known. The number of species in each is shown under the corresponding circles in the top row. These categories were then grouped in a manner appropriate to the states of each model. For example, there are 34 species that are self-compatible and of unknown ploidy; these are coded as either *D* or *P* in the D/P models (uncertain, or consistent with either state), as *C* in the I/C models, and as either *CD* or *CP* in the ID/CD/CP models. In all cases, species were coded as either *A* or *B* in the hidden state models. The models depicted in the different rows use state spaces that are not comparable with one another. For example, we cannot test whether the D/P model fits better than the I/C model because they use states that are not the same and are not 'lumpable' (Tarasov 2018).

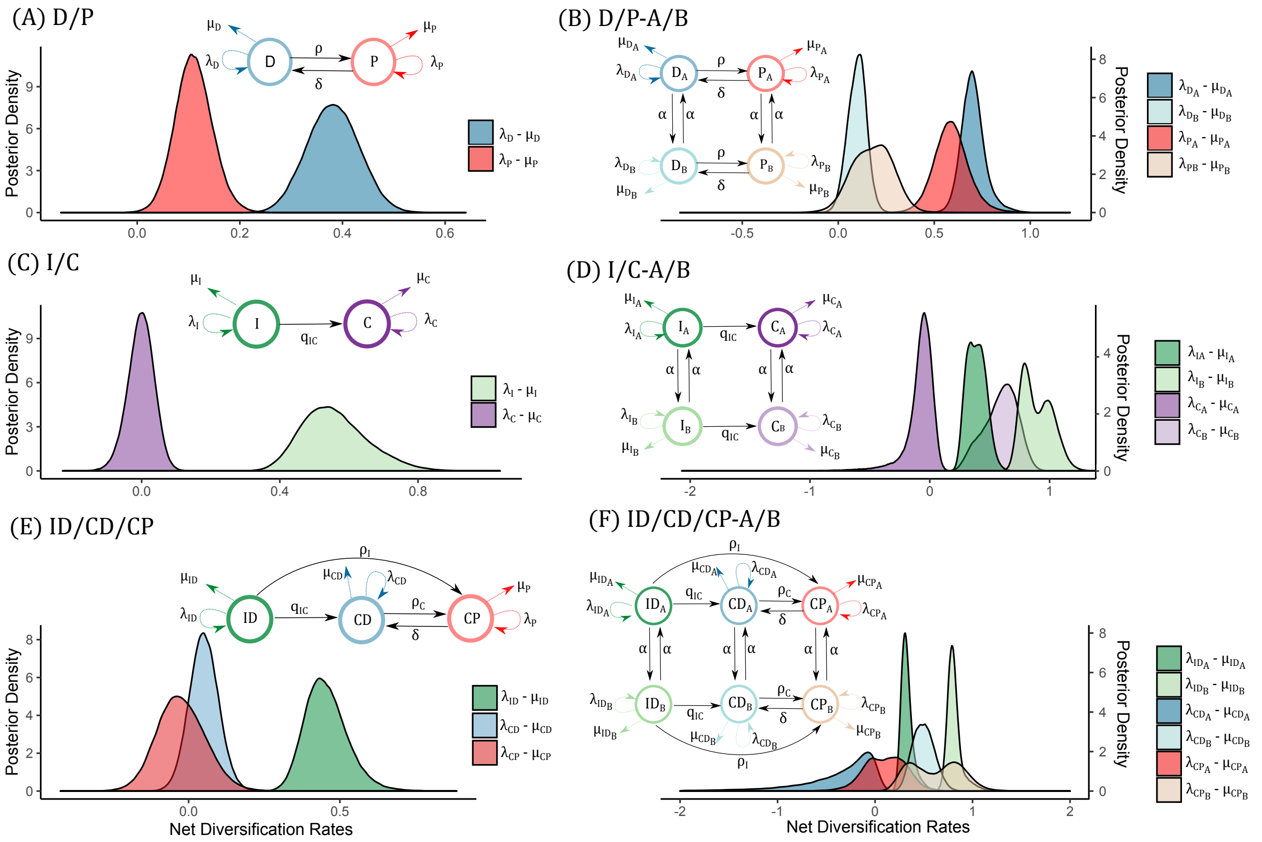
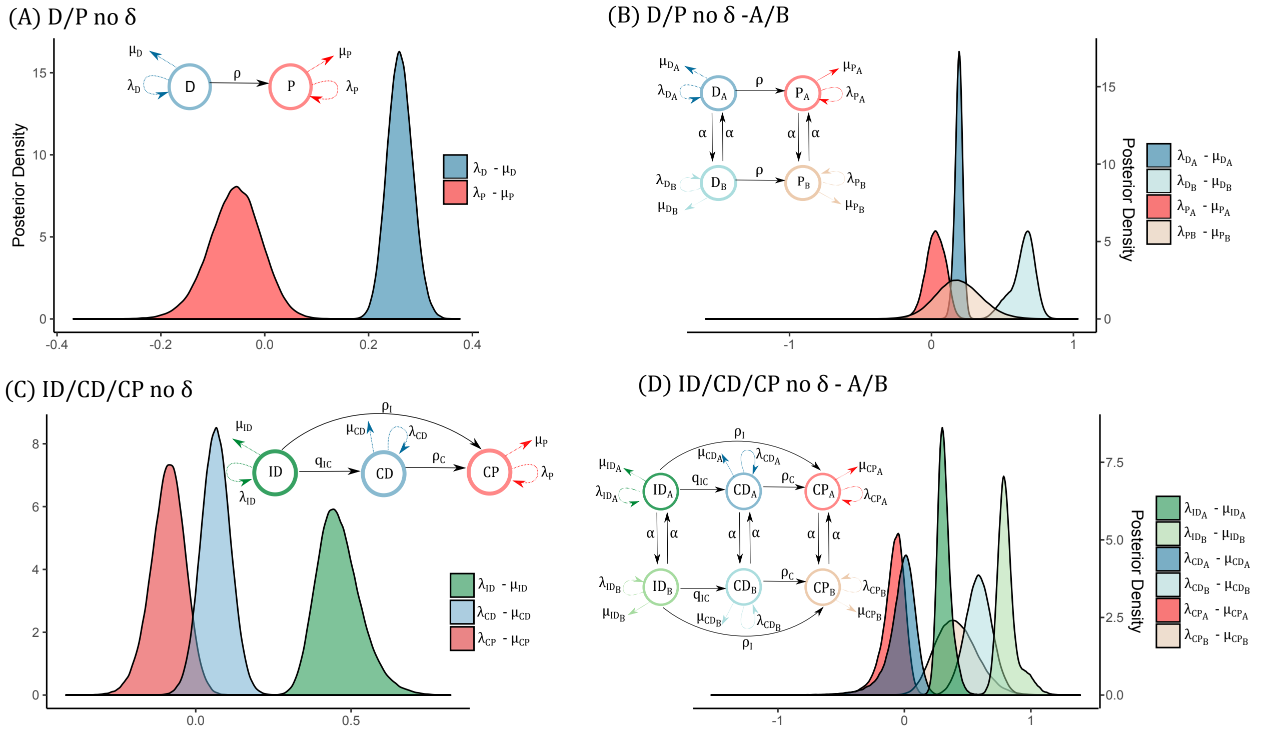


Figure 2: Net diversification rates for all models that include diploidization.



(A) D/P no δ

(B) D/P no δ -A/B

(C) ID/CD/CP no δ

(D) ID/CD/CP no δ - A/B

Figure 3: Net diversification rates for all models that do not include diploidization.

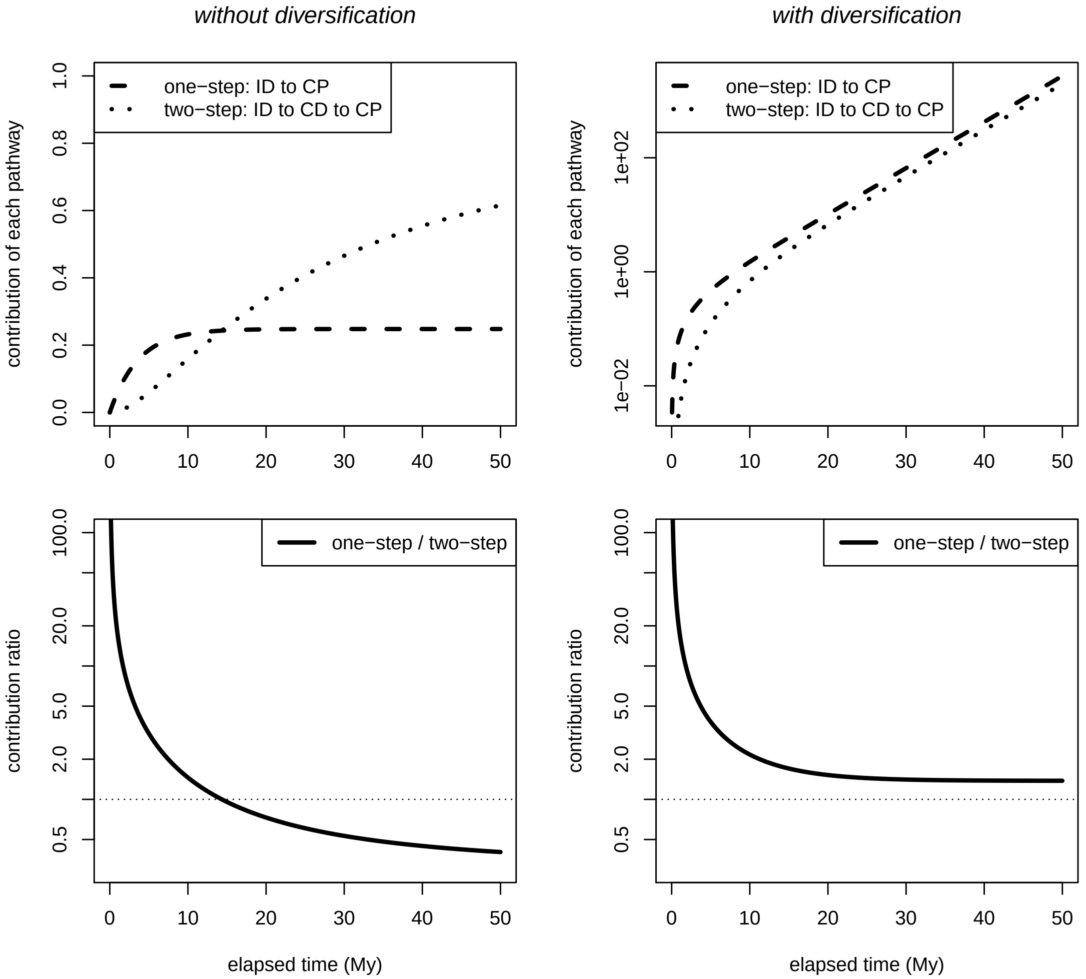


Figure 4: Contributions of the two pathways to polyploidy, the one-step ID→CP transitions, and the two-step, ID→CD→CP transitions. Panel (a) shows the contributions when considering only transition rates among the states (ignoring the diversification rate parameters). The one-step pathway dominates on short timescales and the two-step on long timescales. Panels (b) shows the contributions when including diversification of lineages in each state. The one-step pathway, in which polyploidy breaks down SI, dominates over any timescale. Panels (c) and (d) plot the ratio of the pathway contributions in the upper panels, (a) and (b), respectively.

the one-step pathway, from ID to CP

Model	Ploidy	Diploidization	Breeding System	Hidden State	Num Parameters	Marginal Log-Likelihood
1. D/P	Yes	Yes	No	No	6	-1182.93
2. D/P no δ	Yes	No	No	No	5	-1193.66
3. D/P+A/B	Yes	Yes	No	Yes	11	**-1145.69**
4. D/P+A/B no δ	Yes	No	No	Yes	10	-1150.99
5. I/C	No	No	Yes	No	5	-1194.80
6. I/C+A/B	No	No	Yes	Yes	10	**-1155.37**
7. ID/CD/CP	Yes	Yes	Yes	No	10	-1344.50
8. ID/CD/CP no δ	Yes	No	Yes	No	9	-1345.87
9. ID/CD/CP+A/B	Yes	Yes	Yes	Yes	16	**-1300.35**
10. ID/CD/CP+A/B no δ	Yes	No	Yes	Yes	15	-1303.55

Table 1: The ten models and their marginal likelihoods. Values in bold are for the best models within each class that are comparable (see Table 2). Abbreviations are D: diploid, P: polyploid, I: self-incompatible, C: self-compatible, A: one state of hidden trait, B: other state of hidden trait, δ: diploidization.

Ploidy Models					Breeding System Models			Ploidy and Breeding System Models				
	1	2	3	4		5	6		7	8	9	10
1. D/P	·	10.72	-37.24	-31.94	5. I/C	·	-39.43	7. ID/CD/CP	·	1.36	-44.15	-40.95
2. D/P no δ	·	·	-47.97	-42.66	**6. I/C+A/B**	·	·	8. ID/CD/CP no δ	·	·	-45.51	-42.31
3. D/P+A/B	·	·	·	5.30				**9. ID/CD/CP+A/B**	·	·	·	3.2
4. D/P+A/B no δ	·	·	·	·				10. ID/P/CD no δ-A/B	·	·	·	·

Table 2: Bayes factors for model comparisons. Each of the three boxes contains models that can be compared with one another, based on the character states they include (see Fig. 1). Models are numbered in as Table 1. Bayes factors are reported on the natural log scale, so numbers greater than $+2$ mean that the model in the row label has 'positive' support relative to the model in the column label; numbers less than -2 mean that model in the column label is the preferred one. Conventional thresholds for 'strong' and 'very strong' support are 6 and 10, respectively. The best model in each set is written in bold. In each case, it is the most complex model of the set.

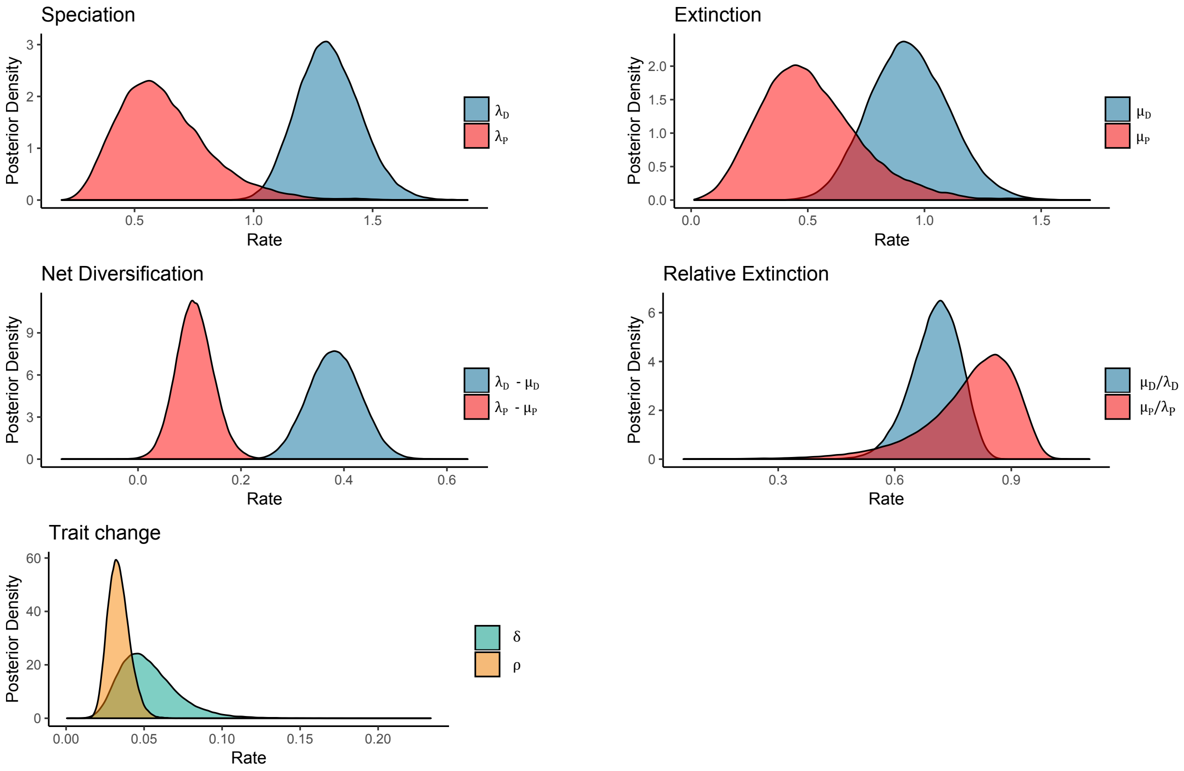


Figure S1: Posterior distribution for each of the parameters in the D/P, polyploidy model

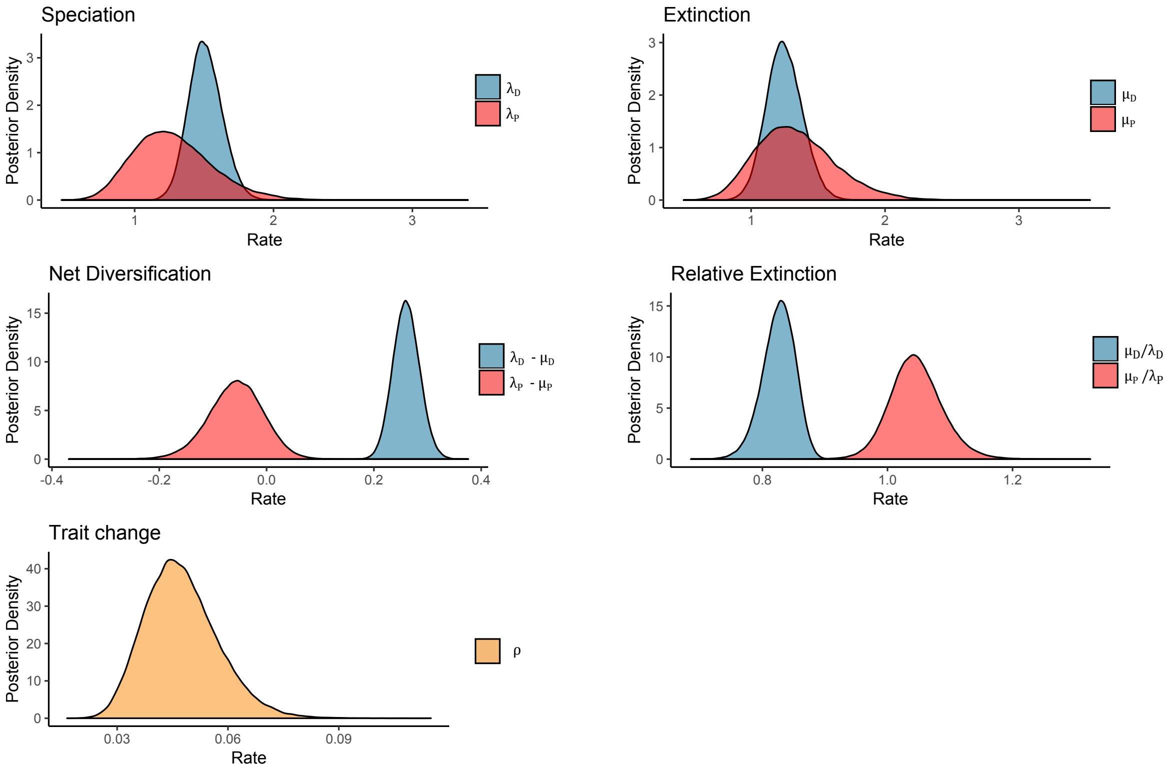


Figure S2: Posterior distribution for each of the parameters in the D/P no δ, polyploidy model

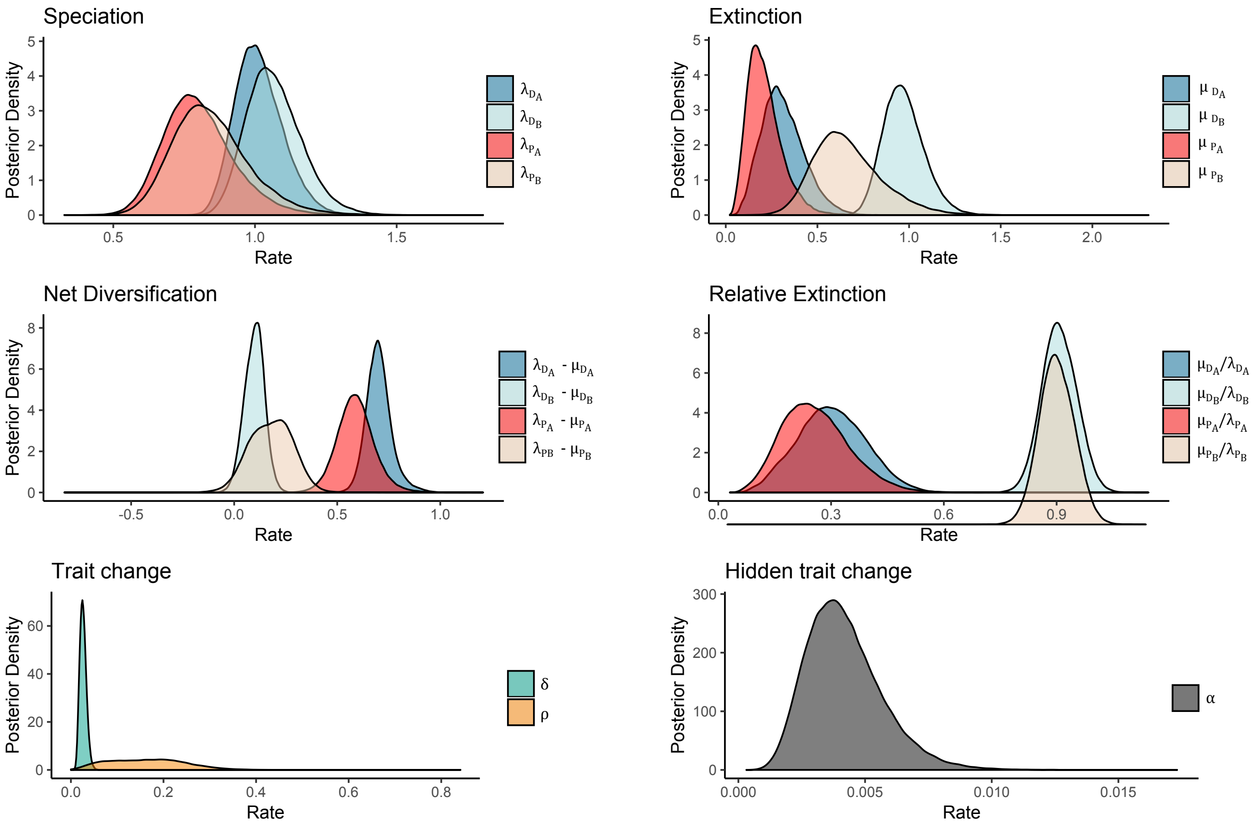


Figure S3: Posterior distribution for each of the parameters in the D/P+A/B, polyploidy model. The axis is offset in one location so that the two overlapping distributions can be seen.

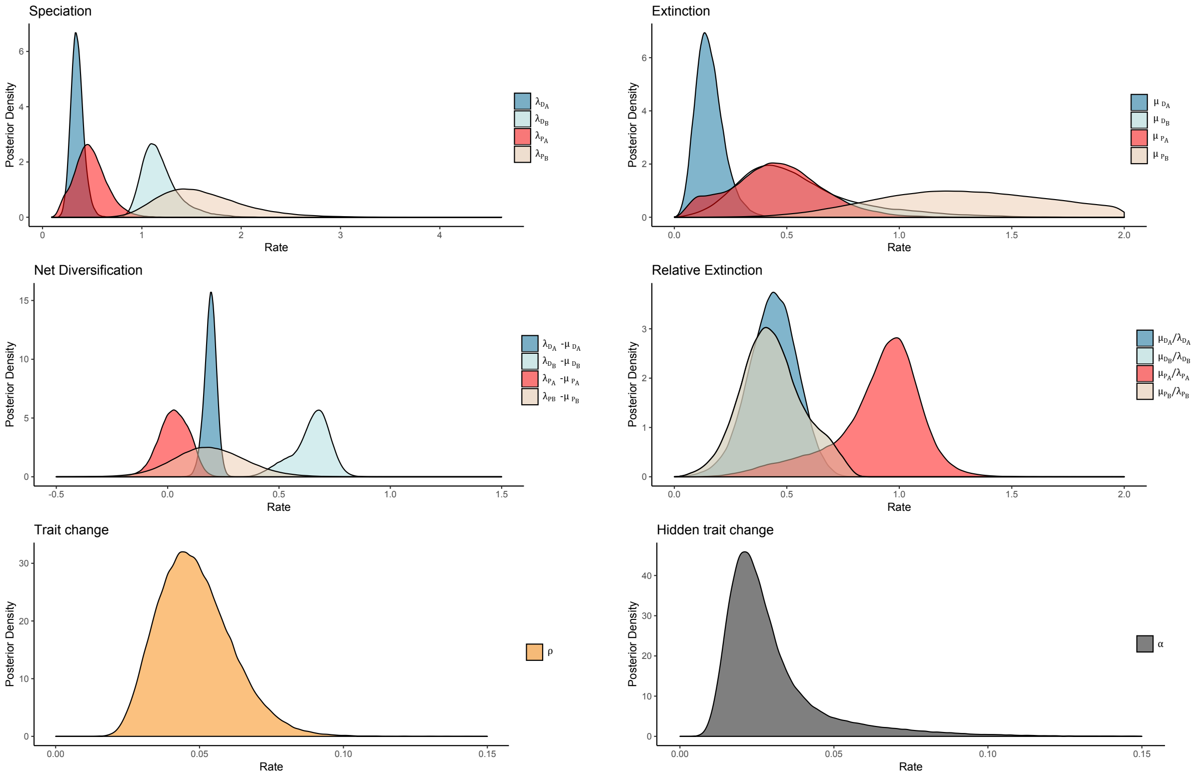


Figure S4: Posterior distribution for each of the parameters in the D/P no δ+A/B, polyploidy model

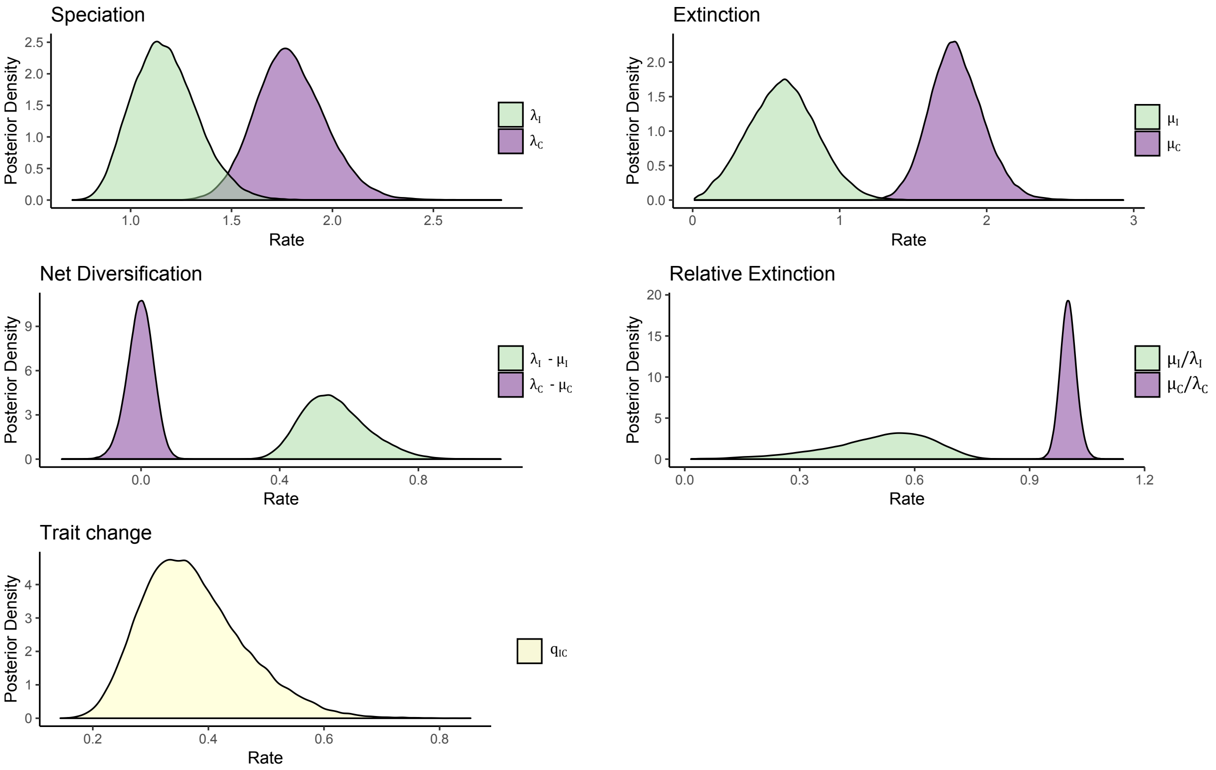


Figure S5: Posterior distribution for each of the parameters in the I/C, breeding system model

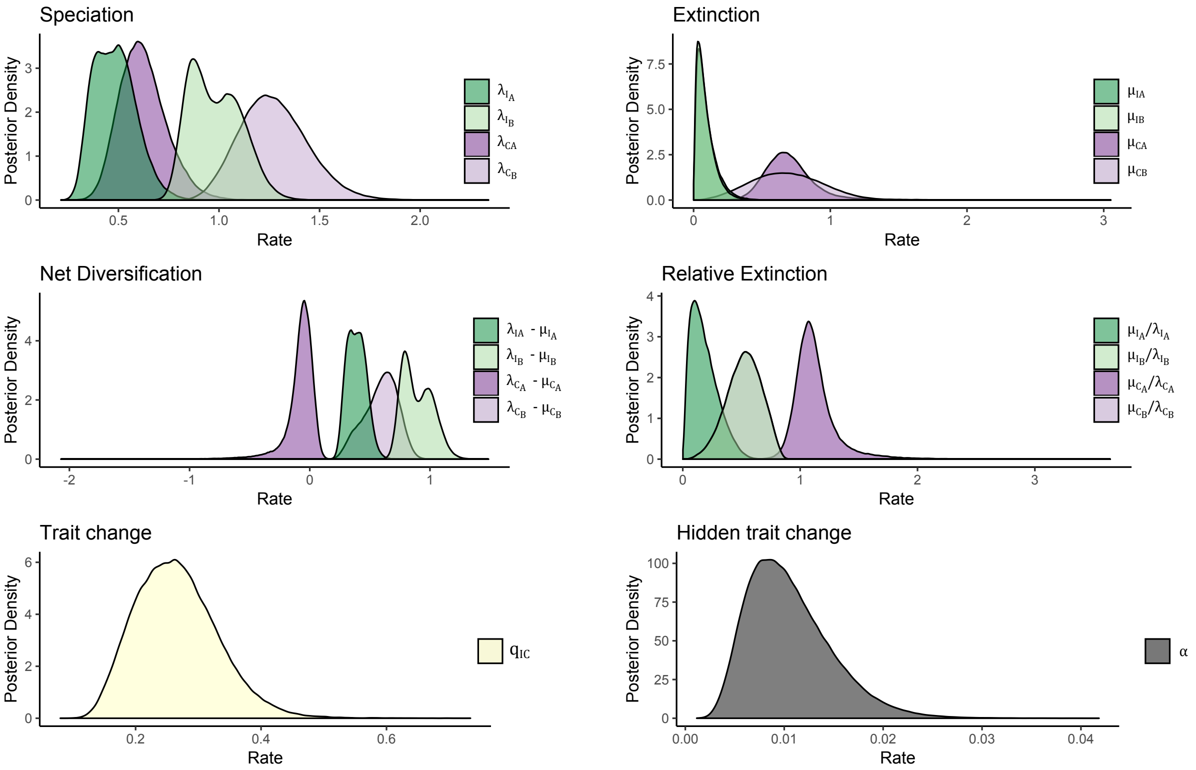


Figure S6: Posterior distribution for each of the parameters in the I/C+A/B, breeding system model

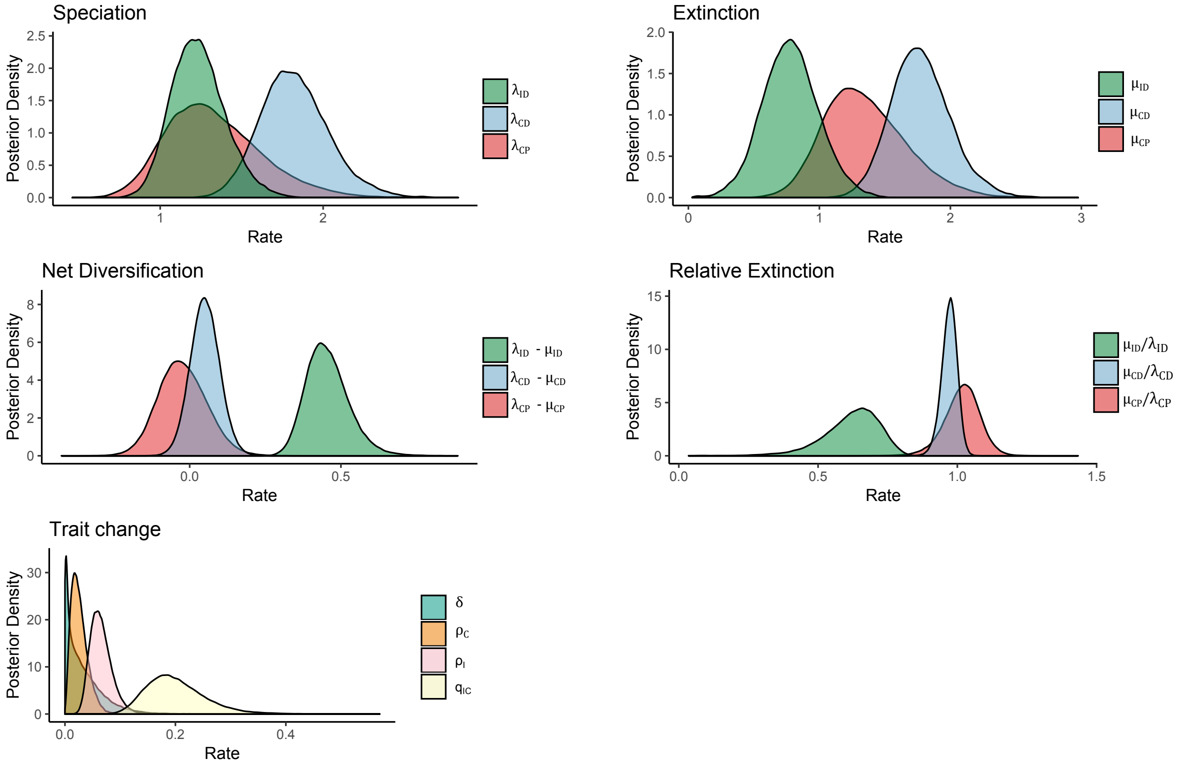


Figure S7: Posterior distribution for each of the parameters in the ID/CD/CP, polyploidy and breeding system model

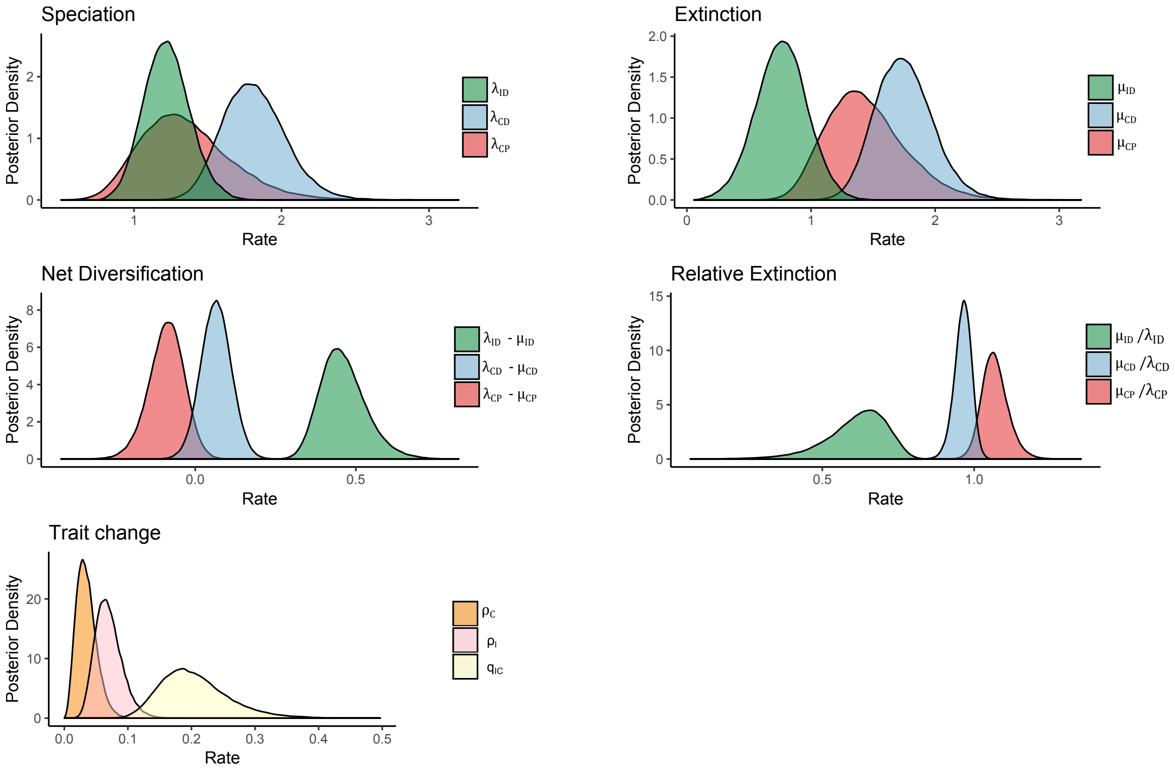


Figure S8: Posterior distribution for each of the parameters in the ID/CD/CP no δ, polyploidy and breeding system model

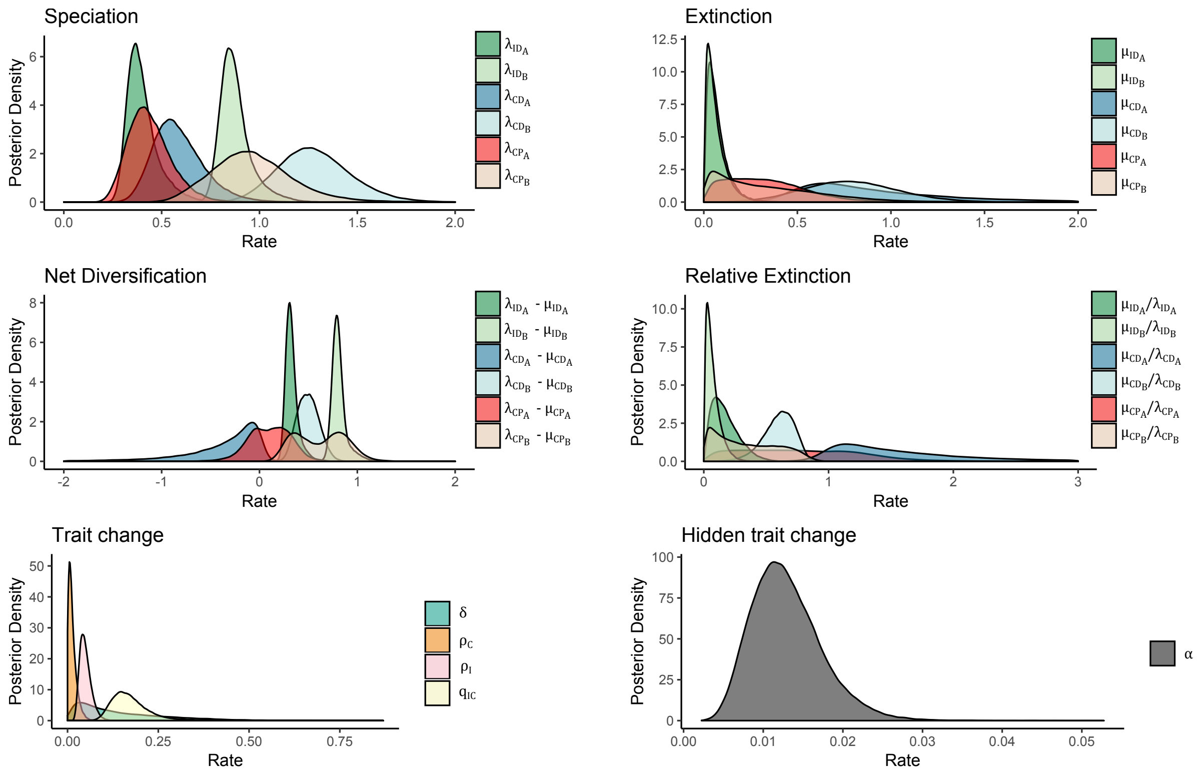


Figure S9: Posterior distribution for each of the parameters in the ID/CD/CP+A/B, polyploidy and breeding system model

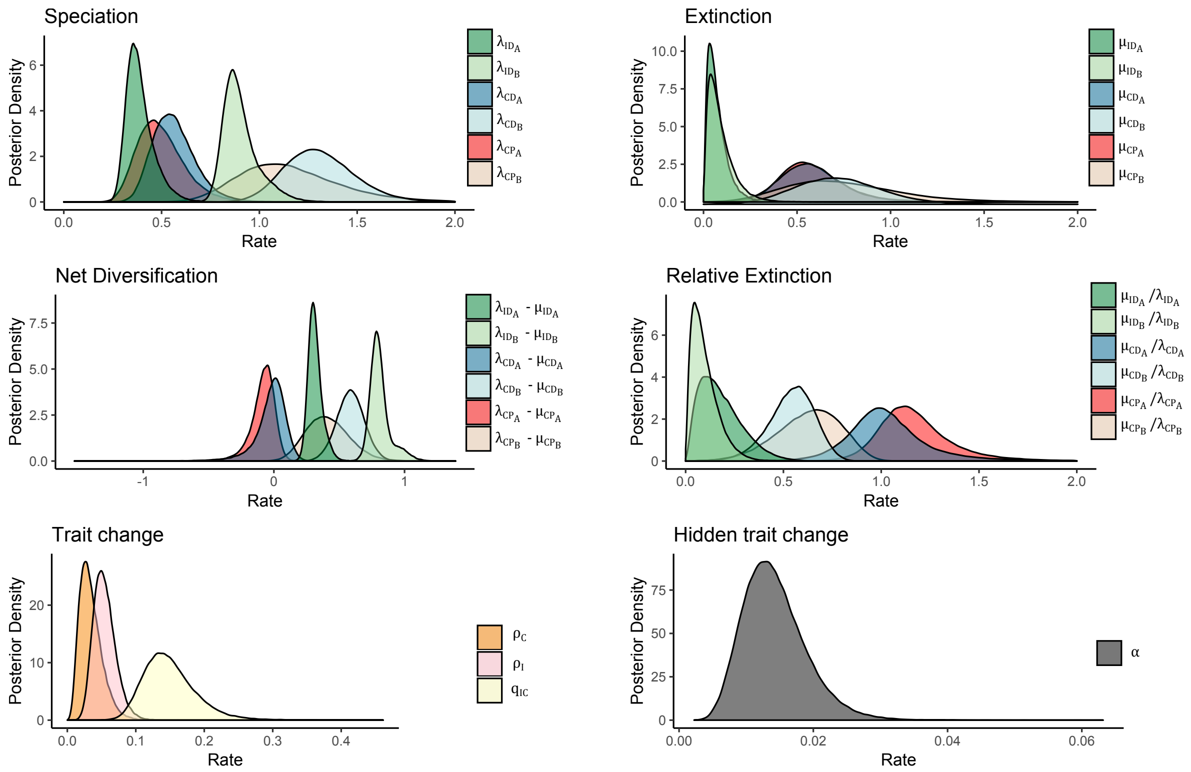


Figure S10: Posterior distribution for each of the parameters in the ID/CD/CP no δ+A/B polyploidy and breeding system model

	D/P	D/P+A/B	ID/CD/CP	ID/CD/CP+A/B
r_D	0.260	0.193, 0.658	—	—
r_P	-0.056	0.030, 0.187	—	—
r_I	—	—	—	—
r_C	—	—	—	—
r_{ID}	—	—	0.455	0.309, 0.797
r_{CD}	—	—	0.065	-0.006, 0.587
r_{CP}	—	—	-0.088	-0.074, 0.403
ρ	0.047	0.047	—	—
ρ_I	—	—	0.067	0.053
ρ_C	—	—	0.033	0.032
q_{IC}	—	—	0.198	0.145

	D/P	D/P+A/B	I/C	I/C+A/B	ID/CD/CP	ID/CD/CP+A/B
r_D	0.382	0.698, 0.100	—	—	—	—
r_P	0.109	0.587, 0.182	—	—	—	—
r_I	—	—	0.550	0.386, 0.877	—	—
r_C	—	—	-0.001	-0.059, 0.606	—	—
r_{ID}	—	—	—	—	0.449	0.318, 0.789
r_{CD}	—	—	—	—	0.050	-0.248, 0.494
r_{CP}	—	—	—	—	-0.027	0.110, 0.634
ρ	0.033	0.026	—	—	—	—
ρ_I	—	—	—	—	0.063	0.047
ρ_C	—	—	—	—	0.024	0.011
δ	0.050	0.162	—	—	0.022	0.107
q_{IC}	—	—	0.364		0.194	0.164

Table S1: Median rate estimates for all fitted models. Units are per million years. Two comma-separated numbers refer to the A and B hidden states, and — means the parameter was not present in the model. Net diversification rates (r) are subscripted with trait state initials (Diploid, Polyploid, Incompatible, Compatible). Transition rates are ρ (polyploidization), subscripted with background breeding system state; δ (diploidization); and q_{IC} (loss of self-incompatibility). The upper section is for models without diploidization, and the lower section is for models with diploidization. The supplemental figures show the corresponding distributions of parameter estimates.