When your favorite trait is not enough to explain diversification: An example of diversification linked to polyploidy and breeding system in Solanaceae

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

[1]University of Minnesota Twin Cities

[2]University of Florida

[3]23 and me

[4]University of Illinois

[5]Tel Aviv University

[6]University of Minnesota Twin Cities

Author for correspondence: Rosana Zenil-Ferguson

Running head: Polyploidy and breeding systems Solanaceae

Keywords: Polyploidy, Breeding System, Diversification, SSE models

Abstract

Abstract text. . .

Introduction

Linking changes in diversification to polyploidy and breeding system has been one of the key botany questions for the past n number of years. The prevalence of polyploidy, and the multiple evolutionary changes of breeding systems in important angiosperm clades, leads to the inevitable question of the evolutionary changes.

- Diversification and polyploidy

An important debate regarding the diversification of angiosperms has been ongoing since the publication of Mayrose et al. (2011). The authors discovered that using the latest diversification model linked to two states diploid and polyploid, the net diversification of polyploids was much slower than the net diversification rate than polyploids. This result was surprising and an sparked a discussion about the long-term evolutionary consequences of polyploidy. **?** questioned if polyploidy should be regarded as an evolutionary dead-end since the net diversification of polyploids was negative, despite overwhelming evidence about the incidence of polyploidy especially at the root of highly diverse angiosperm claims (ref here). A year later, diversification models were re-tested and corrected, and still found the same pattern (**?**), to the disappointment of plenty of botanists there was no denying of the weak trend of diversification that polyploids left behind.

There are two key questions that at the time of the polyploidy debate were difficult to ask. The first is if the models used to measure the diversification of polyploids were correct, and the second question is if the models have potentially included more evidence and potential traits that are not polyploidy or other lines of evidence that could be driving the patterns. At the time, in a different context Beaulieu and O'Meara (2016) were finding an alternative solution to the first question, coming up with a new model that could represent the broad heterogeneity of the diversification process and parse out the signal between the trait of interest and the noise in diversification. Their model, the hidden state speciation and extinction model is a key component to detect whether polyploidy our something else unknown but related to it is driving the speciation and extinction patterns that we see in angiosperms.

The second question (talk here about breeding system and polyploidy, and how this could be different across clades and this is one of the reasons why we focused on Solanaceae). -Diversification and breeding systems

Goldberg and Igic 2012, different perspectives

-What we did

35 We investigate the relative importance of both b

Debate of Mayrose

Methods

Data

Chromosome number data was obtained from the Chromosome Count Database (CCDB, Rice et al. (2015))

40 for all taxa from Solanaceae family. The records obtained were first cleaned using CCDBcurator R pack-
age (Zenil-Ferguson et al. 2017), to those we aggregated independent data of chromosome numbers and
breeding systems recorded and curated by Emma Goldberg and Boris Igić (citations for original sources can
be found in the supplementary information). In total we obtained 14,000 records of chromosome number
and breeding system than were curated semi automatically by CCDBcurator and by hand to score the chro-

45 mosome number and the breeding system with the most reconciliation. Furthermore, for the majority of
taxa ploidy was assigned according to information from original sources and Kew Royal Botanic Gardens
C-value DNA (bennett citation here dataset and the original sources agreggated by Goldberg and Igić. For
taxa without ploidy information but with information about chromosome number we assigned ploidy based
on multiplicity of chromosomes within the genus. For example, *Solanum betaceum* did not include informa-

50 tion about ploidy level but its number of chromosomes is 24, since $x = 12$ is the base chromosome number
of the *Solanum* genus we assigned *Solanum betaceum* as diploid. Species with more than one ploidy level
were assigned smallest most frequent ploidy level recorded. Domesticated taxa (i.e. *Solanum lycopersicon*)
were removed to avoid since they contained multiple ploidy levels due to domestication and not spontaneous
polyploidy. The remainder of taxa was then matched to largest time-calibrated phylogeny of Solanaceae and

55 this process resulted in 595 taxa used in each of the diversification models.

Binary or three state classification of ploidy and breeding system for the 595 taxa was done accord-
ing to Table **??**

Models

We first defined a binary state speciation and extinction model (BiSSE, Maddison et al. (2007)) for

60 polyploidy evolution where taxa was classified as diploid (D=0) and polyploid (P=1) **??**. 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 ($v_D = \frac{\mu_D}{\lambda_D}, v_D = \frac{\mu_D}{\lambda_D}$)
just as previously explored in Mayrose et al. (2011). The cladogenetic changes explored in BiSSE were
polyploidization rate are represented by parameter ρ and by diploidization parameter rate δ.

4

As a second step, we fitted a hidden state speciation and extinction model (HiSSE, Beaulieu and O'Meara (2016)) to evaluate whether the differences in diversification rates were found due to a hidden trait associated to polyploidy. As discussed by Beaulieu and O'Meara (2016) BiSSE-like models suffer from a large type I error because they fail to account as part of the null hypothesis heterogeneous diversification rate changes that do not depend on the trait of interest. By including a hidden trait in a state speciation and extinction model, HiSSE-like models address the heterogeneous background diversification changes while also parsing out the possible signal of diversification due to the trait of interest. Therefore, HiSSE model for polyploidy linked to diversification has four states: diploid and polyploid subdivided by a binary hidden trait with states A and B for which we estimated the posterior probability distributions of speciation rates $(\lambda_{D_A}, \lambda_{D_B}, \lambda_{P_A}, \lambda_{P_B})$, extinction rates $(\mu_{D_A}, \mu_{D_B}, \mu_{P_A}, \mu_{P_B})$, net diversification rates $(r_{D_A}, r_{D_B}, r_{P_A}, r_{P_B})$, and relative extinction$(\nu_{D_A}, \nu_{D_B}, \nu_{PA}, \nu_{P_B})$. The cladogenetic changes assumed by the fitted model are polyploidization rate ρ, diploidization rate ρ, and we assumed that the changes between every hidden state are symmetrical with rate α.

For breeding system both BiSSE and a HiSSE models were fitted independently by coding data as self-compatible (SC=0) or self-incompatible (SI=1) **??**. The BiSSE model fitted to breeding system data was done with the goal of investigating if the effect of self-incompatibility in diversification was similar to the pattern found by Goldberg and Igić (2012). However, the HiSSE model on breeding system parses out if the effect of breeding system in diversification is significant by minimizing type I error as mentioned above. For these models we assumed that self-compatibility is irreversible as discussed by Igić and Busch (2013), so we defined the transition rate q_{IC} as the parameter for changes from self-incompatible to self-compatible state.

Next, we proposed a multivariate speciation and extinction model (MuSSE, FitzJohn (2012)) to investigate the link amongst diversification, breeding system, and polyploidy simultaneously. The MuSSE model is defined using three states self-compatible diploids (SC-D=0), polyploids that are always self-compatible (ref for this??) (P=1), and self-incompatible diploids (SI-D=2, see table **??**) and contains ten parameters, six defining diversification $(\lambda_{SD}, \mu_{SD}, \lambda_P, \mu_P, \lambda_{ID}, \mu_{ID})$ and the other four key cladogenetic changes that are: polyploidization of self-compatible diploids ρ_{SD}, diploidization δ, polyploidization of self-incompatible diploids ρ_{ID}, and self-incompatible to self-compatible rate q_{IC}.

Since the null hypothesis of the MuSSE model is that the diversification is equal and constant for all three states defined, it is possible that MuSSE can also suffer from large type I errors. In order to account for heterogeneity of the diversification rates and parse the signal in diversification coming from breeding system and polyploidy we extended MuSSE model to account for a hidden state, as Beaulieu and O'Meara (2016) did in the bivariate case. The model we proposed is a multivariate and hidden states speciation and extinction stochastic process (MuHiSSE) that allowed us to account simultaneously for the diversification rates linked to breeding system and polyploidy but also the presence of some more heterogeneity in the process. This model is analogous to the GeoHiSSE model proposed by Caetano et al. (2018) in a biogeographical context. The full MuHiSSE model has 26 parameters, however, our goal was to look for diversification rate differences so we fitted a simplified version of 17 parameters by fixing the rates amongst hidden states to be equal with parameter α and the transition rates amongst breeding system and polyploidy as defined in the MuSSE model (ρ_{SD}, δ, ρ_{ID}, q_{IC}) despite the hidden state. Using MuHiSSE we estimated twelve speciation and extinction rates ($\lambda_{SD_A}, \mu_{SD_A}, \lambda_{P_A}, \mu_{P_A}, \lambda_{ID_A}, \mu_{ID_B}, \lambda_{SD_B}, \mu_{SD_B}, \lambda_{P_B}, \mu_{P_B}, \lambda_{ID_B}, \mu_{ID_B}$), and the net diversification and relative extinction rates associated with them.

All the models were performed using RevBayes (Höhna et al. 2016) software that performs Bayesian inference via MCMC using Metropolis-Hastings algorithm. A correction for sampling bias was done in all models by assuming that Solanaceae family has approximately 3,000 species ($s = 595/3000$) as the Solanaceae Source project indicates (Source 2011). In all models, speciation and extinction parameters used log-normal prior distributions that with mean the expected net diversification rate ($\frac{(number of taxa)/2}{root age}$) and standard deviation 0.5. Prior distributions for parameters defining cladogenetic changes were gamma distributed with parameters $k = 0.5$ and $\theta = 1$. MCMC was performed for 96 hours in the cluster at Minnesota Supercomputing Instute which allowed for 5,000 generation of burn-in and a minimum of 200,000 generations of MCMC for each of the 6 models. Convergence of the MCMC was tested using R package coda (ref) and software Tracer (ref) to assess convergence and mixing (see supplementary information).

Type	Total of Taxa	Models		
		BiSSE/HiSSE Polyploidy	BiSSE/HiSSE Breeding System	MuSSE/MuHiSSE
Diploid Self Compatible	152	0	0	0
Diploid Self Incompatible	97	0	1	2
Diploid with unknown breeding system	219	0	(0,1)	(0,2)
Polyploid	81	1	0	1
Unknow ploidy and self compatible	34	(0,1)	0	(0,1)
Unknown ploidy and self incompatible	12	0	1	2

Results

Discussion

125

7

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. URL `http://dx.doi.org/10.1093/sysbio/syw022`.

FitzJohn, R. G., 2012. Diversitree : comparative phylogenetic analyses of diversification in r. Methods Ecol Evol 3:1084–1092. URL `http://doi.wiley.com/10.1111/j.2041-{210X}.2012.00234.x`.

Goldberg, E. E. and B. Igić, 2012. Tempo and mode in plant breeding system evolution. Evolution 66:3701–3709. URL `http://dx.doi.org/10.1111/j.1558-5646.2012.01730.x`.

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. URL `http://dx.doi.org/10.1093/sysbio/syw021`.

Igić, B. and J. W. Busch, 2013. Is self-fertilization an evolutionary dead end? New Phytol 198:386–397. URL `http://dx.doi.org/10.1111/nph.12182`.

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. URL `http://dx.doi.org/10.1080/10635150701607033`.

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. URL `http://dx.doi.org/10.1126/science.1207205`.

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. URL `http://dx.doi.org/10.1111/nph.13191`.

Source, S., 2011. A worldwide taxonomic monograph of all species in the genus solanum.

Zenil-Ferguson, R., J. M. Ponciano, and J. G. Burleigh, 2017. Testing the association of phenotypes with polyploidy: An example using herbaceous and woody eudicots. Evolution 71:1138–1148. URL `http://dx.doi.org/10.1111/evo.13226`.