RESEARCH ARTICLE

AMERICAN JOURNAL OF BOTANY

INVITED PAPER

For the Special Issue: The Evolutionary Importance of Polyploidy

Phylogenetic evidence for cladogenetic polyploidization in land plants[1]

Shing H. Zhan[2], Michal Drori[3], Emma E. Goldberg[4], Sarah P. Otto[2], and Itay Mayrose[3,5]

PREMISE OF THE STUDY: Polyploidization is a common and recurring phenomenon in plants and is often thought to be a mechanism of "instant speciation". Whether polyploidization is associated with the formation of new species (cladogenesis) or simply occurs over time within a lineage (anagenesis), however, has never been assessed systematically.

METHODS: We tested this hypothesis using phylogenetic and karyotypic information from 235 plant genera (mostly angiosperms). We first constructed a large database of combined sequence and chromosome number data sets using an automated procedure. We then applied likelihood models (ClaSSE) that estimate the degree of synchronization between polyploidization and speciation events in maximum likelihood and Bayesian frameworks.

KEY RESULTS: Our maximum likelihood analysis indicated that 35 genera supported a model that includes cladogenetic transitions over a model with only anagenetic transitions, whereas three genera supported a model that incorporates anagenetic transitions over one with only cladogenetic transitions. Furthermore, the Bayesian analysis supported a preponderance of cladogenetic change in four genera but did not support a preponderance of anagenetic change in any genus.

CONCLUSIONS: Overall, these phylogenetic analyses provide the first broad confirmation that polyploidization is temporally associated with speciation events, suggesting that it is indeed a major speciation mechanism in plants, at least in some genera.

KEY WORDS anagenesis; cladogenesis; ClaSSE; polyploidy; speciation

Polyploidization, or whole-genome duplication, has been a rampant and ongoing process contributing to plant evolution. Building on the early work by Stebbins (1938), the latest estimates suggest that 35–40% of extant flowering plant species are recent polyploids, or "neopolyploids" (Wood et al., 2009; Scarpino et al., 2014), with genomes that have doubled since the initial divergence of their genus. Deeper in time, all seed plants are thought to have undergone a polyploidization event some time during their evolutionary history (Jiao et al., 2011). The frequency of polyploidization, along with the common observation of reproductive incompatibilities between polyploids and related diploids (triploid block; Ramsey and Schemske, 1998), has led to the view that polyploidization is a mechanism of "instant speciation" and a relatively easy path to sympatric speciation, particularly in plants (Coyne and Orr, 2004). Previous phylogenetic estimates of the rate of polyploidy have not, however, assessed whether polyploidization is indeed coupled in time with speciation itself. Rather, prior work has focused on methods that estimate the rate of polyploidization per unit time (anagenesis) or on methods that do not distinguish when ploidy shifts occur (Stebbins, 1938; Grant, 1963; Masterson, 1994; Wood et al., 2009; Mayrose et al., 2011; Scarpino et al., 2014). It is indeed possible that transitions in ploidy occur either without full reproductive isolation ever evolving (i.e., without speciation) and/or by simple displacement of diploids by polyploid descendants. Here, we ask whether there is a phylogenetic signal that polyploidization is coupled in time with speciation events (cladogenesis), using recent phylogenetic methods that tease apart anagenetic and cladogenetic processes.

In addition to initiating reproductive incompatibilities, polyploidization is thought to be a driver of speciation because newly formed polyploids often differ from their diploid ancestors in morphological, physiological, and life history characteristics (e.g.,

[1] Manuscript received 9 March 2016; revision accepted 12 July 2016.

[2] Department of Zoology, 4200-6270 University Boulevard, University of British Columbia, Vancouver, British Columbia V6T 1Z4 Canada;
[3] Department of Molecular Biology and Ecology of Plants, George S. Wise Faculty of Life Sciences, Tel Aviv University, P. O. Box 39040, Tel Aviv 69978, Israel; and
[4] Department of Ecology, Evolution, & Behavior, University of Minnesota, 140 Gortner Laboratory, 1479 Gortner Avenue, St. Paul, Minnesota 55108 USA
[5] Author for correspondence (e-mail: itaymay@post.tau.ac.il), phone: +972-3-640-7212, fax: +972-3-640-9380

doi:10.3732/ajb.1600108

Levin, 1983; Ramsey and Schemske, 2002). Polyploidy therefore may serve as an important mechanism for niche differentiation and ecological diversification, which may contribute to the successful establishment of new polyploid species (Levin, 1983; Otto, 2007). Establishing the causative link between polyploidization and speciation is challenging, however. For example, ecological differences between related diploid and polyploid taxa may have occurred independently of the polyploidization event (before or after). Similarly, it is difficult to determine whether polyploidization itself was a major early driver of reproductive isolation or occurred later in the speciation process.

Indeed, it is known that polyploidy does not always lead to immediate reproductive isolation. For example, Slotte et al. (2008) found that polyploidy does not terminate gene flow between the diploid parent and its polyploid progeny in *Capsella*. Furthermore, extensive intraspecific variation in ploidy levels (Stebbins, 1971; Wood et al., 2009; Rice et al., 2015) and evidence of multiple origins in many polyploid lineages (Soltis and Soltis, 1999) suggest that multiple cytotypes often segregate within species. Gene flow between diploids and polyploids remains possible via a number of mechanisms (Ramsey and Schemske, 1998, 2002), including the occasional production of viable seeds from triploid intermediates ("triploid bridge"), from crosses involving unreduced gametes produced by diploids, or from genome reduction yielding offspring bearing half the genome size of their polyploid parents (polyhaploids). Evidence for gene flow between diploids and polyploids has been found in the genomes of several plants, particularly between crops and their wild relatives (reviewed by Chapman and Abbott, 2010). These observations demonstrate that the speciation of polyploid lineages may be a dynamic—rather than instantaneous—process, which generates and maintains genetic variation within species for some time (Thompson and Lumaret, 1992).

Recent advances in methods to analyze trait evolution across phylogenetic trees allow researchers to infer rates of anagenetic vs. cladogenetic change in a trait and to assess the degree to which a change in a trait, such as polyploidization, occurs concurrently with the formation of species. These methods build upon the BiSSE (binary state speciation and extinction) model (Maddison et al., 2007; FitzJohn et al., 2009), which describes the evolution of a two-state trait (i = diploid or polyploid in this study, for example) that can affect speciation rate (λ_i) and extinction rates (μ_i). BiSSE can be used in Bayesian or maximum likelihood (ML) analyses to assess the parameter combinations that best account for both the present-day trait distribution and the shape of the phylogeny, thus providing a framework within which character-dependent macroevolutionary hypotheses may be statistically tested (e.g., Goldberg et al., 2010; Hugall and Stuart-Fox, 2012; Beaulieu and Donoghue, 2013; Zhan et al., 2014; Sabath et al., 2016). As originally formulated, the trait evolves over time from state i to state j at rate q_{ij}, assuming that only anagenetic changes are possible. Subsequent work also allowed for trait evolution during cladogenesis, modeled either as the probability that speciation generates daughter species whose traits differ from the parent (BiSSE-ness; Magnuson-Ford and Otto, 2012) or estimating the rate at which speciation with trait change occurs (ClaSSE; Goldberg and Igić, 2012). The models are interchangeable in a likelihood framework but have different natural prior distributions when used in Bayesian analyses. Here, we used ClaSSE with a uniform prior on the fraction of trait changes that are cladogenetic, ϕ (see Appendix S1 in Supplemental Data with online version of this article).

In the current study, we tested the main prediction of the hypothesis that polyploidization is a major speciation mechanism: ploidy shifts should coincide with speciation events (either at internal nodes of the phylogeny or at "hidden speciation nodes" along the branches due to subsequent extinction of a daughter lineage). To do so, we applied the ClaSSE model in both Bayesian and ML frameworks to a large cohort of plant genera (mostly angiosperms) for which adequate sequence data and chromosome number data are available. Our study provides the first broad confirmation that polyploidization is frequently cladogenetic in plants.

MATERIALS AND METHODS

Database construction—For this study, we assembled a database of plant genera exhibiting variation in ploidy levels. We created 223 angiosperm genus data sets, which are collectively referred to as PloiDB (Table S1 in Appendix S2 with online Supplemental Data), by retrieving and combining sequence and karyotypic data from various public data sources. Phylogenetic trees for each data set were reconstructed as similarly described in Sabath et al. (2016). Briefly, ultrametric Bayesian phylogenies were inferred using sequence data available at NCBI GenBank (www.ncbi.nlm.nih.gov/genbank). Sequences were binned by locus using the program OrthoMCL v2.0.3 (Li et al., 2003). An appropriate outgroup, which was used to root the phylogeny, was selected and added to the list of sequences, which were aligned using the program MAFFT v7.149b (Katoh and Standley, 2013). GUIDANCE v1.41 (Penn et al., 2010) was applied to the resulting multiple sequence alignment (MSA) of each cluster to discard sequences and positions that reduce the MSA reliability. The best-supported model of sequence evolution was determined for each locus independently using the program jModelTest v2.1.7 (Guindon and Gascuel, 2003; Darriba et al., 2012). The MSAs for multiple clusters were concatenated to form a multilocus MSA. Phylogenies were estimated by applying the program MrBayes v3.2.1 (Ronquist et al., 2012) using two independent runs, each with one cold and three heated chains of 2,000,000 generations each (the average standard deviation of split frequencies of the majority of the runs was below 0.05), and their results were then combined. In each run, the best-supported nucleotide model determined for each locus was used, and branch lengths were allowed to vary according to a birth–death relaxed clock model (Thorne et al., 1998). Finally, the outgroup species were pruned from all resulting trees.

Chromosome numbers were taken from the Chromosome Counts Database v1.1 (ccdb.tau.ac.il; Rice et al., 2015), a database that houses chromosome numbers from multiple compendia. Using 100 randomly sampled MrBayes trees combined with chromosome numbers, we inferred ploidy levels (diploid or polyploid) using the program ChromEvol v2.0 (Mayrose et al., 2010; Glick and Mayrose, 2014). The reliability of estimated ploidy levels was assessed by comparing ploidy inferences across phylogenies and by using a simulation-based approach (Glick and Mayrose, 2014). For each genus, the ML parameter estimates inferred using ChromEvol were used to simulate ploidy levels across each of the 100 trees, after which ploidy levels were inferred again using ChromEvol. The simulation reliability score was defined for each species as the fraction of accurate ChromEvol inferences out of 100 simulations, while the phylogenetic reliability score was defined as the fraction of phylogenies with the same ploidy inference as the majority rule as defined

in the ChromEvol manual (http://www.tau.ac.il/~itaymay/cp/chromEvol/). A taxon was considered uncertain and treated as "data not available" (NA) if (1) chromosome number data are available for it and its phylogenetic reliability score was below 0.95, or (2) its combined reliability score (across trees and simulations) was below 0.95 when chromosome number data are not available.

Although the automated procedure included all taxa with sequence information (including infraspecific taxa, such as subspecies and varieties), we chose a single representative in the following analyses to focus on diversification at the species level. For species with multiple infraspecific entries present, we randomly selected one representative and pruned out the remainder.

The above procedure produced over 1000 genus data sets, but only 223 data sets that met the following criteria were retained: (1) the phylogeny contained at least 30 taxa; (2) at most 50% of taxa had uncertain ploidy assignment (NA); (3) at least 20% of the taxa had chromosome number data; and (4) at least one taxon was polyploid and one was diploid. These PloiDB data sets are available for download at the Dryad Data Repository (doi:10.5061/dryad.gr732).

Additionally, we analyzed a previously assembled database (Mayrose et al., 2011) encompassing 63 genus-level data sets (hereafter, referred to as M2011), but dropping two genera (*Cuphea* and *Cerastium*) because of errors in the data (Soltis et al., 2014; Mayrose et al., 2015). The same criteria described above to filter out data sets with low coverage were applied to the M2011 data sets, thereby retaining 29 M2011 data sets (16 are in common with the PloiDB data sets, and *Dryopteris* is replicated in M2011 but is not found in PloiDB, therefore yielding a total of 235 unique genus-level data sets). We used the same set of MrBayes trees and ChromEvol ploidy estimates as was obtained from Dryad repository (http://dx.doi.org/10.5061/dryad.6hf21).

Models of polyploid evolution—Because nearly all plant species descend from a polyploid ancestor if we trace their evolutionary history back far enough in time (Jiao et al., 2011), we cannot examine recent polyploidization events without using a reference point (Mayrose et al., 2015). Therefore, we defined a polyploid lineage with respect to the base of the genus, as we did previously in Mayrose et al. (2011; see also Stebbins, 1938). Thus, in this study a species is denoted as polyploid if it was detected by ChromEvol to have undergone a polyploidization event over its evolutionary history since divergence from the root of the genus phylogeny, regardless of whether its genome subsequently diploidized. While exceptions exist (Mandáková et al., 2016), this assumption is also consistent with the notion that polyploidy is largely an irreversible process over relatively short evolutionary time scales (Meyers and Levin, 2006; Scarpino et al., 2014).

To estimate the mode of ploidy transitions, we employed the ClaSSE model (Goldberg and Igić, 2012) with the following trait-dependent parameters (D for diploid and P for polyploid): diploid and polyploid speciation rates without a change in state (λ_D and λ_P), diploid and polyploid extinction rates (μ_D and μ_P), the rate of polyploidization along a branch ("anagenesis", q_{DP}), and the rate of speciation coupled with a ploidy shift in one of the daughter species ("cladogenesis", λ_{DDP}). We refer to this full six-parameter model as the "dual" model, which allows for both cladogenetic and anagenetic ploidy shifts.

The "dual" model makes several assumptions about the directionality and symmetry of ploidy level transitions. First, we assumed that diploid-to-polyploid transitions do not reverse within the evolutionary history of a genus. This assumption is consistent with the definition of polyploidy used in this study. Because all ploidy shifts are measured relative to the genus ancestor, we fixed the ancestral state of each genus to "diploid". Finally, we assumed that cladogenesis causes a trait shift in only one daughter species. (BiSSE-ness and ClaSSE allow the possibility that both daughter species may differ from the parent, which one might observe with niche or range traits.)

Estimates of speciation and extinction rates will be biased if one does not account for missing taxa. Thus, in all analyses detailed below, we used the "skeletal tree" method of FitzJohn et al. (2009) to adjust the likelihoods for missing data, which assumes that the taxa on the tree are randomly sampled from all taxa of the same state in the clade. The skeletal tree method requires an estimate of the size of a genus, which we obtained from The Plant List v1.1 database (TPL; http://www.theplantlist.org/) by counting the number of accepted species, without regard to the confidence level and excluding entries with infraspecific ranks. (In cases where the TPL count was less than the observed number of species that were represented on the phylogeny, a sampling fraction of 100% was assumed.) For the unsampled species, we assumed the same fraction of polyploid vs. diploid species as among the observed taxa.

Bayesian analysis—First, we took a Bayesian approach to measure the relative proportion of cladogenetic vs. anagenetic ploidy shifts, defined as $\phi = \lambda_{DDP}/(\lambda_{DDP} + q_{DP})$. Values of ϕ close to 1 imply that polyploidy occurs cladogenetically more often than anagenetically, and values close to 0 imply the reverse.

For each PloiDB data set, a single Markov chain Monte Carlo (MCMC) was run using the "dual" model (denoted as M_D) on 20 randomly chosen trees for 2000 generations each, discarding the first 50% as burn-in—thereby resulting in 20,000 generations (trace plots showed that the MCMC chain moved across the parameter space rapidly, with effective sample sizes ranging between 400 and 8600 for 220 of 223 genera). The same MCMC procedure was conducted for each M2011 data set, except that 50 randomly selected MrBayes trees were used. For each MCMC run, a heuristic starting point was calculated based on a character-independent birth–death model, and exponential priors were placed on the model parameters using the following rates: $1/2r$ (for λ_D), $1/2r$ (for λ_{DDP}), $1/r$ (for λ_P), $1/2r$ (for μ_D), $1/2r$ (for μ_P), and $1/2r$ (for q_{DP}), where r = ln(number of taxa)/tree length. As shown in the *Mathematica* file (see Appendix S1), these prior choices on the parameters led to a uniform prior distribution for ϕ.

To assess support for one transition mode over the other, we report the 95% highest posterior density (HPD) interval of ϕ. In any one genus, strong support for a preponderance of cladogenetic change was inferred if the entire HPD interval fell above 0.5 and for anagenetic change if it fell below 0.5. We also examined the distribution of HPD intervals across genera to detect departures from a uniform posterior distribution. HPD intervals of ϕ were constructed by pooling together values of ϕ calculated from the MCMC samples from all MrBayes trees analyzed.

Maximum likelihood analysis—We also used likelihood ratio tests to identify the best-fitting model. In addition to M_D, we analyzed two reduced models, one permitting only cladogenetic shifts (denoted as M_C, with $q_{DP} = 0$) and the other permitting only anagenetic shifts (denoted as M_A, with $\lambda_{DDP} = 0$). By comparing data fits to

these three models, we were able to test whether there was significant evidence for the presence of cladogenesis (M_A rejected in favor of M_D) and/or whether there was significant evidence for anagenesis (M_C rejected in favor of M_D).

For the PloiDB data sets, ML fitting was performed on each of the 20 MrBayes trees analyzed in the MCMC analysis. Ten starting points were randomly drawn from the MCMC samples (described above), and the parameter set that yielded the maximum likelihood of the data across the 10 attempts was kept. This procedure was conducted for each of M_A, M_C, and M_D. To summarize the results across trees, we calculated, for each tree, twice the difference in the maximum log likelihood values (2ΔlnLik) between the "dual" model (M_D) and a reduced model (M_A or M_C) (i.e., 2 × [lnLik of M_D – lnLik of M_A or M_C]), and then took the median over all trees. M_A (or M_C) was rejected in favor of M_D when the median 2ΔlnLik was greater than $\chi^2_{\alpha=0.05} = 3.841$. For the M2011 data sets, we performed ML fitting to 50 MrBayes trees (those used in the MCMC analysis) instead of 20, using the same procedure as for the PloiDB data sets.

Implementation—The MCMC and ML analyses were performed in the R statistical computing environment (R Core Team, 2015) using some phylogenetic utilities in the package *ape* v3.4 (Paradis et al., 2004) and the ClaSSE model and statistical methods in the package *diversitree* v0.9-8 (FitzJohn, 2012). The HPD intervals were computed using the package *coda* v0.18-1. The R scripts implementing the analysis procedures are available in Dryad Data Repository.

RESULTS

In this study, we tested at a broad phylogenetic scale whether the mode of polyploid transition in plants is mainly cladogenetic (i.e., coinciding with branching events) or anagenetic (i.e., arising along branches). To this end, we assembled a large database of angiosperm genus data sets (PloiDB) using an automated procedure, and then combined it with manually curated data sets from a previous study (M2011). The sampling fraction of the PloiDB data sets ranged from 5 to 100% (median of 50%) and the percentage of polyploids from 1 to 98% (median of 18%) (Table S1 in Appendix S2). Using the PloiDB and M2011 data, we took two statistical approaches (MCMC and ML) to examine whether the mode of polyploid transition, cladogenetic or anagenetic, can be identified in each genus-level data set.

In the Bayesian analysis of the PloiDB data sets, we performed an MCMC procedure to determine the 95% HPD interval of the proportion of ploidy shifts that are cladogenetic (ϕ). The lower bound of the HPD interval of ϕ was greater than 0.5 in three genera (*Allium*, *Artemisia*, and *Taraxacum*), consistent with the cladogenesis hypothesis. The upper bound of the HPD interval of ϕ, however, was never less than 0.5, indicating no support for the hypothesis that anagenesis is the main mode of polyploid transition. The posterior distribution for ϕ revealed a slight but significant trend toward higher values of ϕ, with the median lying above 0.5 for 129 out of 223 genera ($P = 0.0226$, exact two-tailed binomial test with $N = 223$ and $p = 0.5$; Fig. 1A and Table S2 in Appendix S2). Similarly, in the Bayesian analysis of the M2011 data sets, the HPD intervals supported cladogenesis in a single genus (*Achillea*) and anagenesis in none of the data sets examined. The overall posterior distribution of ϕ was also shifted upward, with a median above 0.5 in 21 of 29 genera ($P = 0.0241$, exact two-tailed binomial test with $N = 29$ and $p = 0.5$; Fig. 1B and Table S3 in Appendix S2), again suggesting that cladogenesis is the dominant mode of polyploid transition.

In the ML analysis of the PloiDB data sets, we performed likelihood ratio tests to determine whether reduced models (M_A and M_C, permitting only either anagenetic or cladogenetic transitions,

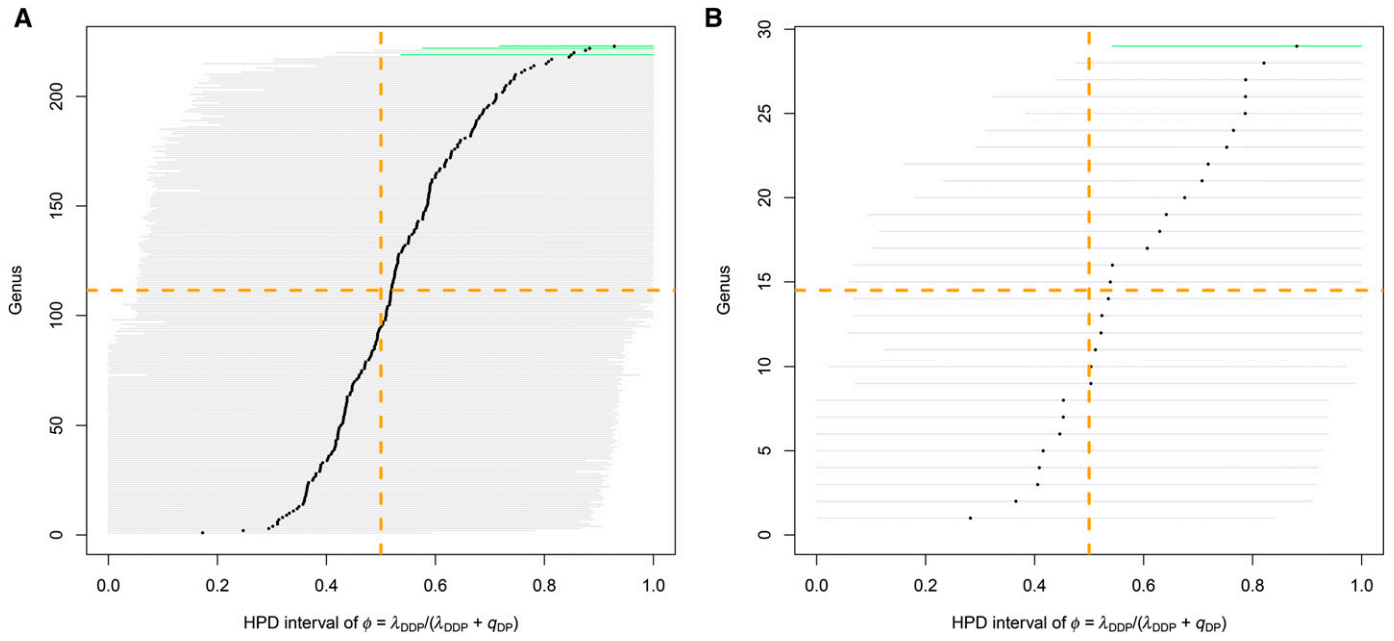

FIGURE 1 Distribution of 95% highest posterior density (HPD) intervals of the relative proportion of cladogenetic transitions (ϕ) in the (A) PloiDB and (B) M2011data sets. HPD intervals that lie entirely above 0.5 are highlighted in green. The median of the posterior distribution of ϕ is marked by a black dot, notably more of which lie above $\phi = 0.5$ (right of intersection of orange dashed lines), in favor of the cladogenetic shift hypothesis.

respectively) could be rejected in favor of the "dual" model (M_D, where both types of transitions are possible). Among the 223 PloiDB data sets, M_A was rejected in favor of M_D in 29 genera, supporting the inclusion of cladogenesis in the model (Table S2 in Appendix S2). Conversely, M_C was rejected in favor of M_D in only six genera, supporting the inclusion of anagenesis in the model. In one genus (*Veronica*), both M_A and M_C were rejected in favor of M_D. We discovered a consistent result from the ML analysis of the M2011 data sets, with M_A rejected in favor of M_D in 11 data sets (including both the replicated *Dryopteris* data sets) and M_C in only one genus (*Physalis*) (Table S3 in Appendix S2). For genera in both data sets, results were generally concordant or not significant. In *Penstemon*, the same model was rejected significantly in both data sets. In three cases (*Achillea*, *Arisaema*, and *Erodium*), the same model was rejected, but significance was reached for only one of the two data sets. In three cases (*Mimulus*, *Physalis*, and *Solanum*), different models received support, but significance was only reached for one data set. Only in one case (*Campanula*) were different models rejected significantly. Such discrepancies could be explained by (1) higher coverage of PloiDB data sets compared with M2011 (*Campanula* and *Solanum*), (2) different percentages of taxa with unreliable ploidy estimates (*Physalis*; 16% in PloiDB and 35% in M2011), or (3) different inferred percentages of polyploids (*Mimulus*; 30% in PloiDB and 45% in M2011). Taken together, these likelihood analyses indicated that the anagenesis hypothesis is rejected significantly more often than the cladogenesis hypothesis (35 vs. three genera, excluding one of the duplicated results for *Dryopteris* and for *Penstemon*, the four genera supporting different models, and *Veronica* in which both models were significantly rejected; $P = 7 \times 10^{-8}$, exact two-tailed binomial test with $N = 38$ and $p = 0.5$).

DISCUSSION

The tempo and mode by which traits evolve over time is one of the most enduring questions in evolutionary biology (Simpson, 1944). In this study, we investigated the tempo and mode by which the genome evolves by considering the pattern of polyploidization events across the phylogenetic trees for 235 unique genus-level clades (223 in PloiDB and 29 in M2011, including 16 common clades between PloiDB and M2011 plus *Dryopteris* duplicated within M2011). To this end, we used a likelihood method (ClaSSE; Goldberg and Igić, 2012) that estimates the extent to which trait changes are concentrated at speciation events or occur at a rate proportional to time. Anagenesis produces trees where the number of polyploidization events is proportional to the amount of time spent as a diploid, whereas cladogenesis produces trees where polyploidization events are proportional to the number of speciation events that diploids have undergone (which may or may not leave a node in the phylogeny of extant species, depending on subsequent extinctions). In addition, cladogenetic change is more likely when very closely related sister species differ in ploidy, because the ploidy difference is more probable if speciation itself led to a polyploid daughter (cladogenesis) than if the polyploidization event followed speciation across the very short branch to the present (anagenesis).

In the ML analysis, we found that models with only anagenetic ploidy shifts were rejected in significantly more genera (35 genera) than were models with only cladogenetic shifts (three genera), providing a strong indication that ploidy shifts are associated with speciation events in many genera. Using a Bayesian approach, we also found that the HPD interval of ϕ was consistent with a preponderance of cladogenesis (HPD falling entirely above 0.5) in four genera (across both the PloiDB and M2011 data sets), but never indicated a preponderance of anagenesis (HPD falling entirely below 0.5).

The majority of genera fail to provide a strong enough signal for ClaSSE to distinguish cladogenetic and anagenetic trait changes, which is not surprising given that polyploidization may have occurred only once or a few times in some genera and the signal in any one genus may be very weak. Simulations conducted by Magnuson-Ford and Otto (2012) demonstrated that power to detect cladogenesis, when it does occur, increases substantially with clade size, and that these methods are, if anything, conservative (type I error rates were less than 5% for an α value of 0.05). Power to detect cladogenesis is likely to be substantially reduced in groups with a high extinction rate or a high fraction of species with missing data, as these would obscure the timing of ploidy transitions. For example, extinction or unsampled species can cause nodes in the complete tree to be lost and appear as branches in the inferred tree, making cladogenetic and anagenetic events harder to distinguish.

Nevertheless, considering the PloiDB data sets, the fact that we rejected the M_A model in favor of the M_D model that includes cladogenesis in 29 of 223 likelihood ratio tests (13%) is substantially more often than expected by chance ($P = 3 \times 10^{-6}$; exact two-tailed binomial test with $N = 223$ and $p = 0.05$). These 29 genus data sets are of better quality than the other 194, on average, having more taxa (145 vs. 75) and higher sampling fraction (0.60 vs. 0.51), but similar percentage of taxa with uncertain ploidy estimates (~22%), suggesting more power in the higher quality data sets. By contrast, the M_C model was rejected in favor of the M_D model that includes anagenesis in only six of 223 of likelihood ratio tests (~3%), which is lower than expected but not significantly so ($P = 0.1245$; exact two-tailed binomial test with $N = 223$ and $p = 0.05$).

These results indicate that, at least in some genera, there is a strong signal that polyploidization is associated temporally with speciation events. Correlation does not, however, imply causation. Thus, while our data are consistent with polyploidization as an important mechanism leading to the formation of new species, it must be kept in mind that the direction of causality may be reversed: that speciation may lead to polyploidization. For example, hybrids often produce unreduced gametes at a higher rate (Harlan and deWet, 1975; Ramsey and Schemske, 1998), which implies that newly formed species may hybridize and generate polyploid descendants at a higher rate, leading to a temporal association without polyploization directly causing speciation. For hybridization to lead to a false signal of cladogenesis, however, requires a short time frame within which hybridization remains likely (temporally associated with the speciation event), which might not be the case (Levin, 2012).

Another caveat that must be considered is that likelihood models can only detect processes included within the model and may be sensitive to factors not included that may leave similar signals (see, e.g., FitzJohn, 2012; Rabosky and Goldberg, 2015). Although we do not know exactly what signals may mislead inferences about cladogenesis vs. anagenesis in ClaSSE (or BiSSE-ness), one potential issue is if taxonomists elevate intraspecific ploidy variants to species status more readily than they would for variants exhibiting the same amount of reproductive isolation without ploidy differences. Such taxonomic splitting may cause an excess of recently diverged species pairs to differ in ploidy, providing a misleading signal in favor of cladogenesis. Tree-building artifacts may also be an issue, particularly if they cause artificially short branch lengths between

diploid and polyploid sister species. Conversely, taxonomists may ignore differences displayed by newly formed polyploid species (particularly, autopolyploids; Soltis et al., 2007), lumping together recently diverged diploids and polyploids. This delay in recognizing polyploid species may obscure signals of cladogenesis.

With the caveats mentioned, this study contributes to our understanding of the role of polyploidy in speciation by providing statistical evidence that polyploidization events are synchronized over evolutionary time with the formation of new species in many groups of plants.

ACKNOWLEDGEMENTS

We thank Sean W. Graham and Michael S. Barker for helpful discussions and for their insightful comments on an early working draft of this manuscript. We are also grateful to two anonymous reviewers for helpful suggestions and to Lior Glick for helping to obtain genus diversity counts from The Plant List database. Finally, we thank Compute Canada, Fusion Genomics Corp., and UBC Zoology Computing Unit for providing access to computational resources that facilitated this research. This study was supported by the Israel Science Foundation (1265/12) to I.M., by the United States–Israel Binational Science Foundation (2013286) to I.M. and E.E.G., by the Natural Sciences and Engineering Research Council of Canada to S.P.O., and by the Canadian Institutes of Health Research Doctoral Research Award to S.H.Z.

LITERATURE CITED

Beaulieu, J. M., and M. J. Donoghue. 2013. Fruit evolution and diversification in campanulid angiosperms. *Evolution* 67: 3132–3144.

Chapman, M. A., and R. J. Abbott. 2010. Introgression of fitness genes across a ploidy barrier. *New Phytologist* 186: 63–71.

Coyne, J. A., and H. A. Orr. 2004. Speciation. Sinauer, Sunderland, Massachusetts, USA.

Darriba, D., G. L. Taboada, R. Doallo, and D. Posada. 2012. jModelTest 2: More models, new heuristics and parallel computing. *Nature Methods* 9: 772.

FitzJohn, R. G. 2012. Diversitree: Comparative phylogenetic analyses of diversification in R. *Methods in Ecology and Evolution* 3: 1084–1092.

FitzJohn, R. G., W. P. Maddison, and S. P. Otto. 2009. Estimating trait-dependent speciation and extinction rates from incompletely resolved phylogenies. *Systematic Biology* 58: 595–611.

Glick, L., and I. Mayrose. 2014. ChromEvol: Assessing the pattern of chromosome number evolution and the inference of polyploidy along a phylogeny. *Molecular Biology and Evolution* 31: 1914–1922.

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.

Grant, V. 1963. The origin of adaptations. Columbia University Press, New York, New York, USA.

Guindon, S., and O. Gascuel. 2003. A simple, fast and accurate method to estimate large phylogenies by maximum-likelihood. *Systematic Biology* 52: 696–704.

Harlan, J. R., and J. M. J. deWet. 1975. On Ö. Winge and a prayer: The origins of polyploidy. *Botanical Review* 41: 361–390.

Hugall, A. F., and D. Stuart-Fox. 2012. Accelerated speciation in colour-polymorphic birds. *Nature* 485: 631–634.

Jiao, Y., N. J. Wickett, S. Ayyampalayam, A. S. Chanderbali, L. Landherr, P. E. Ralph, L. P. Tomsho, et al. 2011. Ancestral polyploidy in seed plants and angiosperms. *Nature* 473: 97–100.

Katoh, K., and D. M. Standley. 2013. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Molecular Biology and Evolution* 30: 772–780.

Levin, D. A. 1983. Polyploidy and novelty in flowering plants. *American Naturalist* 122: 1–25.

Levin, D. A. 2012. The long wait for hybrid sterility in flowering plants. *New Phytologist* 196: 666–670.

Li, L., C. J. Stoeckert Jr., and D. S. Roos. 2003. OrthoMCL: Identification of ortholog groups for eukaryotic genomes. *Genome Research* 13: 2178–2189.

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.

Magnuson-Ford, K. S., and S. P. Otto. 2012. Linking the investigations of character evolution and species diversification. *American Naturalist* 180: 225–245.

Mandáková, T., A. D. Gloss, N. K. Whiteman, and M. A. Lysak. 2016. How diploidization turned a tetraploid into a pseudotriploid. *American Journal of Botany* 103: Advance Access published 14 April 2016, doi:10.3732/ajb.1500452.

Masterson, J. 1994. Stomatal size in fossil plants: Evidence for polyploidy in majority of angiosperms. *Science* 264: 421–424.

Mayrose, I., M. S. Barker, and S. P. Otto. 2010. Probabilistic models of chromosome number evolution and the inference of polyploidy. *Systematic Biology* 59: 132–144.

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*. *New Phytologist* 206: 27–35.

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

Meyers, L. A., and D. A. Levin. 2006. On the abundance of polyploids in flowering plants. *Evolution* 60: 1198–1206.

Otto, S. P. 2007. The evolutionary consequences of polyploidy. *Cell* 131: 452–462.

Paradis, E., J. Claude, and K. Strimmer. 2004. APE: Analyses of phylogenetics and evolution in R language. *Bioinformatics* 20: 289–290.

Penn, O., E. Privman, G. Landan, D. Graur, and T. Pupko. 2010. An alignment confidence score capturing robustness to guide tree uncertainty. *Molecular Biology and Evolution* 27: 1759–1767.

R Core Team. 2015. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Rabosky, D. L., and E. E. Goldberg. 2015. Model inadequacy and mistaken inferences of trait-dependent speciation. *Systematic Biology* 64: 340–355.

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.

Ramsey, J., and D. W. Schemske. 2002. Neopolyploidy in flowering plants. *Annual Review of Ecology and Systematics* 33: 589–639.

Rice, A., L. Glick, S. Abadi, M. Einhorn, N. M. Kopelman, A. Salman-Minkov, J. Mayzel, et al. 2015. The Chromosome Counts Database (CCDB)–a community resource of plant chromosome numbers. *New Phytologist* 206: 19–26.

Ronquist, F., M. Teslenko, P. van der Mark, D. L. Ayres, A. Darling, S. Höhna, B. Larget, et al. 2012. MrBayes 3.2: Efficient Bayesian phylogenetic inference and model choice across a large model space. *Systematic Biology* 61: 539–542.

Sabath, N., E. E. Goldberg, L. Glick, M. Einhorn, T. L. Ashman, R. Ming, S. P. Otto, et al. 2016. Dioecy does not consistently accelerate or slow lineage diversification across multiple genera of angiosperms. *New Phytologist* 209: 1290–1300.

Scarpino, S. V., D. A. Levin, and L. A. Meyers. 2014. Polyploid formation shapes flowering plant diversity. *American Naturalist* 184: 456–465.

Simpson, G. G. 1944. Tempo and mode in evolution. Columbia University Press, New York, New York, USA.

Slotte, T., H. Huang, M. Lascoux, and A. Ceplitis. 2008. Polyploid speciation did not confer instant reproductive isolation in *Capsella* (Brassicaceae). *Molecular Biology and Evolution* 25: 1472–1481.

Soltis, D. E., M. C. Segovia-Salcedo, I. Jordon-Thaden, L. Majure, N. M. Miles, E. V. Mavrodiev, W. Mei, et al. 2014. Are polyploids really evolutionary dead-ends (again)? A critical reappraisal of Mayrose et al. (2011). *New Phytologist* 202: 1105–1117.

Soltis, D. E., and P. S. Soltis. 1999. Polyploidy: Recurrent formation and genome evolution. *Trends in Ecology & Evolution* 14: 348–352.

Soltis, D. E., P. S. Soltis, D. W. Schemske, J. F. Hancock, J. N. Thompson, B. C. Husband, and W. S. Judd. 2007. Autopolyploidy in angiosperms: Have we grossly underestimated the number of species? *Taxon* 56: 13–30.

Stebbins, G. L. 1938. Cytological characteristics associated with the different growth habits in the dicotyledons. *American Journal of Botany* 25: 189–198.

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

Thompson, J. D., and R. Lumaret. 1992. The evolutionary dynamics of polyploid plants: Origins, establishment and persistence. *Trends in Ecology & Evolution* 7: 302–307.

Thorne, J. L., H. Kishino, and I. S. Painter. 1998. Estimating the rate of evolution. *Molecular Biology and Evolution* 15: 1647–1657.

Wood, T. E., N. Takebayashi, M. S. Barker, I. Mayrose, P. B. Greenspoon, and L. H. Rieseberg. 2009. The frequency of polyploid speciation in vascular plants. *Proceedings of the National Academy of Sciences, USA* 106: 13875–13879.

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.