Testing the association of phenotypes with polyploidy: An example using herbaceous and woody eudicots

Article *in* Evolution · March 2017

DOI: 10.1111/evo.13226

CITATIONS
8

READS
94

3 authors, including:

José Miguel Ponciano
University of Florida
71 PUBLICATIONS **1,309** CITATIONS

SEE PROFILE

doi:10.1111/evo.13226

Testing the association of phenotypes with polyploidy: An example using herbaceous and woody eudicots

Rosana Zenil-Ferguson,[1,2] José M. Ponciano,[3] and J. Gordon Burleigh[3]

[1] *Department of Biological Sciences, University of Idaho, Idaho 83844*

[2] *E-mail: rosanaz@uidaho.edu*

[3] *Department of Biology, University of Florida, Florida 32611*

Received August 30, 2016
Accepted February 17, 2017

Although numerous studies have surveyed the frequency with which different plant characters are associated with polyploidy, few statistical tools are available to identify the factors that potentially facilitate polyploidy. We describe a new probabilistic model, BiChroM, designed to associate the frequency of polyploidy and chromosomal change with a binary phenotypic character in a phylogeny. BiChroM provides a robust statistical framework for testing differences in rates of polyploidy associated with phenotypic characters along a phylogeny while simultaneously allowing for evolutionary transitions between character states. We used BiChroM to test whether polyploidy is more frequent in woody or herbaceous plants, based on tree with 4711 eudicot species. Although polyploidy occurs in woody species, rates of chromosome doubling were over six times higher in herbaceous species. Rates of single chromosome increases or decreases were also far higher in herbaceous than woody species. Simulation experiments indicate that BiChroM performs well with little to no bias and relatively little variance at a wide range of tree depths when trees have at least 500 taxa. Thus, BiChroM provides a first step toward a rigorous statistical framework for assessing the traits that facilitate polyploidy.

KEY WORDS: **Chromosome number, evolution, herbaceous, models, polyploidy, woody.**

EVOLUTION
INTERNATIONAL JOURNAL OF ORGANIC EVOLUTION

Polyploidy, or whole genome duplication, has occurred frequently throughout the evolutionary history of plants, greatly influencing the structure and content of plant genomes (Otto and Whitton 2000; Soltis et al. 2010; Soltis et al. 2014; Wendel 2015; Barker et al. 2016). Polyploidy is associated with shifts in ecology and distributions (Levin 1983; Brochmann et al. 1992, Brochmann and Elven 1992; Comai 2005; Parisod 2012; Arrigo et al. 2016; Laport et al. 2016), physiology (Stebbins 1950; Cavalier-Smith 1978), and other key innovations (Soltis et al. 2014; Soltis and Soltis 2016) in plants, and may play an important role in plant speciation (e.g., Wood et al. 2009; Zhan et al. 2016). Yet the factors that potentially facilitate polyploidy in plants have rarely been rigorously tested (Soltis et al. 2010).

The percentage of polyploid species and rates of polyploidy vary greatly among plant clades (e.g., Stebbins 1950; Otto and Whitton 2000; Wood et al. 2009, Arrigo and Barker 2012; Husband et al. 2013) and in different environments (e.g., Stebbins 1971; Oberlander et al. 2016), and they may be associated with various plant traits. For example, polyploidy appears to be associated with many aspects of breeding system. More perennial species are polyploid than annual species (Müntzing 1936; Stebbins 1938; Gustafsson 1948; Baquar 1976). This may be because perennial taxa have more opportunities to undergo a somatic duplication than taxa with shorter life spans (Stebbins 1938; Grant 1956, 1981). Also, for polyploids that form through the union of unreduced gametes, the chances of a successful cross with another individual with unreduced gametes may be much higher for species that breed over multiple years (Otto and Whitton 2000). Similarly, many long-lived plants that undergo vegetative or asexual propagation are polyploid (Stebbins 1938; Gustafsson 1948; Grant 1981).

Polyploidy also appears to be more common in selfing than outcrossing species, or at least annuals (e.g., Grant 1956, 1981; Stebbins 1971). The chances of a successful fertilization

involving unreduced gametes may be much greater for self-fertilizing plants than outcrossers, which would require multiple plants with unreduced gametes (Levin 1975; Husband et al. 2008). In contrast, polyploidy is also associated with dioecy or sexual dimorphism in plants (Baker 1984; Miller and Venable 2000; Ashman et al. 2013; Glick et al. 2016), but it is unclear whether polyploidy causes transitions to sexual dimorphism or if transitions to sexual dimorphism are a consequence of polyploidy.

Polyploidy is also linked to plant growth form. Numerous surveys have noted a greater proportion of herbaceous than woody polyploids species (Müntzing 1936; Stebbins 1938, 1985; Levin and Wilson 1976; Levin 1983), although this trend is not always clear. Stebbins (1938) found that woody genera of dicots had lower frequencies of polyploid taxa than herbaceous genera, but this difference was not statistically significant. Genera with only woody species also had higher chromosome numbers than genera with only herbaceous species, suggesting a history of ancient polyploidy in the woody taxa (Stebbins 1938; Levin and Wilson 1976). Stebbins (1938) proposed that the vascular cambium in woody plants may restrict cell size, and thus inhibit polyploidy, but he later proposed that ecological and historical factors are more important in determining frequency of woody and herbaceous polyploids (Stebbins 1971). Otto and Whitton (2000) used the ratio of species with even versus odd haploid chromosome numbers to estimate the incidence of polyploidy in woody and herbaceous dicots, and in contrast to herbaceous species, they found little evidence of whole genome duplications in woody species. However, Vamosi and Dickinson (2006) did not find a correlation between the herbaceous growth habit and polyploidy in Rosaceae.

Although many studies have assessed the frequency of polyploidy among plant taxa with different traits, these studies do not necessarily reveal if the traits potentially promote or facilitate polyploidy. Without a phylogenetic context, it is difficult to infer how many times polyploidy evolved in a plant group based only on the number or proportion of polyploid taxa. The purported high frequency of herbaceous polyploids also could be due to the frequency of whole genome duplications in herbaceous plants, or it could be due to the frequency with which woody polyploids become herbaceous. There have been few direct attempts to compare rates of polyploidy associated with different traits. In this study, we introduce a new probabilistic model to associate the frequency of polyploidy and chromosomal change with a binary character in a phylogeny. Our new model provides a robust statistical framework for testing differences in rates of polyploidy associated with phenotypic characters along a phylogeny while simultaneously allowing for evolutionary transitions between character states. We use this model to test whether polyploidy occurs at a higher rate in woody or herbaceous plants throughout the eudicots.

Methods

Our primary goal was to develop a statistical framework to examine the association of a binary trait with rates of whole genome duplication across a phylogeny. We first describe a continuous time Markov chain (CTMC) called BiChroM (Binary state and Chromosomal change Model), which is a probabilistic model that allows for transitions in chromosome numbers associated with a binary state. In this study, the binary states we chose represent either a herbaceous (H) or a woody (W) growth form, and we test for differences in chromosome doubling rates (as a proxy of whole genome duplication) between herbaceous and woody eudicots. The input for the model is an ultrametric, bifurcating phylogenetic tree, and for each leaf taxon in the tree, the number of chromosomes, and whether the leaf taxon is herbaceous or woody (see example in supporting information). The transitions in chromosome number in BiChroM are similar to those allowed in ChromEvol models (Mayrose et al. 2010; Glick and Mayrose 2014), but we allow different rates of chromosome change to be associated with each binary character state. In BiChroM, chromosome numbers for herbaceous and woody species can increase by one, decrease by one, or double, and there also can be transitions between the woody and herbaceous growth forms (Fig. 1).

We first formally describe the BiChroM model, and then we describe the inferential procedure using BiChroM to test for differences in rates of chromosome doubling associated with woody and herbaceous states. Because BiChroM requires the use of a large and sparse infinitesimal probability matrix defined by transition rates, we also performed simulation experiments to evaluate the estimates of BiChroM using phylogenetic trees of different sizes and depths.

MODEL

In our description of the model, we use woody (W)/herbaceous (H) as the binary character trait; however, the model can work with any binary character. BiChroM is defined through the use of an instantaneous transition probability matrix (Q) with ten parameters that represent rates of change per million years (Fig. 1). The parameters λ_H and λ_W represent the addition of one chromosome in herbaceous and woody lineages respectively; μ_H and μ_W represent the loss of one chromosome in herbaceous and woody lineages, respectively. The parameters ρ_H and ρ_W represent rates of polyploidy, or more specifically, the rate of doubling of the number of chromosomes, in herbaceous (H) and woody (W) plants, respectively. The parameter q_{HW} represents the transition rate from the herbaceous to woody state, and q_{WH} represents the transition rate from the woody to herbaceous state. To limit the overall size of the transition matrix, we define the last chromosome number state as 50+ (i.e., more than fifty chromosomes). We used two ancillary parameters, ε_W and ε_H, to represent transitions in the binary trait state in lineages with more than 50 chromosomes

for woody and herbaceous taxa, respectively. Estimation of parameters ε_W and ε_H can be difficult since our sample of taxa that have more than 50 haploid chromosomes is small. These parameters were added as a mathematical tool to separate the dynamics of chromosomal and binary character change for taxa with more than 50 chromosomes, because we wanted to prevent biases in estimation of the other eight parameters that describe the changes of the majority of the sample.

The states at which BiChroM is defined consist of different combinations of two quantities: $X(t)$, which represents the number of chromosomes of a taxon in a haploid cell at time t, and $Y(t)$, which represents the binary trait of the taxon at time t ($Y(t) = H$ for herbaceous, or $Y(t) = W$ for woody; Fig. 1). BiChroM is a continuous time Markov chain defined in the infinite state space $\{1, 2, 3, \ldots, 50, 50+\}$ x $\{H, W\}$, the state space was organized by listing first all the chromosome numbers associated with herbaceous taxa followed by the chromosome number states for woody taxa.

LIKELIHOOD CALCULATION

The instantaneous change in the probabilistic model BiChroM was defined by an infinitesimal probability Q matrix, whose elements q_{ij} are defined as the instantaneous transition rates (see Karlin and Taylor 1975, p. 151). This matrix allows for the calculation of probabilities of chromosomal and binary trait changes along the branches of a phylogeny denoted by T as follows:

$$Q_{ik,\,jl} = \begin{cases} \lambda_k & \text{if} \quad i \leq 50,\ j = i+1,\ k = l \\ \mu_k & \text{if} \quad i > 1,\ j = i-1,\ k = l \\ \rho_k & \text{if} \quad j = 2i \leq 50 \text{ or if } 2i > 50,\ j = 50+;\ k = l \\ q_{kl} & \text{if} \quad j = i < 50,\ k \neq l \\ \varepsilon_k & \text{if} \quad j = i > 50, k \neq l \\ -\left(\sum_{l=H}^{W} \sum_{j=1, j\neq i}^{50+} Q_{ik,jl}\right) & \text{if} \quad i = j,\ k = l. \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

where $i, j = 1, 2, \ldots, 50, 50+$ are chromosome numbers, and $k, l = W, H$ are growth forms. Note that Q matrix defined above has a dimension of 102×102, and it is highly sparse (i.e., many entries are zeros). This is significantly larger and more complicated than most matrices used in phylogenetic comparative methods (but see Landis et al. 2013 for a discussion on large, but not necessarily sparse, Q matrices).

The probability of changes of chromosome and growth form is $P(t) = e^{Qt}$ along a branch with length t (2). The calculation of the exponential of the matrix was done numerically using the package expm in R with Higham's algorithm (Higham 2009; Goulet et al. 2013). Later, we calculated the negative log likelihood function based on the probabilities calculated in equation (2) using a pruning algorithm (Felsenstein 1981) by modifying the internal pruning calculation used in the corHMM R package (Beaulieu et al. 2013) to obtain the probabilities defined in BiChroM's Q matrix. We calculated the probabilities at the root of the tree using a discrete uniform initial probability over the state space of BiChroM. We also tested the distribution proposed by Fitzjohn et al. (2009) for the root of the tree, which assigns probabilities for the states based on weighted averages of conditional likelihoods.

LIKELIHOOD RATIO TESTS

To test whether herbaceous and woody taxa have different rates of polyploidy (i.e., chromosome number doubling), we defined a reduced model from BiChroM by making the rates of chromosome doubling for herbaceous and woody taxa equal (i.e., $\rho_H = \rho_W = \rho$). This reduced model has nine parameters, instead of the ten parameters in the full model. We calculated the log likelihood and performed optimizations to obtain the maximum likelihood estimates for the reduced model using the same approach as the full BiChoM. Using the log likelihood values of the full $l(full\ BiChroM)$ and reduced $l(reduced\ BiChroM)$ models, we tested the hypothesis $H_0: \rho_W = \rho_H$ of equal rates of chromosome doubling for woody and herbaceous taxa by calculating the statistic $D = -2\,l(reduced\ BiChroM) + 2\,l(full\ BiChroM)$ (3), which has an asymptotic distribution of $\chi^2_{(1)}$ (Kalbfleisch 1985).

We also tested if the rates of single chromosomal changes were different for herbaceous and woody taxa by optimizing a reduced model in BiChroM. The null hypothesis $H_0: \lambda_H = \lambda_W$ and $\mu_H = \mu_W$ was tested using the same statistic D in equation (3) using the likelihood of the reduced model with eight parameters and comparing it to a χ^2 distribution with two degrees of freedom.

DATASET

To test for differences in rates of polyploidy in woody and herbaceous eudicots using BiChroM, we used the ultrametric maximum likelihood tree from Zanne et al. (2014). We initially

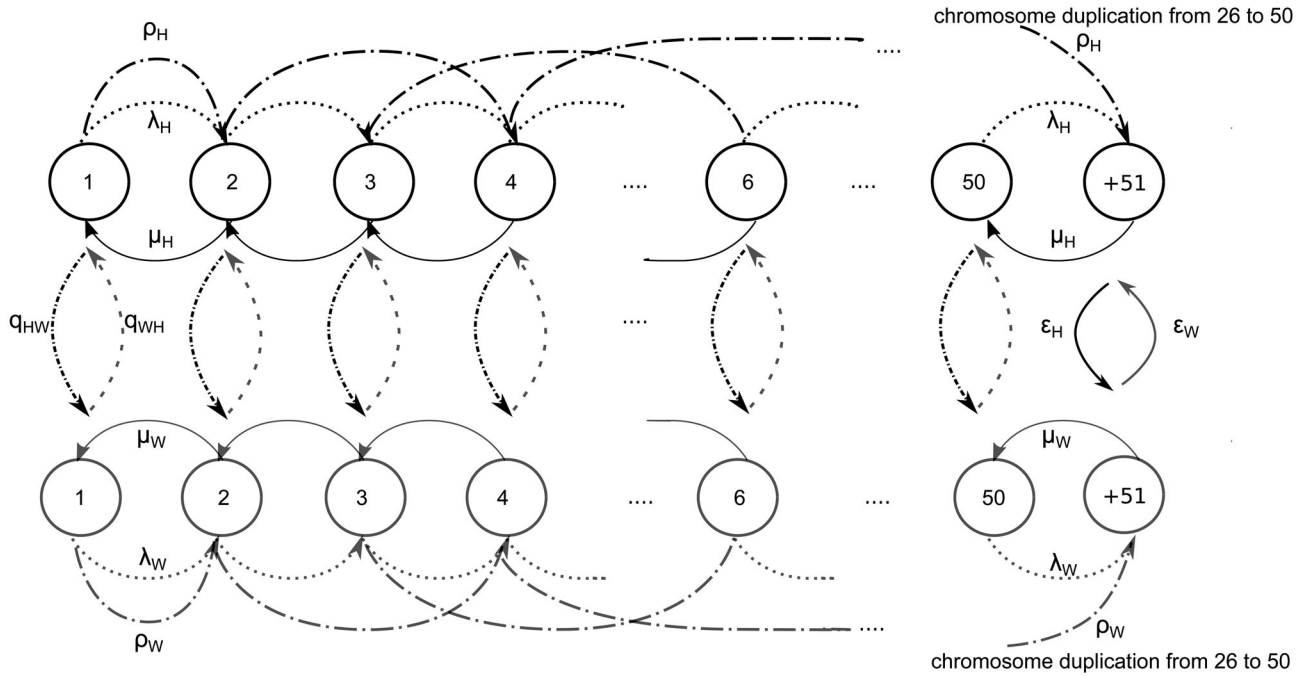

Figure 1. Binary trait and chromosomal change model (BiChroM) with ten parameters. Six parameters represent changes in chromosome number: $\rho_{H/W}$, doubling of chromosome number (herbaceous/woody species), $\lambda_{H/W}$, gain of one chromosome (herbaceous/woody species), $\mu_{H/W}$ loss in one chromosome (herbaceous/woody species). The transition rate from woody to herbaceous is q_{WH} and from herbaceous to woody is q_{HW}. Two ancillary parameters that accumulate the binary trait changes that happen in taxa with more than 50 chromosomes were added to prevent biases in the estimation of the other eight parameters.

pruned all species from the tree except eudicots using Newick utilities (Junier and Zdobnov 2010). We then downloaded all the information of angiosperm chromosome numbers for the eudicots from the Chromosome Counts Database (Rice et al. 2014) using the R package chromer (Pennell 2015). We cleaned the chromosome count records using a custom R package CCDBcurator, (www.github.com/roszenil/CCDBcurator), which outputs all the possible chromosome numbers associated with each taxon. To reduce the size of the Q matrix, we used the haploid chromosome number instead of the raw chromosome counts as input for the BiChoM model, so the resulting Q matrix has dimensions 102×102. To obtain the haploid chromosome numbers, we divided chromosome numbers by two. This did not result in an integer for 130 taxa, and for those taxa we defined their haploid number as smallest preceding integer (e.g., if the chromosome number divided by two was 4.5, we reported the haploid chromosome number as 4). Taxa with a haploid chromosome number greater than 50 were scored as "50+" (e.g., of BiChroM states in Supporting Information).

We identified the eudicot species with unambiguous chromosome counts that were in the phylogenetic tree, and we pruned the phylogenetic tree so that it only included these species. Finally, we scored whether each of the remaining species in the tree was woody or herbaceous. We obtained woody/herbaceous data for some of these species from the Global Woodiness Database (Zanne et al. 2014), and we scored many species manually. If the state of the species was ambiguous or our personal scoring differed from the scoring in the Global Woodiness database, that species was pruned from the tree and excluded from the analyses. After all of the pruning, the tree had 4711 eudicot species with associated haploid chromosome number and woody/herbaceous character states (available at Dryad Digital Repository https://doi.org/10.5061/dryad.6g2c7)

ANALYSES

To explore the likelihood surface for the eudicot woody/herbaceous dataset, we performed 500 optimizations on the high performance cluster at the University of Florida. We drew a Latin hypercube random sample (lhs package; Carnell 2009) for the parameter space $(0,1)^{10}$ that served as starting points for the minimization of the negative log-likelihood via a Nelder-Mead algorithm (Nelder and Mead 1965), which was restricted to 500 iterations. The initial exploration took approximately 60 hours per optimization (i.e., for 500 optimizations, a total of 30,000 hours exploring the parametric space), and the resulting values were ranked from smallest to largest negative log-likelihood. The three points that resulted in the smallest (best) negative log-likelihood values were used as initial points of a new optimization

Table 1. Maximum likelihood estimates for parameters in BiChroM models.

Model	Parameter	Estimate
BiChroM full	λ_H	0.126
	λ_W	0.001
	μ_H	0.249
	μ_W	0.002
	ρ_H	0.036
	ρ_W	0.006
	q_{HW}	0.040
	q_{WH}	0.022
	Negative Log-likelihood	13,045.998
BiChoM reduced	λ_H	0.082
	λ_W	0.010
	μ_H	0.114
	μ_W	0.024
	$\rho = \rho_H \rho = \rho_W$	0.020
	q_{HW}	0.012
	q_{WH}	0.006
	Negative Log-Likelihood	13,159.1184

using a subplex algorithm (nloptr package, Liu and Nocedal 1989; Ypma 2014), where the convergence criterion was based on the changes in the precision of the negative log likelihood of no more of 10^{-8}. Each of the three optimizations took approximately 48 hours and resulted in the maximum likelihood estimates for BiChroM shown in Table 1. The implementation described here is available as an example in an R package available at www.github.com/roszenil/chromploid, and raw R file optimizations can be found at https://doi.org/10.5061/dryad.6g2c7 (see Supporting Information for description of the files).

We also calculated relative profile likelihoods $R(\rho_H)$ and $R(\rho_W)$. For parameter ρ_H the profile likelihood was calculated on the grid (0.03000, 0.03005, ..., 0.04195, 0.04200), and for parameter ρ_W it was calculated in the grid (0.004, 0.00405 ..., 0. 006, 0.007, ..., 0.010, 0.020) (Fig. 2). Profile likelihoods for every parameter are useful diagnostic tools for examining the variation in parameters of interest by accounting for the uncertainty that the rest of the parameters add to the full likelihood (Pawitan 2001). For a given parameter, the maximum of the profile likelihood is located at the MLE, and its curvature indicates how much information the data contains to estimate this parameter. Numerically flat profile likelihoods around the MLE indicate poor estimability. Also, if the profile likelihood is perfectly flat for a given parameter value, the parameter is said to be nonidentifiable (see Ponciano et al. 2012 and citations therein for an extensive discussion on this topic). The relative part of the profile likelihoods refers to the profile likelihoods being rescaled to the maximum likelihood value at the MLEs of the parameters (see Supporting Information for a detailed description of the likelihood calculations and plot of log-likelihoods). We used the asymptotic approximation of the likelihood function to the $\chi^2_{(1)}$ distribution to calculate profile confidence intervals (Fig. 2).

To investigate a possible correlation between the estimators of chromosome doubling rate and the phenotypic change rate we also calculated the relative bivariate profile likelihoods $R(\rho_H, q_{HW})$ and $R(\rho_W, q_{WH})$ (Fig. 3). Optimizations for the bivariate profile likelihoods were done with a reduced precision of 10^{-5}.

PERFORMANCE OF BiChroM

BiChroM involves the optimization of ten parameters, which can result in biased or inaccurate estimates (Moler and Van Loan 2003; Landis et al. 2013). Therefore, it is important to investigate the performance of the BiChroM model under different scenarios. We performed seven different simulation experiments. In four of these experiments, we varied the number of tips in the trees (Fig. 4A and B), and in the other three experiments we varied the tree height (Fig. 4C and D). For the four simulation experiments that varied the number of taxa n, we used the function sim.tree() from the geiger R package (Harmon et al. 2008) to simulate 100 trees using a pure birth process with number of taxa $n = 100, 500, 1000,$ and 2500, respectively, and average tree height older than 100 million years (since the height of the eudicot tree used in our estimations is 136 million years). Next, we modified the function sim.char() in geiger to calculate the exponential of the Q matrix using the exponential algorithm of Higham (2009) with rate parameter values to simulate changes chromosome numbers and growth form along the simulated trees (available in the R package www.github.com/roszenil/chromploid). In all simulations, we assumed that the root of the tree is fixed with five haploid chromosomes and a woody character state. We chose parameter values for the simulation that are approximately the same as those estimated in the eudicot tree ($\lambda_H = 0.12$, $\lambda_W = 0.001$, $\mu_H = 0.25$, $\mu_W = 0.002$, $\rho_H = 0.036$, $\rho_W = 0.006$, $q_{HW} = 0.04$, $q_{WH} = 0.02$, $\varepsilon_H = 1.79$, $\varepsilon_W = 1.57x10^{-14}$). The simulated trees and tip values were used as input for BiChroM and optimized using nloptr (100 optimizations per sample size, limited to 24 hours; Liu and Nocedal 1989; Ypma 2014) to calculate MLEs of parameters.

For the three tree height simulation experiments, we generated the 100 trees for each tree height using the sim.tree() birth only algorithm, with the stopping criterion being total height $l = 25, 50,$ or 75 million years and average sample size more than 1000 taxa. We simulated chromosome number and growth form along trees using a modified version of sim.char() and the same parameter values as in the tree size simulations. The simulated tip values and trees were used as input for an

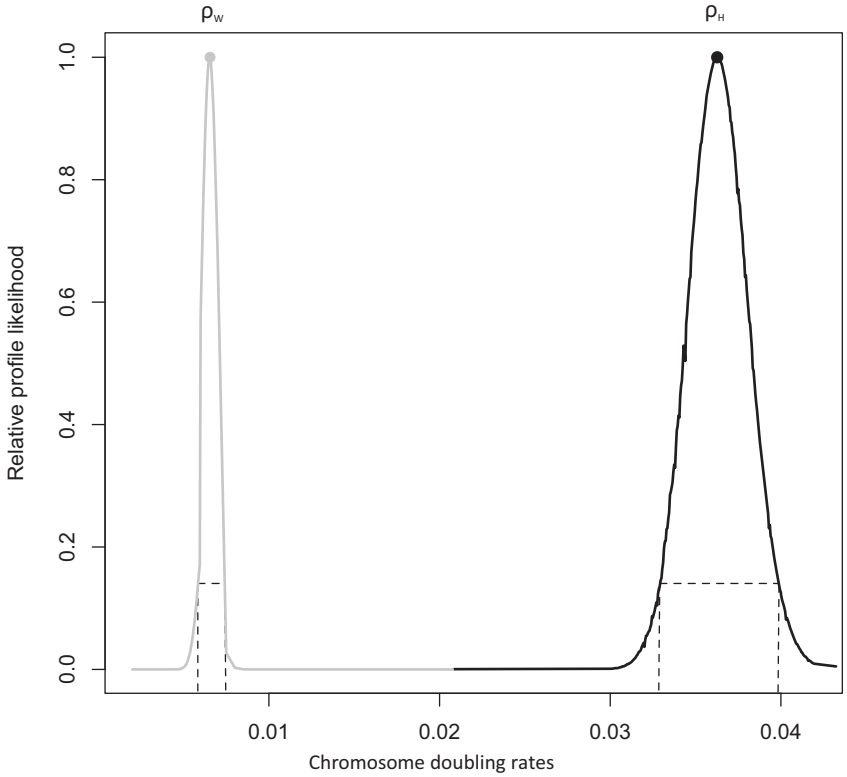

Figure 2. Relative profile likelihoods for rates of chromosome doubling, ρ_W for woody eudicots (gray) and ρ_H for herbaceous eudicots (black). Since the two profile likelihoods do not overlap, there is strong evidence that the rates are different. Dotted lines represent likelihood-confidence intervals for each parameter.

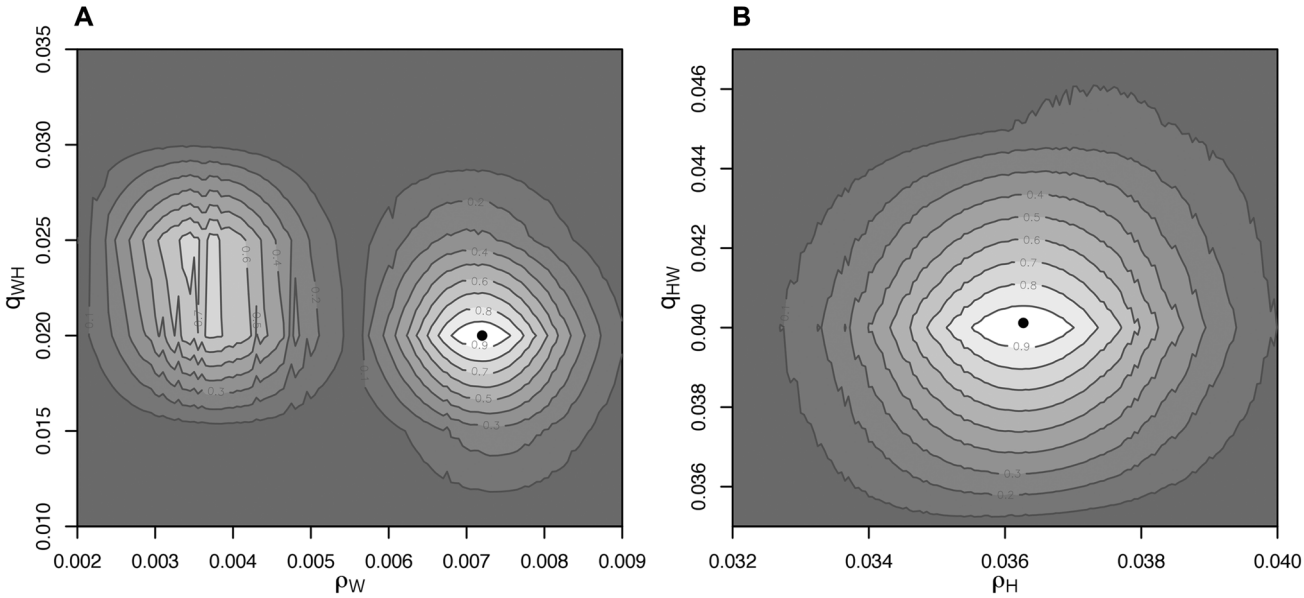

Figure 3. Bivariate relative profile likelihoods for rates of chromosome doubling and binary trait change for woody (A) and herbaceous (B) eudicots. Both bivariate likelihoods show the lack of correlation between estimates of chromosome doubling and binary trait change rate estimates. For woody eudicots (A) optimizations show the presence of multiple maxima. Also at a reduced optimization tolerance (10^{-5}) the MLE for ρ_W is shifted to the right (by 0.0009) from the MLE found in the global optimization.

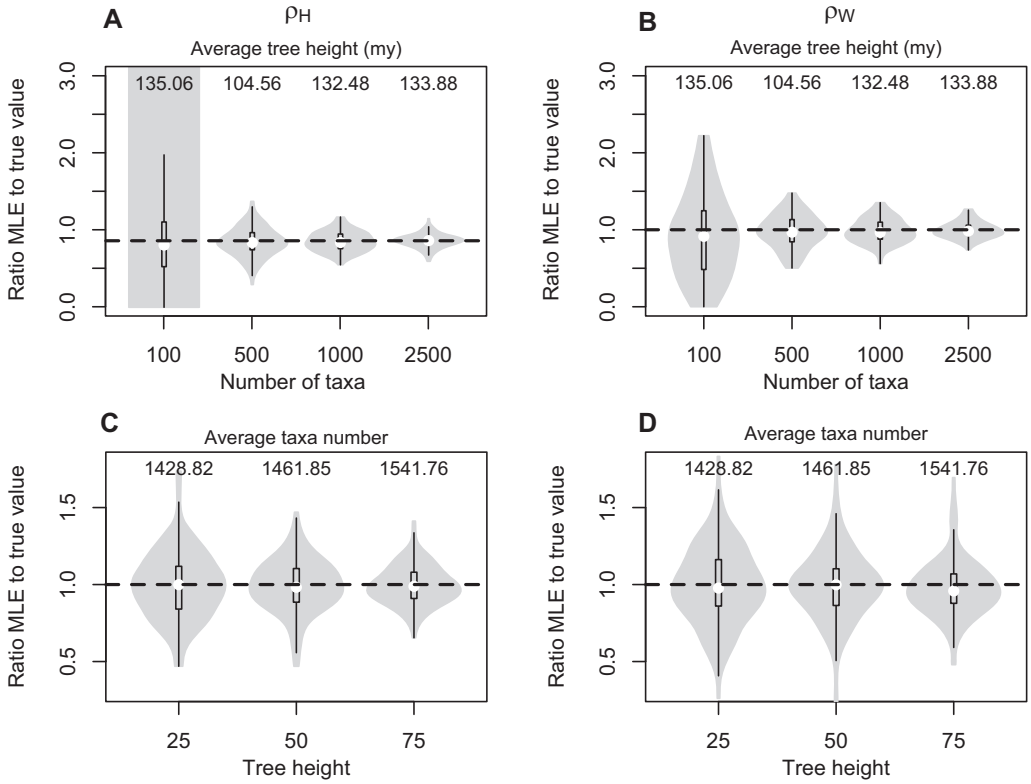

Figure 4. Violin plots of ratio of maximum likelihood estimates to true value used for simulations. Maximum likelihood estimates were obtained from simulated data and trees with fixed number of taxa for parameters (A) ρ_H and (B) ρ_W; and from simulated data and trees with fixed total height for (C) ρ_H and (D) ρ_W. Chromosome number and growth form changes were simulated given known polyploidy rates of herbaceous ($\rho_H = 0.036$) and woody plants ($\rho_W = 0.006$). Dotted lines are placed at 1 and represent maximum likelihood estimates that recovered the truth. Increasing number of taxa in the tips of the tree greatly decreases variation in the estimation of the chromosome doubling rate. Increasing tree height moderately decreases variation in the estimation of chromosome doubling.

optimization procedure that allowed us to optimize MLEs for rate parameters (100 optimizations per tree height lasting less than 24 hours). Simulation files can be found in the Supporting Information.

Results

RATES OF CHROMOSOME CHANGE IN WOODY AND HERBACEOUS SPECIES

Woody eudicots show far more stability in chromosome number than herbaceous eudicots. Rates of gaining one chromosome and losing one chromosome are 100 times larger in herbaceous taxa ($\lambda_H = 0.126$; $\mu_H = 0.249$) than rates associated with woody plants ($\lambda_W = 0.001$; $\mu_W = 0.002$), and the rate of chromosome doubling is six times larger for herbaceous plants ($\rho_H = 0.036$) than for woody plants ($\rho_W = 0.006$; Table 1). Although, the rate of chromosome doubling for woody plants is small, it is significantly greater than zero ($CI^{95\%}(\rho_W) = (0.005,0.007)$), indicating that chromosome doubling occurs in woody taxa. By comparing the negative log-likelihood values of the full and reduced BiChroM models (Table 1), we calculated the likelihood ratio test $D = 226.239$ (from eq. 3) to see if we can reject the null hypothesis that herbaceous and woody taxa have equal rates of chromosome doubling. The P-value in the $\chi^2_{(1)}$ distribution is $P < 1x10^{-41}$, strongly rejecting the null hypothesis. The relative profile likelihoods show further evidence for the difference in chromosome doubling rates between woody and herbaceous taxa (Fig. 2). The likelihood-confidence intervals for the estimates of chromosome doubling in woody and herbaceous species do not overlap (Fig. 2). The likelihood ratio test of equal rates of gain and loss of a single chromosome in herbaceous and woody species ($H_0: \lambda_0 = \lambda_1$ and $\mu_0 = \mu_1$) also strongly rejected this hypothesis ($D = 500.657$, $P < 1x10^{-108}$).

In the full model, the transition rate from herbaceous to woody ($q_{HW} = 0.040$) is almost two times larger than the transition rate from woody to herbaceous ($q_{WH} = 0.022$). The ancillary parameters in the full BiChroM (ε_H, ε_W) were difficult to estimate due to the small number of taxa (32 out of 4711) with more than 50 chromosomes. Still, we found no evidence that these parameters affect the estimation of the other parameters, since the minimum

negative log-likelihood and the maximum likelihood estimates of the other parameters were consistent across optimizations (see Supporting Information for description of optimization files in http://doi.org/10.5061/dryad.6g2c7).

Bivariate relative profile likelihoods showed no correlation between the estimates of the chromosome doubling rates and the binary trait change rates (Fig. 3). For woody plants, optimizing the bivariate likelihood was especially difficult, since multiple optima are present and rates of chromosome doubling are slow and close to zero (Fig. 3A).

All the calculations presented in Table 1 assume discrete uniform probability distributions at the root of the tree. Optimizations assuming the Fitzjohn et al. (2009) initial distribution at the root of the tree resulted in the same maximum likelihood estimates, but the optimizations took 1.5 times longer (i.e., approximately 100 hours per optimization) because more calculations are needed at the root using this algorithm. There were no differences in the parameter estimates using the uniform distribution and the approach of Fitzjohn et al. (2009) at the root, indicating that our sample is sufficiently large that the uncertainty at the root of our tree does not affect the final inference. However, this might not be the case with smaller datasets.

SIMULATIONS

The simulation experiments demonstrate that the sample size of the tree affects the resulting parameter estimates. Although there is little evidence of bias in the estimates (i.e., the mean of the estimates remains very close to true value), for all eight parameters (λ_H, λ_W, μ_H, μ_W, ρ_H, ρ_W, q_{HW}, q_{WH}), increasing sample size reduced variance in the parameter estimates (see Supporting Information for violin plots of all parameters. Simulation results are fully listed in Supporting Information. For rate ρ_H, increasing sample size is especially important for decreasing estimation variance when assuming that the true rate $\rho_H = 0.036$ is fast (Fig. 4A and B). However, for simulations of tree height, given a large sample size (ave. 1400 taxa) the depth of the tree has little effect on the variances in estimation of the parameters. In general, variances for chromosome doubling rate estimates are small even for trees with heights of 25 million years (between (0.01, 0.06) for ρ_H, and (0.002, 0.010) for ρ_W for all tree heights, Fig. 4C and D. See Supporting Information for violin plots of all parameters).

Discussion

While numerous studies have tried to associate polyploidy with life-history traits in plants by comparing the number of recent polyploidy species (e.g., Stebbins 1938; Otto and Whitton 2000; Vamosi and Dickinson 2006), little is known about character-associated rates of polyploidy across plants. We describe a new statistical approach to estimate and compare rates of chromosome doubling associated with binary phenotypic characters. Although in our sample of 4711 eudicot species herbaceous species have both fewer chromosomes on average and less variation in chromosome numbers (mean = 12.33, SE = 7.34) than woody eudicots (mean = 15.17, SE = 9.94), the rate of chromosome doubling associated with herbaceous eudicots is far greater than the rate of chromosome doubling for woody eudicots. The lack of association between chromosome doubling rate estimates and transition rate estimate between woody and herbaceous states also show that for this particular dataset both parameters are estimable. As specified, our model is not targeted to reveal whether polyploidy increases or decreases according to a binary trait (Fig. 3). To test the biological association between polyploidy and transition rates it is necessary to specify a model of the joint variation of these parameters, for example by means of a hierarchical model. Finding an a priori hypothesis (model) relating these two quantities is not a trivial task. As such, this problem is material for future studies. The present study, however highlights the value of using a phylogenetic, modeling-based approach to reveal links between chromosome number evolution and phenotypic traits.

While our analyses do not necessarily imply a mechanistic link between growth form and polyploidy, they demonstrate how chromosome doubling rates can differ dramatically based on life-history traits. The reason that woody plants have lower rates of polyploidy is not clear. While the slower rates of chromosome doubling in woody plants could be due to their longer average generation time or slower rates of molecular evolution (e.g., Smith and Donoghue 2008), there is also evidence that annual herbaceous plants a have a much lower frequency of polyploidy than perennials (e.g., Stebbins 1938; Otto and Whitton 2000). Polyploidy is perceived to be common among angiosperms (e.g., Wood et al. 2009), but our analyses suggest that polyploidy events might be relatively uncommon in many species (i.e., woody species). They also suggest that the extreme heterogeneity in the polyploid rate among taxa should be incorporated in future studies estimating overall rates of polyploidy.

BiChroM builds upon the ChromEvol models that estimate the probability of chromosome number changes along a phylogeny (Mayrose et al. 2010; Glick and Mayrose 2014). BiChroM considers some of the same chromosomal changes defined by ChromEvol models, but in contrast, it associates them with the state of the binary character, enabling tests for differences in the rates of chromosomal change given the value of the binary trait. ChromEvol includes models that allow for additional types of chromosomal changes that are not considered in the current version of BiChroM (e.g., model M4 in ChromEvol has demiploidization). BiChroM can be expanded to allow for more types of chromosomal changes, or potentially more trait associations

(e.g., either multistate discrete characters or multiple discrete characters). However, these changes will increase the parameter space and computational complexity of the analyses, likely leading to more uncertainty in the parameter estimates. The amount of added uncertainty would depend on the size of the dataset. Our simulations suggest that trees with at least 500 taxa may be necessary to obtain parameter estimates with low variance using the current version of BiChroM (Fig. 4), and trees would likely have to be larger for more complex models. Our study also shows the difficulty in estimating slow rates of evolution, like the chromosome doubling rate for woody eudicots, despite having a large sample (Fig. 3A). In any case, simulation experiments can be used to ensure that it is possible to accurately estimate values of parameters from more complex models.

Detecting the amount of uncertainty surrounding parameter estimates is also critical for assessing model performance. The properties of maximum likelihood estimation enable a practical calculation of reliable profile-confidence intervals for BiChroM parameters (Fig. 2). Although computationally intensive to produce, these profile-confidence intervals allow us to assess the amount of information in the phylogeny and uncertainty in parameter estimates, and they have been shown to have much better asymptotic coverage than the traditional Wald intervals (Pawitan 2001). The symmetry and the small variation of the profile likelihoods for our dataset suggest that the sample is large enough to approximate the true value of the chromosome doubling rates and that calculation of likelihood-confidence intervals is appropriate for the polyploidy parameters.

One limitation of the BiChroM model is that it assumes a single value for the transition rates between all the woody and the herbaceous states, as well as between chromosome numbers, across the entire tree. This is a central limitation shared by all complex models of trait evolution: while a large phylogeny may be necessary to obtain reliable parameter estimates (e.g., Fig. 4A and B), there often is extensive heterogeneity in the patterns of trait evolution across a large tree, necessitating the use of a more complex model. Yet this heterogeneity is seldom accounted for in phylogenetic comparative analyses. Beaulieu et al. (2013) addressed variation in transition rates throughout a phylogeny using a hidden rates model (HRM) of binary trait evolution. Similar to our study (Table 1), within Campanulids, they found that the transition rate from herbaceous to woody is approximately twice the woody to herbaceous transition rate when constant transition rates throughout the tree are assumed. This is surprising in light of paleobotanical evidence suggesting that woody species are decreasing in number while herbaceous species are expanding (Chaney 1940). However, when a HRM is used, in many clades woody to herbaceous transitions happen faster than herbaceous to woody transitions (Beaulieu et al. 2013).

Another possible source of bias in the estimation of binary trait and chromosomal change comes from diversification rates. Our model implicitly assumes that the shape of the phylogeny is independent of the evolution of the characters, but polyploidy may affect diversification rates (e.g., Mayrose et al. 2011, 2014; Scarpino et al. 2014), as may the woody or herbaceous states. Such character-associated diversification processes can bias the parameter estimates (Maddison 2006). For example, if polyploidy increases extinction rates (e.g., Mayrose et al. 2011), this could cause underestimation of chromosome doubling parameters and possibly also underestimation of transitions to the herbaceous state. However, the effects of polyploidy on diversification are still debated (e.g., Mayrose et al. 2014), and both chromosome number and the woody or herbaceous state could affect diversification. Therefore, it is difficult to predict what these biases might be. Missing taxa in the trees also may lead to similar biases. If the missing data are randomly distributed on the tree, it may cause a general decrease of power, without biasing the estimates, but if specific chromosome numbers, ploidies, or herbaceous/woody states are undersampled, estimates could be biased just like character-associated diversification can affect final estimations.

Modeling approaches that simultaneously consider diversification and polyploidy have proven difficult (Wood et al. 2009), as have approaches that associate binary traits with diversification (Maddison and Fitzjohn 2014; Rabosky and Goldberg 2015). However, as more new statistical approaches to detect diversification linked to binary traits are being developed (e.g., Rabosky and Huang 2015), linking multiple changes of chromosomes, a binary trait change, and diversification rates may be possible. Models considering chromosome doubling, phenotypic change, and diversification simultaneously will be highly computationally intensive and require very large datasets, but BiChroM can be expanded to include diversification rates linked to a binary trait changes or chromosome doubling events. A similar likelihood approach could enable investigations of correlations between net diversification and chromosomal change for the different phenotypes.

Plant scientists have been studying patterns associated with polyploidy for many years. Recent advances in phylogenetic knowledge, phylogenetic comparative methods, and computational power enable us to test hypotheses generated long ago based on careful observations and surveys across vascular plants. Although the relationship between woodiness and rates of polyploidy is likely complex and tied to other characters, our analyses strongly support the general trend of lower rates of polyploidy, and more stable chromosome numbers, in woody than herbaceous eudicots, and for this large dataset a correlation between estimates of chromosome doubling and binary trait change estimates could not be found. Our study further demonstrates how de novo statistical

models can be defined to address large-scale macroevolutionary questions linking for instance polyploidy to life history traits. This model and framework can be adjusted to seek a better understanding of other traits like life cycles (e.g., annual, perennial), breeding system (monoecy, dioecy), or habitat (tropical, temperate), that may be associated with polyploidy rates. Our model also can be expanded to deal with more complex or multiple characters. With this work, we hope to provide a practical, first step approach toward defining a rigorous statistical framework for studying the evolution of life history traits and ploidy.

AUTHOR CONTRIBUTIONS

R. Zenil-Ferguson designed research, wrote article, organized data, performed statistical analyses, and programmed R package chromploid. J.M. Ponciano performed statistical analyses, and wrote article. J.G. Burleigh designed research, organized data, and wrote article.

ACKNOWLEDGMENTS

The authors thank Dr. Jeremy Beaulieu for helpful advice in the implementation of pruning algorithm, Dr. Nick Matzke for suggestions about the use of large matrices and Drs. Luke Harmon and François Michonneau who offered advice about chromploid R package implementation. We also thank Dr. Scott A. McKinley for the suggestion of exploring the performance of matrix exponentials, and Lauren Suarez for scoring many of the woody and herbaceous character states. This manuscript was greatly improved thanks to the critical comments of associate editor Dr. Cécile Ané and two anonymous reviewers. Support for R. Zenil-Ferguson provided by NSF DDIG DEB-1501547 and NSF DEB-1208912.

DATA ARCHIVING

The doi for our data is 10.5061/dryad.6g2c7.

LITERATURE CITED

Arrigo, N., and M. S. Barker. 2012. Rarely successful polyploids and their legacy in plant genomes. Curr. Opin. Plant Biol. 15:140–146.

Ashman, T.-L., A. Kwok, and B. C. Husband. 2013. Revisiting the dioecy-polyploidy association: alternate pathways and research opportunities. Cytogent. Genome Res. 140:241–255.

Baker, H. G. 1984. Some functions of dioecy in seed plants. Am. Nat. 124:149–158.

Baquar, S. R. 1976. Polyploidy in the flora of Pakistan in relation to latitude, life form, and taxonomic groups. Taxon 25:621–627.

Barker, M. S., B. C. Husband, and J. C. Pires. 2016. Spreading Winge and flying high: the evolutionary importance of polyploidy after a century of study. Am. J. Bot. 103:1139–1145.

Beaulieu, J. M., B. C. O'Meara, and M. J. Donoghue. 2013. Identifying hidden rate changes in the evolution of a binary morphological character: the evolution of plant habit in campanulid angiosperms. Syst. Biol. 62:725–737.

Brochmann, C., and R. Elven. 1992. Ecological and genetic consequences of polyploidy in arctic *Draba* (Brassicaceae). Evol. Trends Plants 6:111–124.

Brochmann, C., P. S. Soltis, and D. E. Soltis. 1992. Recurrent formation and polyphyly of nordic polyploids in *Draba* (Brassicaceae). Am. J. Bot. 79:673–688.

Carnell, R. 2009. Latin hypercube samples. Package "Lhs." Version 0.5. CRAN Repository.

Cavalier-Smith, T. 1978. Nuclear volume control by nucleoskeletal DNA, selection for cell volume and cell growth rate, and the solution of the DNA C-value paradox. J. Cell Sci. 34:247–278.

Chaney, R. W. 1940. Tertiary forests and continental history. Geo. Soc. Am. Bull. 51:469–488.

Comai, L., 2005. The advantages and disadvantages of being polyploid. Nat. Rev. Genet. 6:836–846.

Felsenstein, J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. J. Mol. Evol. 17:368–376.

Fitzjohn, R. G., W. P. Maddison, and S. P. Otto. 2009. Estimating trait-dependent speciation and extinction rates from incompletely resolved phylogenies. Syst. Biol. 58:595–611.

Glick, L., and I. Mayrose. 2014. Chromevol: assessing the pattern of chromosome number evolution and the inference of polyploidy along a phylogeny. Mol. Biol. Evol. 31:1914–1922.

Glick, L., N. Sabath, T.-L. Ashman, E. Goldberg, and I. Mayrose. 2016. Polyploidy and sexual system in angiosperms: is there an association? Am. J. Bot. 103:1223–1235.

Grant, V. P. 1956. The influence of breeding habit on the outcome of natural hybridization in plants. Am. Nat. 90:319–322.

Grant, V. P. 1981. Plant speciation, 2nd ed. Columbia Univ. Press, New York.

Goulet, V., C. Dutang, M. Maechler, D. Firth, M. Shapira, and M. Stadelmann. 2013. Expm: matrix exponential. R Package Version 0.99–0. CRAN Repository.

Gustafsson, Å. 1948. Polyploidy, life-form and vegetative reproduction. Hereditas 34:1–22.

Harmon, L. J., J. T. Weir, C. D. Brock, R. E. Glor, and W. Challenger. 2008. GEIGER: investigating evolutionary radiations. Bioinformatics 24:129–131.

Higham, N. J. 2009. The scaling and squaring method for the matrix exponential revisited. SIAM Rev. 51:747–764.

Husband, B. C., S. J. Baldwin, and J. Suda. 2013. The incidence of polyploidy in natural plant populations: major patterns and evolutionary processes. Pp. 255–276 *in* I. J. Leitch, J. Greilhuber, J. Dolezel, and J. F. Wendel, eds. Plant genome diversity, vol. 2, Springer, Vienna.

Husband, B. C., B. Ozimec, S. L. Martin, and L. Pollock. 2008. Mating consequences of polyploid evolution in flowering plants: current trends and insights from synthetic polyploids. Int. J. Plant Sci. 169:195–206.

Junier, T., and E. M. Zdobnov. 2010. The Newick utilities: high-throughput phylogenetic tree processing in the UNIX shell. Bioinformatics 26:1669–1670.

Kalbfleisch, J. G. 1985. Probability and statistical inference II. Springer-Verlag, New York.

Karlin, S., and H. M. Taylor. 1975. A first course in stochastic processes. Academic, San Diego, CA.

Landis, M. J., N. J. Matzke, B. R. Moore, and J. P. Huelsenbeck. 2013. Bayesian analysis of biogeography when the number of areas is large. Syst. Biol. 62:789–804.

Laport, R. G., R. L. Minckley, and J. Ramsey. 2016. Ecological distributions, phonological isolation, and genetic structure in sympatric and parapatric populations of the *Larrea tridentata* polyploidy complex. Am. J. Bot. 103:1358–1374.

Levin, D. A. 1975. Minority cytotype exclusion in local plant populations. Taxon 24:35–43.

———. 1983. Polyploidy and novelty in flowering plants. Am. Nat. 122:1–25.

Levin, D. A., and A. C. Wilson. 1976. Rates of evolution in seed plants: net increase in diversity of chromosome numbers and species numbers through time. Proc. Natl. Acad. Sci. USA 73:2086–2090.

Liu, D. C., and J. Nocedal. 1989. On the limited memory BFGS method for large scale optimization. Math. Program. 45:503–528.

Maddison, W. P. 2006. Confounding asymmetries in evolutionary diversification and character change. Evolution 60:1743–1746.

Maddison, W. P., and R. G. Fitzjohn. 2014. The unsolved challenge to phylogenetic correlation tests for categorical characters. Syst. Biol. 64:127–136.

Marchant, D. B., D. E. Soltis, and P. S. Soltis. 2016. Patterns of abiotic niche shifts in allopolyploids relative to their progenitors. New Phytol 212:708–718.

Mayrose, I., M. S. Barker, and S. P. Otto. 2010. Probabilistic models of chromosome number evolution and the inference of polyploidy. Syst. Biol. 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. (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 333:1257–1257.

Miller, J. S., and D. L. Venable. 2000. Polyploidy and the evolution of gender dimorphism in plants. Science 289:2335–2338.

Moler C., and C. Van Loan. 2003. Nineteen dubious ways to compute the exponential of a matrix, twenty-five years later. SIAM Rev. 45:3–49.

Müntzing, A. 1936. The evolutionary significance of autopolyploidy. Hereditas 21:363–378.

Nelder, J. A., and R. Mead. 1965. A simplex method for function minimization. Comput. J. 7:308–313.

Oberlander, K. C., L. L. Dreyer, P. Goldblatt, J. Suda, and H. P. Linder. 2016. Species-rich and polyploidy poor: insights into the evolutionary role of whole-genome duplication from the Cape flora biodiversity hotspot. Am. J. Bot. 103:1336–1347.

Otto, S. P., and J. Whitton. 2000. Polyploid incidence and evolution. Annu. Rev. Genet. 34:401–437.

Parisod, C. 2012. Polyploids integrate genomic changes and ecological shifts. New Phytol. 193:297–300.

Pawitan, Y. 2001. In all likelihood: statistical modeling and inference using likelihood. Oxford Univ. Press, Oxford, U. K.

Pennell, M. W. 2015. Package Chromer. Available at https://github.com/ropensci/chromer.

Ponciano, J. M., J. G. Burleigh, E. L. Braun, and M. L. Taper. 2012. Assessing parameter identifiability in phylogenetic models using data cloning. Syst. Biol. 61:955–972.

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

Rabosky, D. L., and H. Huang. 2016. A robust semi-parametric test for detecting trait-dependent diversification. Syst. Biol. 65:181–193.

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.

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

Soltis, D. E., R. J. A. Buggs, J. J. Doyle, and P. S. Soltis. 2010. What we still don't know about polyploidy. Taxon 59: 1387–1403.

Soltis, D. E., C. J. Visger, and P. S. Soltis. 2014. The polyploidy revolution then . . . and now: Stebbins revisited. Am. J. Bot. 101:1057–1078.

Soltis, P. S., and Soltis, D. E. 2016. Ancient WGD events as drivers of key innovations in angiosperms. Curr. Opin. Plant Biol. 30:159–165.

Soltis, P. S., J. G. Burleigh, A. S. Chanderbali, M.-J. Yoo, and D. E. Soltis. 2010. Gene and genome duplication in plants. Pp. 269–298 *in* K. Dittmar and D. Liberles, eds. Evolution after gene duplication. Wiley-Blackwell, Hoboken, NJ.

Soltis, P. S., X. Liu, D. B. Marchant, C. J. Visger, and D. E. Soltis. 2014. Polyploidy and novelty: Gottlieb's legacy. Phil. Trans. Roy. Soc. B 369:20130351.

Stebbins, G. L. 1938. Cytological characteristics associated with the different growth habits in the dicotyledons. Am. J. Bot. 25:189–198.

———. 1950. Variation and evolution in higher plants. Adison-Wesley, London, U. K.

———. 1971. Chromosomal evolution in higher plants. Edward Arnold, London, U. K.

Vamosi, J. C., and T. A. Dickinson. 2006. Polyploidy and diversification: a phylogenetic investigation in Rosaceae. Int. J. Plant Sci. 167:349–358.

Wendel, J. F. 2015. The wondrous cycles of polyploidy in plants. Am. J. Bot. 102:1753–1756.

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. Proc. Nat. Acad. Sci. USA 106:13875–13879.

Ypma, J. 2014. Nloptr: R Interface To Nlopt. R Package Version, 1. CRAN Repositories.

Zanne, A. E., D. C. Tank, W. K. Cornwell, J. M. Eastman, S. A. Smith, R. G. Fitzjohn, D. G. Mcglinn, B. C. O'Meara, A. T. Moles, P. B. Reich, et al. 2014. Three keys to the radiation of angiosperms into freezing environments. Nature 506:89–92.

Zhan, S. H., M. Drori, E. E. Goldberg, S. P. Otto, and I. Mayrose. 2016. Phylogenetic evidence for cladogenic polyploidization in land plants. Am. J. Bot. 103:1252–1258.

Associate Editor: C. Ané
Handling Editor: P. Tiffin

Supporting Information

Additional Supporting Information may be found in the online version of this article at the publisher's website:

Table S1. Sample dataset of chromosome numbers and binary trait. The dataset contains taxon names, chromosome numbers, and a binary state (1 = herbaceous, 0 = woody).

Table S2. Transformed dataset needed for BiChroM analyses. Transformations can be done using function bichrom_dataset() in the chromploid R package.

Figure S1. Profile log-likelihood values for ρ_w(grey) and ρ_H(black). For woody eudicots the tails of the loglikelihood for parameter ρ_w decay faster than for parameter ρ_w. Optimizations for ρ_H are harder to perform when closer to zero.

Figure S2. Violin plots for the ratio between maximum likelihood estimates and true value of the parameters of BiChroM parameters when sample size (number of taxa in the tree) increases.

Figure S3. Violin plots for the ratio between maximum likelihood estimates and the true value of BiChrom parameters in simulation experiments when tree height increases.