New Phytologist

Research

Methods

A non-homogeneous model of chromosome-number evolution to reveal shifts in the transition patterns across the phylogeny

Anat Shafir[1], Keren Halabi[1], Marcial Escudero[2] and Itay Mayrose[1]

[1]School of Plant Sciences and Food Security, George S. Wise Faculty of Life Sciences, Tel Aviv University, Tel Aviv 69978, Israel; [2]Department of Plant Biology and Ecology, University of Seville, Reina Mercedes, ES-41012 Seville, Spain

Author for correspondence:
Itay Mayrose
Email: itaymay@tauex.tau.ac.il

Received: *14 November 2022*
Accepted: *6 February 2023*

New Phytologist (2023)
doi: 10.1111/nph.18805

Key words: chromosome-number evolution, convergent evolution, dysploidy, evolutionary models, polyploidy, rate heterogeneity.

Summary

- Changes in chromosome numbers, including polyploidy and dysploidy events, play a key role in eukaryote evolution as they could expediate reproductive isolation and have the potential to foster phenotypic diversification. Deciphering the pattern of chromosome-number change within a phylogeny currently relies on probabilistic evolutionary models. All currently available models assume time homogeneity, such that the transition rates are identical throughout the phylogeny.
- Here, we develop heterogeneous models of chromosome-number evolution that allow multiple transition regimes to operate in distinct parts of the phylogeny. The partition of the phylogeny to distinct transition regimes may be specified by the researcher or, alternatively, identified using a sequential testing approach. Once the number and locations of shifts in the transition pattern are determined, a second search phase identifies regimes with similar transition dynamics, which could indicate on convergent evolution.
- Using simulations, we study the performance of the developed model to detect shifts in patterns of chromosome-number evolution and demonstrate its applicability by analyzing the evolution of chromosome numbers within the Cyperaceae plant family.
- The developed model extends the capabilities of probabilistic models of chromosome-number evolution and should be particularly helpful for the analyses of large phylogenies that include multiple distinct subclades.

Introduction

Chromosome number is a central feature of eukaryotic genomes, providing, perhaps, the most basic description of the genomic organization of a lineage. For nearly a century, a large body of evolutionary and ecological research was devoted to deciphering chromosome-number dynamics in plant phylogenies and linking these changes to the underlying characteristics of the examined clades (Guerra, 2008). The most recognizable chromosome-number change is through a whole-genome duplication (WGD), or more generally polyploidization, which describes the acquisition of one or more complete chromosome sets to the genome. Polyploidization can spur genome reorganization (Chen & Ni, 2006; Tayalé & Parisod, 2013), increase evolvability (Martin & Husband, 2012), may cause shifts in life history traits and ecological tolerances (Parisod *et al.*, 2010; Tsuda *et al.*, 2013; Segraves, 2017; Baniaga *et al.*, 2020; Moura *et al.*, 2021; van de Peer *et al.*, 2021), and could alter diversification dynamics (Mayrose *et al.*, 2011; Eric Schranz *et al.*, 2012; Tank *et al.*, 2015; Levin, 2019). Aside from polyploidization, single-chromosome changes represent another common pathway underlying chromosome-number variation. While a gain or a loss of an entire chromosome content (i.e. aneuploidy transition) is considered rare evolutionary events (Guerra, 2008; although exceptions exist; Weiss-Schneeweiss & Schneeweiss, 2013), a variety of processes may lead to an increase or a decrease by a single-chromosome number (ascending and descending dysploidy, respectively), while preserving most of the genomic content (reviewed in Mayrose & Lysak, 2020). Recent evidence suggests that dysploidy transitions are prevalent in many plant lineages (Carta *et al.*, 2020) with possible implications on macroevolutionary and genomic processes (Escudero *et al.*, 2014; Dodsworth *et al.*, 2016; Mandáková & Lysak, 2018).

The importance of polyploidy and dysploidy transitions to plant evolutionary and ecological research has motivated the development of advanced computational methodologies for deciphering the dynamics of chromosome numbers in plant phylogenies and to link these changes to the underlying characteristics of the examined clades. Perhaps the most widely used method is ChromEvol (Mayrose *et al.*, 2010; Glick & Mayrose, 2014).

ChromEvol is a likelihood-based method that relies on a continuous-time Markov process. The process is determined based on several parameters that define the rate of change for different types of chromosome-number transitions (e.g. WGD or dysploidy, see the Description section). Given a specified phylogeny, and an assignment of chromosome numbers to tip taxa, ChromEvol allows a range of inferential tasks to be carried out. This includes ancestral states reconstruction, estimation of the expected number of polyploidy and dysploidy transitions occurring along each branch of the phylogeny, assignment of ploidy levels to extant taxa, and hypothesis testing by comparing the fit of different model variants to a particular dataset (e.g. Cusimano *et al.*, 2012; Pellicer *et al.*, 2013; McCann *et al.*, 2016). Several studies extended the ChromEvol framework to allow for more sophisticated analyses. Freyman & Höhna (2018) developed ChromoSSE to account for a possible association between chromosome-number evolution and diversification dynamics. The BiChroM model (Zenil-Ferguson *et al.*, 2017, 2018) allows for possible associations between binary phenotypic traits and chromosome-number evolution. Such an association was also allowed in the chromeplus R package (Blackmon *et al.*, 2019), but in the latter model, states of the binary trait could be additionally linked to different rates of diversification.

In ChromEvol models developed to date, it is assumed that the transition pattern is identical through time and across all parts of the phylogeny. This time-homogeneity assumption is unlikely to hold, especially when applied to large phylogenies that include several distinct subclades. Indeed, Carta *et al.* (2020) applied the ChromEvol framework on a large angiosperm megaphylogeny and concluded that further progress in modeling chromosomal evolution requires the development of models that include heterogeneous dynamics across a phylogeny. Moreover, Rice & Mayrose (2021) demonstrated that when large phylogenies are analyzed, the best-fitted ChromEvol models are frequently determined as inadequate. Several studies have recently tested for shifts in the pattern of chromosome-number evolution in large phylogenies, finding multiple rate shifts in the plant family Cyperaceae and pervasive rate heterogeneity in multiple large insects phylogenies (Márquez-Corro *et al.*, 2019; Ruckman *et al.*, 2020; Sylvester *et al.*, 2020). For example, Márquez-Corro *et al.* (2019) examined the hypothesis that shifts in chromosome-number evolution coincide with alternations in diversification rates in Cyperaceae. In that study, the Cyperaceae phylogeny was partitioned prior to analysis, fitting a model to each subclade independently of the others. Such an approach results in loss of information as it does not consider the branches that connect the different clades, treating them as independent and disregarding the fact that ancestral state probabilities at the root of each subtree are affected by the evolutionary path leading to it. Furthermore, such an approach is restricted to the small subset of subtrees predetermined by the researcher, whereas the many possible alternative shifts remain unexplored. Lastly, the approach is unable to perform a more delicate analysis, in which only some parameters are clade specific (e.g. the polyploidy rate), while others are shared across the phylogeny (e.g. the ascending dysploidy rate).

Here, we present a new model that allows for multiple shifts in the transition pattern, such that different parts of the phylogeny evolve according to different dynamics of chromosome-number change. This modeling scheme allows the identification of subclades that exhibit a unique evolutionary pattern, for example, subclades that are hotspots of polyploidizations, those in which the diploidization process is particularly rapid, or those that exhibit a unique monoploid base number relative to the rest of the phylogeny. Using a sequential model-testing approach – similar in nature to branch-site codon models that aim to identify episodic adaptive evolution along certain lineages (e.g. Anisimova & Yang, 2007) or for identifying groups with altered diversification patterns along the tree (Alfaro *et al.*, 2009) – the method allows the identification of the most plausible shifting points in chromosome-number dynamics, without a restriction to a certain taxonomic classification or to any *a priori* belief. A second sequential testing phase, inspired by the approach developed for continuous traits by Ingram & Mahler (2013), was implemented to identify whether model complexity may be reduced by testing whether independent shifts converge into a similar transition pattern. Such findings could inform on clade-wide evolutionary convergence that could occur due to, for example, similar habitats or ecological contexts (Stebbins, 1938; Gustafsson, 1948). Following a detailed description of the method, we examine its performance using simulations and present its utility by testing for shifts in chromosome-number dynamics in Cyperaceae and compare our results to those previously obtained by Márquez-Corro *et al.* (2019).

Description

The baseline ChromEvol model

The ChromEvol probabilistic framework, first presented in Mayrose *et al.* (2010), is formulated as a continuous-time Markov process. Current models assume time homogeneity, such that the transition process is identical throughout the phylogeny. This process is represented by an instantaneous rate matrix **Q**, which describes the transition rates from a genome with i haploid chromosomes to a genome with j haploid chromosomes. Entries in this matrix are determined based on a combination of model parameters that defines the rate of change for different types of transitions, through which chromosome numbers can change. Specifically, the model in Eqn 1 allows for several types of transitions: an increase by a single-chromosome number (ascending dysploidy; with rate parameter λ) and a decrease by a single-chromosome number (descending dysploidy; rate parameter δ). For both types of dysploidy transitions, in addition to the constant terms (λ and δ), rate modifiers allow for the possibility that rates of ascending and descending dysploidy are linearly dependent on the current number of chromosomes (parameters λ_l and δ_l, respectively). Three types of transitions allow for polyploidization events: (1) an exact duplication of the number of chromosomes, with rate ρ; (2) *demi-polyploidy*, occurring at rate μ, which accounts for possible multiplications of the number of chromosomes by 1.5. This allows, for example, the generation of a hexaploid from a tetraploid lineage via the fusion of reduced and unreduced gametes; and (3)

base number transition. In this transition type, let β represent the monoploid base number of a focal clade; then, the addition of any multiplication of β is allowed at rate ν. For example, assuming $\beta = 10$, the transitions $10 \rightarrow 20$, $10 \rightarrow 30$, or $12 \rightarrow 22$, are allowed in a single step. Combining these possibilities, the most general rate matrix is defined as follows:

$$\mathbf{Q}_{ij} = \begin{cases} \lambda + \lambda_1(i-1) & j = i+1 \\ \delta + \delta_1(i-1) & j = i-1 \\ \rho & j = 2i \\ \mu & j = 1.5i \\ \nu & (j-i) \text{ is divisible by } \beta \\ 0 & \text{otherwise} \end{cases} \quad \text{Eqn 1}$$

The diagonal entries are determined by the constraint that each row in $\mathbf{Q}$ sums to zero. The rate matrix described above allows the likelihood function to be computed, given a specified phylogeny and assignments of chromosome numbers to the tip taxa, as described previously in Mayrose *et al.* (2010). We note that the model relies on a given phylogenetic tree with its associated branch lengths, which may be in units that are proportional to time or to the amount of genetic change (e.g. nucleotide substitutions). The units of $\mathbf{Q}$ are then inversely proportional to these units and are used to transform the given units of branch lengths to units of chromosome-number transitions by a factor that is uniform across the phylogeny.

An initial application of the above model to the Cyperaceae dataset revealed that the model is highly inadequate (following the model adequacy procedures described in Rice & Mayrose, 2021), most probably since genome duplications are allowed to occur at the same rate in genomes with small or large numbers of chromosomes, while in reality the latter are expected to be rarer (Wood *et al.*, 2009). We thus extended the above model by allowing rates of genome duplications to depend on the current number of chromosomes. To this end, two functions of dependencies were implemented: linear ($\rho + \rho_1 (i - 1)$) as defined in Eqn 1 for dysploidy transitions and exponential ($\rho \times e^{\rho_e(i-1)}$), as was previously implemented for dysploidy transitions in ChromoSSE (Freyman & Höhna, 2018). Applying these two variants to the Cyperaceae dataset indicated that the exponential function is more adequate (detailed description of the model adequacy tests is given in Supporting Information Notes S1). Thus, the most general time-homogeneous rate matrix used in this study is:

$$\mathbf{Q}_{ij} = \begin{cases} \lambda + \lambda_1(i-1) & j = i+1 \\ \delta + \delta_1(i-1) & j = i-1 \\ \rho \times e^{\rho_e(i-1)} & j = 2i \\ \mu & j = 1.5i \\ \nu & (j-i) \text{ is divisible by } \beta \\ 0 & \text{otherwise} \end{cases} \quad \text{Eqn 2}$$

A novel heterogeneous model

Our new model uses a similar parameterization to the models described in Eqn 2, but allows the transition pattern to vary across different parts of the phylogeny. This is obtained using different rate matrices for different parts of the phylogeny. For simplicity, in the description below we assume that only three types of transitions are possible: ascending dysploidy, descending dysploidy, and WGD but, in practice, all transitions defined in Eqn 2 are allowed. Accordingly, a certain clade may be assigned with its own rate category k, with a respective set of parameters, λ_k, δ_k, and ρ_k (Eqn 3):

$$\mathbf{Q}^k_{ij} = \begin{cases} \lambda_k & j = i+1 \\ \delta_k & j = i-1 \\ \rho_k & j = 2i \\ 0 & \text{otherwise} \end{cases} \quad \text{Eqn 3}$$

The categorization to k distinct models can be either determined *a priori* or inferred in a sequential testing approach, in which the number and locations of shifts in the transition pattern are determined using the forward and backward algorithms detailed below.

Automatic identification of regime shifts To automatically detect the most plausible locations of shifts in chromosome-number dynamics, without a restriction to a certain taxonomic classification or to any *a priori* belief, we implemented a sequential search that was inspired by the algorithm detailed in Ingram & Mahler (2013). The scan is composed of two phases. The forward phase identifies the most likely shift locations in the transition pattern and partitions the phylogeny into different transition regimes. The backward phase then detects regimes with similar transition dynamics and unifies them, thereby reducing model complexity. See Notes S2 for detailed descriptions of these two phases.

Model comparison In each iteration of the forward and backward phases, model comparison is conducted to test whether the increase in the likelihood scores computed using the more complex models justifies the additional free parameters. In the automatic search algorithm, the more complex model (e.g. M_k with k transition regimes) is chosen over the simpler model (e.g. M_{k-1}) if the difference between their corrected AICc values exceeds a certain threshold, termed ΔAICc*. While users may choose a prespecified value of ΔAICc* (e.g. 0 or 2), we followed the simulation-based procedure proposed by Ingram & Mahler (2013) to determine its value for a given dataset. Accordingly, a large number of simulated datasets are generated based on the parameter values of the optimized homogeneous model, which serves as the null model. Then, for each simulated dataset, two models are fitted to the data: the homogeneous model with a single transition regime and the best model with a single shift (i.e. a model with two transition regimes), as inferred by a single

iteration of the forward search phase. Given these two inferences, the ΔAICc value per simulated dataset is computed. A distribution of the ΔAICc values is obtained by applying this procedure across all simulated datasets. Since any inference of non-homogeneous model is a false-positive result, the ΔAICc* value is chosen as the 95th percentile, although other percentiles may be similarly chosen. For all our analyses performed here, 100 simulations were generated and the ΔAICc threshold was determined as the 95th percentile.

Assessing performance using simulations Simulations were used to investigate the performance of the developed heterogeneous model in terms of statistical power and accuracy to identify the subclade with altered transition pattern. The simulations were based on a given set of model parameters and given a fixed phylogeny. To depict realistic data scenarios, the phylogenies and model parameters used in the simulations were based on two empirical datasets, representing the Cyperaceae and Solanaceae plant families, each characterized by a distinct pattern of chromosome-number change. While the Cyperaceae phylogeny exhibits high rate of dysploidy transitions, low rate of polyploidizations, and was previously suggested to exhibit rate heterogeneity (Márquez-Corro *et al.*, 2019), Solanaceae is characterized by relatively conserved rates of chromosome evolution (Wu & Tanksley, 2010) and is less likely to present a strong signal of rate heterogeneity. The Cyperaceae phylogeny was obtained from Márquez-Corro *et al.* (2019), while the Solanaceae phylogeny was obtained from the mega seed plant phylogeny reconstructed in Smith & Brown (2018), trimmed to include only taxa from Solanaceae with available chromosome numbers extracted from the Chromosome Counts Database (CCDB; Rice *et al.*, 2015). For taxa with multiple entries, the median was used as a representative number. The resulting Cyperaceae phylogeny included 825 taxa and the Solanaceae phylogeny included 412 taxa. To avoid exhaustive running times and to allow fair comparisons between the different simulation scenarios, the model phylogeny of each family was obtained by retaining a random sample of 400 tip taxa. Given these trees and chromosome numbers for tip taxa, model parameters for each family were inferred using ChromEvol. The homogeneous model used for the optimization procedure was the one detailed in Eqn 2. The maximum likelihood estimates of the model parameters as inferred using the homogeneous model for each family are summarized in Table S1, and the obtained functions for modeling the dependency of the transition rates on the number of chromosomes are plotted in Fig. S1.

For each of these two sets of model parameters, we generated simulated data while varying the magnitude of rate heterogeneity among the foreground and background clades, as well as varying the size of the foreground clade. The foreground clade is defined to include all lineages descendant from the branch in which rate shift had occurred, and the background clade contains all other lineages, including the root node. Specifically, we define f as the rate multiplication factor, such that the rates of all transition types simulated for the foreground clade are multiplied by f, relative to their values in the background clade. Four f values were examined: 0.25, 0.5, 2, and 4. To retain the number of simulated transitions similar across different values of f, we rescaled the trees such that the expected number of transitions is constant (multiplying the branches of the foreground clade by f has the same effect as multiplying the transition rate matrix). To test the effect of the foreground-clade size, for each simulation we randomly selected clades that include *c.* 20, 40, 80, and 160 taxa. In total, 32 simulation scenarios were generated this way: a combination of four f values, four different sizes for the foreground clade, and the two sets of model parameters and trees, corresponding to the two plant families (Cyperaceae and Solanaceae). For each simulation scenario, 50 independent simulated datasets were generated.

Another set of simulations was conducted with parameter values as inferred using a heterogeneous model with two rate regimes (M_2) for the Cyperaceae and Solanaceae datasets detailed above, each with 400 species. The simulated parameters are detailed in Table S2, and the differences in the dysploidy and polyploidy transition rates between the foreground and background clades as a function of the number of chromosomes in the genome are shown in Fig. S2. Similar to the simulations detailed above, four different sizes for the foreground clade were examined (20, 40, 80, and 160). In total, eight simulation scenarios were generated this way. For each simulation scenario, 50 datasets were generated.

For each simulated dataset, ChromEvol inferences were conducted either using the homogeneous model with a single rate regime, or the heterogeneous model with two regimes and the minimum foreground-clade size set to 5. To allow for efficient computations, the parallelization option was used with 20 CPUs per dataset. For a given simulated dataset, the heterogeneous model was selected over the homogeneous model if the difference in the AICc value between the two models exceeded ΔAICc* (determined as detailed above). For performance evaluation, the statistical power was defined as the fraction of simulations in which the model with two rate regimes was chosen over the homogeneous model. Accuracy in detecting the correct clade was defined as the overlap, in the number of tip taxa, between the true and inferred foreground clades. Specifically, let C_t and C_i denote the set of taxa belonging to the true and inferred foreground clade, respectively. The accuracy in inferring the clade where rate shift had occurred is defined as the intersection between these two sets divided by their union: $|C_t \cap C_i| / |C_t \cup C_i|$. Note that accuracy was computed only if the existence of rate heterogeneity was supported according to the ΔAICc* threshold. To summarize accuracy across simulations with the same parameter values, we required that the existence of rate heterogeneity is determined for at least 10 simulated datasets. To achieve a sufficient number of simulated datasets ($n > 10$) for which accuracy can be computed, we executed additional simulations for each parameter set until reaching 10 datasets for which rate heterogeneity was detected. For datasets with clade sizes 20 and 40, however, the statistical power to detect rate heterogeneity was very low for most simulated scenarios, such that too many simulations were needed to reach the required number of $n = 10$. Therefore, accuracy is presented only for simulated scenarios with clade sizes 80 and 160.

The Cyperaceae empirical data analysis The Cyperaceae phylogeny used here was originally published by Márquez-Corro *et al.* (2019). This phylogeny included 1057 species out of the *c.* 5500 circumscribed to Cyperaceae and was reconstructed based on a supermatrix alignment consisting of two nuclear ribosomal spacer regions (ETS and ITS), four plastid genes (*mat*K, *ndh*F, *rbc*L, and *ycf*6) and the plastic spacer region *trn*C-*ycf*6. Given the assembled alignment, a maximum likelihood phylogeny was reconstructed using RAxML (Stamatakis, 2006). The phylogeny was then dated using penalized likelihood (Smith & O'Meara, 2012) with a total of 11 calibrations points. Chromosome numbers were collected from the literature and online databases (for details see Márquez-Corro *et al.*, 2019). In total, chromosome counts were obtained for 825 taxa included in the phylogeny. Species in the phylogeny without known chromosome numbers were pruned.

Implementation and availability The heterogeneous model described here was implemented in C++ using the Bio++ phylogenetic library (Guéguen *et al.*, 2013) and was integrated within the ChromEvol software. The program is available as an open-source program at GitHub: https://github.com/anatshafir1/chromevol. Installation instructions and user guidelines are provided in README.md.

Results

Assessing performance using simulations

Simulations with different magnitudes of rate heterogeneity Simulations were used to investigate the statistical power and accuracy of the method to detect rate heterogeneity in simulated datasets in which a single shift in the transition pattern had occurred. We simulated data according to two baseline patterns of transition rates – following those inferred for the Cyperaceae and Solanaceae plant families. Chromosome-number evolution in Cyperaceae is characterized by very high dysploidy rates and much lower rates of polyploidization, while Solanaceae is characterized by moderate rates of both types of transitions (Table S1). For each baseline transition pattern, we varied the magnitude of rate heterogeneity among the foreground clade (the clade descending from the branch where rate shift had occurred) and the background clade (the rest of the phylogeny) according to four different values of the rate multiplication factor $f = (0.25, 0.5, 2, 4)$, the former two values representing rate deceleration in the foreground clade and the latter two representing rate acceleration. In addition, we also varied the size of the foreground clade to encompass *c.* 20, 40, 80, and 160 taxa out of the 400 taxa in each phylogeny.

As expected, the power to correctly reject the null hypothesis (i.e. a homogeneous model) increased with the magnitude of rate heterogeneity simulated. This was true for both simulated patterns and for both rate acceleration and deceleration (Fig. 1). In addition, statistical power was noticeably lower as the size of the foreground clade decreased. This may be expected since small clades contain insufficient information, in terms of the number of chromosome-number transitions, to allow a robust differentiation from the background transition pattern. For example, in simulations with $f = 4$, in 94% of the Cyperaceae simulations and 96% of the Solanaceae simulations, the null model was rejected when the foreground-clade size was 160, but statistical power was reduced to 80% and 42% for the Cyperaceae and Solanaceae, respectively, when the foreground-clade size was 80. While there was some general trend of higher statistical power in detecting rate acceleration compared with rate deceleration (i.e. comparing results obtained with $f = 4$ to $f = 0.25$ and $f = 2$ to $f = 0.5$), this trend was inconsistent across different simulation scenarios (Fig. 1).

For those datasets in which the existence of a shift in the transition pattern was inferred, we assessed the accuracy of the method to detect the correct placement of the rate-shift location. Accuracy was measured as the overlap between the true (i.e. simulated) and inferred foreground clades (see the Description section). Because many of the simulations with clade size of 20 and 40 did not support the existence of rate heterogeneity (and thus accuracy could not be computed), the results in Fig. 2 are presented only for clade sizes 80 and 160. Full results are presented in Table S3. Generally, the trend obtained for the accuracy was in line with those discussed above for the statistical power. Accuracy was high when the magnitude of rate variation was large (i.e. $f = 0.25$ and $f = 4$; Fig. 2) and was higher when the foreground-clade size was large (compare clade size of 160–80; Fig. 2). However, very high accuracies were also obtained for smaller clades. For example, in Solanaceae, accuracy was above 0.68 for all simulation scenarios with a foreground-clade size of 40 or more when simulating with high extent of rate variation ($f = 0.25$ and $f = 4$; Table S3). These results indicate that in case the existence of rate heterogeneity is determined, the location in which rate shift had occurred is inferred rather accurately.

Simulations with empirically inferred parameters for heterogeneous model The simulations detailed above-examined cases of rate acceleration or deceleration, such that the rates of all transition types are multiplied by the same factor in the foreground clade relative to the background clade. Many other patterns of rate heterogeneity exist in which each rate parameter is shifted to a new value, independent of the other parameters. While the number of such combinations is infinite, here we focused on two sets of parameters, each representing a realistic rate-shift scenario as inferred using a heterogeneous model with two rate regimes for the Cyperaceae and Solanaceae datasets (Fig. S2; Table S2). The estimated parameters of the Cyperaceae dataset represent a scenario of a sharp shift in the transition pattern from a background model with very high dysploidy rates and mild rates of polyploidizations to a foreground model with no dysploidy transitions and many polyploidization events. The Solanaceae dataset represents a milder shift in the transition pattern, where the polyploidy rate is similar for the foreground and the background models, but the rates of both ascending and descending dysploidy are lower at the foreground clade compared with the background clade.

New Phytologist

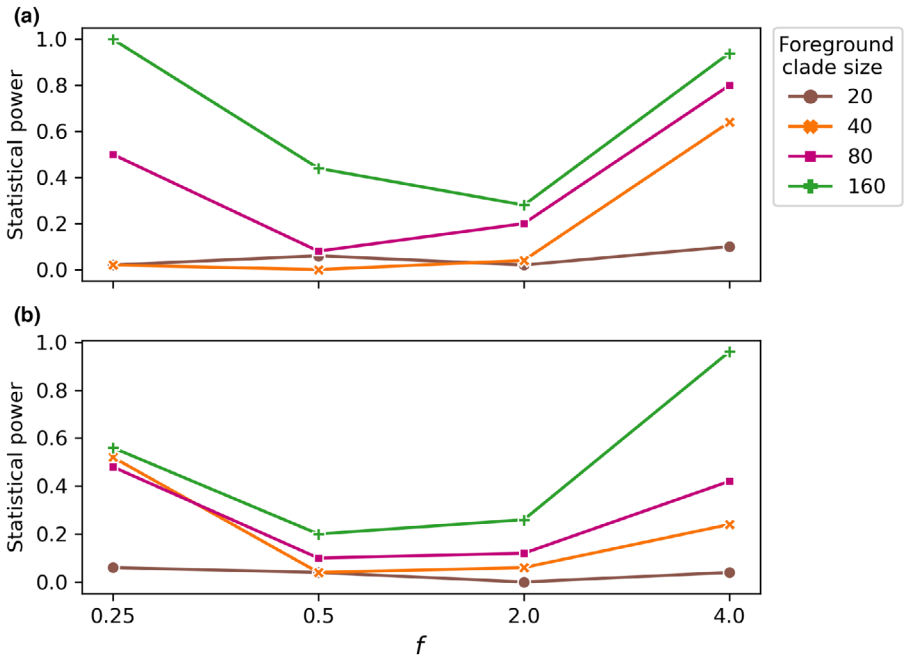

Fig. 1 Statistical power for the Cyperaceae and Solanaceae simulated scenarios. Statistical power (defined as the fraction of simulations in which a model with two rate regimes was chosen over the homogeneous model) as a function of different rate multiplications factors (*f*): 0.25, 0.5, 2.0, and 4.0, and using different sizes of the foreground clade: 20 (brown), 40 (orange), 80 (magenta), and 160 (green) taxa. The analyses are presented for simulated parameters inferred from (a) Cyperaceae and (b) Solanaceae datasets.

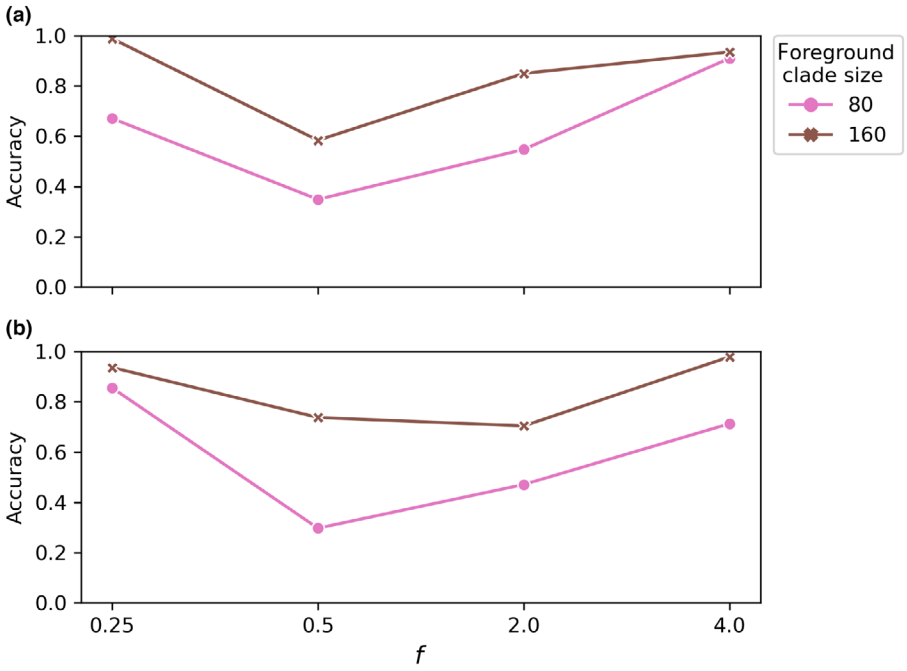

Fig. 2 Accuracy of the inferred clades for the Cyperaceae and Solanaceae simulated scenarios. Accuracy in detecting the foreground clade as a function of different rate multiplications factors (*f*): 0.25, 0.5, 2.0, and 4.0, and using different sizes of the foreground clade: 80 (pink), and 160 (brown) taxa. Two simulation parameters are presented, based on (a) Cyperaceae and (b) Solanaceae empirical datasets. Accuracy is defined as the overlap, in the number of tip taxa, between the true and inferred foreground clades. Note that the number of simulated datasets is not identical for all simulated scenarios since accuracy is computed only for those datasets for which rate heterogeneity was determined, but in all cases presented accuracy was computed for > 10 datasets (see the Description section).

Simulations based on the transition pattern of the Cyperaceae dataset indicated very high statistical power and high accuracy. Specifically, in 100% of the simulations the null model of rate homogeneity was rejected. This high sensitivity was obtained regardless of the size of the simulated foreground clade (Table 1). Still, the observed ΔAICc between the homogeneous and the heterogeneous models was larger in simulations in which the foreground-clade size were larger, thus providing a stronger support for the heterogeneous model. The accuracy to correctly infer the clade with an altered rate regime was very high in all cases (average accuracy above 0.9 for all sizes of the foreground clade).

For the Solanaceae dataset, the statistical power to detect a rate shift was generally lower (Table 1). Power was particularly low when the foreground clade was small: when the foreground clade contained 20 taxa, only 0.04 of the simulations correctly rejected the null model. However, statistical power steadily increased with the size of the foreground clade, reaching a rejection rate of 1.0 when the foreground clade contained 160 taxa. A similar pattern was observed for the accuracy measure, although here, accuracy increased rapidly from 0.06 in simulations with small foreground clades of size 20 to 0.763, 0.858, and 0.996 when the foreground-clade size was 40, 80, and 160, respectively.

Table 1 Statistical power and accuracy using simulated parameters as inferred for the Cyperaceae and Solanaceae empirical datasets.

Foreground clade size	Power	Accuracy	Median ΔAICc value
Cyperaceae			
20	1.0	0.965	39.7
40	1.0	0.912	109.9
80	1.0	0.996	268.0
160	1.0	0.998	349.3
Solanaceae			
20	0.04	0.062	−5.25
40	0.28	0.763	1.84
80	0.58	0.858	6.96
160	1.0	0.996	28.8

Power was defined as the fraction of simulations in which rate heterogeneity was detected and accuracy was defined as the overlap between the inferred and simulated foreground clade. The significance of the inferred shift in chromosome-number rates was determined according to the ΔAICc threshold value, determined using parametric bootstrapping, comparing the null model and a model with one shift in chromosome-number evolution (for details see 'Model comparison' in the Description section). The inferred ΔAICc threshold value for the Solanaceae dataset was 5.66. For the Cyperaceae dataset, the ΔAICc threshold value was negative, and we therefore decided to conservatively set it to 0.

Analysis of the Cyperaceae empirical dataset

We exemplify the use of the developed model on Cyperaceae, being a large plant family that had been previously explored in the context of inferring shifts in the pattern of chromosome-number evolution. We first provide some background on this clade and on the previous analysis by Márquez-Corro *et al.* (2019) and then present our analysis with the model developed here.

Cyperaceae is a highly diverse family, containing *c.* 5600 species that are distributed across the globe and grow in a wide range of habitats, with the greatest diversity in the humid and semihumid tropics (Goetghebeur, 1998). All species in the family are characterized by the presence of holocentric chromosomes, whose centromere is diffused and the kinetochore assemble along the entire chromosome. As a consequence of frequent chromosome fission and fusion events, dysploidy transitions are very common in Cyperaceae, and especially the genus *Carex* where fission and fusion rates are the highest (Escudero *et al.*, 2012; Márquez-Corro *et al.*, 2019). Polyploidy, on the contrary, is more frequent in the genera *Schoenus* and *Eleocharis* (Elliott *et al.*, 2022) and less common in *Carex* than in other Cyperaceae genera (Roalson, 2008). Polyploidy was hypothesized to be rare in *Carex* due to the disintegration of three nuclei after meiosis (pseudomonads instead of tetrads) making unreduced gametes much less likely to be formed (Heilborn, 1932), although it is unclear whether this cytogenetic peculiarity is unique to *Carex* or shared with other Cyperaceae species (Simpson *et al.*, 2003). The known differences in the mode of chromosome evolution in this family makes this lineage ideal to test heterogeneous patterns of chromosome evolution. Márquez-Corro *et al.* (2019) previously examined whether shifts in chromosome-number evolution coincide with alternations in diversification rates. To this end, they partitioned the phylogeny to four partially nested subtrees (Fig. 3) that were inferred as exhibiting acceleration in diversification rates (Escudero & Hipp, 2013; Spalink *et al.*, 2016): the SDC + FAEC clade ($n = 791$), which encompasses the SDC clade (abbreviated for the included tribes: Scirpeae, Dulichieae, and Cariceae) and the FAEC clade (abbreviated for the included tribes: Fuireneae, Abildgaardieae, Eleocharideae, and Cypereae); the FAEC clade ($n = 168$ taxa); the C_4 Cyperus clade (characterized by the C_4 photosynthetic pathway) that is nested within the FAEC clade ($n = 37$), and the species-rich non-Siderostictae Carex clade that is nested within SDC ($n = 583$) and includes the clades of all subgenera (*Carex*, *Vignea*, *Uncinia*, *Euthyceras*, and *Psyllophorae*) but subgenus *Siderostictae* (Villaverde *et al.*, 2020). The authors then independently applied the homogeneous ChromEvol model to each of these trees and compared whether the fit of each of these models provides a better fit compared with a homogeneous model whose parameters were inferred based on the entire phylogeny. Their analysis indicated a shift in the transition pattern for three of the four clades examined: in the FAEC clade, the C_4 Cyperus clade, and in the non-Siderostictae Carex clade, while the transition pattern inferred for the SDC + FAEC clade was nonsignificant compared with the pattern inferred based on the entire phylogeny.

Here, we applied the heterogeneous ChromEvol model to the same phylogeny and chromosome-number data assembled by Márquez-Corro *et al.* (2019). First, we examined whether shifts in chromosome-number dynamics occur at the four branches as specified in Márquez-Corro *et al.* (2019). Unlike that study, in our analysis the entire phylogeny was evaluated rather than partitioned to subtrees. To this end, we compared the fit of the null homogeneous model to a heterogeneous model with a single shift: at each of the four preassigned branches. Our analysis indicated that the AICc of the homogeneous model was lower (i.e. better) than the heterogeneous model with a shift at the branch leading to the SDC + FAEC clade (ΔAICc = −20.2). On the contrary, a better fit was obtained for a heterogeneous model with single shift placed at the branches leading to either the FAEC clade (ΔAICc = 36.13), the non-Siderosticta Carex clade (ΔAICc = 38.03), and to the C_4 Cyperus clade (ΔAICc = 65.45). Notably, a model that included all these three shifts was found inferior to a model that included only two of them (the non-Siderosticta Carex clade and the C_4 Cyperus clade). The AICc scores of all models examined are provided in Table S4.

The above analysis that identified shifts in the transition pattern could be driven by more prominent shifts occurring elsewhere in the phylogeny (i.e. nearer or further away from the root). We thus applied the automatic forward-phase algorithm (see the Description section) to identify the number and most probable shift locations. The AICc cutoff above which a shift in the transition regime was considered as significant was 14.31, as determined using parametric bootstrapping (see the Description section). The optimal model inferred in this analysis identified three shifts that partition the phylogeny to four distinct transition regimes (Fig. 3; Table 2): (1) The most significant shift included a small subtree of 29 species, spanning most of the species from *Carex* subgenus *Uncinia*, that was not explored in Márquez-Corro *et al.* (2019). This clade is characterized by a moderate rate

New Phytologist

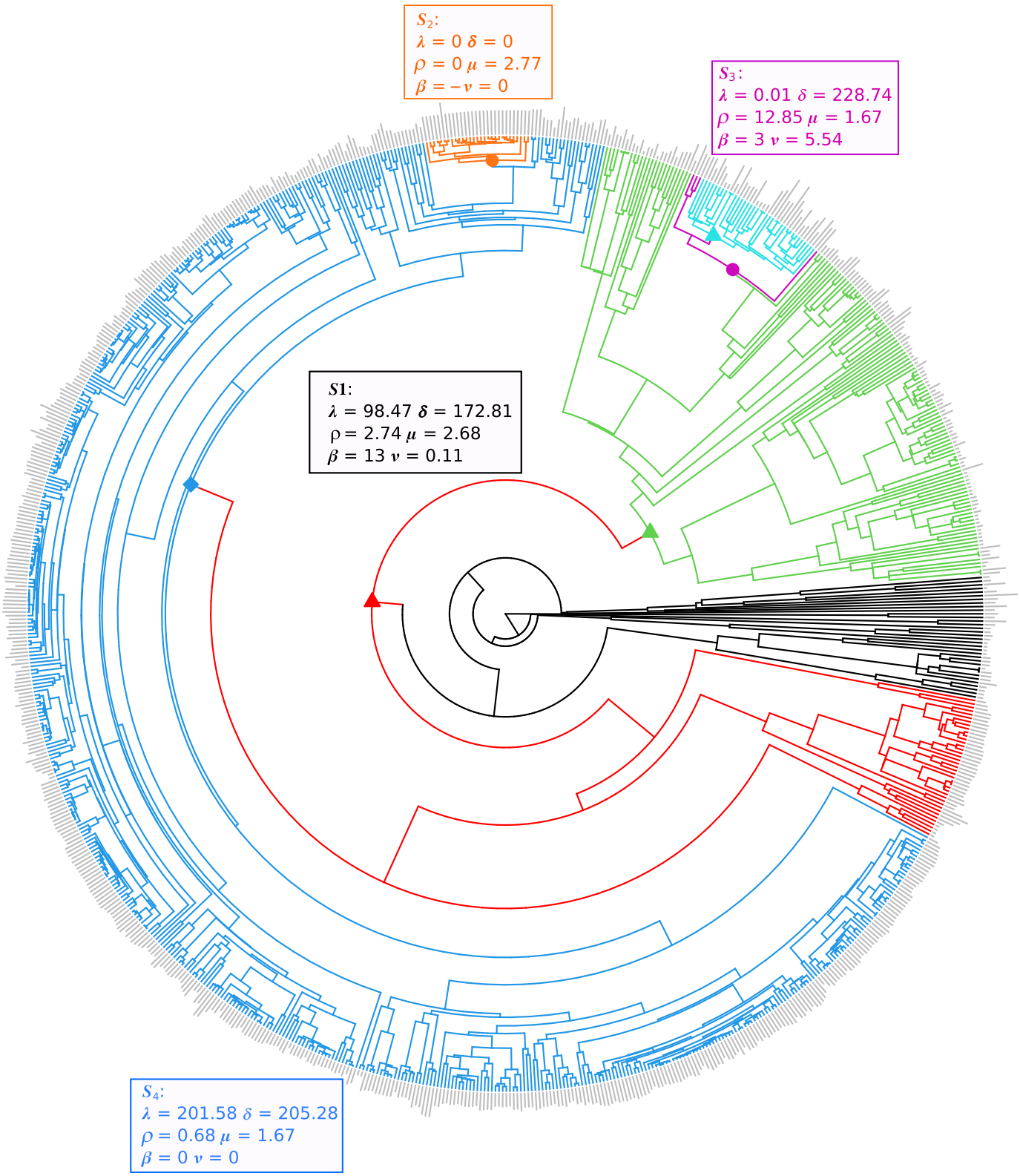

Fig. 3 Cyperaceae phylogeny and its partition to clades. The entire phylogeny of the Cyperaceae plant family and its partition to subclades: the red triangle (SDC + FAEC clade), the green triangle (FAEC clade), the cyan triangle (C_4 Cyperus clade), and the blue diamond (non-Siderostictae Carex clade) indicate the locations of four clades analyzed in Márquez-Corro *et al*. (2019). In that previous study, significant shifts were detected in three of these four clades (SDC + FAEC clade was the exception). The orange circle (S_2: a clade that includes most of the *Carex* subgenus *Uncinia*), the magenta circle (S_3: a Cyperus clade that includes the C_4 Cyperus clade), and the blue diamond (S_4: non-Siderostictae Carex clade) indicate the locations of significant shifts in rates of chromosome-number evolution inferred in this study using the automatic forward search approach; the inferred set of model parameters for each regime is indicated by the respective color; Branches belonging to the background rate regime (S_1) are colored in black. The grey bars extending from the tips are indicative of the respective chromosome numbers.

Table 2 Inferred transition regimes and their corresponding transition rates obtained for the heterogeneous model following the application of the forward search phase on the Cyperaceae dataset.

Regime	Clade size	λ	δ	ρ	μ	β	ν	ΔAICc
S_1	201	98.47	172.81	2.74	2.68	13	0.11	–
S_2	29	0	0	0	2.77	–	0	78.06
S_3	41	0.01	228.74	12.85	1.67	3	5.54	70.66
S_4	554	201.58	205.28	0.68	1.67	–	0	15.42

The transition regimes are presented in Fig. 3 and roughly correspond to: S_2, a subclade of the *Carex* subgenus *Uncinia*; S_3, a subclade of the genus *Cyperus*; S_4, the non-Sidorosticta Carex clade, S_1 is the background clade. Clade size refers to the number of terminal taxa in each regime. To allow comparison, the rates are computed as the expected number of events for each transition type under a given regime, divided by the sum of branch lengths belonging to the given regime. The exact inferred rates are given in Supporting Information Table S5. λ, δ, ρ, and μ represent the expected rates of ascending dysploidy, descending dysploidy, genome duplication, and demi-polyploidization, respectively; β and ν denote the base number and base number transition rate, respectively. ΔAICc score is shown for each model with respect to the previous one (e.g. the ΔAICc between a model with the two specified shifts and a model with one shift is 70.66).

of demi-polyploidy and no other types of polyploidy events. In stark contrast to the rest of the Cyperaceae phylogeny, no dysploidy transitions were inferred in this clade. (2) A clade containing 41 species, belonging to the genus *Cyperus*, which well overlaps (37/41 species) with the C_4 Cyperus clade tested by Márquez-Corro *et al.* (2019). This clade is characterized by extreme rates of descending dysploidy, no ascending dysploidy, and relatively high rates of polyploidy. (3) The large non-Sidorosticta Carex clade that was identical to the shift location tested by Márquez-Corro *et al.* (2019) (although here the regime did not include the 29 species from *Carex* subgenus *Uncinia* belonging to the first regime identified). This clade is characterized by extreme rates of dysploidy (both ascending and descending) and moderate rates of polyploidy. (4) All other branches in the phylogeny (with $n = 201$ species) represent the background regime with very high rates of dysploidy, although lower compared with that inferred for non-Sidorosticta Carex, and moderate rates of polyploidy. Given this model with four transition regimes, running the backward search phase did not identify any further convergence in the transition regimes.

Discussion

In this study, we presented the development of models of chromosome-number evolution that account for rate heterogeneity across a phylogeny. The use of this model should allow researchers to infer clades with distinct rates of evolution and to possibly link shifts in polyploidy and dysploidy rates with other evolutionary processes, such as diversification dynamics or the evolution of phenotypic traits. One possible use of the model is to manually specify several clades of interest and to assign a separate model for each. This is similar to the analysis conducted by Márquez-Corro *et al.* (2019), but the model developed here enables the analysis of the entire phylogeny, thereby incorporating the maximal amount of information without the need to partition the phylogeny to separate trees. As an alternative, the analysis may take an exploratory mode, letting the data inform the researcher on the most plausible shifting points in chromosome-number dynamics without a restriction to a certain taxonomic classification or to any *a priori* belief. To this end, we implemented a stepwise AICc approach based on the two-step algorithm. The forward phase divides the phylogeny into distinct regimes of chromosome-number evolution. The backward phase then identifies cases of convergent evolution by testing whether independent shifts converge into a similar transition pattern. It should be noted that the power of the automatic identification approach could be lower compared with that of the manual approach, due to the need to correct for multiple testing. On the contrary, and as was the case for the Cyperaceae analysis, the automatic forward phase has the potential to uncover additional rate-shifting clades, beyond those envisioned by the researcher.

Our results on simulated data indicated that, as might be expected, the accuracy and statistical power to correctly infer rate heterogeneity increases with the size of the foreground clade, as well as with the extent of dissimilarity in the transition pattern between the foreground and background clades (Figs 1, 2). For example, the statistical power was very high (≥ 0.94) when the foreground-clade size was large (160 taxa) and a fourfold difference between the foreground and background rates was simulated. However, the statistical power dropped to below 0.5 when the foreground-clade size was 80 and a twofold difference between the foreground and background rates was simulated (Fig. 1). Nevertheless, given that a shift was detected, our results indicated that the identified clade was generally overlapped closely with the true one (Fig. 2). We also expected that the statistical power would be generally lower in scenarios where there is deceleration in the dysploidy and polyploidy rates compared with rate acceleration of the same magnitude. In case of slowdown in the transition pattern, there may not be a sufficient number of transitions to allow for robust inferences. This pattern was apparent for some (e.g. Fig. 1a), but not all, simulated scenarios. Notably, in our simulations we rescaled the total tree length so that the number of simulated transitions remains similar across simulation scenarios. We conjecture that this procedure eroded the above-expected signal.

In the simulations mentioned above, we examined the scenario where all the transition rates change in the same direction and magnitude. Still, we foresee that a more typical application of the model would result in shifts in the transition pattern where the change in one transition type is not necessarily consistent with those inferred for other transition types (e.g. the ascending dysploidy would increase by fourfold while the polyploidy rate would decrease by tenfold). We thus conducted another set of simulations based on model parameters inferred from empirical data using a heterogeneous model with two rate regimes. The results differ considerably between the two sets of empirical data simulated. When the simulated parameters were derived from the Cyperaceae dataset, the statistical power and accuracy were exceptionally high (statistical power of 1.0, and accuracy above 0.9; Table 1), but were much lower using model parameters derived from the Solanaceae dataset.

These results were in accordance with the extent of rate heterogeneity simulated. The Cyperaceae dataset exhibited a clear pattern of shift in the transition regime from a background regime with very high dysploidy rates and low rates of polyploidization to a foreground regime with dysploidy rates approaching zero and moderate polyploidization rates. The shift in the Solanaceae dataset was subtler with moderate differences in the dysploidy rates between the foreground and background clades and no difference in the rate of polyploidization.

Notably, the most time-consuming step of our implementation is to determine the ΔAICc threshold, above which a detected shift can be considered statistically significant. Naively, one could choose a predefined ΔAICc threshold, under the assumption that models that obtain lower AICc values better fit the data at hand. Unfortunately, differences in AICc are not directly translated to statistical significance. In addition, for the hypothesis test considered here, multiple tests are performed in every iteration of the search algorithm (e.g. in the forward phase, one for every subtree that contains more than N_{min} taxa) and these tests are not independent of each other, thus complicating the decision of the exact ΔAICc* to be used. To this end, we followed a parametric bootstrap approach and created a null distribution of the ΔAICc values by simulating a large number of datasets using the maximum likelihood estimates of the null homogeneous model. This procedure requires to optimize two models for each simulated dataset: a heterogeneous model with a single shift and the null homogeneous model. Thus, assuming that at least 100 simulated datasets are needed to obtain a reasonable null distribution of ΔAICc values, the running time of generating the null distribution is several folds longer than the optimization of the models to the empirical data. This could be highly time-consuming, particularly if the range of chromosome counts and the number of taxa in the phylogeny are large.

One should bear in mind that the parametric bootstrapping approach is constructed based on the comparison between the homogeneous null model and a heterogeneous model with a single shift. Theoretically, a separate distribution should be constructed to determine the ΔAICc for the comparison between each two consecutives heterogeneous models (e.g. a model with a single shift vs a model with two shifts, and so on). The creation of multiple null distributions created this way would entail exceedingly long running time and was not attempted here. We note that the number of multiple tests decreases as models with higher number of shifts are examined, and thus, it is expected that the use of the null distribution constructed for testing the presence of the first shift is conservative when used to test for subsequent shifts. Thus, in case one of the higher-order shifts detected had an ΔAICc value close to the critical threshold, it might be beneficial to simulate a null distribution of ΔAICc values specific to the number and locations of the shifts inferred so far.

A possible direction to markedly reduce the time and computational resources needed to generate a null distribution could be to predict a suitable ΔAICc threshold per dataset using a machine learning approach. Accordingly, the ΔAICc threshold values are learned based on a large training data consisting of datasets of different phylogeny sizes, and other features, such as the number of transitions of each transition type and the variation of chromosome numbers at the tips. Given a trained predictive model, the determination of the ΔAICc threshold value that fits an examined dataset would only require the computation of the set of informative features. On the contrary, using such an approach requires substantial computational resources to construct an informative learning dataset, consisting of many independent clades and their respective ChromEvol inferences that include the inferred parametric bootstrap distribution. While such an endeavor is challenging, we believe it would be possible with the accumulation of more empirical datasets analyzed by the community using the developed heterogeneous model.

Acknowledgements

We thank Laurent Guéguen for his help in implementing ChromEvol within Bio++. We also thank the associate editor, Angelino Carta, and the two other anonymous reviewers for providing insightful comments. This study was supported by PhD fellowships from the Edmond J. Safra Center for Bioinformatics at Tel Aviv University (to AS and KH), by a TAD fellowship from the Data Science & AI Center at TAU (KH), by the Israel Science Foundation (grant nos. 961/17 and 1843/21 to IM) and by MICINN-FEDER, Project DiversiChrom (PID2021-122715 NB-I00 to ME).

Competing interests

None declared.

Author contributions

IM and AS conceived the study. AS developed the new version of ChromEvol and conducted the simulations and data analyses. KH participated in the software development and assisted in the data analyses. ME carried out the empirical data analysis on the Cyperaceae. All the authors wrote the manuscript. IM supervised the study.

ORCID

Marcial Escudero https://orcid.org/0000-0002-2541-5427
Keren Halabi https://orcid.org/0000-0001-6009-1598
Itay Mayrose https://orcid.org/0000-0002-8460-1502
Anat Shafir https://orcid.org/0000-0001-8059-3544

Data availability

The method developed here is freely available as an open-source program at GitHub: https://github.com/anatshafir1/chromevol.

References

Alfaro ME, Santini F, Brock C, Alamillo H, Dornburg A, Rabosky DL, Carnevale G, Harmon LJ. 2009. Nine exceptional radiations plus high

turnover explain species diversity in jawed vertebrates. *Proceedings of the National Academy of Sciences, USA* **106**: 13410–13414.

Anisimova M, Yang Z. 2007. Multiple hypothesis testing to detect lineages under positive selection that affects only a few sites. *Molecular Biology and Evolution* **24**: 1219–1228.

Baniaga A, Marx H, Arrigo N, Barker M. 2020. Polyploid plants have faster rates of multivariate niche differentiation than their diploid relatives. *Ecology Letters* **23**: 68–78.

Blackmon H, Justison J, Mayrose I, Goldberg EE. 2019. Meiotic drive shapes rates of karyotype evolution in mammals. *Evolution* **73**: 511–523.

Carta A, Bedini G, Peruzzi L. 2020. A deep dive into the ancestral chromosome number and genome size of flowering plants. *New Phytologist* **228**: 1097–1106.

Chen ZJ, Ni Z. 2006. Mechanisms of genomic rearrangements and gene expression changes in plant polyploids. *BioEssays* **28**: 240–252.

Cusimano N, Sousa A, Renner SS. 2012. Maximum likelihood inference implies a high, not a low, ancestral haploid chromosome number in Araceae, with a critique of the bias introduced by '*x*'. *Annals of Botany* **109**: 681–692.

Dodsworth S, Chase MW, Leitch AR. 2016. Is post-polyploidization diploidization the key to the evolutionary success of angiosperms? *Botanical Journal of the Linnean Society* **180**: 1–5.

Elliott T, Zedek F, Bruhl J, Escudero M, Hroudová Z, Joly S, Larridon I, Luceño M, Márquez-Corro JI, Martin-Bravo S *et al.* 2022. Chromosome size matters: genome evolution in the cyperid clade. *Annals of Botany* **130**: 999–1014.

Eric Schranz M, Mohammadin S, Edger PP. 2012. Ancient whole genome duplications, novelty and diversification: the WGD radiation lag-time model. *Current Opinion in Plant Biology* **15**: 147–153.

Escudero M, Hipp A. 2013. Shifts in diversification rates and clade ages explain species richness in higher-level sedge taxa (Cyperaceae). *American Journal of Botany* **100**: 2403–2411.

Escudero M, Hipp AL, Waterway MJ, Valente LM. 2012. Diversification rates and chromosome evolution in the most diverse angiosperm genus of the temperate zone (*Carex*, Cyperaceae). *Molecular Phylogenetics and Evolution* **63**: 650–655.

Escudero M, Martín-Bravo S, Mayrose I, Fernández-Mazuecos M, Fiz-Palacios O, Hipp AL, Pimentel M, Jiménez-Mejías P, Valcárcel V, Vargas P *et al.* 2014. Karyotypic changes through dysploidy persist longer over evolutionary time than polyploid changes. *PLoS ONE* **9**: e85266.

Freyman WA, Höhna S. 2018. Cladogenetic and anagenetic models of chromosome number evolution: a Bayesian model averaging approach. *Systematic Biology* **67**: 195–215.

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

Goetghebeur P. 1998. Cyperaceae. In: Kubitzki K, ed. *Flowering plants. Monocotyledons, vol. 4*. The families and genera of vascular plants. Berlin, Germany: Springer, 141–190.

Guéguen L, Gaillard S, Boussau B, Gouy M, Groussin M, Rochette NC, Bigot T, Fournier D, Pouyet F, Cahais V *et al.* 2013. Bio++: efficient extensible libraries and tools for computational molecular evolution. *Molecular Biology and Evolution* **30**: 1745–1750.

Guerra M. 2008. Chromosome numbers in plant cytotaxonomy: concepts and implications. *Cytogenetic and Genome Research* **120**: 339–350.

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

Heilborn O. 1932. Lethal gene-combinations and pollen sterility in diploid apple varieties: a critique and a theory. *Hereditas* **16**: 1–18.

Ingram T, Mahler DL. 2013. SURFACE: detecting convergent evolution from comparative data by fitting Ornstein-Uhlenbeck models with stepwise Akaike information criterion. *Methods in Ecology and Evolution* **4**: 416–425.

Levin DA. 2019. Why polyploid exceptionalism is not accompanied by reduced extinction rates. *Plant Systematics and Evolution* **305**: 1–11.

Mandáková T, Lysak MA. 2018. Post-polyploid diploidization and diversification through dysploid changes. *Current Opinion in Plant Biology* **42**: 55–65.

Márquez-Corro JI, Martín-Bravo S, Luceño M, Spalink D, Luceño M, Escudero M. 2019. Inferring hypothesis-based transitions in clade-specific models of chromosome number evolution in sedges (Cyperaceae). *Molecular Phylogenetics and Evolution* **135**: 203–209.

Martin SL, Husband BC. 2012. Whole genome duplication affects evolvability of flowering time in an autotetraploid plant. *PLoS ONE* **7**: e44784.

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

Mayrose I, Lysak MA. 2020. The evolution of chromosome numbers: mechanistic models and experimental approaches. *Genome Biology and Evolution* **13**: 1–15.

Mayrose I, Zhan SH, Rothfels CJ, Magnuson-Ford K, Barker MS, Rieseberg LH, Otto SP. 2011. Recently formed polyploid plants diversify at lower rates. *Science* **333**: 1257.

McCann J, Schneeweiss GM, Stuessy TF, Villaseñor JL, Weiss-Schneeweiss H. 2016. The impact of reconstruction methods, phylogenetic uncertainty and branch lengths on inference of chromosome number evolution in American Daisies (Melampodium, Asteraceae). *PLoS ONE* **11**: e0162299.

Moura RF, Queiroga D, Vilela E, Moraes AP. 2021. Polyploidy and high environmental tolerance increase the invasive success of plants. *Journal of Plant Research* **134**: 105–114.

Parisod C, Holderegger R, Brochmann C. 2010. Evolutionary consequences of autopolyploidy. *New Phytologist* **186**: 5–17.

van de Peer Y, Ashman TL, Soltis PS, Soltis DE. 2021. Polyploidy: an evolutionary and ecological force in stressful times. *Plant Cell* **33**: 11–26.

Pellicer J, Kelly LJ, Magdalena C, Leitch IJ, Bainard J. 2013. Insights into the dynamics of genome size and chromosome evolution in the early diverging angiosperm lineage Nymphaeales (water lilies). *Genome* **56**: 437–449.

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

Rice A, Mayrose I. 2021. Model adequacy tests for probabilistic models of chromosome-number evolution. *New Phytologist* **229**: 3602–3613.

Roalson EH. 2008. A synopsis of chromosome number variation in the cyperaceae. *Botanical Review* **74**: 209–393.

Ruckman SN, Jonika MM, Casola C, Blackmon H. 2020. Chromosome number evolves at equal rates in holocentric and monocentric clades. *PLoS Genetics* **16**: 1–13.

Segraves KA. 2017. The effects of genome duplications in a community context. *New Phytologist* **215**: 57–69.

Simpson DA, Furness CA, Hodkinson TR, Muasya AM, Chase MW. 2003. Phylogenetic relationships in Cyperaceae subfamily Mapanioideae inferred from pollen and plastid DNA sequence data. *American Journal of Botany* **90**: 1071–1086.

Smith SA, Brown JW. 2018. Constructing a broadly inclusive seed plant phylogeny. *American Journal of Botany* **105**: 302–314.

Smith SA, O'Meara BC. 2012. TREEPL: divergence time estimation using penalized likelihood for large phylogenies. *Bioinformatics* **28**: 2689–2690.

Spalink D, Drew BT, Pace MC, Zaborsky JG, Starr JR, Cameron KM, Givnish TJ, Sytsma KJ. 2016. Biogeography of the cosmopolitan sedges (Cyperaceae) and the area-richness correlation in plants. *Journal of Biogeography* **43**: 1893–1904.

Stamatakis A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22**: 2688–2690.

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

Sylvester T, Hjelmen CE, Hanrahan SJ, Lenhart PA, Johnston JS, Blackmon H. 2020. Lineage-specific patterns of chromosome evolution are the rule not the exception in Polyneoptera insects: patterns of chromosome evolution. *Proceedings of the Royal Society B: Biological Sciences* **287**: 20201388.

Tank DC, Eastman JM, Pennell MW, Soltis PS, Soltis DE, Hinchliff CE, Brown JW, Sessa EB, Harmon LJ. 2015. Nested radiations and the pulse of angiosperm diversification: increased diversification rates often follow whole genome duplications. *New Phytologist* **207**: 454–467.

Tayalé A, Parisod C. 2013. Natural pathways to polyploidy in plants and consequences for genome reorganization. *Cytogenetic and Genome Research* **140**: 79–96.

Tsuda H, Kunitake H, Yamasaki M, Komatsu H, Yoshioka K. 2013. Production of intersectional hybrids between colchicine-induced tetraploid

shashanbo (*Vaccinium bracteatum*) and highbush blueberry 'Spartan'. *Journal of the American Society for Horticultural Science* **138**: 317–324.

Villaverde T, Jiménez-Mejías P, Luceño M, Waterway MJ, Kim S, Lee B, Rincón-Barrado M, Hahn M, Maguilla E, Roalson EH *et al.* 2020. A new classification of *Carex* (Cyperaceae) subgenera supported by a HybSeq backbone phylogenetic tree. *Botanical Journal of the Linnean Society* **194**: 141–163.

Weiss-Schneeweiss H, Schneeweiss GM. 2013. Karyotype diversity and evolutionary trends in angiosperms. In: Greilhuber J, Dolezel J, Wendel J, eds. *Plant genome diversity, vol. 2*. Vienna, Austria: Springer, 209–230.

Wood TE, Takebayashi N, Barker MS, Mayrose I, Greenspoon PB, Rieseberg LH. 2009. The frequency of polyploid speciation in vascular plants. *Proceedings of the National Academy of Sciences, USA* **106**: 13875–13879.

Wu F, Tanksley SD. 2010. Chromosomal evolution in the plant family Solanaceae. *BMC Genomics* **11**: 182.

Zenil-Ferguson R, Burleigh JG, Ponciano JM. 2018. CHROMPLOID: an R package for chromosome number evolution across the plant tree of life. *Applications in Plant Sciences* **6**: e1037.

Zenil-Ferguson R, Ponciano JM, Burleigh JG. 2017. Testing the association of phenotypes with polyploidy: an example using herbaceous and woody eudicots. *Evolution* **71**: 1138–1148.

Supporting Information

Additional Supporting Information may be found online in the Supporting Information section at the end of the article.

Fig. S1 Representation of the transition rates as a function of the current number of chromosomes in the optimized homogeneous model in the Cyperaceae and Solanaceae datasets.

Fig. S2 Representation of the transition rates as a function of the current number of chromosomes in the optimized one-shift heterogeneous model in the Cyperaceae and Solanaceae datasets.

Notes S1 Model adequacy tests.

Notes S2 Automatic identification of regime shifts.

Table S1 Maximum likelihood estimates of the model parameters as inferred using the homogeneous M_e model for Cyperaceae and Solanaceae families.

Table S2 Maximum likelihood estimates of the model parameters inferred for the heterogeneous model with a single shift (termed M_2) for the Cyperaceae and Solanaceae families.

Table S3 Mean accuracy of correctly inferring the foreground clades for all examined simulated foreground-clade sizes (20, 40, 80, and 160), and f = 0.25, 0.5, 2.0, and 4.0 for the Cyperaceae and Solanaceae datasets.

Table S4 AICc and log-likelihood values of the homogeneous model and the heterogeneous models applied on the Cyperaceae dataset with specified number of shifts for the clade partitions as specified in Márquez-Corro *et al.* (2019).

Table S5 Rates inferred from the forward-phase analysis of the Cyperaceae empirical dataset.

Please note: Wiley is not responsible for the content or functionality of any Supporting Information supplied by the authors. Any queries (other than missing material) should be directed to the *New Phytologist* Central Office.