Methods

Model adequacy tests for probabilistic models of chromosome-number evolution

Anna Rice and Itay Mayrose

[1]School of Plant Sciences and Food Security, George S. Wise Faculty of Life Sciences, Tel Aviv University, Tel Aviv 69978, Israel

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

Received: *2 August 2020*
Accepted: *18 November 2020*

New Phytologist (2021) **229:** 3602–3613
doi: 10.1111/nph.17106

Key words: ChromEvol, chromosome number, dysploidy, model adequacy, model selection, model testing, phylogenetics, polyploidy.

Summary

- Chromosome number is a central feature of eukaryote genomes. Deciphering patterns of chromosome-number change along a phylogeny is central to the inference of whole genome duplications and ancestral chromosome numbers. ChromEvol is a probabilistic inference tool that allows the evaluation of several models of chromosome-number evolution and their fit to the data. However, fitting a model does not necessarily mean that the model describes the empirical data adequately. This vulnerability may lead to incorrect conclusions when model assumptions are not met by real data.
- Here, we present a model adequacy test for likelihood models of chromosome-number evolution. The procedure allows us to determine whether the model can generate data with similar characteristics as those found in the observed ones.
- We demonstrate that using inadequate models can lead to inflated errors in several inference tasks. Applying the developed method to 200 angiosperm genera, we find that in many of these, the best-fitting model provides poor fit to the data. The inadequacy rate increases in large clades or in those in which hybridizations are present.
- The developed model adequacy test can help researchers to identify phylogenies whose underlying evolutionary patterns deviate substantially from current modelling assumptions and should guide future methods development.

Introduction

Chromosome number is widely recognized as a key feature of eukaryote genomes. Its popularity in cyto-taxonomical and evolutionary studies has been attributed to its ability to provide a concise description of the karyotype, the ease by which it can be recorded, and its stable phenotype across repeated measurements. Processes that lead to changes in chromosome numbers have direct consequences on central evolutionary developments related to reproductive isolation and speciation, thus providing important information for species determination and phylogenetic relationships (Guerra, 2008; Weiss-Schneeweiss & Schneeweiss, 2013). Although chromosome numbers generally exhibit strong phylogenetic signal (e.g. Vershinina & Lukhtanov, 2017; Carta *et al.*, 2018), they also are highly dynamic. This variability has been particularly well acknowledged in plants, with counts ranging from $n = 2$ to $n = 720$ (Khandelwal, 1990; Ruffini Castiglione & Cremonini, 2012), and records showing intraspecific variation in 23% of angiosperm species (Rice *et al.*, 2015). Understanding the underlying processes that gave rise to these changes allows inference of major genomic events that have occurred in the history of a clade of interest and the processes that have shaped its diversification.

Of the various mechanisms underlying chromosome-number change, polyploidy, or whole genome duplication (WGD) has received significant attention because of the profound impacts such an event has on the organism. Polyploids often differ markedly from their progenitors in morphological, physiological or life-history characteristics, which may contribute to their establishment in novel ecological settings (Stebbins, 1971; Levin, 1983; Ramsey & Schemske, 2002; Soltis *et al.*, 2007; Leitch & Leitch, 2008; Ramsey & Ramsey, 2014; Spoelhof *et al.*, 2017; Rice *et al.*, 2019). Polyploidy is thus recognized as one of the major processes that has driven and shaped the evolution of higher organisms. A more subtle change in chromosome number is dysploidy, leading to step-wise changes in the number of chromosomes, but typically this does not immediately alter the genomic content. Dysploidy occurs via several types of genome rearrangements, leading to ascending or descending dysploidy through chromosome fission or fusion (Weiss-Schneeweiss & Schneeweiss, 2013). Deciphering the pattern of chromosome-number change within a clade allows inference of the number and type of transitions that have occurred along branches of a phylogeny, estimation of ancestral chromosome numbers and categorization of extant species as diploids or polyploids.

In the last decade, several tools that infer changes in chromosome numbers along a phylogeny were developed (Mayrose *et al.*, 2010; Hallinan & Lindberg, 2011; Glick & Mayrose, 2014; Freyman & Höhna, 2017; Zenil-Ferguson *et al.*, 2017, 2018; Blackmon *et al.*, 2019). Among these, the ChromEvol probabilistic framework (Mayrose *et al.*, 2010) was the first to incorporate a continuous time Markov process that describes the instantaneous rate of change from a genome with i haploid chromosomes to a genome with j haploid chromosomes via specific types of dysploidy and polyploidy transitions. Further development of this framework allowed for more intricate types of chromosome-number transitions (Glick & Mayrose, 2014), to differentiate between transitions that coincide with speciation events and those that occur continuously in time along branches of the phylogeny (Freyman & Höhna, 2017), and to associate patterns of chromosome-number change with the evolution of a discrete character trait (Zenil-Ferguson *et al.*, 2017; Blackmon *et al.*, 2019).

In the ChromEvol model, each type of transition is represented by a parameter describing its rate of change. The inclusion (or exclusion) of different parameters entails different hypotheses regarding the pathways by which the evolution of chromosome number proceeded in the clade under study. In a regular application of the ChromEvol framework, a model selection procedure is first employed in which different models are fitted to the data and the best one is chosen by comparing the relative fit of each model to the data at hand. This can be done using a number of alternative model selection criteria, such as the likelihood ratio test or Akaike Information Criterion (AIC; Akaike, 1974). In reality, however, no empirical dataset will meet all ofthe assumptions of any model and thus relying on the best model (or set of models) may be vulnerable to incorrect conclusions in datasets whose underlying evolutionary process deviate substantially from current modelling assumptions (Brown & Thomson, 2018). Such instances could be identified using a model adequacy analysis that simulates datasets under a specified model and tests whether the generated data are similar to the data at hand. Although the general model adequacy procedure is well-developed (reviewed in Brown & Thomson, 2018), the comparison between the empirical and simulated data is not trivial and should reflect the specificities of the type of data analyzed. Here, we develop a model adequacy test that allows us to determine whether a given model of chromosome-number evolution provides a realistic description of the evolutionary process for reliable inferences.

Several assumptions made by existing models of chromosome-number evolution may be violated when empirical data are analyzed. For example, all models rely on a memory-less Markovian process, in which the transition rates are dictated only by the current number of chromosomes of the lineage. Thus, for example, the transition rate from $n = 10$ to $n = 9$ is not affected by the duration of time for which the lineage possessed 10 chromosomes, nor by the sequence of events that had led to it. However, because rates of descending dysploidy may increase following WGD (Wood *et al.*, 2009; Wendel, 2015; Soltis *et al.*, 2016), the transition from $n = 10$ to $n = 9$ is more probable if $n = 5$ was the ancestral state compared to $n = 11$. Additionally, most models assume that the transition rates are similar across the phylogeny, although in practice the transition patterns may be rather different in some subclades compared to others, as has been demonstrated, for example, in Cyperaceae (Márquez-Corro *et al.*, 2019). Finally, all current models are based on a phylogenetic structure and thus ignore the possibility of hybridizations. Notably, allopolyploidy, one of the main types of polyploidy, is defined by such reticulate evolutionary events and the biases caused by their presence is rather unexplored.

One aspect of understanding the reliability of a model and interpreting its results is to quantify its adequacy for the data and the question at hand. The aim of model adequacy tests is to determine the absolute fit of a model to the data, rather than to compare its relative fit among a set of models. With some variations, the general procedure of such tests is composed of several steps: first, given an empirical dataset, obtain the best-fitting model and its parameter values. Next, use that model to generate multiple simulated datasets. Then, compute several test statistics that describe various characteristics of the data on each simulated dataset and on the empirical dataset. If the empirical values of the test statistics fall outside the range of variation encompassed by the simulated data, then it may be concluded that the model cannot provide an adequate description of the data at hand. To date, model adequacy approaches are established for several types of data and inference tasks, including those related to sequence evolution (Bollback, 2002; Brown, 2014; Duchêne *et al.*, 2015; Chen *et al.*, 2019), and for continuous and discrete organismal traits (Slater & Pennell, 2013; Beaulieu *et al.*, 2013; Blackmon & Demuth, 2014; Pennell *et al.*, 2015). However, these are inappropriate for data and analyses concerning the evolution of chromosome numbers as the former rely on statistics derived from many sites, whereas the latter rely on Brownian motion statistics.

In the following, we first provide the details of the developed model adequacy framework for likelihood models of chromosome-number evolution. We then use simulations to assess the type I error rate and to explore the consequences of using inadequate models in several common inference tasks, such as ancestral reconstructions of chromosome numbers and ploidy-level inference. Finally, we apply the developed procedure to a large cohort of angiosperm genera, as well as to clades that are expected to violate model assumptions.

Description

Model adequacy framework for chromosome-number evolution

Given chromosome count data (denoted as D) and a compatible phylogeny, ChromEvol can be used to assess the fit of various models (M_1, M_2,..., M_N; N denotes the number of models) to D. Each model differs with respect to the included rate parameters or the constraints placed on them [$\theta(M_1)$, $\theta(M_2)$, ..., $\theta(M_N)$]. The most general model considered here includes six free parameters (Glick & Mayrose, 2014) and assumes that five types of events are possible: a single chromosome-number increase (ascending dysploidy with rate λ) or decrease (descending dysploidy with rate δ), whole genome duplication (WGD)

(i.e. exact duplication of the number of chromosomes with rate ρ), demi-polyploidy (multiplications of the number of chromosomes by 1.5 with rate μ) and base-number transitions (the addition to the genome by any multiplication of an inferred base number, where β is the inferred base number and ν is its respective transition rate). A combination of these parameters allows a range of models to be evaluated (Table 1 shows the various models considered here). We note that the ChromEvol software also allows the ascending and descending dysploidy rates to depend on the current number of chromosomes, but this option was not evaluated here.

In a common application of ChromEvol, several models are fitted to D, the optimal model is selected based on its relative fit using established model selection criteria (e.g. Aikake Information Criterion, AIC), and subsequent inference tasks are performed based on this model. The model adequacy test can be carried out on any model of interest, whether or not it is the most fitted one. The general aim of this test is to examine whether a specified model, M_x, is able to generate data that are similar to D. Our model adequacy procedure is based on parametric bootstrapping (Goldman, 1993; Efron & Tibshirani, 1994), where the observed data are compared to a background distribution generated from simulations. These simulations are generated under the specified model, whose parameters, $\hat{\theta}(M_x)$, were optimized with respect to D and the respective probabilities of chromosome numbers inferred at the root of the phylogeny (exact details of the simulation procedure are given in the Supporting Information Methods S1). Comparing true and simulated data is performed using a set of test statistics, which reflects various characteristics of the data. First, the test statistics (T_1, T_2,..., T_m; m denotes for the number of statistics) are computed for the true data D. Second, multiple datasets are simulated under the specified model and its inferred parameters. For each simulated dataset, the same set of test statistics is computed, resulting in a distribution for each test statistic (T_{s1}, T_{s2},..., T_{sm}). We then calculate the midpoint two-tailed P-value of each statistic as described in Höhna *et al.* (2018). If $P > 0.05$ the model is considered capable of generating data similar to the original ones and is thus inferred as adequate; otherwise it is inferred as inadequate. A schematic illustration of the developed model adequacy test is presented in Fig. 1.

Table 1 The set of ChromEvol models examined in this study, together with their rate parameters.

Model	Model parameters[1]	Description	Nested in[2]
D_{ys}	λ, δ	Dysploidy (descending or ascending)	$D_{ys}D_{up}$, $D_{ys}D_{up}D_{em}$*, $D_{ys}D_{up}D_{em}$, $D_{ys}B_{num}$, $D_{ys}D_{up}B_{num}$
$D_{ys}D_{up}$	λ, δ, ρ	Dysploidy and duplication	$D_{ys}D_{up}D_{em}$*, $D_{ys}D_{up}D_{em}$, $D_{ys}D_{up}B_{num}$
$D_{ys}D_{up}D_{em}$*	λ, δ, $\rho = \mu$	Dysploidy, constraining equal rates of duplication and demi-polyploidy	$D_{ys}D_{up}D_{em}$
$D_{ys}D_{up}D_{em}$	λ, δ, ρ, μ	Dysploidy, duplication, and demi-polyploidy	
$D_{ys}B_{num}$	λ, δ, β, ν	Dysploidy and base number transition	$D_{ys}D_{up}B_{num}$
$D_{ys}D_{up}B_{num}$	λ, δ, ρ, β, ν	Dysploidy, base number transition, and duplication	

[1]The model parameters are the base number (β), and rates of ascending dysploidy (λ), descending dysploidy (δ), duplication (ρ), demi-duplication (μ) and base number transition (ν).
[2]In case all parameters of the model are a subset of other models, the more complex models are indicated.

In our implementation, four test statistics were calculated given the chromosome-number data of extant taxa and the corresponding phylogeny.

(**1**) Variance ($\sum(x_i - \bar{x})^2/n$): higher values in the simulated data relative to the observed ones may point to some constraints that were not accounted for by the model (e.g. hard bounds on the number of chromosomes in the genome), or to errors in the parameter estimation process.

(**2**) Shannon's entropy (Shannon, 1948): lower entropy of the observed data than predicted by the model is indicative of higher-than-expected concentration of genomes with certain haploid numbers. This could be due to selective constraints, or to a very low variability exhibited in certain subclades of the phylogeny, such that specific states are clumped into large blocks of the tree more than expected.

(**3**) Parsimony score: the most parsimonious number of character transitions across the phylogeny is calculated based on Fitch (1971). If the parsimony scores of the observed data are lower than expected it means that the model assumes more transitions than actually occurred. This could occur due to rate heterogeneity across the tree; for example, if chromosome-number transitions occur more frequently in one subclade relative to the rest of the phylogeny, this could be accommodated by inferring higher values of the transition rates.

(**4**) Parsimony vs time ($Pars^{Time}$): the parsimonious number of transitions is computed per branch using the accelerated transformation criterion (Acctran; Farris, 1970). The regression line between the divergence times (computed from the root to the end of the branch) and their parsimony scores is calculated, and the slope of this line is taken as the test statistic. This statistic is similar in spirit to that employed by Pennell *et al.* (2015) for testing the adequacy of models for continuous trait evolution. Under a time-homogenous model, as implemented in ChromEvol, we expect no relationship between the divergence times and the number of transitions. Violations of this assumption suggest that transitions are either concentrated around the root or occur more frequently towards the tips. We note that aside from these four statistics, two additional ones were computed (the range and the number of unique counts). These two statistics were found to be highly correlated with the other test statistics ($r^2 = 0.74$ between range and variance and $r^2 = 0.66$ between unique counts and entropy, when computed over the 200 empirical datasets; detailed below), and thus we chose to discard them from further analyses. The coefficient of determinations between all pairs of

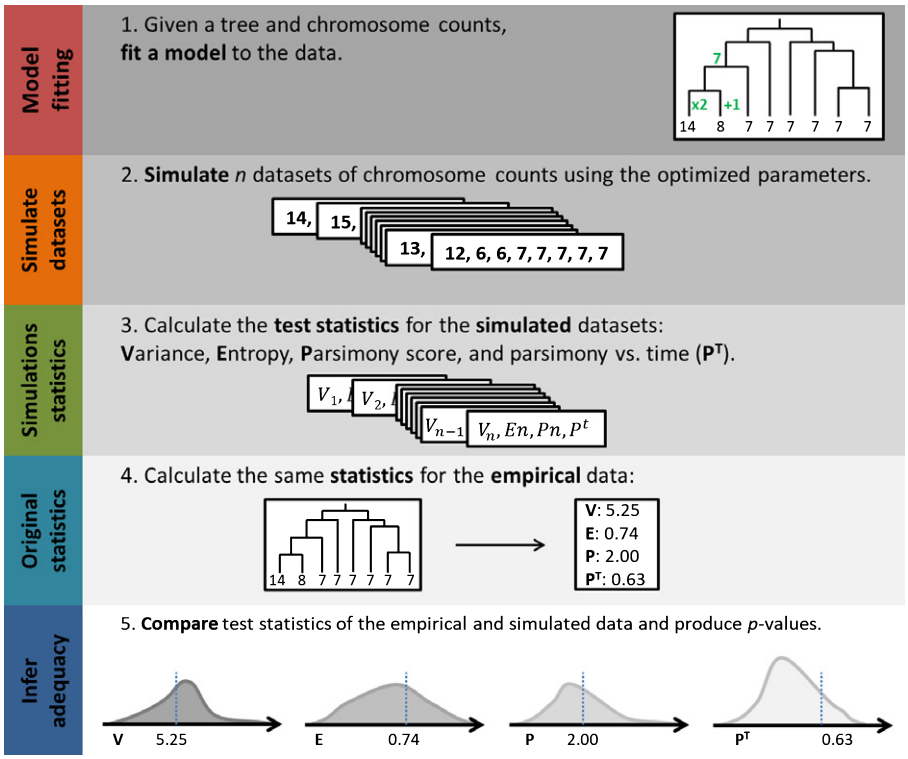

Fig. 1 A schematic illustration of the model adequacy framework for likelihood models of chromosome-number evolution. In the case illustrated here, the model is adequate because all *P*-values are greater than 0.05.

the four remaining test statistics was < 0.36 (Supporting Information Table S1). Because the four test statistics are not independent and researchers might be interested in revealing the specific aspects of the data that differ from expectations, we followed Pennell *et al.* (2015) and did not apply a multiple testing correction. Thus, in all analyses presented here a model is considered as adequate only if all four statistics have $P > 0.05$ for the simulated distribution.

Performance assessment using simulations

Simulations were conducted to examine the performance of the model adequacy procedure. Given an input phylogeny and a set of model parameters, simulated chromosome numbers were generated as described previously in Mayrose *et al.* (2010). As the number of simulation conditions is infinite, we concentrated on eight scenarios that vary in terms of data size (the number of tips in the phylogeny and the observed chromosome-number distribution) and the inferred pattern of chromosome-number change (Table 2). The phylogenies, chromosome counts and model parameters were taken from empirical datasets analyzed previously using ChromEvol (Glick *et al.*, 2016; Rice *et al.*, 2019), thus representing realistic data characteristics. In all simulation scenarios considered, the input phylogenies were assumed to be time-calibrated (i.e. ultrametric). For each simulation scenario, a total of 100 replicates were generated. Each simulated dataset was

Table 2 The eight simulation scenarios examined in this study.

Genus	Number of taxa	Generating model	Total branch lengths	Model parameters[1]				
				λ	δ	ρ	β	ν
Aloe	120	$D_{ys}D_{up}$	0.57	0 (0)	0.34 (1)	2.61 (8)		
Phacelia	53	$D_{ys}D_{up}$	0.61	0.20 (2)	2.33 (21)	0.67 (6)		
Lupinus	77	D_{ys}	0.47	0.85 (7)	9.53 (76)			
Hypochaeris	38	D_{ys}	0.86	1.14 (5)	0.43 (2)			
Brassica	36	$D_{ys}B_{num}$	0.32	1.24 (11)	0.70 (6)		8	0.55 (5)
Pectis	49	$D_{ys}B_{num}$	0.26	0 (0)	0.40 (2)		12	0.55 (3)
Crepis	81	$D_{ys}D_{up}B_{num}$	1.33	2.41 (19)	0.99 (8)	0.26 (2)	8	0.18 (1)
Hordeum	36	$D_{ys}D_{up}B_{num}$	1.26	0 (0)	0 (0)	1.80 (5)	7	1.36 (4)

[1]In parentheses: average number of simulated events across the tree. Shaded cells mark parameters that are not included in the generating model.

then fitted to a set of four models: D_{ys}, $D_{ys}D_{up}$, $D_{ys}B_{num}$, $D_{ys}D_{up}B_{num}$. For each simulation scenario, one of these models was the generating model (i.e. the model that was used to simulate the data) and three were nongenerating models. We note that these models share both common and distinct aspects of the parameter space, such that some – but not all – models are nested within each other (Table 1). Finally, the adequacy of each model to the simulated data was assessed.

Inference errors of adequate and inadequate models

Aside from the inferred model parameters, the ChromEvol software inherently allows several inference tasks to be carried out. These include ancestral state reconstruction and the inferred number of each type of transition (dysploidy, WGD, demi-polyploidization and base-number transition) occurring along each branch of the phylogeny. Additionally, a follow-up analysis provides an explicit categorization of each tip taxon as either diploid or polyploid, relative to the group in question (see Glick & Mayrose, 2014). Naturally, these inference tasks are directly affected by the model in question. The consequences of using an adequate vs inadequate model were evaluated by comparing the errors of four common inference tasks:

(**1**) The chromosome number at the root of the phylogeny; calculated as the deviation from 1.0 of the posterior probability assigned to the true (i.e. simulated) chromosome number at the root.

(**2**) The total number of dysploidy events across the phylogeny; calculated as the relative error between the inferred and simulated number of events: $2 \times (|x_1 - x_2|/x_1 + x_2)$, where x_1 and x_2 are the simulated and inferred number of dysploidy events, respectively. In case both x_1 and x_2 equal zero, the error was assigned as zero.

(**3**) The total number of polyploidization events across the phylogeny; the relative error was calculated similar to the total number of dysploidy events. Duplication events, demi-duplications and base-number transitions were regarded as polyploidization events.

(**4**) Ploidy-level assignments; the ploidy-level inference of tip taxa, as either diploids or polyploids, was based on the procedure described in Glick & Mayrose (2014). The assignments of all tips were compared between the inferred and true values. The number of falsely inferred taxa, divided by the total number of taxa, was used as the error measure.

In this analysis, six of the eight simulation scenarios in Table 2 were examined. The two scenarios excluded were those generated under the simple D_{ys} model, for which not all inference tasks are relevant. To eliminate possible confounding effects between the specific model used for inference and the magnitude of the error, in this evaluation a single nongenerating model ($D_{ys}D_{up}$ or $D_{ys}B_{num}$) was fitted to the data per simulation scenario (Table S2). For each simulation scenario, 300 replicates were generated. For each replicate, the phylogeny and the simulated chromosome counts were given as input to the model adequacy test and the dataset was determined as either adequate or inadequate. A one-sided Student's *t*-test was conducted to determine whether the error of a certain inference task was significantly larger in the inadequate set compared to the adequate set.

Application to empirical datasets

In order to demonstrate the usability of the model adequacy framework, we applied it to a dataset of 200 angiosperm genera, which were selected randomly from a large database consisting of thousands of plant genera, excluding genera with no variations in chromosome numbers as well as those with five or fewer species with both phylogenetic and chromosome-number information. The initial database was used, in part or as a whole, in several previous analyses (e.g. Glick *et al.*, 2016; Salman-Minkov *et al.*, 2016; Zhan *et al.*, 2016; Rice *et al.*, 2019). From this database we also selected 40 angiosperm genera that each contain at least one allopolyploid species, based on data from Barker *et al.* (2016). As a consequence of overlaps between these two sets, a total of 233 unique datasets were analyzed. Full details of the reconstruction of the original database are described in Rice *et al.* (2019). Briefly, for each genus, the OneTwoTree pipeline (Drori *et al.*, 2018) was used to automatically reconstruct the phylogeny using publicly available sequence data as they appear in GenBank (Benson *et al.*, 2013). Chromosome numbers for all species were retrieved from the Chromosome Counts Database (CCDB; Rice *et al.*, 2015). These data were given as input to ChromEvol, which was executed on the six models detailed in Table 1. In addition, we applied similar procedures to seven clades of higher taxonomical ranks, including five families, one subfamily and one tribe. The evolution of chromosome numbers in these clades using ChromEvol was examined previously in several studies (Table S3).

Implementation and availability

The model adequacy procedure was implemented in Python and R (R Core Team, 2013). The source codes and running instructions are available at https://github.com/MayroseLab/chromEvol_model_adequacy. The obligatory inputs were three files obtained through a ChromEvol run of the examined model: the summary results file, the tree with the inferred ancestral reconstruction in a Newick format, and the original counts file in Fasta format. The program outputs, for each test statistic examined, were its midpoint two-sided *P*-value computed from the simulated distribution of the test statistics. The model adequacy test also is available for on-line use through the ChromEvol webserver (http://chromevol.tau.ac.il/). We suggest that any standard ChromEvol analysis should be followed by the model adequacy procedure to confirm the absolute fit of the model and to determine whether any deviations of the model are of concern to the specific inference task.

Results

In this work we developed a statistical framework for testing the adequacy of likelihood models of chromosome-number evolution. In essence, the method tests whether a specified model is

capable of generating data that are similar to the data at hand. If not, the model is considered as providing inadequate description of the data, suggesting that other processes than those modelled have driven the evolution of chromosome numbers along the examined phylogeny. We first evaluated the performance of the model adequacy framework using simulations. We then applied it to a large number of real datasets derived from dozens of angiosperm genera, as well as to seven clades of higher taxonomic ranks, that together vary greatly in their extent of divergence time and patterns of chromosome number variation.

Framework validation

Simulations were used to validate the developed model adequacy approach. Several simulation scenarios were examined, whose phylogenies and simulated parameters were derived from real data analyses and cover various data characteristics (Table 2). In each scenario, a single model was used to generate the data. Given the simulated data, the generating model and three additional models were fitted to the data, and their adequacies were examined. The four examined models are indicated by the type of transitions they allow for: D_{ys}, $D_{ys}D_{up}$, $D_{ys}B_{num}$ and $D_{ys}D_{up}B_{num}$ (Table 1). In total, eight different simulation scenarios were examined; two for each type of generating model.

We first examined the type I error rate: inferring the generating model as inadequate. Our results indicated that when considering a single test statistic independently, the error rate was generally below the expected value of 0.05 (average = 0.02, across the eight simulation scenarios and four test statistics; Table S4). The four test statistics considered showed the pattern described by Gelman *et al.* (2013), with the distributions of the *P*-values being more concentrated around 0.5 than expected under ideal conditions (Fig. S1). The statistic closest to the expected 5% was $Pars^{Time}$, which rejects on average 4.4% of the simulations. The *P*-values obtained using this statistic were more uniformly distributed than the other three statistics. Combining multiple test statistics together, we considered a model as inadequate if the *P*-value of the simulated distribution of one or more of the statistics is ≤ 0.05 (see the Description section). Under this definition, the percentage of generating models that were inferred as inadequate varied between 0.04 and 0.17 across the eight simulation scenarios (average 8%; Table 3). When Bonferroni correction for multiple testing was applied, the type I error rate dropped to an average of 0.03, making the test conservative. We note, however, that the four test statistics were not independent, violating the assumption of this correction.

We next examined the capability of the adequacy test to detect models that deviate from that of the generating models. Three types of model misspecification were examined: overparameterization, underparameterization, and misparameterization. In the case of overparameterization, the tested model allowed for additional types of chromosome-number change (as represented by extra free parameters) than those used to generate the data. This corresponds to cases where the generating model was nested within the tested model (e.g. generating model $D_{ys}D_{up}$, tested model $D_{ys}D_{up}B_{num}$). Our results indicated that the performances of overparameterized models were very similar to those of the generating models (Table 3). The few discrepancies were the result of either: (1) inaccurate parameter estimates of the more general model due to the extra degrees of freedom; (2) the optimization procedure reaching suboptimal regions of the parameter space (we note that although ChromEvol allows for more thorough likelihood optimization search, which should reduce such instances, this was not attempted here due to the large number of simulations employed); and (3) very similar parameter estimates obtained using the two models, but slight deviations of the test statistics leading one model to be inferred as inadequate and the other adequate.

In the case of underparameterized models, the tested model allowed for fewer types of transitions than the generating model (e.g. generating model $D_{ys}D_{up}$, tested model D_{ys}). As may be expected, in all simulation scenarios the underparameterized models were more frequently inferred as inadequate compared to the generating models. The adequacy rate was very low when the

Table 3 The inadequacy rates of the four tested models in the various simulation scenarios examined (100 simulations per tested model per scenario).

Simulation scenario	Generating model	Tested models[1]			
		$D_{ys}D_{up}$	D_{ys}	$D_{ys}B_{num}$	$D_{ys}D_{up}B_{num}$
Aloe	$D_{ys}D_{up}$	0.06	1.00	0.04	0.08
Phacelia	$D_{ys}D_{up}$	0.04	0.99	0.04	0.03
Lupinus	D_{ys}	0.06	0.09	0.06	0.07
Hypochaeris	D_{ys}	0.03	0.06	0.04	0.05
Brassica	$D_{ys}B_{num}$	0.18	0.98	0.05	0.06
Pectis	$D_{ys}B_{num}$	0.86	1.00	0.12	0.07
Crepis	$D_{ys}D_{up}B_{num}$	0.28	0.95	0.11	0.07
Hordeum	$D_{ys}D_{up}B_{num}$	0.78	1.00	0.29	0.17

The diagonal (white cells) are cases where the generating model is also the tested model. Dark grey represents over-parametrized models, light grey under-parametrized models, and patterned cells miss-parametrized models.

tested model allowed only for dysploid transitions whereas in reality polyploid transitions (either WGD and/or base-number transitions) have occurred (Table 3; all cases where the tested model was D_{ys}). The adequacy rates were higher when the generating model allowed for multiple types of polyploid transitions (i.e. $D_{ys}D_{up}B_{num}$ allowing for both exact duplications and base-number transitions), whereas the tested model allowed for a subset of these ($D_{ys}D_{up}$ and $D_{ys}B_{num}$ that allow only for duplications or base-number transitions, respectively). Comparing the adequacy of the two underparametrized models ($D_{ys}D_{up}$ and $D_{ys}B_{num}$), the $D_{ys}B_{num}$ model that incorporated base-number transitions had higher adequacy rates compared to the $D_{ys}D_{up}$ model that allowed for exact duplications, as the former allowed for several transitions that frequently include also exact duplications (e.g. in case the base number is 8, both 8→16 and 8→24 transitions are allowed).

In the case of misparametrization, the tested and generated models were not nested within each other and thus their parameters only overlapped partially. For the set of models examined here, this fits the case with generating model $D_{ys}D_{up}$ and tested model $D_{ys}B_{num}$, or vice versa. When the tested model was $D_{ys}B_{num}$, it obtained similar adequacy rates to those of the generating $D_{ys}D_{up}$ model. By contrast, and similar to the results detailed in the case of underparameterized models, the $D_{ys}D_{up}$ model was inferred as inadequate a large number of times when the generating model was $D_{ys}B_{num}$.

Inference errors of adequate and inadequate models

A central usage of probabilistic models of chromosome number evolution is their inference capabilities, such as ancestral reconstructions of chromosome numbers, or predicting the branches in which dysploidy and polyploidy events have most likely occurred. Still, it is unclear whether the use of inadequate models would deteriorate the performance of such inference tasks. To this end, simulations were used to compare the errors of the following four common inference tasks when adequate and inadequate models are employed: (1) the chromosome number at the root of the phylogeny; (2) the total number of inferred dysploidy events; (3) the total number of inferred polyploidization events; (4) inferring the ploidy level of tip taxa as either diploid or polyploid (see Description section for details regarding the error computed for each inference task).

Our results demonstrated that the use of inadequate models frequently led to larger inference errors, although under some simulation scenarios the inference errors of inadequate models were similar to those obtained using adequate models. For example, the error in the inference of the root chromosome number was significantly larger in the case of inadequate models under two simulation scenarios, but was nonsignificantly different in the other four (Fig. 2). Likewise, in two out of the six simulation scenarios, the error of inferring the ploidy level of extant taxa was significantly larger when computed using inadequate vs adequate models. In this case, the magnitude of the error was relatively low whether adequate or inadequate models were applied: when inadequate models were applied, the mean error was 7.3% across all simulation scenarios, reaching up to 25% under the *Brassica* simulation scenario. In comparison, the mean error was 2% when adequate models were applied, reaching up to 7% of erroneous inferences under the *Hordeum* simulation scenario. Larger differences in the errors between adequate and inadequate models were observed in inferring the total number of polyploidizations, and even more so in inferring the total number of dysploidy events. For both of these inference tasks, significant differences between adequate and inadequate models were obtained for three of the six simulation scenarios. Generally, the relative error in inferring the total number of dysploidy events was larger compared to that of inferring the total number of polyploidizations (the mean relative error was roughly twice for dysploidy compared to polyploidy transitions, both in the adequate set and the inadequate set; Fig. 2).

Application to empirical datasets

We applied the model adequacy framework to 200 datasets, each corresponding to a single randomly-selected angiosperm genus. First, we performed a standard model selection procedure based on the AIC (Akaike, 1974) to evaluate the relative fit of each of the six ChromEvol models to the data. In 24% of the datasets, the simple D_{ys} model, which allows for dysploid transitions only, was selected. The model that was most frequently selected was $D_{ys}D_{up}$ (28%), whereas models that allow for demi-polyploidy transitions and those that allow for base-number transitions were selected in 27% and 21% of the datasets, respectively (Fig. 3a). Next, we applied the model adequacy test to the best model identified for each dataset. We found that in 70% of the genera, the model that was chosen as best by the AIC was inferred to provide an adequate description of the data. Examining the inadequacy rates of each test statistics revealed that the test statistic with the highest inadequacy rates was $Pars^{Time}$, whereas that with the lowest inadequacy rates was the Parsimony statistic (inadequacy rates 0.095, 0.13, 0.075 and 0.17 for variance, entropy, parsimony and $Pars^{Time}$, respectively). Applying the model adequacy test to all six models per dataset (whether or not selected as best), we found that models that allow for fewer types of transitions were more frequently predicted as inadequate (Fig. 3b). For example, the D_{ys} model that allowed only for dysploidy transitions was adequate in only 23% of the 200 datasets; models that additionally allowed for one type of polyploidy, either duplication or base-number transition, were adequate in 58% and 63% of the cases, respectively; whereas the three models that incorporated two types of polyploidy transitions ($D_{ys}D_{up}D_{em}$, $D_{ys}D_{up}D_{em}$* and $D_{ys}D_{up}B_{num}$) were inferred as adequate most frequently. The adequacy rates of all models were generally related to the complexity of the model that was selected as optimal. Thus, when the most complex models were selected ($D_{ys}D_{up}D_{em}$ and $D_{ys}D_{up}B_{num}$), the adequacy rates of all models – including that of the chosen model – were low (32% and 46%, respectively), whereas when the least complex model was selected, the adequacy rates of all models was high (68%; Table S5).

Next, we examined the model adequacy procedure in groups that have evolved via reticulate evolution at some point in their histories. In these clades, the underlying assumption of the

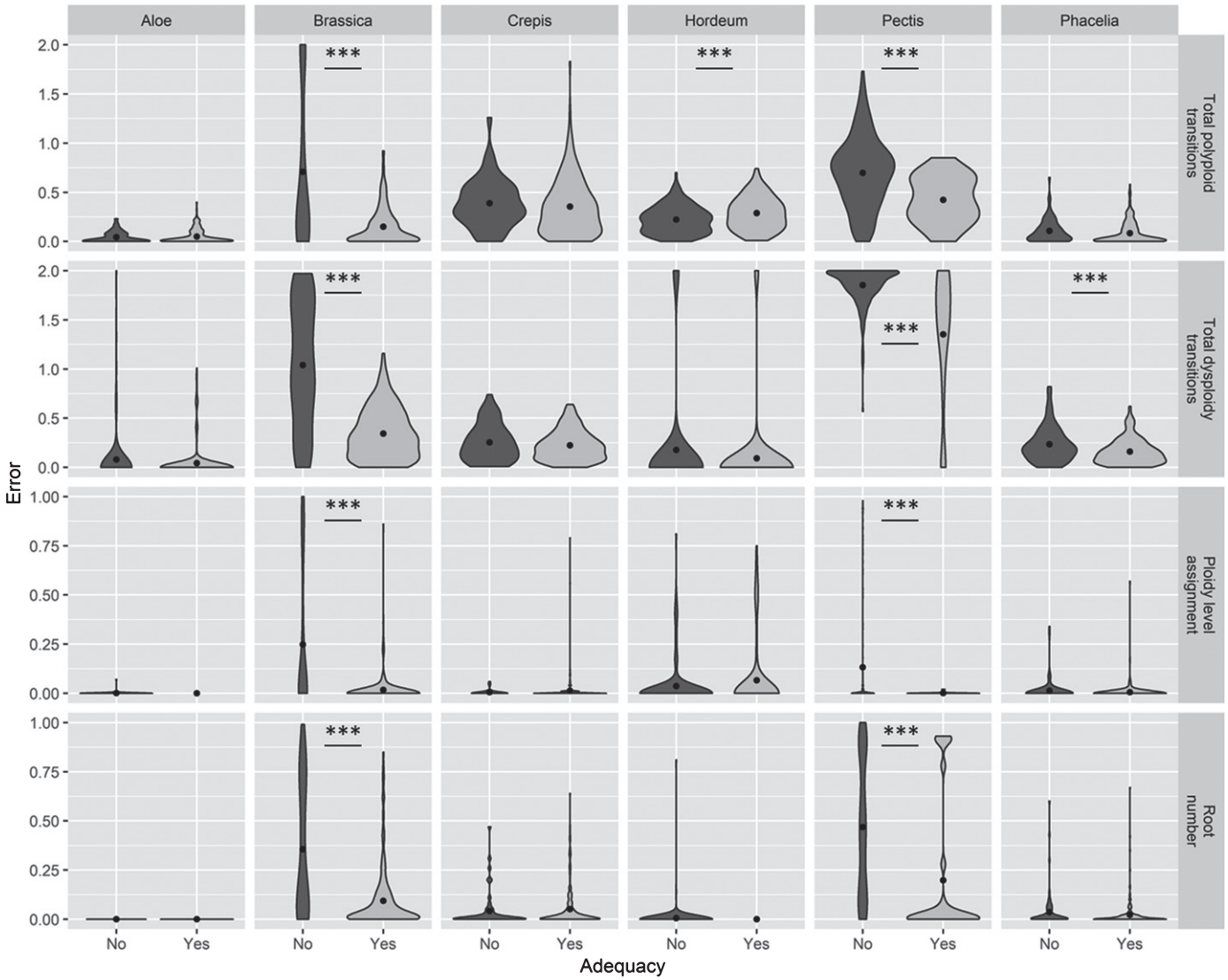

Fig. 2 The mean inference errors obtained under adequate and inadequate models for each simulated scenario. Each row presents the error of a different inference task. From top to bottom: inferring the total number of polyploid events across the tree, inferring the total number of dysploid events across the tree, ploidy-level assignments of extant taxa, the probability of the chromosome number at the root of the phylogeny. Each column denotes a different simulation scenario. For each scenario, 300 simulations were conducted and runs were partitioned to adequate and inadequate models. The violin plots represent the distribution of the errors obtained for the adequate (light grey, right) and inadequate (dark grey, left) sets. The black dot within each distribution denotes its mean. Asterisk indicates significant difference between the two groups (***, $P < 0.01$).

ChromEvol framework, in which evolution proceeds along a phylogenetic structure, is violated, at least to some extent. This analysis was performed on 40 genera that were identified in the literature to include allopolyploid species, and thus hybridizations were reported to occur (data taken from Barker *et al.*, 2016). In the majority of these genera (23 of 40), the model that was selected as optimal according to the AIC was found by our model adequacy procedure to be inadequate. This adequacy rate was significantly lower ($P << 0.05$; χ^2-test) compared to a random set of 193 genera in which allopolyploidy was not reported (the 200 genera analyzed above, omitting seven that include a reported allopolyploid species).

Finally, we evaluated the model adequacy procedure on a set of seven groups whose taxonomic rank was higher than the genus level, thus representing clades whose divergence time is generally older than those inspected above. The evolution of chromosome numbers in these clades likely violates the time homogeneity assumption of ChromEvol, in which the transition pattern is similar across the phylogeny. For five of these seven clades, the model that was chosen as optimal according to AIC did not provide adequate description of the data according to the model adequacy test (Table S3). Taken together, the last two analyses indicated that the model adequacy procedure can identify cases in which the evolution of chromosome numbers is driven by processes that deviate from the basic modelling assumptions of the ChromEvol framework.

Discussion

For over a century, the determination of chromosome numbers has played a vital role in studying evolutionary and genomic

processes in plants. Probabilistic models of chromosome-number change are a relatively recent addition to the research toolbox available to study the evolution of major genomic processes. As the usage of such models increases, so does the need to assess their validity when applied to real data. Here, we developed a model adequacy test for likelihood models of chromosome-number evolution. We focused our analysis on those models implemented in the ChromEvol software (Glick & Mayrose, 2014), but the procedures are general and can be implemented in other platforms that use variations to the ChromEvol model (Freyman & Höhna, 2017; Zenil-Ferguson *et al.*, 2017; Blackmon *et al.*, 2019). The developed test was based on the parametric bootstrapping approach (Goldman, 1993; Efron & Tibshirani, 1994) in which observed data are compared to a simulated distribution generated by the examined model. Using multiple test statistics that describe various characteristics of the data, the test allows us to determine whether the model can generate data that are similar to those found in the observed ones. The described methodology was implemented under the maximum-likelihood paradigm, following the implementation of the ChromEvol software, in order to naturally link the two implementations. Additionally, the model adequacy procedure can be implemented under a Bayesian inference scheme, using posterior predictive simulations. Although the overall implementation is similar in both approaches, the Bayesian approach inherently accounts for uncertainty in the parameter estimates, and possibly the phylogeny, by sampling them from the inferred posterior distribution.

The model adequacy test developed here was based on several data-based statistics (i.e. statistics that depend on the data alone, independently of the model being fitted). As an alternative, inference-based statistics, which depend both on the data and the fitted model, may be used (see Brown & Thomson, 2018; Höhna *et al.*, 2018). These include, for example, comparison of the ancestral states reconstructed by the examined model to the simulated ones, or comparison between the inferred and simulated number of polyploidization and dysploidization transitions. On the one hand, the computation of these inference-based test statistics is far more demanding than the data-based test statistics, because the evaluated model has to be fitted to each simulated dataset being generated, rendering the entire process several orders of magnitude slower than the present approach (e.g. Höhna *et al.*, 2018). On the other, these inference-based test statistics allow a more meaningful assessment of model adequacy by directly examining whether the model at hand is adequate for a desired inference task. Certainly, incorporating additional meaningful test statistics is a promising direction for future research.

Our simulation results indicate that the model adequacy framework had an acceptable type I error rate (i.e. inferring as inadequate a model that was used to generate the data). However, higher type I errors were found in models that allowed for base-number transitions ($D_{ys}B_{num}$ and $D_{ys}B_{num}D_{up}$). This suggests that these models might not be appropriate in all cases. The current implementation of such models assumes the same rate for all possible base-number transitions (e.g. given a base number of $\beta = 7$, the additions of 7, 14 or 21 chromosomes are equally likely). Alternatively, it may be more appropriate to place a probability distribution over the possible base-number transitions.

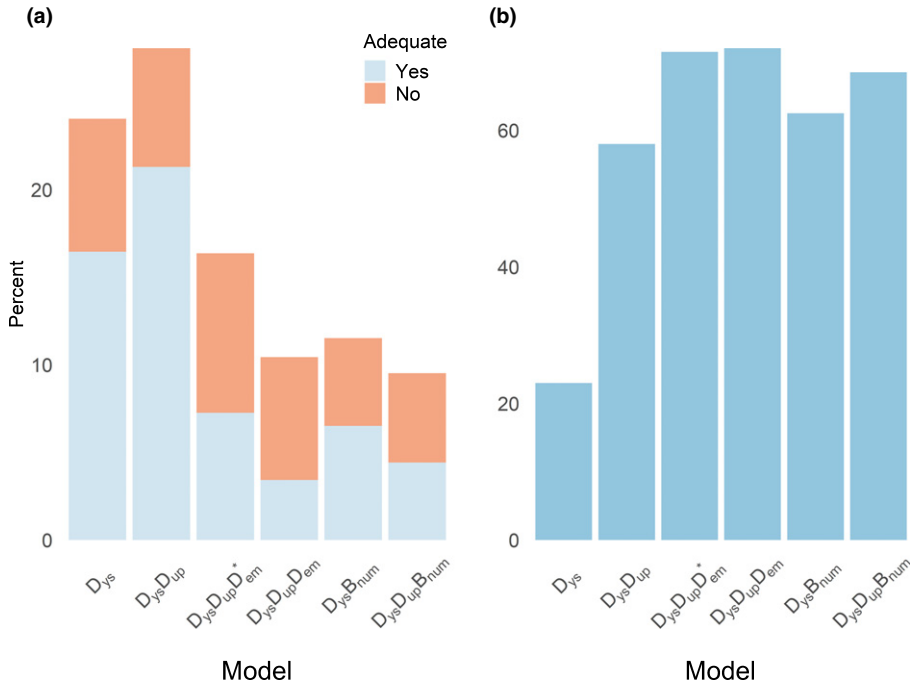

Fig. 3 Application of the model adequacy test to 200 angiosperm genera. (a) A bar plot representing the frequency of selection according to the Aikake Information Criterion (AIC) of each of the six tested models in the 200 examined angiosperm genera. The height of each bar is partitioned according to the percentage of genera that were determined as adequate (light blue) or inadequate (red). (b) The adequacy rate of each model when applied to all genera, regardless of whether the model was selected (n = 200).

This will allow, for the example of $\beta = 7$, higher rates for additions by 7 chromosomes compared to those by 21.

Our simulation results also demonstrated that the adequacy rate of overparameterized models, which allows for more types of transitions than those that truly occurred, is similar to that of the generating models. Although it is expected that the accuracy of inferring the model parameters will decrease as overly-complexed models are evaluated, in many cases the auxiliary parameters were optimized to very low values, resulting in a process nearly identical to the generating model. Thus, it seems that the flexibility offered by complex models does not necessarily lead to their disadvantage, at least for some inference tasks, as has been demonstrated recently for models of nucleotide sequence evolution (Abadi *et al.*, 2019). In other cases of model violations, either for underparameterized or misparameterized models, when the rate parameters deviated substantially from the original ones (e.g. dysploidy rates an order of magnitude larger than the simulated rates), the model adequacy framework detected such cases as inadequate. This suggests that the adequacy test was capable of detecting models that are completely wrong. In other cases, the nature of model misspecification affected the outcome. In the simulations examined here, $D_{ys}B_{num}$ was more frequently adequate than $D_{ys}D_{up}$, both in the case of underparameterization (i.e. when the generating model was $D_{ys}D_{up}B_{num}$ such that both models miss one type of transition) and misparameterization. Nevertheless, we note that the $D_{ys}B_{num}$ model may not fit well in large phylogenies with high dysploidy rates. In its current implementation, the model assumes that a single base number typifies a clade. However, if there is a high dysploidy rate, each subclade of the phylogeny may be characterized by its own base number or by multiple base numbers, which will necessitate more complex modelling options.

We further tested the consequences of using an inadequate model by examining the errors of several inference tasks. First, we found that the difference in inference error between adequate and inadequate models depended on the simulation scenarios: in some simulation scenarios the use of inadequate models resulted in significantly inflated inference errors compared to the use of adequate models, in some scenarios it affected only certain inference tasks and not others, whereas in others the difference was negligible for all tasks. Second, we found that some inference tasks were much more sensitive to model misspecification than others. The assignment of extant taxa as diploids or polyploids was the inference task that was least affected from using an inadequate model, and in general, the error of this inference task was very low. This indicates that determining the ploidy levels of extant taxa is generally robust to model misspecification. However, the error of determining the number of events that had occurred – either dysploid or polyploid transitions – can be substantial when inadequate models are employed.

Applying the model adequacy test to hundreds of angiosperm genera, we found that in the majority of the cases the best-fitted model provided sufficient approximation to the evolutionary processes underlying the data and was determined as adequate. However, in roughly a quarter of the examined genera, this selection turned out to be inadequate, suggesting that there is ample room for future modelling improvements. Examining the inadequacy rates of each test statistic across the 200 genera showed that $Pars^{Time}$ was the test statistic that rejected the best-fitted model most frequently. This statistic tests whether the number of transitions is distributed homogeneously in time across the phylogeny. Deviation of this statistic from model expectations thus indicates that transitions in chromosome numbers occurred either closer to the root of the tree, or located more near the tips. In a further analysis we found high rates of model inadequacy when applying the developed procedures to two types of clades that are expected to violate basic modelling assumptions: (1) clades in which allopolyploidy events are known to occur, thus violating the assumption that evolution proceeds via a phylogenetic structure; and (2) large and diverse clades in which a single transition process is fitted to the entire phylogeny, thus violating the time homogeneity assumption. Together, these results indicate that promising future developments would be to focus on analytical procedures based on phylogenetic networks (Nakhleh, 2010), rather than on bifurcating phylogenies, and to further incorporate time-heterogeneous processes.

It was shown previously that the power of model adequacy tests tends to increase with data size (Brown & Thomson, 2018). Indeed, in our analysis of 200 angiosperm genera we found that the size of the tree, approximated by its number of tips, was larger when the best selected model was determined as inadequate compared to when it was determined as adequate (36 tips on average in the inadequate set, compared to 20 tips on average in the adequate set; $P < 0.01$ Student's *t*-test). In addition, there were inverse correlations between the tree size and the *P*-values of all test statistics, suggesting that our test is more powerful when larger trees are analyzed. Nevertheless, larger trees may exhibit lower adequacy rates due to genuine rate heterogeneity present in the data, thus violating one of the assumptions underlying the ChromEvol model. However, the association between data size and the test statistic that specifically reflects rate heterogeneity across the phylogeny ($Pars^{Time}$) was lower than two of the other test statistics ($r^2 = 0.012$, 0.13 and 0.14 for $Pars^{Time}$, entropy and variance, respectively), suggesting that the tree size influences the power of the test, irrespective of whether rate heterogeneity is present.

Phylogenetic model adequacy tests previously have been developed for other data types and inference tasks, although their use has not been adopted widely. This could be a consequence of the apparent limited benefit offered to a researcher if all examined models are deemed inadequate when applied to a clade of interest. We argue, however, that model adequacy tests are of practical use to methods developers and end users alike, and should thus be practiced regularly as part of a broader model assessment routine. For researchers interested in data analysis, inadequate models can hint at errors in the input data, which should thus be more carefully inspected. In the case studied here, possible sources of errors include those in the assumed phylogenetic hypothesis, in the collection of chromosome counts or in taxa sampling. Inadequacy also could point to additional attributes that should be considered in the analysis. For example, if all models that assume a time-homogenous transition process fail, it

could suggest that patterns of chromosome-number change are clade-specific (Márquez-Corro *et al.*, 2019) or depend on an organismal trait (e.g. the plant growth form) The analysis thus could be enhanced if such factors are accounted for using more complex models (e.g. Zenil-Ferguson *et al.*, 2017; Blackmon *et al.*, 2019; see Zenil-Ferguson *et al.*, 2019 for discussion about the advantages of using complex models). For researchers interested in large-scale analyses that include multiple datasets, where the in-depth examination of each inadequate dataset is not feasible, the filtration of such clades is one obvious possible direction. For some inference tasks, such as the identification of ploidy levels of extant taxa, the effect of using an inadequate model is rather negligible, indicating that the treatment of the flagged clades should be tuned to the analysis in question. For developers, the frequent application of model adequacy tests should provide interesting test cases on which new models are trained. Moreover, when a model is deemed inadequate, the test statistics that fail to align may point to processes absent from existing models, which could be included in the future. Model adequacy should thus take a vital part in this recurrent chain of scientific progress in which new methods are developed, regularly used, and then replaced by more advanced alternatives.

Acknowledgements

The authors wish to thank Sebastian Höhna, two anonymous reviewers and the Associate Editor for their insightful comments. AR is supported by a fellowship from the Edmond J. Safra Center for Bioinformatics at Tel Aviv University and by the NA'AMAT Professional Scholarship. This study was supported by grant no. 961/17 from the Israel Science Foundation and National Science Foundation-BSF (1655478) to IM.

Author contributions

IM and AR conceived the study; AR built the tool and analyzed the data; AR and IM wrote the manuscript; and IM supervised the study.

ORCID

Itay Mayrose https://orcid.org/0000-0002-8460-1502
Anna Rice https://orcid.org/0000-0002-2213-0688

References

Abadi S, Azouri D, Pupko T, Mayrose I. 2019. Model selection may not be a mandatory step for phylogeny reconstruction. *Nature Communications* **10**: 1–11.

Akaike H. 1974. A new look at the statistical model identification. *IEEE Transactions on Automatic Control* **19**: 716–723.

Barker MS, Arrigo N, Baniaga AE, Li Z, Levin DA. 2016. On the relative abundance of autopolyploids and allopolyploids. *New Phytologist* **210**: 391–398.

Beaulieu JM, O'Meara BC, Donoghue MJ. 2013. Identifying hidden rate changes in the evolution of a binary morphological character: The evolution of plant habit in campanulid angiosperms. *Systematic Biology* **62**: 725–737.

Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. 2013. GenBank. *Nucleic Acids Research* **41**: 36–42.

Blackmon H, Demuth JP. 2014. Estimating tempo and mode of Y chromosome turnover: explaining Y chromosome loss with the fragile Y hypothesis. *Genetics* **197**: 561–572.

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

Bollback JP. 2002. Bayesian model adequacy and choice in phylogenetics. *Molecular Biology and Evolution* **19**: 1171–1180.

Brown JM. 2014. Detection of implausible phylogenetic inferences using posterior predictive assessment of model fit. *Systematic Biology* **63**: 334–348.

Brown JM, Thomson RC. 2018. Evaluating model performance in evolutionary biology. *Annual Review of Ecology, Evolution, and Systematics* **49**: 95–114.

Carta A, Bedini G, Peruzzi L. 2018. Unscrambling phylogenetic effects and ecological determinants of chromosome number in major angiosperm clades. *Scientific Reports* **8**: 1–14.

Chen W, Kenney T, Bielawski J, Gu H. 2019. Testing adequacy for DNA substitution models. *BMC Bioinformatics* **20**: 349.

Drori M, Rice A, Einhorn M, Chay O, Glick L, Mayrose I. 2018. OneTwoTree: an online tool for phylogeny reconstruction. *Molecular Ecology Resources* **18**: 1492–1499.

Duchêne DA, Duchêne S, Holmes EC, Ho SYW. 2015. Evaluating the adequacy of molecular clock models using posterior predictive simulations. *Molecular Biology and Evolution* **32**: 2986–2995.

Efron B, Tibshirani RJ. 1994. *An introduction to the bootstrap*. Boca Raton, FL, USA: CRC Press.

Farris JS. 1970. Methods for computing Wagner trees. *Systematic Biology* **19**: 83–92.

Fitch WM. 1971. Toward defining the course of evolution: minimum change for a specific tree topology. *Systematic Zoology* **20**: 406.

Freyman WA, Höhna S. 2017. Cladogenetic and anagenetic models of chromosome number evolution: a Bayesian model averaging approach. *Systematic Biology* **67**: 195–215.

Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB. 2013. *Bayesian data analysis, 3rd edn*. Boca Raton, FL, USA: Taylor & Francis.

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.

Glick L, Sabath N, Ashman T-L, Goldberg E, Mayrose I. 2016. Polyploidy and sexual system in angiosperms: is there an association? *American Journal of Botany* **103**: 1223–1235.

Goldman N. 1993. Statistical tests of models of DNA substitution. *Journal of Molecular Evolution* **36**: 182–198.

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

Hallinan NM, Lindberg DR. 2011. Comparative analysis of chromosome counts infers three paleopolyploidies in the mollusca. *Genome Biology and Evolution* **3**: 1150–1163.

Höhna S, Coghill LM, Mount GG, Thomson RC, Brown JM. 2018. P3: Phylogenetic posterior prediction in RevBayes. *Molecular Biology and Evolution* **35**: 1028–1034.

Khandelwal S. 1990. Chromosome evolution in the genus Ophioglossum L. *Botanical Journal of the Linnean Society* **102**: 205–217.

Leitch AR, Leitch IJ. 2008. Genomic plasticity and the diversity of polyploid plants. *Science* **320**: 481–483.

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

Márquez-Corro JI, Martín-Bravo S, 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.

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

Nakhleh L. 2010. Evolutionary phylogenetic networks: models and issues. In: Heath LS, Ramakrishnan N, eds. *Problem solving handbook in computational biology and bioinformatics*. Boston, MA, USA: Springer, 125–158.

Pennell MW, FitzJohn RG, Cornwell WK, Harmon LJ. 2015. Model adequacy and the macroevolution of angiosperm functional traits. *The American Naturalist* **186**: E33–E50.

R Core Team. 2013. *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Ramsey J, Ramsey TS. 2014. Ecological studies of polyploidy in the 100 years following its discovery. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* **369**: 1–20.

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

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, Šmarda P, Novosolov M, Drori M, Glick L, Sabath N, Meiri S, Belmaker J, Mayrose I. 2019. The global biogeography of polyploid plants. *Nature Ecology & Evolution* **3**: 265–273.

Ruffini Castiglione M, Cremonini R. 2012. A fascinating island: $2n = 4$. *Plant Biosystems* **146**: 711–726.

Salman-Minkov A, Sabath N, Mayrose I. 2016. Whole-genome duplication as a key factor in crop domestication. *Nature Plants* **2**: 1–4.

Shannon CE. 1948. A mathematical theory of communication. *Bell System Technical Journal* **27**: 379–423.

Slater GJ, Pennell MW. 2013. Robust regression and posterior predictive simulation increase power to detect early bursts of trait evolution. *Systematic Biology* **63**: 293–308.

Soltis D, Soltis P, Schemske D. 2007. Autopolyploidy in angiosperms: have we grossly underestimated the number of species? *Taxon* **56**: 13–30.

Soltis DE, Visger CJ, Marchant DB, Soltis PS. 2016. Polyploidy: pitfalls and paths to a paradigm. *American Journal of Botany* **103**: 1146–1166.

Spoelhof JP, Soltis PS, Soltis DE. 2017. Pure polyploidy: closing the gaps in autopolyploid research. *Journal of Systematics and Evolution* **55**: 340–352.

Stebbins GL. 1971. *Chromosomal evolution in higher plants*. London, UK: Edward Arnold Ltd.

Vershinina AO, Lukhtanov VA. 2017. Evolutionary mechanisms of runaway chromosome number change in Agrodiaetus butterflies. *Scientific Reports* **7**: 1–9.

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

Wendel JF. 2015. The wondrous cycles of polyploidy in plants. *American Journal of Botany* **102**: 1753–1756.

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.

Zenil-Ferguson R, Burleigh JG, Freyman WA, Igić B, Mayrose I, Goldberg EE. 2019. Interaction among ploidy, breeding system and lineage diversification. *New Phytologist* **224**: 1252–1265.

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.

Zhan SH, Drori M, Goldberg EE, Otto SP, Mayrose I. 2016. Phylogenetic evidence for cladogenetic polyploidization in land plants. *American Journal of Botany* **103**: 1252–1258.

Supporting Information

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

Fig. S1 The distribution of *P*-values across the simulation scenarios assessing the type I error.

Methods S1 Model adequacy simulations.

Table S1 Pearson's *r* coefficient between each pair of statistics.

Table S2 The generating and fitted model for each simulation scenario used in the comparison of inference error between adequate and inadequate models.

Table S3 Details of the seven plant clades, whose taxonomic rank is above the genus level, examined in this study.

Table S4 Type I error rates for each test statistic per simulation scenario.

Table S5 Adequacy rates of all models, including those of the chosen models.

Please note: Wiley Blackwell are 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.