MBE Advance Access published April 25, 2014

ChromEvol: Assessing the Pattern of Chromosome Number Evolution and the Inference of Polyploidy along a Phylogeny

Lior Glick[1] and Itay Mayrose*[,1]

[1]Department of Molecular Biology and Ecology of Plants, Tel Aviv University, Tel Aviv, Israel

***Corresponding author:** E-mail: itaymay@post.tau.ac.il.

Associate editor: Matthew Hahn

Abstract

We announce the release of chromEvol version 2.0, a software tool for inferring the pattern of chromosome number change along a phylogeny. The software facilitates the inference of the expected number of polyploidy and dysploidy transitions along each branch of a phylogeny and estimates ancestral chromosome numbers at internal nodes. The new version features a novel extension of the model accounting for general multiplication events, other than doubling of the number of chromosomes. This allows the monoploid number (commonly referred to as *x*, or the base-number) of a group of interest to be inferred in a statistical framework. In addition, we devise an inference scheme, which allows explicit categorization of each terminal taxon as either diploid or polyploid. The new version also supports intraspecific variation in chromosome number and allows hypothesis testing regarding the root chromosome number. The software, alongside a detailed usage manual, is available at http://www.tau.ac.il/~itaymay/cp/chromEvol/.

Key words: **chromosome numbers, polyploidy, whole-genome duplication, software, chromEvol, base number.**

Chromosome number is a remarkably variable feature of eukaryotic genomes with variations in the somatic chromosome number known to exist at all levels of taxonomic resolution. The most recognizable chromosome number transition is whole-genome duplication (WGD), or more generally polyploidy, which due to the recent finding that many seemingly diploid species have in fact undergone recurrent episodes of genome duplication (Furlong and Holland 2004; Aury et al. 2006; Cui et al. 2006; Albertin and Marullo 2012), has arguably been the focal point of interest of many evolutionary biologists and genome sequencing projects. Although perhaps less appreciated, single chromosome number change is another common mechanism by which the evolution of chromosome numbers proceeds. These transitions include gain/loss of entire chromosomes by a process known as aneuploidy and processes such as chromosome fission or fusion (ascending or descending dysploidy, respectively), which change the karyotype but not the genomic content of a lineage. Although variation in chromosome numbers reaches its zenith in plants, interest in polyploidy and dysploidy transitions have been studied extensively in other groups, revealing an important role for chromosome number change on micro and macro evolutionary patterns in diverse groups such as fish (Taylor et al. 2003; Zhan et al. 2014), yeast (Scannell et al. 2006), and butterflies (Kandul et al. 2007).

The remarkable variations of chromosome numbers has drawn botanists, in particular, to evaluate the pattern of chromosome number change, to estimate the base chromosome number (commonly termed *x*), and to infer the frequency of polyploids within a clade of interest. The term base number has been used equivocally in the literature, either as the hypothesized chromosome number at the root of a clade, or alternatively, as the highest common factor of an observed chromosome number distribution (reviewed in Cusimano et al. 2012; Peruzzi 2013). Whatever the exact meaning of *x* may be, it is usually taken to represent the monoploid number of a group and is used to infer the ploidy level of a lineage. Earlier studies employed various threshold techniques to estimate the base number and thereby the occurrences and locations of polyploidy events (Stebbins 1938; Grant 1963; Goldblatt 1980; Wood et al. 2009). For example, assuming $x = 7–9$, Grant (1963) considered an angiosperm species polyploid if it has a haploid number greater than $n = 14$, whereas Goldblatt (1980) argued that this threshold should be lower and set it as $n = 11$. Alternatively, Stebbins (1938) designated a species as polyploid if its haploid number was a multiple of the lowest count found in the genus, whereas Wood et al. (2009) set this multiplication factor to be 1.75. Clearly, such threshold methods suffer from a large degree of extrapolation, do not account for the phylogenetic relationships among the species, and mostly disregard the frequency of chromosome number changes that are not due to polyploidy. More recently, polyploidy was inferred within a phylogenetic context following the maximum parsimony principle (e.g., Schultheis 2001; Ohi-toma et al. 2006). Accordingly, ancestral chromosome numbers are reconstructed and the chromosome number inferred at the most recent common ancestor (MRCA) of the group examined is designated as the base number. A certain lineage is inferred to be polyploid if its chromosome number is larger by a chosen factor compared with the base number.

Notably, most of these analyses have failed to account for the full information contained within the phylogeny of the species studied or for dysploidy transitions. Furthermore, these methods also implicitly assumed that the chromosome numbers at extant taxa must include the chromosome number that was present at the MRCA—an assumption that is particularly problematic if rates of chromosome

Letter

number change within the group are high. Recently, we have formulated a series of likelihood models depicting the pattern of chromosome number change along a phylogeny (Mayrose et al. 2010). These models moved the analysis of chromosome number change into a robust probabilistic inference framework, thereby allowing explicit questions regarding the pattern of chromosome number evolution to be formulated and tested.

Here, we present chromEvol v. 2.0, a software tool for inferring the pattern of chromosome number change along a phylogeny. Using this method, it is possible to assess the fit of several models of chromosome number change along a phylogeny, to infer the expected number of polyploidy and dysploidy transitions along each branch, and to estimate ancestral chromosome numbers at internal nodes of the tree. First, we describe a novel extension of the model that accounts for general multiplication events, other than doubling of the number of chromosomes. Using this extended model, the monoploid number of a group of interest can be inferred in a statistical framework. Second, we devise an inference scheme to allow explicit categorization of each terminal taxon to either diploid or polyploid relative to other taxa in the group examined. In addition, this extended version of chromEvol now supports intraspecific variations in chromosome numbers at tip taxa and permits users to fix the chromosome number at the root, thereby explicitly allowing the comparison between several alternative hypotheses regarding the root ancestral state to be statistically compared.

The General ChromEvol Model

The evolutionary models implemented in chromEvol are based on a continuous time Markov process. The most general model originally considered in Mayrose et al. (2010) includes six parameters and assumes that in an infinitesimal time interval, four types of events are possible: ascending dysploidy (with rate λ), descending dysploidy (rate δ), WGD (i.e., exact duplication of the number of chromosomes, at rate ρ), and demi-polyploidy (rate μ). This last term was first introduced in Mayrose et al. (2010) to account for possible multiplications of the number of chromosomes by 1.5, leading to, for example, triplication events. For example, a hexapolyploid may be formed via a triploid bridge followed by a genome duplication (Ramsey and Schemske 1998). Additional two rate parameters allow the ascending and descending dysploidy rates to depend on the current number of chromosomes. All these factors are captured within the instantaneous rate matrix Q, describing the rate of change from a genome with i haploid chromosomes to a genome with j chromosomes. For $i \neq j$, we define:

$$[Q]_{ij} = \begin{cases} \lambda + \lambda_l \times (i-1) & j = i+1 \text{ (ascending dysploidy)} \\ \delta + \delta_l \times (i-1) & j = i-1 \text{ (descending dysploidy)} \\ \rho & j = 2i \quad \text{(WGD)} \\ \mu & j = 1.5i \text{ (demi-polyploidy)} \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

The diagonal elements are determined by the constraint that each row in Q sums to zero.

Although inclusion of the demi-polyploidy parameter in the transition matrix allows for several alternative genome multiplications to be formulated, it also results in several shortcomings. First, demi-polyploidy transitions are well defined only for even haploid numbers. Thus, for example, a demi-polyploidy transition from $n = 9$ may lead (at equal rates) to either $n = 13$ or $n = 14$. However, neither of these two possibilities can truly represent a triplication event; assuming a subsequent duplication event, a $9 \rightarrow 27$ triplication would then involve an additional chromosome number gain (through the pathway $9 \rightarrow 13 \rightarrow 26 \rightarrow 27$) or loss (through the pathway $9 \rightarrow 14 \rightarrow 28 \rightarrow 27$), thus artificially increasing the rates of dysploidy. In addition, the model depicted in equation (1) still fails to account for other possible additions of the entire chromosome set to the genome. This is particularly true for clades exhibiting a noticeable polyploidy series or for clades in which intercytotype mating is possible. Consider, for example, the plant genus *Chrysanthemum*, which exhibits high variation in chromosome numbers and ploidy levels ($n = 9$, 18, 27, 36, 45) with $x = 9$ representing the hypothesized base number (Liu et al. 2012). According to the allowed transitions described in equation (1), a hypothesized $18 \rightarrow 45$ transition could not be obtained solely by any combination of demi- and polyploidy events and must artificially entail some additional dysploidy events. However, such an event can be reached via, for example, an $n = 18 + n = 27$ inter cytotype hybridization followed by a genome duplication.

To overcome these difficulties, in the new chromEvol implementation we developed a new parameterization of the model that explicitly allows for any multiplication of the monoploid chromosome number to be added to the genome. This is done by including two additional parameters to the model: β, the monoploid (base) number and ν, its respective transition rate. Under this scenario, the Q matrix takes the following form:

$$[Q]_{ij} = \begin{cases} \lambda + \lambda_l \times (i-1) & j = i+1 \\ \delta + \delta_l \times (i-1) & j = i-1 \\ \rho & j = 2i \\ \mu & j = 1.5i \\ \nu & (j-i) \text{ is divisible by } \beta \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

In the most general case, ν, ρ, and μ are all included in the model. Obviously, however, these parameters are somewhat redundant (e.g., assuming $\beta = 9$, an $18 \rightarrow 27$ transition can be reached via either demi-polyploidy or by a monoploid addition), and the user may choose to examine any subset of these parameters.

Using ChromEvol to Determine Polyploid Lineages

Although chromEvol can be used to infer the expected number of ploidy transitions along each branch of the tree, it does not classify extant taxa as diploid or polyploid, which is the aim of many analyses. For example,

in Mayrose et al. (2011), an extant taxon was categorized as a polyploid if the estimated expected number of ploidy transitions from the root to the tip exceeded a certain predefined threshold (i.e., 0.9) and as diploid otherwise. However, by arbitrarily setting a strict (or lenient) threshold for assigning polyploidy, the number of polyploid taxa may be underestimated (or overestimated). This misclassification may be particularly pronounced for groups with sparse chromosome number data. Thus, to prevent misestimating polyploid diversity, we developed a computational pipeline to infer which taxa underwent a polyploidization event since divergence from the MRCA of the group examined. By doing so, we explicitly treat the root of the phylogeny as diploid. We note that this methodology allows categorizing an extant species as polyploid or diploid regardless of whether chromosome number data are available for that specific taxon. Specifically, the pipeline accepts as input a file containing the chromosome numbers in a FASTA format and a file containing the phylogeny in a Newick format. The phylogeny can be either a single tree or a set of trees (obtained using, e.g., MrBayes; Ronquist and Huelsenbeck 2003), thus accounting for phylogenetic uncertainties. The output is the inferred ploidy level of each species or "NA" (not available) in case this could not be reliably inferred for a certain species. This pipeline is supplied as an external perl package, available through the program webpage. The procedure follows these general steps:

1) Find the optimal chromosome number evolutionary model for the given data set using a single tree (in our analyses, we used the maximum a-posteriori probability [MAP] tree). This model is used in all subsequent steps. The main aim of this step is to reduce computation time, so that not all available models are run in each step. The optimal model is selected by means of the Akaike information Criterion (AIC) (Akaike 1974).
2) In case multiple trees are available, randomly choose a prespecified number of trees (e.g., 100) to be used in subsequent steps.
3) The parameters of the model chosen in step 1 are optimized independently for each tree sampled in step 2.
4) Simulate chromosome number evolution along the input tree(s) using the corresponding model parameters as inferred in step 3. By default, 100 simulations are performed. The simulated chromosome numbers at the tips are then used as data input to chromEvol; thereby comparing "true" (simulated) and inferred ploidy transitions. To make the inference step as realistic as possible, simulated chromosome numbers at the tips are retained only for those species with available chromosome number in the original input data and are converted to "unknown" for species with missing data. There are two purposes for this simulation step. The first is to detect two thresholds required for inferring ploidy levels: A species is inferred as polyploid (diploid) if the expected number of polyploidization events from the root to the tip is above (below) the polyploid (diploid) threshold and as NA if this expectation is in between these two thresholds. The optimal thresholds are determined using the Matthews Correlation Coefficient (Matthews 1975) that balances between true/false positives and negatives. The second purpose for using simulations is to detect taxa for which ploidy level could not be reliably inferred according to the underlying model. These are the taxa that suffer from high false-positive/-negative rates (e.g., a taxon was inferred as diploid while it was simulated as polyploid). A taxon is marked as unreliably inferred if its ploidy level according to the optimized threshold was erroneously inferred in more than 5% of the simulations. Usually, these taxa reside in a subtree with a large fraction of species with missing chromosome counts, so that the inferences are based mainly on the evolutionary model and less on the observed data. These taxa are marked as NA.
5) In case multiple trees are given, infer ploidy levels across a sample of trees. This step is based on the thresholds found in step 4. Inferring from a sample of trees rather than from a single tree increases the robustness of inference by accounting for phylogenetic uncertainties. For each species, inferences obtained across the trees are compared and those species which do not exhibit the same inference across 95% or more of the topologies are marked as NA.
6) Combine the reliability estimates from steps 4 and 5 and summarize ploidy levels. When multiple trees are given, summary statistics of the optimized model parameters obtained across trees (median and 95% interval) further allows users to evaluate the consistency of the optimization search, thereby possibly locating multiple optima.

Additional Features

Determining the Root Frequencies

Likelihood calculations also require the assignment of root frequencies. In the previous implementation, root frequencies were determined according to their respective probabilities of giving rise to the extant data, given the model parameters. In the current version, we also allow these frequencies to be set by the user, as well as fixing the root state to a certain number. This allows users to directly compare several alternative hypotheses regarding the ancestral state (often referred to as the base chromosome number or "x") via a statistical model selection criterion. Note that the root state may or may not be equal to the monoploid number, β (eq. 2). Thus, this chromEvol implementation specifically differentiates between these two alternative definitions of the base number.

Intraspecific Chromosome Number Variation

In the initial implementation of chromEvol, each terminal taxon was assumed to possess a unique chromosome number. However, many plant species exhibit more than one chromosome number, and such intraspecific variation (often termed cytotypes) may in fact be very common. For example, 12–13% angiosperm and 17% fern species were estimated to harbor multiple ploidy levels (thus, ignoring other variations caused by dysploidy events;

Table 1. Summary of the Simulated Parameters and Inference Accuracy under Empirical Data Scenarios.

Data Set	Number of Taxa	Simulated Model	Simulated Parameters					Percent of Simulations with Correct β Inference[a]	ν Mean Relative Error[a,b]
			λ	δ	ρ	β	ν		
Sorghum	23	M9	0	0	0	5	0.46	100	0.42
Primula	16	M10	0.32	0.05	0.16	9	0.11	60	0.49
Lippia	23	M9	0.78	3.75	0	12	0.37	46	0.35
Hordeum	35	M10	0	0	0.38	14	0.24	94	0.58

[a]For each data set, the model used for inference was the same as the one used for simulation.

[b]Mean relative error was calculated using the formula $\frac{\sum_{i=1}^{100} |\nu_i - \nu_s| / \nu_s}{100}$, where ν_s is the simulated value of ν, and ν_i is the inferred value of ν in the i-th simulation.

Wood et al. 2009). In the updated chromEvol version, such intraspecific polymorphism was accommodated by allowing several chromosome numbers, together with their respective probabilities, to be set for each tip taxon. For example, a tip taxon having the following counts: 10, 10, 10, 20 will be given probabilities of 0.75 and 0.25 to have $n = 10$ and $n = 20$, respectively. Accordingly, assuming an ancestral state of $n = 10$, the expected number of WGDs leading to this taxon will be 0.25, because this distribution can be explained by one duplication occurring in 25% of the population. Note, that the expected number of WGDs would be 0.5 if we assumed the following distribution: 10,10,10,40 (two duplications leading to $n = 40$). Thus, caution should be taken when using this option, particularly when combining it with the ploidy inference pipeline described above, because higher ploidy counts may cause a taxon to be inferred as polyploid even when the majority of its sampled counts come from diploid plants. An alternative, computationally intensive, strategy that is recommended when high intraploidy variation is present would accommodate intraspecific variation using a sampling strategy whereby a single count for each taxon is iteratively sampled (e.g., 100 times) based on its observed distribution. ChromEvol should subsequently be executed on each sample, and ploidy levels can be inferred based on the inferred ploidy-levels distribution across the 100 chromEvol runs.

Program Implementation

The models and inference methods described here were implemented in C++. The program and source codes are available at www.tau.ac.il/~itaymay/cp/chromEvol (last accessed April 13, 2014). The obligatory inputs to the program are a tree file in a Newick format and a file containing chromosome numbers for extant taxa in a FASTA-like format. Given a rooted phylogenetic tree and given the assignment of chromosome numbers to extant species, the likelihood of the data can be calculated as described previously (Mayrose et al. 2010). A regular application of chromEvol entails obtaining the maximum likelihood (ML) scores of several alternative models, each represented by a different set of parameters. Model comparisons, for example, using the AIC, can then be used to determine the model that best fits a particular data set. For each model analyzed, model parameters are estimated under the ML criterion, ancestral states are inferred using both ML and Bayesian approaches, and an estimation of the number of events for each transition type along each branch of the phylogeny is given. The ancestral reconstruction and transition events may be viewed using any tree visualization software.

Assessing Accuracy via Simulations

Simulations were used to investigate the accuracy by which chromEvol infers the values of β (the base number) and ν (the corresponding transition rate). To restrict the parameter space examined, simulations were performed based on four sets of trees and parameter values as obtained from empirical data sets (table 1; see below). These data sets were chosen as they represent a range of evolutionary scenarios involving base-number transition, and in all these data sets, the best-fitted model included the β and ν parameters. The *Sorghum* data set represented the simplest scenario in which only transitions by base number were simulated. In *Hordeum*, WGD events were also simulated, occurring at a rate that is higher than that of base number transitions ($\rho = 0.38$, $\nu = 0.24$). In *Lippia*, high rate of dysploidy transitions were simulated, but WGD were not explicitly simulated, whereas the *Primula* data set represented the most complex scenario whereby all types of transitions were simulated. A total of 100 simulations were performed on each data set. Simulated data for each data set were prepared by modeling the evolutionary process given a phylogeny (i.e., the MAP tree) and a set of model parameters following the procedure described in Mayrose et al. (2010). Simulated chromosome numbers at the tips of the tree were then used as input for chromEvol.

First, we examined the inferred values of β throughout the simulated data sets (table 1). Highly accurate inferences were obtained for the *Hordeum* and *Sorghum* simulated data sets where the correct base number was inferred in nearly all simulations. The inference of β was less accurate in the *Primula* and *Lippia* data sets that included dysploidy transitions; this was particularly pronounced in the *Lippia* data set where very high dysploidy rates were simulated, resulting in the correct inference of β in only 46% of the simulations. The relative accuracy of inferring the ν parameter throughout the simulated data sets is presented in table 1. Our simulation results demonstrated that a more accurate inference is expected when only base-number transitions are present (as in *Lippia* and *Sorghum*) compared with data sets that additionally include WGD events that occur at a distinct rate (as in *Hordeum* and *Primula*). These simulations further demonstrated that the inference accuracy of the ν parameter

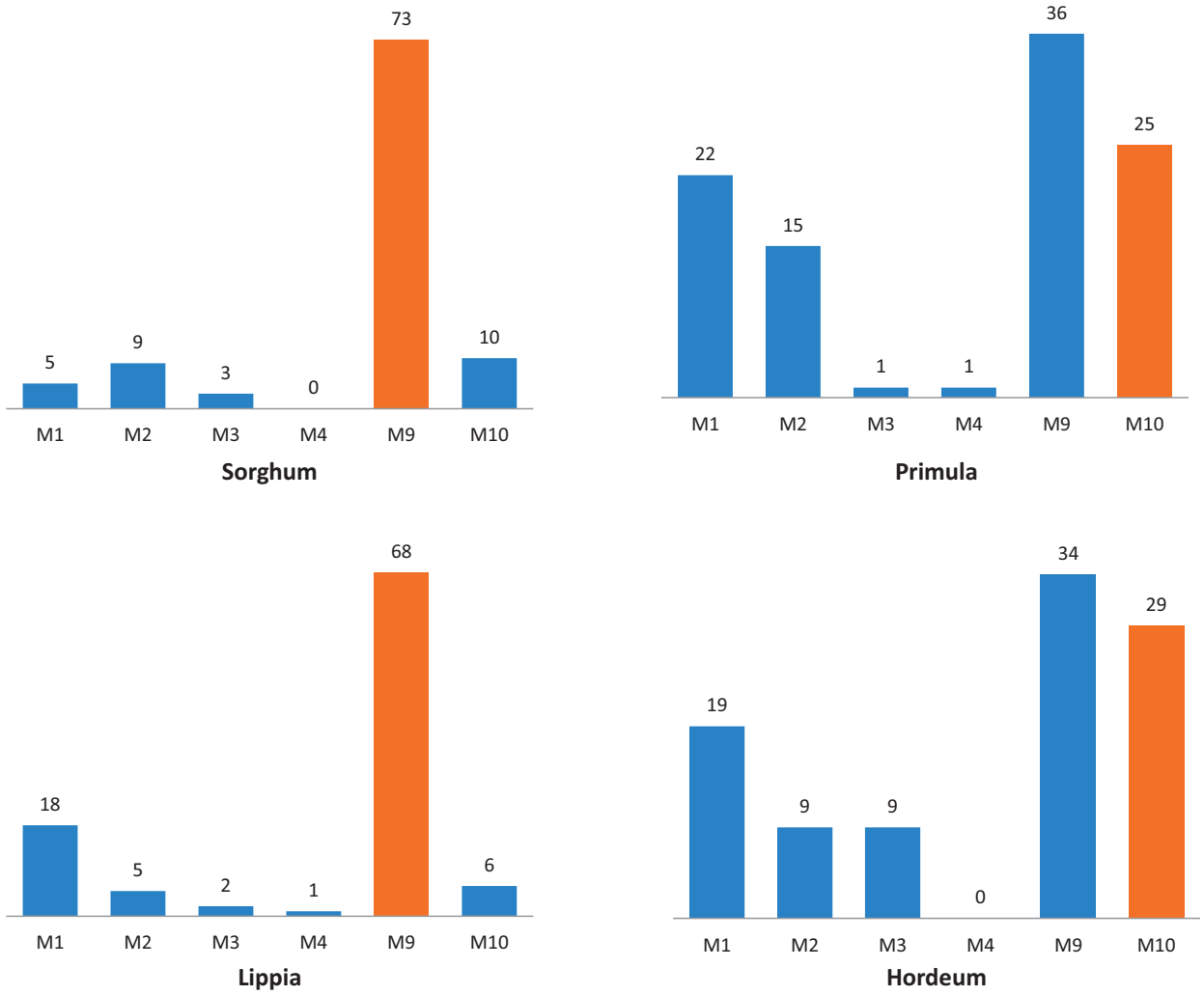

Fig. 1. Best-fitted model for simulated data sets. In this analysis, six models (as defined in table 2; models M5–M8 were omitted because these are the most computationally intensive and are rarely chosen as the best-fitted model as determined from empirical data sets; supplementary fig. S1, Supplementary Material online) were compared using the AIC.

is similar to those obtained for other model parameters (i.e., ρ, λ, and δ; supplementary table S4, Supplementary Material online).

Finally, these simulations allowed us to compare the fit of different chromEvol models and to examine whether the best-fitted model (as determined using the AIC) was the one that was actually simulated. As presented in figure 1, when M10 was simulated (and so a separate rate parameter for WGD was included in addition to other base number transitions that may occur at a different rate), a large fraction of simulated data sets best-supported model M9 (in which WGDs may occur only through base number transitions). In several such cases, this could be explained by the overlap between allowed base number transitions and WGDs, which made inclusion of the ρ parameter unnecessary. In other simulated data sets, dysploidy events were not followed by WGD events and so all polyploidization events were due to transitions by a common factor (representing the base number). Similarly, in a small fraction of the simulated data sets, M10 was best supported, whereas M9 was the one simulated. In these cases, a large number of base-number transitions practically resulted in exact duplication of the number of chromosomes, justifying the inclusion of the extra ρ parameter. In simulations where all base number transitions corresponded only to WGDs, the simple M1 model was preferred over the more complex M9/M10 models that were simulated.

Applying ChromEvol to a Large Number of Plant Groups

To examine the usability of the new models suggested herein, we have used chromEvol to analyze 100 plant genera (supplementary table S5, Supplementary Material online). Phylogenies and chromosome numbers data for each of these genera were acquired as described in supplementary materials, Supplementary Material online; for each genus, a single tree (the MAP tree from a Bayesian phylogenetic reconstruction) was used as the input for chromEvol. All genera analyzed included between 16 and 93 species, with chromosome numbers available for at least half of the species in the

Table 2. Summary of Chromosome Number Evolutionary Models, Their Respective AIC Scores, and Optimized Parameter Values as Inferred in the *Primula* Analysis.

Model	ΔAIC	Parameters							
		λ	λ_1	δ	δ_1	ρ	μ	ν	β
M1: CONST_RATE	31.674	46.3	—	50.2	—	0.80	—	—	—
M2: CONST_RATE_DEMI	9.465	0.39	—	0.04	—		0.60	—	—
M3: CONST_RATE_DEMI_EST	10.259	0.37	—	0.10	—	0.63	0.18	—	—
M4: CONST_RATE_NO_DUPL	33.686	78.7	—	79.7	—	—	—	—	—
M5: LINEAR_RATE	20.689	11.7	0.66	~0	3.06	2.02	—	—	—
M6: LINEAR_RATE_DEMI	20.224	10.0	0.09	~0	2.88		1.68	—	—
M7: LINEAR_RATE_DEMI_EST	22.353	8.9	—0.04	~0	2.94	1.52	2.80	—	—
M8: LINEAR_RATE_NO_DUPL	33.851	44.0	2.70	33.7	3.47	—	—	—	—
M9: BASE_NUM	2.743	0.4	—	0.06	—	—	—	0.16	9
M10: BASE_NUM_DUPL	0	0.32	—	0.054	—	0.17	—	0.11	9

tree. For each genus, the ten evolutionary models summarized in table 2 were fitted to the data, and the AIC was used to determine the model that best fits a particular data set. Results for this analysis are summarized in supplementary figure S1 and table S5, Supplementary Material online. In 71% of the data sets, models allowing for polyploidy events (WGD, demi-duplication or transition by base number) were preferred over models that allow for dysploidy transitions only, whereas very few data sets (2%) supported linear dependency of dysploidy rates on the current number of chromosomes. Models M9 and M10, which allow for chromosome number transitions by a base number, were chosen in 27% of the data sets. This supports the need for the inclusion of such models when analyzing complex patterns of chromosome number evolution using chromEvol. Interestingly, M9 was chosen approximately three times more often than M10, implying that in many cases modeling transitions by a base number alleviate the need to differently model WGD events. Finally, in 21 out of the 27 data sets where M9 or M10 were chosen, the best-supported model apart from these two included demi-polyploidy transitions. This observation suggests that modeling transitions by a base number mostly constitutes an alternative to the somewhat controversial demi-duplication event.

Illustrative Example: Primula Aleuritia Complex

Primula L. (Primulaceae) is a genus of flowering perennial plants comprising approximately 500 species and has a wide geographic distribution (Schmidt-Lebuhn et al. 2012). Species of *Primula* are known for their range of floral morphs (e.g., heterostyly) and for the frequent incidence of polyploidy and hybridization (Davies 1953; Mast and Conti 2006; Li et al. 2011). The arctic alpine Aleuritia complex (*Primula* sect. *Aleuritia* subsect *Aleuritia*) includes 21 described species, with wide karyological variability ($2n = 18, 22, 36, 54, 72$) with $x = 9$ being the hypothesized base number (Richards 2002). The high frequency of polyploidy in this group supposedly arose through secondary contact of differentiated diploid populations that reunited following glacier retreat (Guggisberg et al. 2009).

Here, we exemplify the use of chromEvol using a data set comprising 16 species from sections *Aleuritia* and *Armerina*. These two sections were shown to be placed in the same, largely unresolved, clade (Guggisberg et al. 2009). First, given the MAP tree (reconstructed using MrBayes [Ronquist and Huelsenbeck 2003] as detailed in supplementary materials, Supplementary Material online), the optimized likelihood values were determined for each of ten chromosome number evolutionary models (table 2). These models include eight models described previously (Mayrose et al. 2010) and two new models involving chromosome number transitions by a base number. The best-supported model based on this analysis was the one allowing for separate rate parameters for both WGDs and base-number transitions (model M10; table 2). This model inferred $\beta = 9$ as the most likely monoploid number as was suggested before for this genus (Richards 2002). Subsequently, the ploidy levels of 16 extant taxa were determined based on this model and based on the best-supported model out of those not allowing for base number transitions (model M2; table 2). Notably, the number of transitions inferred under the two models was markedly different (16 and 10 ploidy transitions inferred using model M2 and M10 respectively; supplementary table S3, Supplementary Material online). Similarly, these two models resulted in different ancestral chromosome numbers inferences (fig. 2). For example, using M10, the most probable chromosome number at the root was 9 (posterior probability of 0.9), followed by $x = 8$ ($P = 0.07$). However, using M2, a flatter distribution of probable chromosome numbers were obtained, with $x = 2$ being the most probable root state ($P = 0.26$) and $x = 9$ receiving very low support ($P = 0.02$). This discrepancy can be explained by the fact that under M2, multiple polyploidization events are needed to accommodate a single base-number transition (e.g., from 9 to 36), and thus high rates of chromosome number transitions were inferred under this model. Consequently, the root state inferred under M2 was much lower than that inferred under M10, because multiple transitions were inferred to occur along the two lineages descendent from the root when analyzed using M2, whereas no transitions were inferred using M10. Additionally, there were marked differences in the

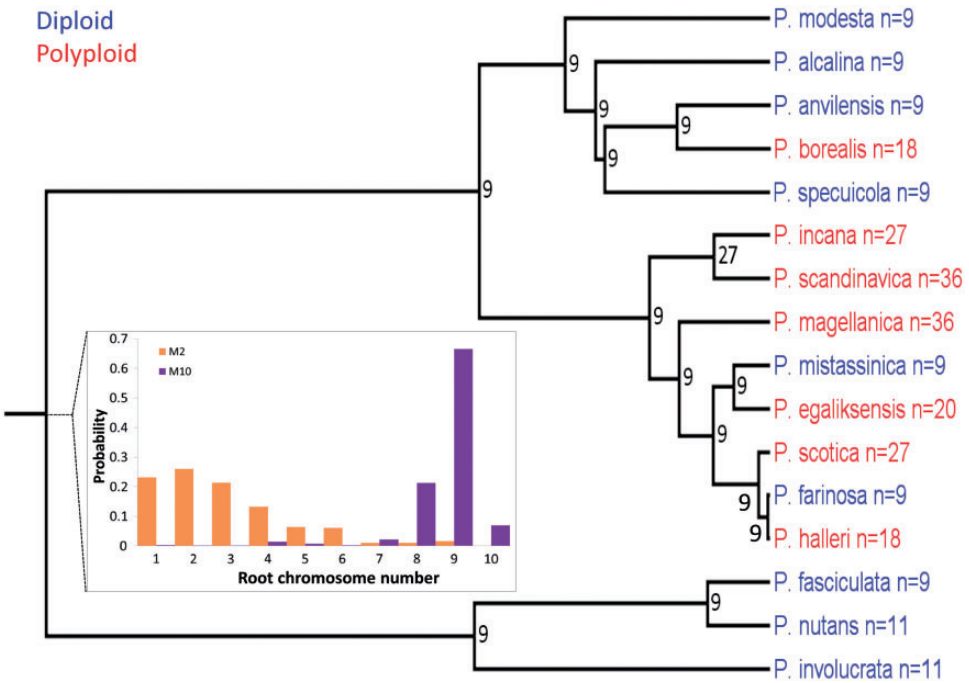

Fig. 2. ChromEvol inferences for *Primula Aleuritia* complex. Ploidy levels inferences for tip taxa based on model M10 are marked as blue (diploids) and red (polyploids). The ML ancestral chromosome numbers estimated using model M10 are given next to the internal nodes. The inset compares the inferred chromosome number distribution at the root using models M2 (orange) and M10 (purple).

inferred ploidy levels under the two models; using model M10, the ploidy-level inference pipeline resulted in nine diploid species and seven polyploid species (fig. 2), whereas the inference under model M2 was deemed unreliable (i.e., the inferred ploidy levels for all species was NA) suggesting model inadequacy.

Conclusion

The chromEvol software implements a variety of models depicting the evolution of chromosome numbers thereby providing users with a statistical tool to elucidate the pattern of chromosome number change along a phylogeny. In the new version of the program presented here, the base number of a group of interest can be inferred in a statistical framework. The concept of the base number is regularly taken to represent "the haploid number present in the initial population of a monophyletic clade" (Guerra 2008). Because this ancestral number is frequently, although not always, equal to the monoploid number of the polyploid series observed in the group, the two terms have been used somewhat intermingly. In chromEvol, we explicitly separated these two concepts. Although the ancestral root number can be computationally inferred or be set by the user, the addition by any multiplication of the monoploid number is now an integral part of the model using the introduced β parameter (eq. 2). Several complexities should be noted regarding this parameter. First, although a single monoploid number may define a large clade, it is also possible that, due to dysploidy, each subclade in an analyzed phylogeny will exhibit a unique polyploid series with the respective unique monoploid number (Guerra 2008). Here, we treated the monoploid number as a single possible value. Allowing for several distinct monoploid numbers can easily be integrated into our models but will come at the expense of additional free parameters that may be warrant only when large clades are considered. Second, in the current implementation, we treated additions by any multiplier of β as equally likely. Other scenarios whereby, for example, single β additions are more likely than higher multiplications can be integrated into the model (but again, increasing model complexity). Finally, although the β parameter was initially included in our models as a means to infer the base number of the group examined, we note that practically its optimized value is not necessarily so, particularly when additional parameters modeling other polyploidization events (i.e., ρ and μ) are considered. For example, in the *Hordeum* data set the best-supported model was M10 with $\beta = 14$ and a high rate of WGDs (table 1). Notably, most duplications involved transitions from 7 to 14 via the ρ parameter. Although seven is an obvious base number of this genus, the β parameter allowed for several $7 \rightarrow 21$ transitions, which otherwise must have been explained by multiple transitions.

The chromEvol program allows users to estimate the branches in which ploidy transitions most probably occurred. However, these estimates cannot trivially be used to determine which lineages are polyploid and which are diploid. To this end, we introduced a simulation-based approach to determine the ploidy levels of extant taxa. This approach can further be used to pinpoint the lineages in which ploidy estimates are deemed unreliable based on the current chromosome number distribution. Importantly, in the current implementation, chromEvol ignores possible association

between transitions in chromosome numbers and transitions in ploidy levels. For example, it is possible that due to diploidization processes (Wolfe 2001), the rate of descending dysploidy will be higher in a polyploid compared with a diploid background. Furthermore, polyploidy is known to have a profound impact on rates of diversification (Fawcett et al. 2009; Soltis et al. 2009; Mayrose et al. 2011), which could confound estimation of chromosome number transition rates and ancestral state reconstruction (Maddison 2006; Goldberg and Igic 2008). One may envision a covarion-like process (Galtier 2001), in which the evolution of chromosome numbers and ploidy levels is jointly modeled. Accordingly, rather than assuming a constant pattern of chromosome number change across the phylogeny, different lineages may evolve under different evolutionary patterns, dictated by their ploidy levels. Such a combined model could also be integrated within a Bayesian framework (i.e., by using a Markov chain Monte-Carlo sampling strategy; Hastings 1970), thereby accounting for uncertainty in parameter estimation and phylogeny reconstruction. Using this formulation, chromEvol may further be extended to allow chromosome number or ploidy levels to influence rates of speciation and extinction under the Binary State Speciation and Extinction (BiSSE) framework (Maddison et al. 2007). Such possible extensions will come at the expense of additional free parameters and modeling complexities that may only be justified when large trees are considered. However, such large trees may be particularly unrealistic for the time-homogeneity assumption (i.e., a single transition matrix across the whole phylogeny). Certainly, exploring the association between patterns of chromosome number change and ploidy levels as well as the range of complexities that can be explored using the chromEvol model are important future directions.

Supplementary Material

Supplementary materials, tables S1–S5, and figure S1 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

The authors are grateful to Sarah Otto for helpful discussions; to Shing Zhan for help in the implementation of the ploidy inference pipeline; and to two anonymous referees and the associate editor for many constructive comments. This study was supported by a Marie Curie Reintegration Grant and by Israel Science Foundation grant number 1265/12.

References

Akaike H. 1974. A new look at the statistical model identification. *IEEE Trans Autom Control.* 119:716–723.

Albertin W, Marullo P. 2012. Polyploidy in fungi: evolution after whole-genome duplication. *Proc Biol Sci.* 279:2497–2509.

Aury JM, Jaillon O, Duret L, Noel B, Jubin C, Porcel BM, Segurens B, Daubin V, Anthouard V, Aiach N, et al. 2006. Global trends of whole-genome duplications revealed by the ciliate *Paramecium tetraurelia*. *Nature* 444:171–178.

Cui L, Wall PK, Leebens-Mack JH, Lindsay BG, Soltis DE, Doyle JJ, Soltis PS, Carlson JE, Arumuganathan K, Barakat A, et al. 2006. Widespread genome duplications throughout the history of flowering plants. *Genome Res.* 16:738–749.

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". *Ann Bot.* 109: 681–692.

Davies EW. 1953. Polyploidy in *Primula farinosa* L. *Nature* 171:659–660.

Fawcett JA, Maere S, Van de Peer Y. 2009. Plants with double genomes might have had a better chance to survive the Cretaceous-Tertiary extinction event. *Proc Natl Acad Sci U S A.* 106:5737–5742.

Furlong RF, Holland PWH. 2004. Polyploidy in vertebrate ancestry: Ohno and beyond. *Biol J Linn Soc Lond.* 82:425–430.

Galtier N. 2001. Maximum-likelihood phylogenetic analysis under a covarion-like model. *Mol Biol Evol.* 18:866–873.

Goldberg EE, Igic B. 2008. On phylogenetic tests of irreversible evolution. *Evolution* 62:2727–2741.

Goldblatt P. 1980. Polyploidy in angiosperms: monocotyledons. In: Lewis WH, editor. Polyploidy: biological relevance. New York: Plenum Press. p. 219–239.

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

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

Guggisberg A, Mansion G, Conti E. 2009. Disentangling reticulate evolution in an arctic-alpine polyploid complex. *Syst Biol.* 58: 55–73.

Hastings WK. 1970. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* 57:97–109.

Kandul NP, Lukhtanov VA, Pierce NE. 2007. Karyotypic diversity and speciation in *Agrodiaetus* butterflies. *Evolution* 61:546–559.

Li J, Webster MA, Smith MC, Gilmartin PM. 2011. Floral heteromorphy in *Primula vulgaris*: progress towards isolation and characterization of the S locus. *Ann Bot.* 108:715–726.

Liu PL, Wan Q, Guo YP, Yang J, Rao GY. 2012. Phylogeny of the genus *Chrysanthemum* L.: evidence from single-copy nuclear gene and chloroplast DNA sequences. *PLoS One* 7:e48970.

Maddison WP. 2006. Confounding asymmetries in evolutionary diversification and character change. *Evolution* 60:1743–1746.

Maddison WP, Midford PE, Otto SP. 2007. Estimating a binary character's effect on speciation and extinction. *Syst Biol.* 56: 701–710.

Mast AR, Conti E. 2006. The primrose path to heterostyly. *New Phytol.* 171:439–442.

Matthews BW. 1975. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim Biophys Acta.* 405: 442–451.

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

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.

Ohi-toma T, Sugawara T, Murata H, Wanke S, Neinhuis C, Murata J. 2006. Molecular phylogeny of *Aristolochia* sensu lato (Aristolochiaceae) based on sequences of rbcL, matK, and phyA genes, with special reference to differentiation of chromosome numbers. *Syst Bot.* 31:481–492.

Peruzzi L. 2013. "x" is not a bias, but a number with real biological significance. *Plant Biosyst.* 147:1238–1241.

Ramsey J, Schemske DW. 1998. Pathways, mechanisms, and rates of polyploid formation in flowering plants. *Annu Rev Ecol Syst.* 29: 467–501.

Richards J. 2002. *Primula*. Vol. 346. London: BT Batsford. p. 40.

Ronquist F, Huelsenbeck JP. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* 19:1572–1574.

Scannell DR, Byrne KP, Gordon JL, Wong S, Wolfe KH. 2006. Multiple rounds of speciation associated with reciprocal gene loss in polyploid yeasts. *Nature* 440:341–345.

Schmidt-Lebuhn AN, de Vos JM, Keller B, Conti E. 2012. Phylogenetic analysis of *Primula* section *Primula* reveals rampant non-monophyly

among morphologically distinct species. *Mol Phylogenet Evol.* 65: 23–34.
Schultheis LM. 2001. Systematics of *Downingia* (Campanulaceae) based on molecular sequence data: implications for floral and chromosome evolution. *Syst Bot.* 26:603–621.
Soltis DE, Albert VA, Leebens-Mack J, Bell CD, Paterson AH, Zheng C, Sankoff D, dePamphilis CW, Wall PK, Soltis PS. 2009. Polyploidy and angiosperm diversification. *Am J Bot.* 96:336–348.
Stebbins GL. 1938. Cytological characteristics associated with the different growth habits in the dicotyledons. *Am J Bot.* 25: 189–198.
Taylor JS, Braasch I, Frickey T, Meyer A, Van de Peer Y. 2003. Genome duplication, a trait shared by 22000 species of ray-finned fish. *Genome Res.* 13:382–390.
Wolfe KH. 2001. Yesterday's polyploids and the mystery of diploidization. *Nat Rev Genet.* 2:333–341.
Wood TE, Takebayashi N, Barker MS, Mayrose I, Greenspoon PB, Rieseberg LH. 2009. The frequency of polyploid speciation in vascular plants. *Proc Natl Acad Sci U S A.* 106:13875–13879.
Zhan S, Glick L, Tsigenopoulos C, Otto S, Mayrose I. 2014. Comparative analysis reveals that polyploidy does not decelerate diversification in fish. *J Evol Biol.* 27:391–403.