Syst. Biol. 59(2):132–144, 2010

DOI:10.1093/sysbio/syp083
Advance Access publication on December 10, 2009

Probabilistic Models of Chromosome Number Evolution and the Inference of Polyploidy

ITAY MAYROSE[1,*], MICHAEL S. BARKER[2,3], AND SARAH P. OTTO[1]

[1]*Department of Zoology, University of British Columbia, Vancouver, BC V6T 1Z4, Canada;*
[2]*Department of Botany, University of British Columbia, Vancouver, BC V6T 1Z4, Canada; and*
[3]*Department of Biology, Indiana University, Bloomington, IN 47405, USA;*
[*]*Correspondence to be sent to: Department of Zoology, University of British Columbia, Vancouver, BC V6T 1Z4, Canada; E-mail: mayrose@zoology.ubc.ca.*

Received 31 March 2009; reviews returned 21 May 2009; accepted 27 October 2009
Associate Editor: Susanne S. Renner

Abstract.—Polyploidy, the genome wide duplication of chromosome number, is a key feature in eukaryote evolution. Polyploidy exists in diverse groups including animals, fungi, and invertebrates but is especially prevalent in plants with most, if not all, plant species having descended from a polyploidization event. Polyploids often differ markedly from their diploid progenitors in morphological, physiological, and life history characteristics as well as rates of adaptation. The altered characteristics displayed by polyploids may contribute to their success in novel ecological habitats. Clearly, a better understanding of the processes underlying changes in the number of chromosomes within genomes is a key goal in our understanding of speciation and adaptation for a wide range of families and genera. Despite the fundamental role of chromosome number change in eukaryotic evolution, probabilistic models describing the evolution of chromosome number along a phylogeny have not yet been formulated. We present a series of likelihood models, each representing a different hypothesis regarding the evolution of chromosome number along a given phylogeny. These models allow us to reconstruct ancestral chromosome numbers and to estimate the expected number of polyploidization events and single chromosome changes (dysploidy) that occurred along a phylogeny. We test, using simulations, the accuracy of this approach and its dependence on the number of taxa and tree length. We then demonstrate the application of the method for the study of chromosome number evolution in 4 plant genera: *Aristolochia*, *Carex*, *Passiflora*, and *Helianthus*. Considering the depth of the available cytological and phylogenetic data, formal models of chromosome number evolution are expected to advance significantly our understanding of the importance of polyploidy and dysploidy across different taxonomic groups. [Chromosome evolution; dysploidy; evolutionary models; genome duplication; polyploidy.]

Chromosome number is a remarkably dynamic feature of eukaryotic evolution. Chromosome numbers have repeatedly increased by doubling (polyploidy), increased by a single chromosome (ascending dysploidy via, e.g., chromosome fission), and decreased by a single chromosome (descending dysploidy via, e.g., chromosome fusion). Of these mechanisms of chromosome number change, polyploidy has received significant attention because the genomes of diverse eukaryotes, including animals (Furlong and Holland 2004), fungi (Kellis et al. 2004), and protozoa (Aury et al. 2006), contain evidence of past genome duplications. However, polyploidy reaches its zenith in plants, where 50–100% of flowering plants are believed to have a polyploid ancestry (Goldblatt 1980; Masterson 1994; Cui et al. 2006; Soltis et al. 2009) and 20–40% of extant flowering plant species thought to be recent polyploids (Stebbins 1971). Polyploids often differ markedly from their progenitors in morphological, physiological, or life history characteristics (Levin 1983; Ramsey and Schemske 2002), and these differences may contribute to the establishment and success of a polyploid species in novel ecological settings. It is thus hypothesized that polyploidy may serve as an important mechanism for ecological diversification, especially in harsh environments (reviewed in Otto 2007). Despite the fundamental role of chromosome number change in eukaryotic evolution, probabilistic models describing the evolution of chromosome number along a phylogeny have not yet been formulated. Considering the availability of cytological and phylogenetic data, formal models of chromosome number evolution could significantly advance our understanding of the importance of polyploidy and dysploidy in eukaryotic evolution.

The precise inference of polyploidy from cytological data, particularly ancient polyploidy, is often a challenging task. Polyploidy is demonstrated by the observation of multivalents for all or most chromosomes during meiosis, or when chromosome numbers among closely related species are entire multiplications of each other. These methods are mainly applicable to recent polyploids, rather than ancient ones, as genome rearrangements and changes in pairing behavior over time gradually erase clear signals of genome duplication. Genomic scans have also been used to search for evidence of past polyploidization by identifying regions of duplicated gene synteny (Vision et al. 2000; Bowers et al. 2003; Jaillon et al. 2007) or peaks in the age distribution of duplicated genes (Lynch and Conery 2000; Bowers et al. 2003; Blanc and Wolfe 2004; Cui et al. 2006; Barker et al. 2008). Although these genomic approaches are powerful, they are confined to species with wholly sequenced genomes or transcriptomes. Traditionally, polyploidy within a genus is estimated by evaluating its chromosome number distribution. Typically, the lowest haploid chromosome numbers for a genus are assumed to represent the nonpolyploidized state, termed the base chromosome number. Species with a chromosome number approximately twice the base number are considered polyploids. Different estimates vary with

regard to the way the base number is determined and the precise threshold employed (reviewed in Otto and Whitton 2000). For example, Stebbins (1938) considered a species polyploid if it has a haploid chromosome number, which is a multiple (or near multiple) of the lowest one found in the genus. Clearly, such methods can only infer recent polyploidization events and do not account for changes in chromosome numbers that are not due to polyploidy (i.e., dysploidy). More recently, chromosome numbers have been examined in an explicit phylogenetic framework. Generally, ancestral chromosome numbers are reconstructed using the maximum parsimony (MP) principle (Schultheis 2001; Baldwin et al. 2002; Mansion and Zeltner 2004; Guggisberg et al. 2006; Hansen et al. 2006; Ohi-toma et al. 2006; Timme et al. 2007). Based on this reconstruction, a certain lineage is inferred to be polyploid if its chromosome number is larger by a chosen factor than the base chromosome number.

Although MP has been widely used to infer ploidy levels and base chromosome numbers, there are significant drawbacks with this approach. First, all types of transitions are unrealistically assigned the same weight. In the MP framework, it is possible to a priori specify a weight for each transition. However, such weighting schemes are subjective as there is no way to determine whether one set of weights is more justified than another, and it is further not clear if polyploidy should be given the same weight as a gain or a loss of a single chromosome. Moreover, there is no clear way to specify these weights as there may be a number of different routes to transition from one chromosome number to another (e.g., does a transition from 5 to 13 chromosomes involve 1 or 2 polyploidizations?). Second, the MP method ignores the possibility of multiple or backward transitions, thus systematically underestimating the number of events. Third, the method ignores any uncertainty in the assignments of ancestral states. Thus, unobserved data on internal nodes are treated in exactly the same way as observed chromosome numbers in extant species, even though ancestral state estimates under MP are expected to have wide confidence limits (Maddison 1995). Fourth, there are often a number of equally parsimonious reconstructions, each of which may suggest different evolutionary scenarios. Finally, there is no way to objectively determine the accuracy of the reconstruction.

All these problems can be addressed by using a probabilistic approach with an explicit model for the changes in chromosome number over time. The evolutionary models developed here assume that such changes are the result of a combination of polyploidy and dysploidy events along branches of a phylogeny. The likelihood approach allows us to determine not only the probability of a given chromosome number at any internal node but also to gain insight into the evolutionary process itself. Model parameters are estimated from the data and can be compared across different taxonomic groups.

In this paper, we present a series of likelihood models, each representing a different hypothesis regarding the pathways by which the evolution of chromosome number proceeds. We then use this framework to reconstruct ancestral chromosome numbers, the base chromosome number of the group, and to estimate the expected number of polyploidy and dysploidy transitions that have occurred. We use simulations to assess the accuracy of our approach and test its dependence on the number of taxa examined and on the tree length. Using our new models, we then re-evaluate data from 4 plant genera for which chromosome number evolution has previously been analyzed: *Aristolochia* (Ohi-toma et al. 2006), *Carex* (Hipp 2007), *Passiflora* (Hansen et al. 2006), and *Helianthus* (Timme et al. 2007). Following standard convention, throughout this paper, we refer to x as the base chromosome number of a lineage, $2n$ as the chromosome number in somatic tissues, and n as the haploid chromosome number observed in gametes.

METHODS

A Likelihood-Based Model for Chromosome Evolution

Our models of chromosome number evolution are represented as a continuous time Markov process defined by the instantaneous rate matrix Q. The first model we consider assumes that in an infinitesimal time interval 3 types of events are possible: polyploidization, a gain of one chromosome (ascending dysploidy) or a loss of one chromosome (descending dysploidy). These factors are captured within the rate matrix Q, which describes the instantaneous rate of change from a genome with an i haploid chromosomes to a genome with j chromosomes. For $i \neq j$, we define:

$$Q_{ij} = \begin{cases} \lambda & j = i+1, \\ \delta & j = i-1, \\ \rho & j = 2i, \\ 0 & \text{otherwise}, \end{cases} \tag{1}$$

where λ, δ, and ρ are the rates of gains, losses, and polyploidizations, respectively. The diagonal elements are determined by the constraint that each row in Q sums to zero:

$$Q_{ii} = -\sum_{i \neq j} Q_{ij}. \tag{2}$$

Hereafter, the model parameters are collectively referred to as θ.

Given the Q matrix, transition probabilities from state i to state j along a branch of length t can be computed by matrix exponentiation:

$$P_{ij}(t) = e^{Qt} = \sum_{m=0}^{\infty} \frac{(Qt)^m}{m!}. \tag{3}$$

For each value of t, we sum the series until the addition of another term does not alter the resulting matrix. The squaring and scaling method (Moler and Van Loan 2003) is used to increase the accuracy of the approximation in cases where $\|Qt\|$ is large.

It is emphasized that in our model, time reversibility is not assumed, and hence, the results are conditional on the location of the root. Furthermore, the model relies on a given phylogenetic tree, defined by the tree topology with its associated branch lengths. For succinctness, we assume throughout that the tree is fixed and is part of the data, and we do not refer to it explicitly in the equations. Note that in our model, the rate matrix is not normalized (i.e., the average rate of change does not equal one). Thus, branch lengths may be in any unit proportional to time, with the units of Q being inversely proportional to this time unit. For example, if branch lengths in the given phylogeny represent average number of nucleotide substitutions, then Q will be the rate of events per nucleotide substitution per site. In this way, branch lengths from amino acid or nucleotide substitution units can be transformed to chromosome number transition units by a factor that determines the ratio between these 2 units. This factor is assumed to be uniform across the topology.

Theoretically, the haploid number of chromosomes within a genome can be any positive integer value. However, to simplify likelihood calculations, we allow a total of C states, corresponding to $\{1, 2, \ldots, C-1, \geq C\}$ chromosomes. The last state represents all chromosome numbers equal to or greater than C. We set C to be large enough that the probability of it being reached is extremely small, so that the error due to truncation is negligible. Practically, C is set to be 10 units larger than the maximal chromosome number observed in the analyzed family. Larger values of C (such as twice as large as the maximal chromosome number) gave identical results.

Likelihood calculations also require the assignment of root frequencies. A usual choice in likelihood calculation would be to assume a stationary distribution at the root (Yang 2006). However, for most Q matrices examined here, the stationary distribution has all its mass at the last state C, meaning that the process is still a transient one from an initially small genome. Thus, we weigh each root state, R, according to its probability of giving rise to the extant data, D, given the model parameters, θ (FitzJohn et al. 2009; see Discussion section). Hence, π_i, the probability that the state at the root is i, is given by the likelihood of the data given that the root is in state i divided by the sum of the likelihoods over all states:

$$\pi_i = \frac{P(D|R=i,\theta)}{\sum_{j=1}^{C} P(D|R=j,\theta)}, \quad (4)$$

where $P(D|R=i,\theta)$ is the likelihood of the observed data given that the root is in state i and the estimated model parameters.

The overall likelihood of the tree is then:

$$L = P(D|\theta) = \sum_{i=1}^{C} \pi_i P(D|R=i,\theta). \quad (5)$$

Given a rooted phylogenetic tree and given the assignment of chromosome numbers to extant species, the likelihood of the data can be calculated using the pruning algorithm of Felsenstein (1981). The model parameters were estimated by maximum likelihood (ML) using Brent's optimization scheme (Press et al. 2002). To avoid getting caught at local maxima, 10 random starting points were used during the optimization process.

Model Variations

Demi-polyploidization.—The model presented above assumes that the number of chromosomes might increase or decrease by one or might double; we refer to this model as M1 (Table 1). However, within a polyploid population, the union of reduced and unreduced gametes may generate a new cytotype of higher ploidy. For example, auto-hexaploids may be produced from a tetraploid by the union of $2x$ (reduced) and $4x$ (unreduced) gametes (Ramsey and Schemske 1998). New odd-ploidy cytotypes could also be produced by this mechanism. For example, it has been suggested that unreduced gamete production in hexaploid *Andropogon gerardii* generated a $9x$ cytotype (Norrmann et al. 1997). Hexaploids may also be formed via alloploidy from hybridization of closely related taxa of $2x$ and $4x$ cytotypes. This process may generate intermediates of odd ploidy (e.g., $3x$), which subsequently produce new even ploidy $6x$ cytotypes (Ramsey and Schemske 1998). We refer to these phenomena as "demi"-polyploidization. Under this scenario, the Q matrix takes the following form:

$$Q_{ij} = \begin{cases} \lambda & j = i+1, \\ \delta & j = i-1, \\ \rho & j = 2i, \\ \mu & j = 1.5i, \\ 0 & \text{otherwise}, \end{cases} \quad (6)$$

where μ is the demi-polyploidization rate. The Q matrix above is valid only for even values of i. For odd values, we set $Q_{ij} = \mu/2$ for the 2 possible adjacent integer values of j ($j = 1.5i$ rounded up and down). The Q matrix of Equation 6 leads to 2 additional models. In Model

TABLE 1. Various models for chromosome number evolution

Model	Number of p[a]	Parameters	Possible events and model assumptions
M0	2	λ, δ	Gains and losses
M1	3	λ, δ, ρ	Gains, losses, and polyploidy
M2	3	λ, δ, ρ	Gains, losses, demi-polyploidy, and polyploidy Rates of polyploidy and demi-polyploidy are equal
M3	4	λ, δ, μ, ρ	Gains, losses, demi-polyploidy, and polyploidy
M4	5	λ, δ, λ_1, δ_1, ρ	Gains, losses, and polyploidy Rates of gain and loss depend linearly on the current chromosome number
M5	4	λ, δ, λ_1, δ_1	Gains and losses Rates depend linearly on the current chromosome number

[a]Free model parameters.

M2, we assume that the rate of demi-polyploidization is equal to that of polyploidization (μ is set to equal ρ), resulting in a model with 3 free parameters: $\theta = \{\lambda, \delta, \rho\}$. In Model M3, μ is treated as an additional free parameter ($\theta = \{\lambda, \delta, \rho, \mu\}$). Finally, Models M2 and M3 allow for a triplication of the chromosome number via a demi-polyploidization event followed by whole genome duplication.

Linear dependency on the current number of chromosomes.—Here, we allow for the possibility that the rates of ascending and descending dysploidy depend on the current number of chromosomes. This dependency results in Model M4 with 5 free parameters:

$$Q_{ij} = \begin{cases} \lambda + \lambda_l \times (i-1) & j = i+1, \\ \lambda + \delta_l \times (i-1) & j = i-1, \\ \rho & j = 2i, \\ 0 & \text{otherwise}, \end{cases} \quad (7)$$

λ_l and δ_l are the linear components of the gain and loss rate, respectively, whereas λ and δ are the corresponding constant factors. The terms λ_l and δ_l are multiplied by $(i-1)$ so that the rates of ascending and descending dysploidy are λ and δ, respectively, for species with one chromosome ($i = 1$). In Models M1–M3, all parameters were constrained to be nonnegative for the Q matrix to be legitimate. For the linear model, we require $\lambda + \lambda_l \times (i-1) \geq 0$ and $\delta + \delta_l \times (i-1) \geq 0$ for $0 \leq i \leq C$. It is thus possible for λ_l to be negative, implying that as the number of chromosomes increases the probability of gaining an additional chromosome decreases. We note that the possibility of demi-polyploidy can also be integrated into the above model but because none of the data sets tested supported this extension, we do not refer to this option in the manuscript.

Null models.—As noted in the Introduction section, one of our main aims in constructing models for the evolution of chromosome number was to infer the role of polyploidization in a given clade. A null model, which does not permit polyploidization, allows statistical testing of the hypothesis that the observed chromosome number distribution resulted solely from chromosome gains and losses. This can be modeled by assuming $\rho = 0$ (and $\mu = 0$). As such, Model M0 is treated as the null model for M1–M3:

$$Q_{ij} = \begin{cases} \lambda & j = i+1, \\ \delta & j = i-1, \\ 0 & \text{otherwise}, \end{cases} \quad (8)$$

whereas M5 is the null model for M4:

$$Q_{ij} = \begin{cases} \lambda + \lambda_l \times (i-1) & j = i+1, \\ \delta + \delta_l \times (i-1) & j = i-1, \\ 0 & \text{otherwise}. \end{cases} \quad (9)$$

Model Comparison

Table 1 summarizes the assumptions and free parameters of all models presented above. As can be seen, not all models are nested within others (e.g., M2 cannot be obtained from M4). Thus, we focus on using the Akaike information criterion (AIC) (Akaike 1974) to determine the model that best fits a particular data set. In addition to general model exploration, where we wish to test whether there is significant evidence for polyploidization, the likelihood ratio test (LRT) can be used to compare nested models with and without the possibility of polyploidy (i.e., M0 vs. M1 or M2; M5 vs. M4). Because the null models represent boundary conditions (setting ρ and μ to 0), the asymptotic distribution of the LRT statistic, $2\Delta \log L$, is approximately distributed according to $\frac{\chi_1^2 + \chi_0^2}{2}$ (Ota et al. 2000).

Inferring Ancestral States

Two methods were used to reconstruct the chromosome numbers at ancestral nodes. First, the joint ML reconstruction of ancestral states was inferred using the dynamic algorithm of Pupko et al. (2000). This method infers the set of ancestral chromosome numbers that maximize the likelihood of the data (the assignment of chromosome numbers to extant taxa) given the phylogeny. Second, a Bayesian approach similar to that of Koshi and Goldstein (1996) was used to obtain, for any ancestral node V, the probability of each chromosome number occurring at that node:

$$P(V = x|D, \theta) = \sum_{y=1}^{C} P(V = x, F(V) = y|D, \theta), \quad (10)$$

where $P(V = x, F(V) = y|D, \theta)$ is the joint probability that the chromosome number assignment at node V is x and the assignment at the parent of node V (closer to the root) is y, given the data and model parameters. Because our model is irreversible, likelihood calculations depend on the assignment at the root. See Stern et al. (2010) for a description on how to obtain $P(V = x, F(V) = y|D, \theta)$ for irreversible models.

Calculating the Expected Number of Events

The likelihood framework allows us to compute the expected number of events along each branch of the phylogeny. Here, we consider 4 types of events: polyploidization, demi-polyploidization, ascending dysploidy, and descending dysploidy. In general, the expected number of transitions from state u to state v along a certain branch can be calculated using the following formula:

$$E(N_{uv}(AB)|D, \theta) = \sum_{y=1}^{C} \sum_{z=1}^{C} P(A = y, B = z|D, \theta) \times E(N_{uv}(AB)|A = y, B = z, \theta), \quad (11)$$

where AB is a branch that starts at node A and ends at node B, and the double summation is taken over all possible state combinations at both branch terminals. The left term of the summation (the joint probability of observing states at the tips of a branch given the data) is calculated as in Equation 10. The right term of the summation, the expected number of transitions given the terminal states at the tips of a branch, was calculated using simulations (see Stern et al. 2010, for details).

To calculate the number of events of a certain type, we sum over all relevant transitions. For example, to calculate the expected number of polyploidization events along branch AB we sum over all u,v, such that $v = 2u$:

$$E(N_{uv}(AB)|D,\theta) = \sum_{u\in\{1,\ldots,C\},v=2u} E(N_{uv}(AB)|D,\theta). \quad (12)$$

The computation is similar for the expected number of gains ($v = u + 1$), losses ($v = u - 1$), and demi-polyploidization events ($v = 1.5u$). The total number of transitions of a certain type throughout the phylogeny is summed over all branches:

$$E(N_{uv}(\text{tree})|D,\theta) = \sum_{\text{branch}\in\text{tree}} (E(N_{uv}(\text{branch})|D,\theta)) \quad (13)$$

Assessing Accuracy via Simulations

Simulations were used to investigate the accuracy of our method with regard to 3 attributes: 1) ancestral state reconstruction, 2) inference of model parameters, and 3) estimation of the expected number of transitions of a certain type. Simulated data were prepared by modeling the evolutionary process given a fixed tree and a given set of model parameters. The simulations were performed using the embedded-discrete-time Markov chain of the rate matrix Q (i.e., the waiting time at state i is exponentially distributed with rate $-Q_{ii}$ and given that a change has occurred, the probability to jump to state j is $-Q_{ij}/Q_{ii}$). In this way, we recorded, for each tree node, the simulated ("true") state and the evolutionary path leading to it from its parent node (e.g., the number of polyploidizations along that branch). The resulting chromosome numbers at the tips of the tree were used as the data input to our method. Model parameters, ancestral states, and expected number of transitions were then inferred and compared with the simulated values.

Random trees were generated according to a birth–death process using the Mesquite program (Maddison W.P. and Maddison D.R. 2008) with default parameters (speciation rate 0.3 and extinction rate 0.1). Simulation scenarios varied with respect to the number of taxa available (10, 20, 30, 40, 50, and 60) and the evolutionary distance from the root to the tips (0.05, 0.1, 0.2, 0.4, and 0.8). All simulations were conducted under Model M1 with parameters $\theta = \{\lambda = 1, \delta = 1, \rho = 1\}$ and M2: $\theta = \{\lambda = 1, \delta = 1, \rho = 1, \mu = 1\}$. For each simulated scenario, 100 independent runs were conducted. The results obtained under the M1 and M2 models were qualitatively similar, and we thus report only those obtained under M1. To verify that our simulations are not biased because of the arbitrary way parameters and trees were generated, simulations were also performed based on empirical trees and parameter values. The results of these sets of simulations largely agree with those of the random trees (see supplementary materials available from http://www.sysbio.oxfordjournals.org).

Implementation and Availability

The models and inference methods described here were implemented in C++. The program and source codes are available at http://www.zoology.ubc.ca/prog/chromEvol.html. 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 format. The program allows users to run all models at once or to specify a single model. For each model, the parameters are estimated, ancestral states are inferred using the ML and Bayesian methods, and an estimation of the number of events along each branch is given.

RESULTS

Assessing Accuracy Using Simulations

Accuracy of ancestral reconstruction.—Simulations were used to investigate the accuracy of our method in reconstructing the true ancestral chromosome numbers as a function of the number of taxa available and with different tree lengths, t (defined as the distance from the root to the tips). We first focus on the ML reconstruction of the most likely set of ancestral chromosome numbers. Figure 1a shows the absolute difference between the true chromosome number and the ML inferred value at the root of the tree as a function of the number of available taxa. As expected, accuracy increases when more taxa are available. In addition, when the evolutionary distance between taxa is short, the reconstruction is very accurate; in other words, when few events occur related species tend to be in the same state and reconstructions are easier. It is expected that for trees with larger evolutionary distances accuracy will decline. Indeed, for the longest trees examined (root to tip distance 0.8), the average difference between the true chromosome number at the root and the inferred one was 1.98 for 10 taxa and 0.75 for 60 taxa; these trees corresponded to an average of 10 and 34 transitions of the 3 types allowed (polyploidization, ascending dysploidy, or descending dysploidy). Thus, fairly accurate reconstructions are obtained for even the longest trees examined. Figure 1b presents the accuracy of the ML reconstruction for all ancestral nodes of the tree, including the root. Noticeably, the average error associated with all internal nodes is much smaller than that of the root simply because of the increased certainty nearer the present. In all simulations considered, the average difference between the true and reconstructed ancestral nodes was less than 0.85.

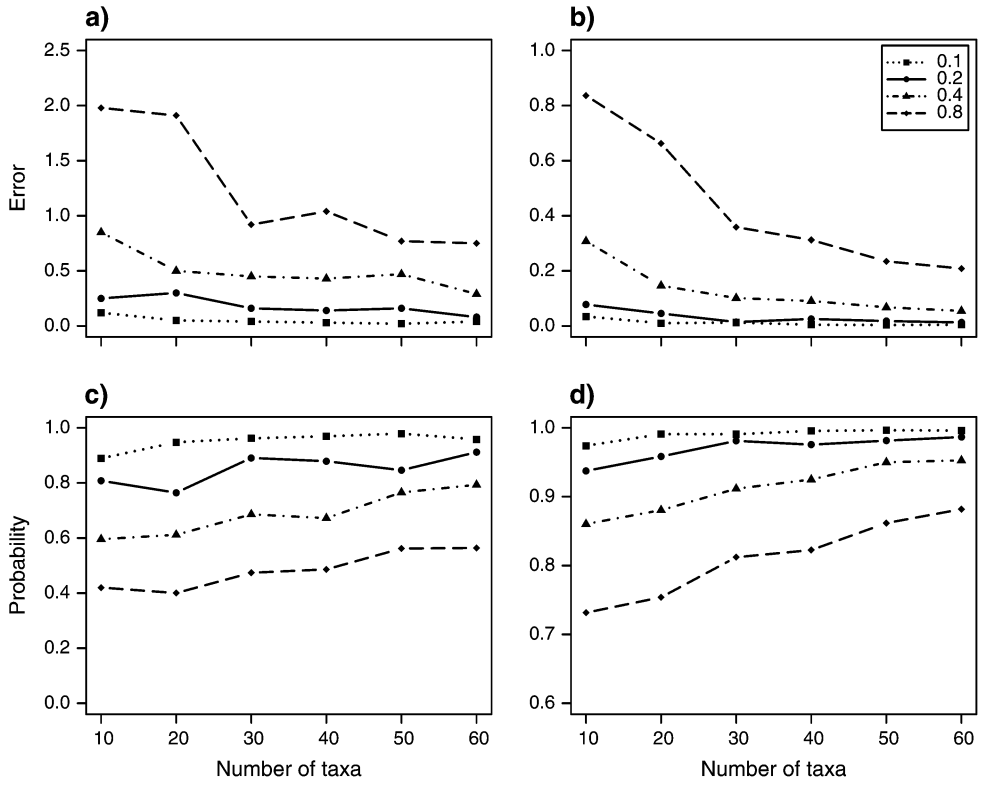

FIGURE 1. Accuracy of ancestral state reconstruction. Error of ML reconstruction of the chromosome number at the a) root only and b) all ancestral nodes (error is computed as the absolute difference between the true chromosome number and the ML inferred value). The probability of obtaining the true chromosome number at the c) root only and d) all ancestral nodes. Several different tree lengths (as measured by the distance from root to tips) are plotted as indicated in the legend inside panel b.

Our reconstruction method allows us to estimate not only the most likely chromosome number existing at an ancestral node but also the probability of each possible chromosome number at that node. Figure 1c presents the average probability assigned to the true ancestral chromosome number at the root. For all simulation conditions, the probability of the correct root chromosome number was significant and higher than 0.5 for all but the longest trees with few taxa ($t = 0.8$, 10–30 taxa). Again, the average accuracy for all ancestral nodes was higher than that for the root node only. Noticeably, the average probability of the true chromosome numbers is higher than 0.7 for all simulations examined (Fig. 1d). As may be expected, the standard deviation (SD) around these average probabilities decreased with inference accuracy, especially when shorter trees are simulated. For example, whereas the SD for the root node was 0.45 for $t = 0.8$ and 10 taxa, it decreased to 0.17 for $t = 0.05$. The SD for all simulations is given in supplementary materials.

Accuracy of parameter estimation.—Besides reconstructing ancestral states, it is useful to examine the accuracy of the method in estimating the model parameters. Figure 2 presents the average and standard error for the inferred rate of polyploidy, ρ, for various numbers of taxa and tree lengths. As can be seen, the average inferred rate was generally close to the simulated value. For shorter trees (length from root to tips 0.05 or 0.1), the variance of the estimated parameter was relatively large. This was particularly true when few taxa were available. For example, with 10 taxa and a tree length of 0.05, either zero or one genome duplication events occurred in the simulations, with an average number of duplications of 0.2. The estimated rate of polyploidy then either approached zero or rose above 4. In such cases, estimating the true value of the parameter cannot be expected, and it is more informative to test inference accuracy with regard to the number of transitions that took place (see below). Estimating polyploidization rate was particularly accurate when several ploidy events have occurred. As can be seen, for longer trees (or shorter ones with a large number of taxa), the average inferred ρ parameter was very close to the simulated one and had a small standard error.

Accuracy of estimating the number of events.—As detailed in the Methods section, we can infer the expected number of each transition type along every branch of the tree. Here, we focus on the expected number of transitions across the whole phylogeny and compare this to the total number of transitions that took place in the simulations. As can be seen in Figure 3a, for all but the very largest trees, the true number of duplication events

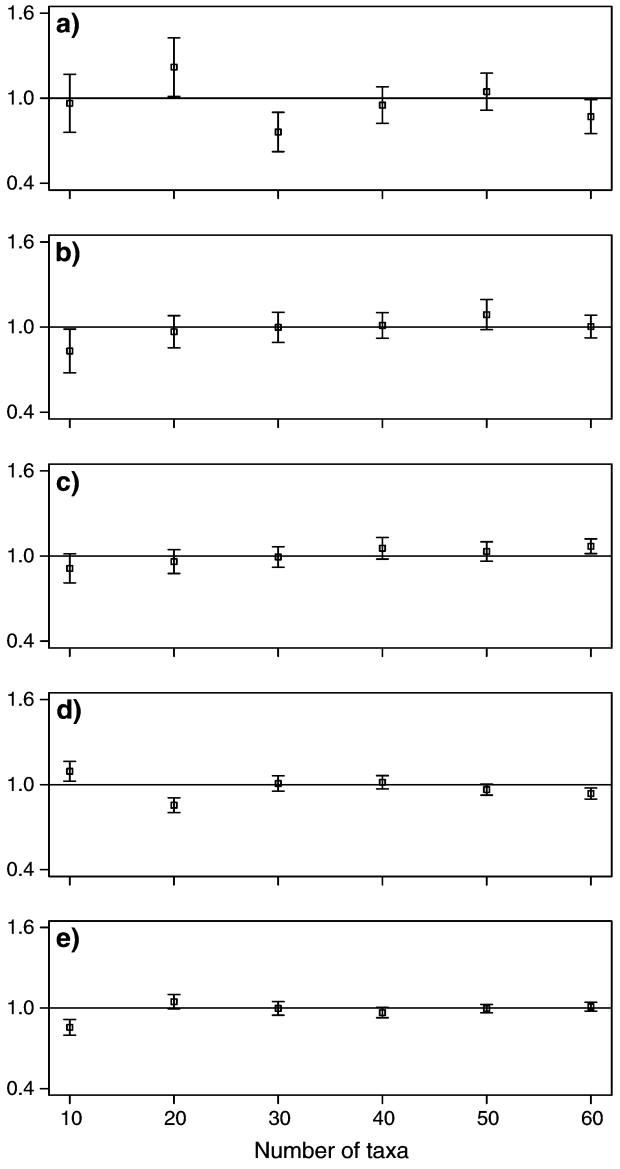

FIGURE 2. Accuracy of the inferred polyploidization rate parameter ρ. The average value of ρ and the standard error are shown as a function of the number of simulated taxa for different tree lengths: a) 0.05, b) 0.1, c) 0.2, d) 0.4, and e) 0.8. The true parameter value in the simulations was $\rho = 1$ (solid line).

can accurately be inferred. When branch lengths were short, our method precisely inferred the true number of duplications. For example, simulating 60 taxa with root to tip length of 0.05, the absolute difference between the inferred numbers of duplications and the simulated number of duplications was less than 0.18 for all 100 simulation runs (with 0, 1, 2, 3, and 4 duplication events occurring in 57, 29, 11, 1, and 2 runs).

Inference inaccuracies were more evident for longer trees with few taxa. When the simulations included 10 taxa with root to tip length of 0.8, the method tended to underestimate the true number of duplications (average difference between the inferred and the simulated number of duplications -0.25). In such a scenario, the tree has a few long branches leading to homoplasy due to polyploidizations on multiple branches. This, in turn, leads to higher inferred chromosome numbers at ancestral nodes and to a lower overall number of estimated transitions. When more taxa are available, the prediction becomes more accurate as convergent evolution events can be more readily detected.

It is expected that the difference between the inferred and the true number of events will increase with the number of events that truly happened (e.g., detecting all polyploidization events is easier when only one event truly happened compared with 10). To account for this bias, we also computed the relative error of prediction, defined as

$$\frac{|N_{PP} - \hat{N}_{PP}|}{\rho \sum_{b \in \text{tree}} L_b}, \tag{14}$$

where N_{PP} and $\hat{N}_{PP}$ are the true and estimated number of polyploidization events, L_b is the length of branch b, and the sum is over all branches of the tree multiplied by the true polyploidization rate. The denominator represents the expected number of events along a given tree and is used instead of the true number (N_{pp}) due to the occurrence of zero events in some simulation runs. The relative error for the number of gains and losses was similarly computed. Figure 3b presents the relative error of the number of polyploidizations versus the total tree length for different numbers of simulated taxa. As can be seen, the relative error of prediction was similar for different tree lengths and for different numbers of taxa. Two exceptions are noticeable. First, for short trees, the relative errors were exceptionally small. This is because few events of any type occurred. Second, as may be expected, the relative error was larger for trees with only 10 taxa. In such cases, the relative error tended to increase with the tree length. As noted above, this is the result of homoplasy and multiple transitions taking place on long branches. Inferring the correct number of transitions is particularly challenging in such cases.

Figure 3c,d present the relative error when estimating the number of dysploidy transitions. In general, the error in the estimation of the number of polyploidization events was smaller than that of ascending or descending dysploidy. This is expected as the influence of a polyploidization event is large compared with a change of a single chromosome and can thus be more easily detected.

Biological Data Sets Analyzed

We exemplify the use of the probabilistic method to re-evaluate data from 4 plant genera for which chromosome number evolution has previously been analyzed. Table 2 lists the number of taxa, the total tree length, and the range of chromosome numbers observed in each data set. Table 3 presents the AIC score of each data set under the 6 chromosome evolution models developed here. The ML ancestral reconstructions under the best fitting model are given in supplementary materials.

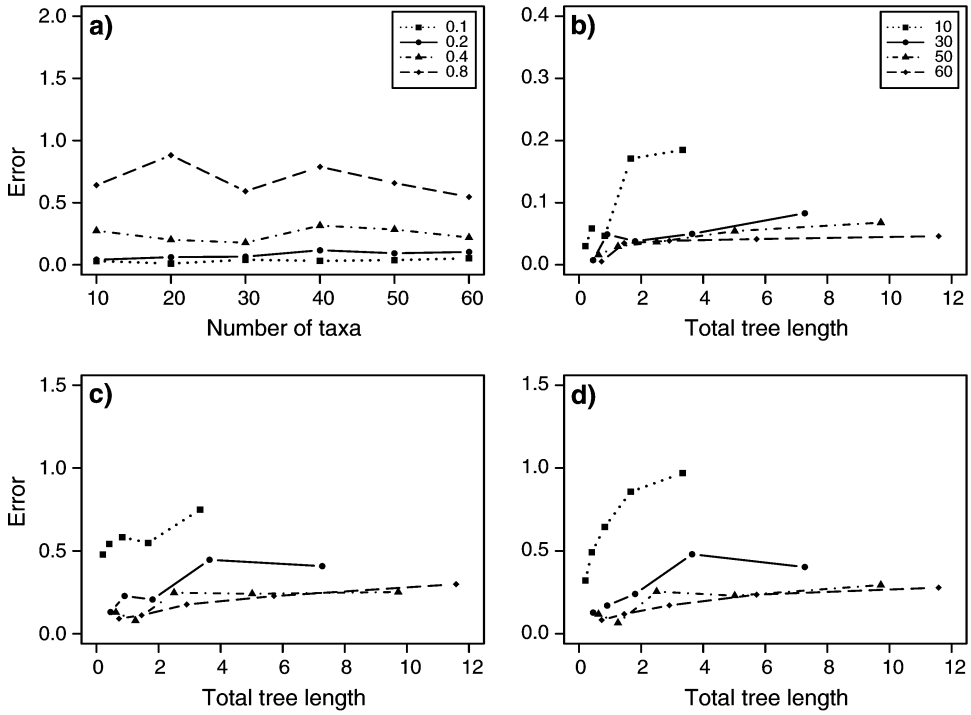

FIGURE 3. Accuracy of the inferred number of transitions. a) The average absolute difference between the estimated number of polyploidization events and the true number as a function of the number of taxa for different tree lengths as specified in the legend. b–d) The relative error of inference, as defined in Equation 14, for the number of b) polyploidizations, c) single chromosome gains, and d) single chromosome losses is plotted as a function of the total tree length (sum over all branches). From left to right, the 5 dots along the line correspond to root to tip length of 0.05, 0.1, 0.2, 0.4, and 0.8. The legend inside panel b specifies different number of simulated taxa and applies also to panels c and d.

Aristolochia.—The genus *Aristolochia s.l.* contains more than 400 species from warm temperate to tropical regions worldwide. Ohi-Toma et al. (2006) presented a molecular phylogeny of the genus based on nucleotide sequences of the chloroplast *rbcL* and *matK* genes (TreeBase study accession S1531). The study of Ohi-toma et al. (2006) also includes chromosome numbers for 15 and 78 species present in the *rbcL* and *matK* phylogenies, respectively. Their results indicated that the genus can be divided into 2 subtribes: Aristolochiinae and Isotrematinae. In their study, 5 different chromosome numbers were observed ($n = 3$, 6, 7, 8, and 16), which were predominantly congruent with the phylogeny: the Aristolochiinae clade shows chromosome numbers of $n = 3$, 6, 7, or 8, whereas the Isotrematinae clade is characterized by $n = 16$.

We used our models to analyze the distribution of chromosome numbers in the genus using the *rbcL* and *matK* gene trees. Despite the difference in the amount of data present in the 2 gene trees, the results are highly congruent. In both cases, M1 is the best supported model (Table 3). The data strongly support the occurrence of polyploidization in this genus with large differences in AIC scores between the models with and without polyploidization (Δ AIC between M1 and M0 of 31.4 and 11.4 for *matK* and *rbcL* gene trees, respectively; nonparametric bootstrap p value <0.005). In contrast, our analysis provides no support for the occurrence of demi-polyploidization events with the estimated demi-polyploidization rate in Model M3 approaching zero. Ancestral state reconstruction strongly supports $n = 8$ at the root of *Aristolochia s.l.* with probability of 0.98, followed by a decrease of one chromosome along the branch leading to Aristolochiinae (probability of 0.97 for $n = 7$ in both gene trees). The model predicts one polyploidization event in the genus (expected numbers of polyploidizations across the whole *matK* and *rbcL* gene trees were 1.01 and 1.06, respectively). This polyploidization most certainly occurred along the branch leading from the root to the Isotrematinae clade (expected number of polyploidizations along this branch was >0.97; probability for $n=16$ was >0.95 at the base of the Isotrematinae clade). This polyploidization event is also supported by the existence of 2 paralogous *phyA* gene copies in several species belonging to the Isotrematinae clade but a single *phyA* gene copy in species of the Aristolochiinae clade (Ohi-toma et al.

TABLE 2. Biological data sets examined

Data set	Number of taxa	Total tree length	Chromosome numbers range
Aristolochia			
matk	78	0.54	3–16
rbcL	15	0.11	3–16
Carex	53	16.3	26–43
Passiflora	58	0.61	6–12
Helianthus	107	0.86	17–51

TABLE 3. AIC scores and parameter estimates for the biological data sets examined[a]

Data set	M0	M1	M2	M3	M4	M5	Best model parameters
Aristolochia							
matk	105.5	**74.1**	75.5	76.1	76.1	104.8	M1 : $\lambda = 2.2, \delta = 14.4, \rho = 1.9$
rbcL	73.0	**61.6**	62.8	63.6	63.1	75.9	M1 : $\lambda = 16, \delta = 84, \rho = 10$
Carex	309.6	311.6	311.6	313.6	306.6	**304.6**	M5 : $\lambda = 70, \delta = 52.4, \lambda_l = -0.8, \delta_l = 1.2$
Passiflora	**59.3**	61.3	**59.7**	61.0	65.1	63.1	M0 : $\lambda = 0, \delta = 23.4$
							M2 : $\lambda = 1.7, \delta = 0, \rho = \mu = 5.0$
Helianthus	1158	637	**208**	210	633	1162	M2 : $\lambda = 0, \delta = 0, \rho = \mu = 14.5$

[a]Bold type indicates model with best AIC score.

2006). The model further predicts that across the whole tree chromosome number more often decreases than increases (~1.5 expected number of single chromosome increases vs. ~8 decreases).

Carex section ovales.—The sedge genus *Carex* L. is one of the most species rich of angiosperm genera, with 2000 species worldwide. *Carex* have unusually diverse chromosome numbers with every number from $n = 6$ to $n = 47$ represented by at least one species, with several counts as high as $n = 66$ (Hipp 2007). Sedges are characterized by holocentric chromosomes that lack localized centromeres. As a consequence, chromosome fragments that arise segregate normally and result in viable gametes that may become stabilized through backcrossing or selfing. Thus, fission and fusion events are very common in this group, whereas polyploidization is thought to be rare (reviewed in Hipp et al. 2009).

Here, we investigated chromosome evolution in *Carex* section *Ovales*, the most species-rich section of the genus *Carex* in the New World using the phylogeny and chromosome data presented in Hipp (2007). For all model variants, the inferred polyploidization rate approaches zero, with high rates of dysploidy transitions. Accordingly, in all comparisons (M0 vs. M1, M2, or M3; M5 vs. M4), the null hypothesis of no polyploidization cannot be rejected. According to the AIC selection criterion, M5 is the best supported model, indicating that the rates of gains and losses are dependent on the current chromosome number. Our ancestral reconstruction predicts that the likely ancestral chromosome number of this section is $x = 61$ (with $C = 63$ being the largest chromosome number allowed), with probability of 0.98 that it is 50 or higher. Notably, this number of chromosomes is higher than those observed in this clade, suggesting that, at least in this section, chromosome evolution has proceeded from higher to lower numbers. This result is consistent with the analysis of Hipp et al. (2007). That study classified taxa within the ENA clade of the *Ovales* section as having "low," "medium," or "high" chromosome numbers. Then, an MP-based reconstruction was performed, and the number of the common ancestor was inferred to have been high.

Passiflora.—The tropical genus *Passiflora* L. contains more than 530 species and is widely distributed from southern Argentina into southern United States with an additional 20 species restricted to the Old World. The species in the genus span a variety of life history strategies (weedy colonizers of secondary vegetation to the large canopy lianas of primary forests). The chromosome number distribution is highly congruent with the phylogeny. The 2 largest lineages in the genus are *Decaloba* and *Passiflora* with typical chromosome numbers of $n = 6$ and $n = 9$, respectively, whereas 2 smaller subgenera (*Astrophea* and *Deidamioides*) have a haploid number of $n = 12$. The base chromosome of the genus is controversial, being proposed at $x = 6$, 9, 10, or 12 (de Melo and Guerra 2003). Hansen et al. (2006; TreeBase study accession S1330) constructed a Bayesian phylogeny of the genus based on molecular chloroplast data for 61 species for which 58 have known chromosome numbers. They then employed a MP methodology (giving all transition types equal weights) to reconstruct ancestral chromosome numbers. The authors hypothesized a chromosome number of $x = 12$ at the base of the genus. This placement implies descending dysploidy (from 12 to 6 and 9) with no polyploidization events in the species analyzed.

Using the probabilistic models developed here, we reexamined chromosome number evolution in the genus based on the consensus Bayesian tree and chromosome numbers as used by Hansen et al. (2006). Using the AIC criterion, the best supported models are M0 (which allows only for gains and losses of single chromosomes) and M2 (which allows also for demi- and polyploidizations), with no power to support one over the other (Δ AIC of 0.4 between the 2 models). Notably, chromosome number evolution in *Passiflora* is highly different under the 2 models (see supplementary materials, available from http://www.sysbio.oxfordjournals.org/). Under M0, the predicted base chromosome number is $x = 12$ with probability of 0.98. The only possible type of event is descending dysploidy with around 14 expected single chromosome decreases across the whole tree. In contrast, under M2, $x = 6$ is highly supported as the base chromosome number of the genus with probability of 0.99. This is followed by approximately 4 polyploidization events along the branches leading to the $n = 12$ groups and approximately 2 demi-polyploidization events along the branches leading to the $n = 9$ subgenera and to *Passiflora microstipula* ($n = 9$), within the subgenus *Decabola*. In light of the completely different possible reconstructions, we conclude that more data, in terms of broader sampling of chromosome numbers, are necessary to infer with confidence the probable evolutionary paths in the genus (but see Discussion section for violation of the models assumptions).

Helianthus.—The sunflower genus, *Helianthus*, has become a model organism for studying diploid and polyploidy speciation. Thirteen species of the genus are known to be polyploid, although it is currently unknown which polyploid species are autopolyploids and which are allopolyploids. Timme et al. (2007) created a high-resolution phylogeny of *Helianthus* based on the external transcribed spacer region of the nuclear 18S–26S ribosomal DNA region. We have used the chromosome numbers and phylogeny reported in Timme et al. (2007) to study chromosome number evolution in this genus. Three haploid chromosome numbers are present in the phylogeny: $n = 17$, 34, and 51. The best supported model is M2 suggesting that both demi- and polyploidization events have occurred in the evolution of this family with rates of chromosome gain and loss approaching zero (Table 3). This model is much better supported than M1, which assumes that demi-polyploidization is not possible or M0, which does not allow for any polyploidization type (ΔlnL = 214 and 476 between M2 vs. M1 and M2 vs. M0, respectively). Furthermore, the difference in lnL between Models M2 and M3 is marginal, meaning that the rates of polyploidization and demi-polyploidization are approximately equal in the genus. Our ancestral reconstruction places $x = 17$ at the root of the genus with probability approaching 1.0. Moreover, chromosome assignments to all ancestral nodes are either 17 or 34, meaning that the demi-ploypoloidization events leading to $n = 51$ occurred only along terminal branches, possibly suggesting a lower speciation rate following such an event.

Discussion

Understanding the role of polyploidy in evolution has been a long standing interest (Stebbins 1938; Grant 1963; Levin 1983; Masterson 1994). The models presented here provide a much needed tool to elucidate the evolution of eukaryotic chromosome numbers. In particular, the likelihood models are the first to provide an explicit probabilistic framework for estimating rates of chromosome number change, polyploid incidence, and the reconstruction of ancestral chromosome numbers and ploidy levels. A significant advance over previous methods of estimating chromosome number evolution is that the likelihood method allows hypothesis testing and provides a measure of confidence in the results. The computed probabilities for ancestral states are especially critical for estimates of base numbers, which have often been inferred without an explicit statistical framework. The simulation results showed that our models are useful with realistic sample sizes and may readily be applied to a large collection of cytological and phylogenetic data available in the literature. Combined with other sources of data, such as genomic inferences of paleopolyploidy, these models provide a foundation for making broad comparisons of chromosome evolution across eukaryotes.

Our simulation results demonstrated that a relatively small set of taxa is sufficient for an accurate inference of ancestral chromosome numbers and of the number of ploidy transitions. The exception would be when the divergence times between species are relatively long coupled with insufficient sampling (i.e., a small number of taxa with large branch lengths). Given the availability of cytological data, especially for plants (Goldblatt and Johnson 1979), obtaining a sufficiently diverse sample is not an insurmountable obstacle. Our statistical framework also allows for models with increasing complexity to be applied when more data are available. For example, the dependence between the dysploidy rates and the current chromosome number may be examined in large data sets where the additional parameters are more accurately inferred. Similarly, the rate of demi-polyploidization (e.g., a transition from tetraploidy to hexaploidy) can be distinguished from the polyploidy rate using a model with an additional free parameter.

In most examples studied here, we assumed that each species possesses a single cytotype. In reality, polymorphism may be present such that different chromosome numbers may be treated as different haplotypes segregating within the population. Such intraspecific polymorphism can be accommodated in the likelihood approach presented here by treating the current state as a vector of probabilities for each chromosome count.

The models developed do not include the possibility of polyploidy reversals. Although a transition from a polyploid taxon to a diploid one (a process termed polyhaploidy) can easily be integrated into our models by an additional rate parameter, such a model was not presented. As reviewed by Ramsey and Schemske (2002), polyhaploid mutants do occasionally occur, but their chance of survival and establishment is particularly low. Thus, polyploidization is regarded as a largely irreversible process leading to an increase in ploidy levels over time (e.g., Meyers and Levin 2006 and references within). This is especially likely to be the case among long established polyploids, which do not necessarily contain 2 complement diploid gene sets (Scannell et al. 2006; but see Gerstein et al. 2006). Nevertheless, our model accommodates the reduction of chromosome number to prepolyploidization levels via several rounds of chromosome losses.

When calculating the overall likelihood of the data given a phylogenetic tree and parameter estimates, the probability distribution of the states at the root needs to be specified. In the current implementation, the root frequencies are determined according to their probability of giving rise to the observed data. Alternatively, the root frequencies may be treated as additional free parameters and estimated using ML. However, this will always result in the most likely chromosome number having probability 1, and all the others having probability 0. Another possibility is to set the root frequencies as the equilibrium distribution of the Markov process. This implies that the process of chromosome evolution has reached stationarity and that it has evolved over a

sufficient amount of time, which is generally not valid for chromosome count data (Meyers and Levin 2006).

In addition, the method developed here assumes that the phylogenetic tree is known with certainty. One could, however, account for phylogenetic uncertainty by applying the method to a set of trees, for example, those generated using either a Bayesian analysis or a bootstrap approach. The set of likelihood models should then be optimized separately for each tree. The model that best fit a particular tree should be used to compute ancestral chromosome numbers and the number of ploidy shifts. Thus, a distribution of AIC scores for each model would be generated as well as additional distributions for ancestral states and the number and location of ploidy shifts. This approach should result in a more robust inference given phylogenetic uncertainty and can assess whether the results are sensitive to specific branches of the phylogeny that are poorly resolved.

The models presented here assume "time-homogeneity"; the transition matrix Q is the same across all parts of the phylogeny. A more realistic approach may be to consider certain branches as hot spots for polyploidization, whereas the polyploidization rate in other parts of the tree would be lower. Rate heterogeneity can be represented by a branch model (e.g., Yang 1998), where one subclade of the phylogeny displays one certain pattern (e.g., low polyploidy rate), whereas the second subclade another (e.g., high polyploidy rate). Alternatively, the rate of polyploidy may switch several times between high and low values. Such heterotachy can be accommodated using 2 (or more) sets of rate parameters and an additional free parameter specifying the transition rate between the 2 rate classes (Galtier 2001). Such a model may be appropriate if, as hypothesized by Grant (1981), chromosome number increases due to polyploidy limit subsequent genome doubling. Accordingly, following a polyploidization event, a lineage may enter a state that inhibits further polyploidization but may switch back to a state more tolerable to polyploidization at a later stage. On the other hand, our inability to distinguish between the various models fitted to the *Passiflora* data set might be due to violation of the time-homogeneity assumption. In *Passiflora*, changes in chromosome number seem to occur mostly at deeper nodes of the phylogeny with near homeostasis of chromosome number in the 2 major lineages of the genus (*Passiflora* having chromosome numbers of $n = 9$ for all species and *Decaloba* having chromosome numbers of $n = 6$ for all species but one). Certainly, exploring rate heterogeneity is a good direction for future work.

Ideally, we would like to track evolutionary changes not only in chromosome numbers but also in ploidy levels using a 2 character-state model (chromosome number and ploidy). This would allow us to differentiate between 2 different kinds of demi-polyploidization events: those leading to odd ploidy levels and those leading to even ones. For example, a diploid species with $n = 12$ chromosomes would become a triploid (odd ploidy level) after a demi-polyploidization event, whereas a tetraploid species with $n = 12$ chromosomes would become a hexaploid (even ploidy level). Transitions leading to odd ploidy levels (e.g., from di- to triploid) are considered to be far more transient than transitions to even ploidy levels, reducing the rate of successful demi-polyploidization events of this type. The development of a 2 character-state model is another important future direction.

In the current implementation, changes in chromosome number are assumed to occur gradually and in proportion to the time available for change. Hence, changes at or near speciation events (i.e., internal nodes) are not assumed more likely than anywhere else along the phylogeny. Because shifts in karyotype are often assumed to be associated with diversification, it may be reasonable to limit their occurrence to internal nodes rather than to be homogenous along the branches of the tree. In such a speciational change model (Mooers et al. 1999), the distance between taxa is proportional to the number of speciation events between them. A third possibility would be to combine the gradual and speciational change models. A fraction of chromosomal shifts may correlate with speciation, whereas another would be in proportion with time, representing transitions due to polyploid series segregating within a population. This, in turn, may allow examining the interplay between speciation and polyploidization within a phylogenetic framework. Again, this is a promising future direction.

Supplementary Material

Supplementary material can be found at: http://www.sysbio.oxfordjournals.org/.

Funding

This work was supported by a Killam postdoctoral fellowship (I.M.) and the Discovery Grant from the National Science and Engineering Research Council, Canada (S.P.O.).

Acknowledgments

We are grateful to Andrew Hipp, Ruth Timme, Tetsuo Ohi-Toma, and Katie Hansen for providing chromosome numbers and phylogenies for the *Carex*, *Helianthus*, *Aristolochia*, and *Passiflora* data sets and to To Adi Stern for helpful discussions on irreversible likelihood models. We thank the associate editor and the reviewers for insightful comments and suggestions.

References

Akaike H. 1974. A new look at the statistical model identification. IEEE Trans. Automat. Contr. 119:716–723.

Aury J.M., Jaillon O., Duret L., Noel B., Jubin C., Porcel B.M., Segurens B., Daubin V., Anthouard V., Aiach N., Arnaiz O.,

Billaut A., Beisson J., Blanc I., Bouhouche K., Camara F., Duharcourt S., Guigo R., Gogendeau D., Katinka M., Keller A.M., Kissmehl R., Klotz C., Koll F., Le Mouel A., Lepere G., Malinsky S., Nowacki M., Nowak J.K., Plattner H., Poulain J., Ruiz F., Serrano V., Zagulski M., Dessen P., Betermier M., Weissenbach J., Scarpelli C., Schachter V., Sperling L., Meyer E., Cohen J., Wincker P. 2006. Global trends of whole-genome duplications revealed by the ciliate *Paramecium tetraurelia*. Nature. 444:171–178.

Baldwin B.G., Wessa B.L., Panero J.L. 2002. Nuclear rDNA evidence for major lineages of Helenioid Heliantheae (Compositae). Syst. Bot. 27:161–198.

Barker M.S., Kane N.C., Matvienko M., Kozik A., Michelmore R.W., Knapp S.J., Rieseberg L.H. 2008. Multiple paleopolyploidizations during the evolution of the Compositae reveal parallel patterns of duplicate gene retention after millions of years. Mol. Biol. Evol. 25:2445–2455.

Blanc G., Wolfe K.H. 2004. Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. Plant Cell. 16:1667–1678.

Bowers J.E., Chapman B.A., Rong J., Paterson A.H. 2003. Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. Nature. 422:433–438.

Cui L., Wall P.K., Leebens-Mack J.H., Lindsay B.G., Soltis D.E., Doyle J.J., Soltis P.S., Carlson J.E., Arumuganathan K., Barakat A., Albert V.A., Ma H., de Pamphilis C.W. 2006. Widespread genome duplications throughout the history of flowering plants. Genome Res. 16:738–749.

de Melo N.F., Guerra M. 2003. Variability of the 5S and 45S rDNA sites in *Passiflora* L. species with distinct base chromosome numbers. Ann. Bot. 92:309–316.

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

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

Furlong R.F., Holland P.W.H. 2004. Polyploidy in vertebrate ancestry: Ohno and beyond. Biol. J. Linn. Soc. 82:425–430.

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

Gerstein A.C., Chun H.J., Grant A., Otto S.P. 2006. Genomic convergence toward diploidy in *Saccharomyces cerevisiae*. PLoS Genet. 2:e145.

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

Goldblatt P., Johnson D.E., editors. 1979. Index to plant chromosome numbers. St. Louis (MO): Missouri Botanical Garden.

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

Grant V. 1981. Plant speciation. New York: Columbia University Press.

Guggisberg A., Mansion G., Kelso S., Conti E. 2006. Evolution of biogeographic patterns, ploidy levels, and breeding systems in a diploid-polyploid species complex of Primula. New Phytol. 171:617–632.

Hansen A.K., Gilbert L.E., Simpson B.B., Downie S.R., Cervi A.C., Jansen R.K. 2006. Phylogenetic relationships and chromosome number evolution in *Passiflora*. Syst. Bot. 31:138–150.

Hipp A.L. 2007. Nonuniform processes of chromosome evolution in sedges (*Carex*: Cyperaceae). Evolution. 61:2175–2194.

Hipp A.L., Rothrock P.E., Reznicek A.A., Berry P.E. 2007. Chromosome number changes associated with speciation in sedges: a phylogenetic study in *Carex* section *Ovales* (Cyperaceae) using AFLP data. Aliso. 23:193–203.

Hipp A.L., Rothrock P.E., Roalson E.H. 2009. The evolution of chromosome arrangements in *Carex* (Cyperaceae). Bot. Rev. 75: 96–109.

Jaillon O., Aury J.M., Noel B., Policriti A., Clepet C., Casagrande A., Choisne N., Aubourg S., Vitulo N., Jubin C., Vezzi A., Legeai F., Hugueney P., Dasilva C., Horner D., Mica E., Jublot D., Poulain J., Bruyere C., Billault A., Segurens B., Gouyvenoux M., Ugarte E., Cattonaro F., Anthouard V., Vico V., Del Fabbro C., Alaux M., Di Gaspero G., Dumas V., Felice N., Paillard S., Juman I., Moroldo M., Scalabrin S., Canaguier A., Le Clainche I., Malacrida G., Durand E., Pesole G., Laucou V., Chatelet P., Merdinoglu D., Delledonne M., Pezzotti M., Lecharny A., Scarpelli C., Artiguenave F., Pe M.E., Valle G., Morgante M., Caboche M., Adam-Blondon A.F., Weissenbach J., Quetier F., Wincker P. 2007. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature. 449:463–467.

Kellis M., Birren B.W., Lander E.S. 2004. Proof and evolutionary analysis of ancient genome duplication in the yeast Saccharomyces cerevisiae. Nature. 428:617–624.

Koshi J.M., Goldstein R.A. 1996. Probabilistic reconstruction of ancestral protein sequences. J. Mol. Evol. 42:313–320.

Levin D.A. 1983. Polyploidy and novelty in flowering plants. Am. Nat. 122:1–25.

Lynch M., Conery J.S. 2000. The evolutionary fate and consequences of duplicate genes. Science. 290:1151–1155.

Maddison W.P. 1995. Calculating the probability distributions of ancestral states reconstructed by parsimony on phylogenetic trees. Syst. Biol. 44:474–481.

Maddison W.P., Maddison D.R. 2008. Mesquite: a modular system for evolutionary analysis. Version 2.5. Available from: http://mesquiteproject.org.

Mansion G., Zeltner L. 2004. Phylogenetic relationships within the new world endemic *Zeltnera* (Gentianaceae-Chironiinae) inferred from molecular and karyological data. Am. J. Bot. 91:2069–2086.

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

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

Moler C., Van Loan C. 2003. Nineteen dubious ways to compute the exponential of a matrix, twenty-five years later. SIAM Rev. 45:3–50.

Mooers A.Ø., Vamosi S.M., Schluter D. 1999. Using phylogenies to test macroevolutionary hypotheses of trait evolution in Cranes (Gruinae). Am. Nat. 154:249–259.

Norrmann G.A., Quarin C.L., Keeler K.H. 1997. Evolutionary implications of meiotic chromosome behavior, reproductive biology, and hybridization in 6x and 9x cytotypes of *Andropogon gerardii* (Poaceae). Am. J. Bot. 84:201–207.

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.

Ota R., Waddell P.J., Hasegawa M., Shimodaira H., Kishino H. 2000. Appropriate likelihood ratio tests and marginal distributions for evolutionary tree models with constraints on parameters. Mol. Biol. Evol. 17:798–803.

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

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

Press W.H., Teukolsky S.A., Vetterling W.T., Flannery B.P. 2002. Numerical recipes in C++: the art of scientific computing. Cambridge (MA): Cambridge University Press.

Pupko T., Pe'er I., Shamir R., Graur D. 2000. A fast algorithm for joint reconstruction of ancestral amino acid sequences. Mol. Biol. Evol. 17:890–896.

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

Ramsey J., Schemske D.W. 2002. Neopolyploidy in flowering plants. Annu. Rev. Ecol. Syst. 33:589–639.

Scannell D.R., Byrne K.P., Gordon J.L., Wong S., Wolfe K.H. 2006. Multiple rounds of speciation associated with reciprocal gene loss in polyploid yeasts. Nature. 440:341–345.

Schultheis L.M. 2001. Systematics of *Downingia* (Campanulaceae) based on molecular sequence data: implications for floral and chromosome evolution. Syst. Bot. 26:603–621.

Soltis D.E., Albert V.A., Leebens-Mack J., Bell C.D., Paterson A.H., Zheng C., Sankoff D., de Pamphilis C.W., Wall P.K., Soltis P.S.

2009. Polyploidy and angiosperm diversification. Am. J. Bot. 96:336–348.
Stebbins G.L. 1938. Cytological characteristics associated with the different growth habits in the dicotyledons. Am. J. Bot. 25: 189–198.
Stebbins G.L. 1971. Processes of organic evolution. 2nd ed. Englewood Cliffs (NJ): Prentice-Hall.
Stern A., Mayrose I., Shaul S., Gophna U., Pupko T. 2010. On the evolution of thymidine synthesis: a tale of two enzymes and a virus. Syst. Biol. (in press).
Timme R.E., Simpson B.B., Linder C.R. 2007. High-resolution phylogeny for *Helianthus* (Asteraceae) using the 18S-26S ribosomal DNA external transcribed spacer. Am. J. Bot. 94:1837–1852.
Vision T.J., Brown D.G., Tanksley S.D. 2000. The origins of genomic duplications in Arabidopsis. Science. 290:2114–2117.
Yang Z. 1998. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Mol. Biol. Evol. 15: 568–573.
Yang Z. 2006. Computational molecular evolution. Oxford: Oxford University Press.