

Version dated: November 15, 2016

RH: CLADOGENETIC AND ANAGENETIC MODELS OF CHROMOSOME
EVOLUTION

Cladogenetic and Anagenetic Models of Chromosome Number Evolution: a Bayesian Model Averaging Approach

WILLIAM A. FREYMAN¹ AND SEBASTIAN HÖHNA^{1,2}

¹*Department of Integrative Biology, University of California, Berkeley, CA, 94720, USA;*

²*Department of Statistics, University of California, Berkeley, CA, 94720, USA*

Corresponding author: William A. Freyman, Department of Integrative Biology,
University of California, Berkeley, CA, 94720, USA; E-mail: freyman@berkeley.edu.

Abstract.— Chromosome number is a key feature of the higher-order organization of the genome, and changes in chromosome number play a fundamental role in evolution. Dysploid gains and losses in chromosome number, as well as polyploidization events, may drive reproductive isolation and lineage diversification. The recent development of probabilistic models of chromosome number evolution in the groundbreaking work by Mayrose et al. (2010, ChromEvol) have enabled the

inference of ancestral chromosome numbers over molecular phylogenies and generated new interest in studying the role of chromosome changes in evolution. However, the ChromEvol approach assumes all changes occur anagenetically (along branches), and does not model events that are specifically cladogenetic. Cladogenetic changes may be expected if chromosome changes result in reproductive isolation. Here we present a new class of models of chromosome number evolution (called ChromoSSE) that incorporate both anagenetic and cladogenetic change. The ChromoSSE models allow us to determine the mode of chromosome number evolution; is chromosome evolution occurring primarily within lineages, primarily at lineage splitting, or in clade-specific combinations of both? Furthermore, we can estimate the location and timing of possible chromosome speciation events over the phylogeny. We implemented ChromoSSE in a Bayesian statistical framework, specifically in the software RevBayes, to accommodate uncertainty in parameter estimates while leveraging the full power of likelihood based methods. We tested ChromoSSE's accuracy with simulations and re-examined chromosomal evolution in *Aristolochia*, *Carex* section *Spirostachyae*, *Helianthus*, *Mimulus* sensu lato (s.l.), and *Primula* section *Aleuritia*, finding evidence for clade-specific combinations of anagenetic and cladogenetic dysploid and polyploid modes of chromosome evolution.

(Keywords: ChromoSSE; chromosome evolution; phylogenetic models; anagenetic; cladogenetic; dysploidy; polyploidy; whole genome duplication; chromosome speciation; reversible-jump Markov chain Monte Carlo; Bayes factors)

1 A central organizing component of the higher-order architecture of the
2 genome is chromosome number, and changes in chromosome number have long
3 been understood to play a fundamental role in evolution. In the seminal work
4 *Genetics and the Origin of Species* (1937), Dobzhansky identified “the raw
5 materials for evolution”, the sources of natural variation, as two evolutionary
6 processes: mutations and chromosome changes. “Chromosomal changes are one of
7 the mainsprings of evolution,” Dobzhansky asserted, and changes in chromosome
8 number such as the gain or loss of a single chromosome (dysploidy), or the
9 doubling of the entire genome (polyploidy), can have phenotypic consequences,
10 affect the rates of recombination, and increase reproductive isolation among
11 lineages and thus drive diversification (Stebbins 1971). Recently, evolutionary
12 biologists have studied the macroevolutionary consequences of chromosome changes
13 within a molecular phylogenetic framework, mostly due to the groundbreaking
14 work of Mayrose et al. (2010, ChromEvol) which introduced likelihood-based
15 models of chromosome number evolution. The ChromEvol models have permitted
16 phylogenetic studies of ancient whole genome duplication events, rapid
17 “catastrophic” chromosome speciation, major reevaluations of the evolution of
18 angiosperms, and new insights into the fate of polyploid lineages (e.g. Pires and
19 Hertweck 2008; Mayrose et al. 2011; Tank et al. 2015).

20 One aspect of chromosome evolution that has not been thoroughly studied
21 in a probabilistic framework is cladogenetic change in chromosome number.
22 Cladogenetic changes occur solely at speciation events, as opposed to anagenetic
23 changes that occur along the branches of a phylogeny. Studying cladogenetic

24 chromosome changes in a phylogenetic framework has been difficult since the
25 approach used by ChromEvol models only anagenetic changes and ignores the
26 changes that occur specifically at speciation events and may be expected if
27 chromosome changes result in reproductive isolation. Reproductive
28 incompatibilities caused by chromosome changes may play an important role in the
29 speciation process, and led White (1978) to propose that chromosome changes
30 perform “the primary role in the majority of speciation events.” Indeed,
31 chromosome fusions and fissions may have played a role in the formation of
32 reproductive isolation and speciation in the great apes (Ayala and Coluzzi 2005),
33 and the importance of polyploidization in plant speciation has long been
34 appreciated (Coyne et al. 2004; Rieseberg and Willis 2007). Recent work by Zhan
35 et al. (2016) revealed phylogenetic evidence that polyploidization is frequently
36 cladogenetic in land plants. However, their approach did not examine the role
37 dysploid changes may play in speciation, and it required a two step analysis in
38 which one first used ChromEvol to infer ploidy levels, and then a second modeling
39 step to infer the proportion of ploidy shifts that were cladogenetic.

40 Here we present models of chromosome number evolution that
41 simultaneously account for both cladogenetic and anagenetic polyploid as well as
42 dysploid changes in chromosome number over a phylogeny. These models
43 reconstruct an explicit history of cladogenetic and anagenetic changes in a clade,
44 enabling estimation of ancestral chromosome numbers. Our approach also identifies
45 different modes of chromosome number evolution among clades; we can detect
46 primarily anagenetic, primarily cladogenetic, or clade-specific combinations of both

47 modes of chromosome changes. Furthermore, these models allow us to infer the
48 timing and location of possible polyploid and dysploid speciation events over the
49 phylogeny. Since these models only account for changes in chromosome number,
50 they ignore speciation that may accompany other types of chromosome
51 rearrangements such as inversions. Our models cannot determine that changes in
52 chromosome number “caused” the speciation event, but they do reveal that
53 speciation and chromosome change are temporally correlated. Thus, these models
54 can give us evidence that the chromosome number change coincided with
55 cladogenesis and so may have played a significant role in the speciation process.

56 A major challenge for all phylogenetic models of cladogenetic character
57 change is accounting for unobserved speciation events due to lineages going extinct
58 and not leaving any extant descendants (Bokma 2002). Teasing apart the
59 phylogenetic signal for cladogenetic and anagenetic processes given unobserved
60 speciation events is a major difficulty. The Cladogenetic State change Speciation
61 and Extinction (ClaSSE) model (Goldberg and Igić 2012) accounts for unobserved
62 speciation events by jointly modeling both character evolution and the phylogenetic
63 birth-death process. Our class of chromosome evolution models uses the ClaSSE
64 approach, and could be considered a special case of ClaSSE. We implemented our
65 models (called ChromoSSE) in a Bayesian framework and use Markov chain Monte
66 Carlo algorithms to estimate posterior probabilities of the model’s parameters.
67 However, compared to most character evolution models, SSE models require
68 additional complexity since they must model extinction and speciation processes.
69 Using simulations, we examined the impact of this additional complexity on our

70 chromosome evolution models' performance.

71 Out of the class of ChromoSSE models described here, it is possible that no
72 single model will adequately describe the chromosome evolution of a given clade.
73 The most parameter-rich ChromoSSE model has 13 independent parameters,
74 however the models that best describe a given dataset (a phylogeny and a set of
75 observed chromosome counts) may be special cases of the full model. For example,
76 there may be a clade for which the best fitting models have no anagenetic rate of
77 polyploidization (the rate = 0.0) and for which all polyploidization events are
78 cladogenetic. To explore the entire space of all possible models of chromosome
79 number evolution we constructed a reversible jump Markov chain Monte Carlo
80 (Green 1995) that samples across models of different dimensionality, drawing
81 samples from chromosome evolution models in proportion to their posterior
82 probability and enabling Bayes factors for each model to be calculated. This
83 approach incorporates model uncertainty by permitting model-averaged inferences
84 that do not condition on a single model; we draw estimates of ancestral
85 chromosome numbers and rates of chromosome evolution from all possible models
86 weighted by their posterior probability. For general reviews of this approach to
87 model averaging see Madigan and Raftery (1994), Hoeting et al. (1999), Kass and
88 Raftery (1995), and for its use in phylogenetics see Posada and Buckley (2004).
89 Averaging over all models has been shown to provide a better average predictive
90 ability than conditioning on a single model (Madigan and Raftery 1994).
91 Conditioning on a single model ignores model uncertainty, which can lead to an
92 underestimation in the uncertainty of inferences made from that model (Hoeting

93 et al. 1999). In our case, this can lead to overconfidence in estimates of ancestral
94 chromosome numbers and chromosome evolution parameter value estimates.

95 Our motivation in developing these phylogenetic models of chromosome
96 evolution is to determine the mode of chromosome number evolution; is
97 chromosome evolution occurring primarily within lineages, primarily at lineage
98 splitting, or in clade-specific combinations of both? By identifying how much of the
99 pattern of chromosome number evolution is explained by anagenetic versus
100 cladogenetic change, and by identifying the timing and location of possible
101 chromosome speciation events over the phylogeny, the ChromoSSE models can help
102 uncover how much of a role chromosome changes play in speciation. In this paper
103 we first describe the ChromoSSE models of chromosome evolution and our
104 Bayesian method of model selection, then we assess the models' efficacy by testing
105 them with simulated datasets, particularly focusing on the impact of unobserved
106 speciation events on inferences, and finally we apply the models to five empirical
107 datasets that have been previously examined using other models of chromosome
108 number evolution.

109 METHODS

110 *Models of Chromosome Evolution*

111 In this section we introduce our class of probabilistic models of chromosome
112 number evolution. We are interested in modeling the changes in chromosome

113 number both within lineages (anagenetic evolution) and at speciation events
114 (cladogenetic evolution). The anagenetic component of the model is a
115 continuous-time Markov process similar to Mayrose et al. (2010) as described
116 below. The cladogenetic changes are accounted for by a birth-death process similar
117 to Maddison et al. (2007) and Goldberg and Igić (2012), except each type of
118 cladogenetic chromosome event is given its own rate. Thus, the birth-death process
119 has multiple speciation rates (one for each type of cladogenetic change) and a single
120 constant extinction rate. Our models of chromosome number evolution can
121 therefore be understood as a specific case of the Cladogenetic State change
122 Speciation and Extinction (ClasSE) model (Goldberg and Igić 2012), which
123 integrates over all possible unobserved speciation events (due to lineages that have
124 gone extinct) directly in the likelihood calculation of the observed chromosome
125 counts and tree shape. To test the importance of accounting for unobserved
126 speciation events we also briefly describe a version of the model that handles
127 different cladogenetic event types as transition probabilities at each observed
128 speciation event and ignores unobserved speciation events, similar to the
129 dispersal-extinction-cladogenesis (DEC) models of geographic range evolution (Ree
130 and Smith 2008).

131 Our models contain a set of 6 free parameters for anagenetic chromosome
132 number evolution, a set of 5 free parameters for cladogenetic chromosome number
133 evolution, an extinction rate parameter, and the root frequencies of chromosome
134 numbers, for a total of 13 free parameters. All of the 11 chromosome rate
135 parameters can be removed (fixed to 0.0) except the cladogenetic no-change rate

136 parameter. Thus, the class of chromosome number evolution models described here
137 has a total of $2^{10} = 1024$ nested models of chromosome evolution.

138 Our implementation assumes chromosome numbers can take the value of
139 any positive integer, however to limit the transition matrices to a reasonable size
140 for likelihood calculations we follow Mayrose et al. (2010) in setting the maximum
141 chromosome number C_m to $n + 10$, where n is the highest chromosome number in
142 the observed data. Note that we allow this parameter to be set in our
143 implementation. Hence, it is easily possible to test the impact of setting a specific
144 value for the maximum chromosome count.

145 *Chromosome evolution within lineages.*—

146 Chromosome number evolution within lineages (anagenetic change) is
147 modeled as a continuous-time Markov process similar to Mayrose et al. (2010). The
148 continuous-time Markov process is described by an instantaneous rate matrix Q
149 where the value of each element represents the instantaneous rate of change within
150 a lineage from a genome of i chromosomes to a genome of j chromosomes. For all
151 elements of Q in which either $i = 0$ or $j = 0$ we define $Q_{ij} = 0$. For the off-diagonal

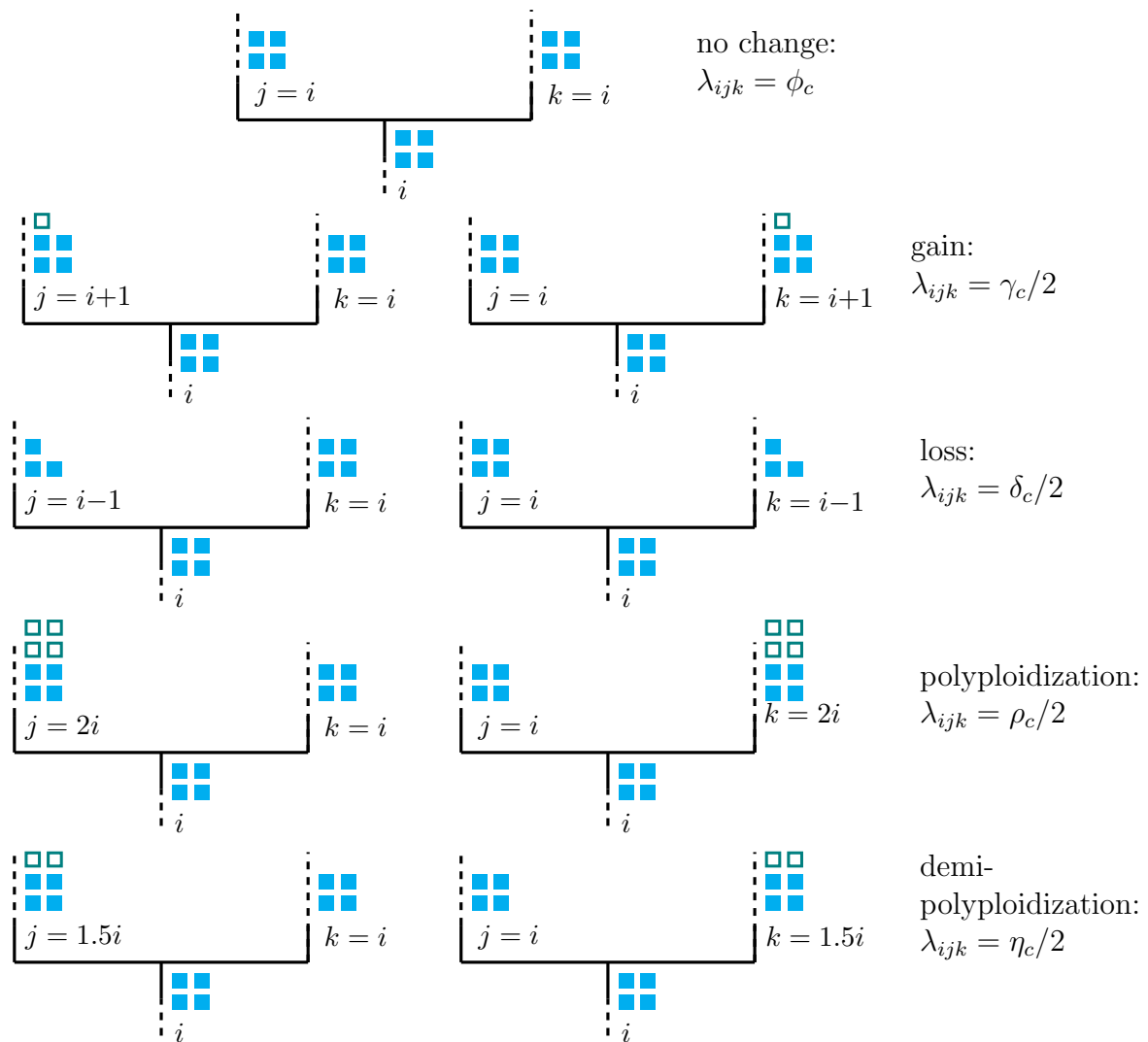


Figure 1: **Modeled cladogenetic chromosome evolution events.** At each speciation event 9 different cladogenetic events are possible. The rate of each type of speciation event is λ_{ijk} where i is the chromosome number before cladogenesis and j and k are the states of each daughter lineage immediately after cladogenesis. The dashed lines represent possible chromosomal changes within lineages that are modeled by the anagenetic rate matrix Q .

152 elements $i \neq j$ with positive values of i and j , Q is determined by:

$$Q_{ij} = \begin{cases} \gamma_a e^{\gamma_m(i-1)} & j = i + 1, \\ \delta_a e^{\delta_m(i-1)} & j = i - 1, \\ \rho_a & j = 2i, \\ \eta_a & j = 1.5i, \\ 0 & \text{otherwise,} \end{cases} \quad (1)$$

153 where γ_a , δ_a , ρ_a , and η_a are the rates of chromosome gains, losses, polyploidizations,
154 and demi-polyploidizations. γ_m and δ_m are rate modifiers of chromosome gain and
155 loss, respectively, that allow the rates of chromosome gain and loss to depend on
156 the current number of chromosomes. This enables modeling scenarios in which the
157 probability of fusion or fission events is positively or negatively correlated with the
158 number of chromosomes. If the rate modifier $\gamma_m = 0$, then $\gamma_a e^{0(i-1)} = \gamma_a$. If the
159 rate modifier $\gamma_m > 0$, then $\gamma_a e^{\gamma_m(i-1)} \geq \gamma_a$, and if $\gamma_m < 0$ then $\gamma_a e^{\gamma_m(i-1)} \leq \gamma_a$.
160 These two rate modifiers replace the parameters λ_l and δ_l in Mayrose et al. (2010),
161 which in their parameterization may result in negative transition rates. Here we
162 chose to exponentiate γ_m and δ_m to ensure positive transition rates, and avoid ad
163 hoc restrictions on negative transition rates that may induce unintended priors.

164 For odd values of i , we set $Q_{ij} = \eta/2$ for the two integer values of j resulting
165 when $j = 1.5i$ was rounded up and down. We define the diagonal elements $i = j$ of

166 Q as:

$$Q_{ii} = - \sum_{i \neq j}^{C_m} Q_{ij}. \quad (2)$$

167 The probability of anagenetically transitioning from chromosome number i to j
168 along a branch of length t is then calculated by exponentiation of the instantaneous
169 rate matrix:

$$P_{ij}(t) = e^{-Qt}. \quad (3)$$

170 *Chromosome evolution at cladogenesis events.*—

171 At each lineage divergence event over the phylogeny, nine different
172 cladogenetic changes in chromosome number are possible (Figure 1). Each type of
173 cladogenetic event occurs with the rate $\phi_c, \gamma_c, \delta_c, \rho_c, \eta_c$, representing the
174 cladogenesis rates of no change, chromosome gain, chromosome loss,
175 polyploidization, and demi-polyploidization, respectively. The speciation rates λ for
176 the birth-death process generating the tree are given in the form of a 3-dimensional
177 matrix between the ancestral state i and the states of the two daughter lineages j

178 and k . For all positive values of i , j , and k , we define:

$$\lambda_{ijk} = \begin{cases} \phi_c & j = k = i \\ \gamma_c/2 & j = i + 1 \text{ and } k = i, \\ \gamma_c/2 & j = i \text{ and } k = i + 1, \\ \delta_c/2 & j = i - 1 \text{ and } k = i, \\ \delta_c/2 & j = i \text{ and } k = i - 1, \\ \rho_c/2 & j = 2i \text{ and } k = i, \\ \rho_c/2 & j = i \text{ and } k = 2i, \\ \eta_c/2 & j = 1.5i \text{ and } k = i, \\ \eta_c/2 & j = i \text{ and } k = 1.5i, \\ 0 & \text{otherwise,} \end{cases} \quad (4)$$

179 so that the total speciation rate of the birth-death process λ_t is given by:

$$\lambda_t = \phi_c + \gamma_c + \delta_c + \rho_c + \eta_c. \quad (5)$$

180 Similar to the anagenetic instantaneous rate matrix described above, for odd values
181 of i , we set $\lambda_{ijk} = \eta_c/4$ for the integer values of j and k resulting when $1.5i$ is
182 rounded up and down. The extinction rate μ is constant over the tree and for all
183 chromosome numbers.

184 Note that this model allows only a single chromosome number change event

185 on a maximum of one of the daughter lineages at each cladogenesis event. Changes
186 in both daughter lineages at cladogenesis are not allowed; at least one of the
187 daughter lineages must inherit the chromosome number of the ancestor. The model
188 also assumes that cladogenesis events are always strictly bifurcating and that there
189 are no polytomies.

190 *Likelihood Calculation Accounting for Unobserved Speciation.*—

191 The likelihood of cladogenetic and anagenetic chromosome number evolution
192 over a phylogeny is calculated using a set of ordinary differential equations similar
193 to the Binary State Speciation and Extinction (BiSSE) model (Maddison et al.
194 2007). The BiSSE model was extended to incorporate cladogenetic changes by
195 Goldberg and Igić (2012). Similar to Goldberg and Igić (2012), we define $D_{Ni}(t)$ as
196 the likelihood that a lineage with chromosome number i at time t evolves into the
197 observed clade N . We let $E_i(t)$ be the probability that a lineage with chromosome
198 number i at time t goes extinct before the present, or is not sampled at the present.
199 However, unlike the full ClaSSE model the extinction rate μ does not depend on
200 the chromosome number i of the lineage. The differential equations for these two
201 probabilities is given by:

202

$$\frac{dD_{Ni}(t)}{dt} = - \left(\sum_{j=1}^{C_m} \sum_{k=1}^{C_m} \lambda_{ijk} + \sum_{j=1}^{C_m} Q_{ij} + \mu \right) D_{Ni}(t)$$

203

$$+ \sum_{j=1}^{C_m} Q_{ij} D_{Nj}(t) + \sum_{j=1}^{C_m} \sum_{k=1}^{C_m} \lambda_{ijk} \left(D_{Nk}(t) E_j(t) + D_{Nj}(t) E_k(t) \right) \quad (6)$$

204

205

206

207

$$\begin{aligned} \frac{dE_i(t)}{dt} = & - \left(\sum_{j=1}^{C_m} \sum_{k=1}^{C_m} \lambda_{ijk} + \sum_{j=1}^{C_m} Q_{ij} + \mu \right) E_i(t) \\ & + \mu + \sum_{j=1}^{C_m} Q_{ij} E_j(t) + \sum_{j=1}^{C_m} \sum_{k=1}^{C_m} \lambda_{ijk} E_j(t) E_k(t), \quad (7) \end{aligned}$$

209

210
211 where λ_{ijk} for each possible cladogenetic event is given by equation 4, and the rates
212 of anagenetic changes Q_{ij} are given by equation 1.

213 The differential equations above have no known analytical solution.

214 Therefore, we numerically integrate the equations for every arbitrarily small time
215 interval moving along each branch from the tip of the tree towards the root. When
216 a node l is reached, the probability of it being in state i is calculated by combining
217 the probabilities of its descendant nodes m and n as such:

$$D_{li}(t) = \sum_{j=1}^{C_m} \sum_{k=1}^{C_m} \lambda_{ijk} D_{mj}(t) D_{nk}(t), \quad (8)$$

218 where again λ_{ijk} for each possible cladogenetic event is given by equation 4. Letting
219 D denote a set of observed chromosome counts, Ψ an observed phylogeny, and θ_q a
220 particular set of chromosome evolution model parameters, then the likelihood for
221 the model parameters θ_q is given by:

$$P(D, \Psi | \theta_q) = \sum_{i=1}^{C_m} \pi_i D_{0i}(t), \quad (9)$$

222 where π_i is the root frequency of chromosome number i and $D_{0i}(t)$ is the likelihood

223 of the root node being in state i conditional on having given rise to the observed
224 tree Ψ and the observed chromosome counts D .

225 *Initial Conditions.*—

226 The initial conditions for each observed lineage at time $t = 0$ for the
227 extinction probabilities described by equation 7 are $E_i(0) = 1 - \rho_s$ for all i where ρ_s
228 is the sampling probability of including that lineage. For lineages with an observed
229 chromosome number of i , the initial condition is $D_{Ni}(0) = \rho_s$. The initial condition
230 for all other chromosome numbers j is $D_{Nj}(0) = 0$.

231 *Likelihood Calculation Ignoring Unobserved Speciation.*—

232 To test the effect of unobserved speciation events on inferences of
233 chromosome number evolution we also implemented a version of the model
234 described above that only accounts for observed speciation events. At each lineage
235 divergence event over the phylogeny, the probabilities of cladogenetic chromosome
236 number evolution $P(\{j, k\}|i)$ are given by the simplex $\{\phi_p, \gamma_p, \delta_p, \rho_p, \eta_p\}$, where
237 $\phi_p, \gamma_p, \delta_p, \rho_p$, and η_p represent the probabilities of no change, chromosome gain,
238 chromosome loss, polyploidization, and demi-polyploidization, respectively. This
239 approach does not require estimating speciation or extinction rates.

240 Here, we calculate the likelihood of chromosome number evolution over a
241 phylogeny using Felsenstein's pruning algorithm (Felsenstein 1981) modified to
242 include cladogenetic probabilities similar to models of biogeographic range
243 evolution (Landis et al. 2013; Landis in press). Let D again denote a set of
244 observed chromosome counts and Ψ represent an observed phylogeny where node l

245 has descendant nodes m and n . The likelihood of chromosome number evolution at
 246 node l conditional on node l being in state i and θ_q being a particular set of
 247 chromosome evolution model parameter values is given by:

248

$$\begin{aligned}
 &P_l(D, \Psi|i, \theta_q) = \\
 &\underbrace{\sum_{j=1}^{C_m} \sum_{k=1}^{C_m} P(\{j, k\}|i)}_{\text{cladogenetic}} \left[\underbrace{\sum_{j_e=1}^{C_m} P_{jj_e}(t_m) P_m(D, \Psi|j_e, \theta_q)}_{\text{anagenetic}} \right] \left[\sum_{k_e=1}^{C_m} P_{kk_e}(t_n) P_n(D, \Psi|k_e, \theta_q) \right], \\
 & \tag{10}
 \end{aligned}$$

251

252 where the length of the branches between l and m is t_m and between l and n is t_n .
 253 The state at the end of these branches near nodes m and n is j_e and k_e ,
 254 respectively. The state at the beginning of these branches, where they meet at node
 255 l , is j and k respectively. The cladogenetic term sums over the probabilities
 256 $P(\{j, k\}|i)$ of all possible cladogenetic changes from state i to the states j and k at
 257 the beginning of each daughter lineage. The anagenetic term of the equation is the
 258 product of the probability of changes along the branches from state j to state j_e
 259 and state k to state k_e (given by equation 3) and the likelihood of the tree above
 260 node l recursively computed from the tips.

261 The likelihood for the model parameters θ_q is given by:

$$P(D, \Psi|\theta_q) = \sum_{i=1}^{C_m} \pi_i P_0(D, \Psi|i, \theta_q), \tag{11}$$

262 where $P_0(D, \Psi|i, \theta_q)$ is the conditional likelihood of the root node being in state i

263 and π_i is the root frequency of chromosome number i .

264 *Estimating Parameter Values and Ancestral States.*—

265 For any given tree with a set of observed chromosome counts, there exists a
266 posterior distribution of model parameter values and a set of probabilities for the
267 ancestral chromosome numbers at each internal node of the tree. Let $P(s_i, \theta_q | D, \Psi)$
268 denote the joint posterior probability of θ_q and a vector of specific ancestral
269 chromosome numbers s_i given a set of observed chromosome counts D and an
270 observed tree Ψ . The posterior is given by Bayes' rule:

$$P(s_i, \theta_q, | D, \Psi) = \frac{P(D, \Psi | s_i, \theta_q) P(s_i | \theta_q) P(\theta_q)}{\int_{\theta} \sum_{s=1}^{C_m} P(D, \Psi | s, \theta) P(s | \theta) P(\theta) d\theta}. \quad (12)$$

271 Here, $P(s_i | \theta_q)$ is the prior probability of the ancestral states s conditioned on the
272 model parameters θ_q , and $P(\theta_q)$ is the joint prior probability of the model
273 parameters.

274 In the denominator of equation 12 we integrate over all possible values of θ
275 and sum over all possible ancestral chromosome numbers s . Since θ is a vector of
276 13 parameters and s is a vector of $2n - 1$ ancestral states where n is the number of
277 observed tips in the phylogeny, the denominator of equation 12 requires a high
278 dimensional integral and an extremely large summation that is impossible to
279 calculate analytically. Instead we use Markov chain Monte Carlo methods
280 (Metropolis et al. 1953; Hastings 1970) to estimate the posterior probability
281 distribution in a computationally efficient manner.

282 Joint ancestral states are inferred using a two-pass tree traversal procedure
 283 as described in Pupko et al. (2000), and previously implemented in a Bayesian
 284 framework by Huelsenbeck and Bollback (2001) and Pagel et al. (2004). First,
 285 partial likelihoods are calculated during the backwards-time post-order tree
 286 traversal in equations 6 and 7. Joint ancestral states are then sampled during a
 287 pre-order tree traversal in which the root state is first drawn from the marginal
 288 likelihoods at the root, and then states are drawn for each descendant node
 289 conditioned on the state at the parent node until the tips are reached. Again, we
 290 must numerically integrate over a system of differential equations during this
 291 root-to-tip tree traversal. This integration, however, is performed in forward-time,
 292 thus the set of ordinary differential equations must be slightly altered since our
 293 models of chromosome number evolution are not time reversible. Accordingly, we
 294 calculate:

$$\begin{aligned}
 \frac{dD_{Ni}(t)}{dt} = & - \left(\sum_{j=1}^{C_m} \sum_{k=1}^{C_m} \lambda_{ijk} + \sum_{j=1}^{C_m} Q_{ji} + \mu \right) D_{Ni}(t) \\
 & + \sum_{j=1}^{C_m} Q_{ji} D_{Nj}(t) + \sum_{j=1}^{C_m} \sum_{k=1}^{C_m} \lambda_{ijk} \left(D_{Nj}(t) E_k(t) + D_{Nk}(t) E_j(t) \right) \quad (13)
 \end{aligned}$$

$$\begin{aligned}
 \frac{dE_i(t)}{dt} = & \left(\sum_{j=1}^{C_m} \sum_{k=1}^{C_m} \lambda_{ijk} + \sum_{j=1}^{C_m} Q_{ji} + \mu \right) E_i(t) \\
 & - \mu - \sum_{j=1}^{C_m} Q_{ji} E_j(t) - \sum_{j=1}^{C_m} \sum_{k=1}^{C_m} \lambda_{ijk} E_j(t) E_k(t), \quad (14)
 \end{aligned}$$

304 during the forward-time root-to-tip pass to draw joint conditional ancestral states.

305 *Priors.*—

306 Model parameter priors are listed in Table 1. Our implementation allows all
307 priors to be easily modified so that their impact on results can be effectively
308 assessed. Priors for anagenetic rate parameters are given an exponential
309 distribution with a mean of $2/\Psi_l$ where Ψ_l is the length of the tree Ψ . This
310 corresponds to a mean rate of two events over the observed tree. The priors for the
311 rate modifiers γ_m and δ_m are assigned a uniform distribution with the range
312 $-3/C_M$ to $3/C_m$. This sets minimum and maximum bounds on the amount the
313 rate modifiers can affect the rates of gain and loss at the maximum chromosome
314 number to $\gamma_a e^{-3} = \gamma_a 0.050$ and $\gamma_a e^3 = \gamma_a 20.1$, and $\delta_a e^{-3} = \delta_a 0.050$ and
315 $\delta_a e^3 = \delta_a 20.1$, respectively.

316 The speciation rates are drawn from an exponential prior with a mean equal
317 to an estimate of the net diversification rate \hat{d} . Under a constant rate birth-death
318 process not conditioning on survival of the process, the expected number of lineages
319 at time t is given by:

$$E(N_t) = N_0 e^{td}, \quad (15)$$

320 where N_0 is the number of lineages at time 0 and d is the net diversification rate
321 $\lambda - \mu$ (Nee et al. 1994b; Höhna 2015). Therefore, we estimate \hat{d} as:

$$\hat{d} = (\ln N_t - \ln N_0)/t, \quad (16)$$

322 where N_t is the number of lineages in the observed tree that survived to the

323 present, t is the age of the root, and $N_0 = 2$.

324 The extinction rate μ is given by:

$$\mu = r \times \lambda_t = r \times (\phi_c + \gamma_c + \delta_c + \rho_c + \eta_c), \quad (17)$$

325 where λ_t is the total speciation rate and r is the relative extinction rate. The
 326 relative extinction rate r is assigned a uniform (0,1) prior distribution, thus forcing
 327 the extinction rate to be smaller than the total speciation rate. The root
 328 frequencies of chromosome numbers π are drawn from a flat Dirichlet distribution.

Table 1: **Model parameter names and prior distributions.** See the main text for complete description of model parameters and prior distributions. Ψ_l represents the length of tree Ψ and C_m is the maximum chromosome number allowed.

	Parameter	X	$f(X)$
Anagenetic	Chromosome gain rate	γ_a	Exponential($\lambda = \Psi_l/2$)
	Chromosome loss rate	δ_a	Exponential($\lambda = \Psi_l/2$)
	Polyploidization rate	ρ_a	Exponential($\lambda = \Psi_l/2$)
	Demi-polyploidization rate	η_a	Exponential($\lambda = \Psi_l/2$)
	Linear component of chromosome gain rate	γ_m	Uniform($-3/C_m, 3/C_m$)
	Linear component of chromosome loss rate	δ_m	Uniform($-3/C_m, 3/C_m$)
Cladogenetic	No change	ϕ_c	Exponential($\lambda = 1/\hat{d}$)
	Chromosome gain	γ_c	Exponential($\lambda = 1/\hat{d}$)
	Chromosome loss	δ_c	Exponential($\lambda = 1/\hat{d}$)
	Polyploidization	ρ_c	Exponential($\lambda = 1/\hat{d}$)
	Demi-polyploidization	η_c	Exponential($\lambda = 1/\hat{d}$)
Other	Root frequencies	π	Dirichlet(1, ..., 1)
	Relative-extinction	r	Uniform(0, 1)

329 *Model Uncertainty and Selection*

330 *Model Averaging.*—

331 To account for model uncertainty we calculate the posterior density of
332 chromosome evolution model parameters θ without conditioning on any single
333 model of chromosome evolution. For each of the 1024 chromosome models M_k ,
334 where $k = 1, 2, \dots, 1024$, the posterior distribution of θ is

$$P(\theta|D) = \sum_{k=1}^K P(\theta|D, M_k)P(M_k|D). \quad (18)$$

335 Here we average over the posterior distributions conditioned on each model
336 weighted by the model's posterior probability. We assume an equal prior
337 probability for each model $P(M_k) = 2^{-10}$.

338 *Reversible Jump Markov Chain Monte Carlo.*—

339 To sample from the space of all possible chromosome evolution models, we
340 employ reversible jump MCMC (Green 1995). This algorithm draws samples from
341 parameter spaces of differing dimensions, and in stationarity samples each model in
342 proportion to its posterior probability. This permits inference of each model's fit to
343 the data while simultaneously accounting for model uncertainty.

344 Our reversible jump MCMC moves between models of different dimensions
345 using augment and reduce moves (Huelsenbeck et al. 2000; Pagel and Meade 2006;
346 May et al. 2016). The reduce move proposes that a parameter should be removed
347 from the current model by setting its value to 0.0, effectively disallowing that class
348 of evolutionary event. Augment moves reverse reduce moves by allowing the
349 parameter to once again have a non-zero value. Both augment and reduce moves
350 operate on all chromosome rate parameters except for ϕ_c the rate of no

351 cladogenetic change. Thus the least complex model the MCMC can sample from is
352 one in which $\phi_c > 0.0$ and all other chromosome rate parameters are set to 0.0,
353 corresponding to a model of no chromosomal changes over the phylogeny. The prior
354 probability of reducing or augmenting model M_k is $P_r(M_k) = P_a(M_k) = 0.5$.

355 *Bayes Factors.*—

356 In some cases we wish to compare the fit of models to summarize the mode
357 of evolution within a clade. Bayes factors (Kass and Raftery 1995) compare the
358 evidence between two competing models M_i and M_j

$$B_{ij} = \frac{P(D|M_i)}{P(D|M_j)} = \frac{P(M_i|D)}{P(M_j|D)} / \frac{P(M_i)}{P(M_j)}. \quad (19)$$

359 In words, the Bayes factor B_{ij} is given by the ratio of the posterior odds to the
360 prior odds of the two models. Unlike other methods of model selection such as
361 Akaike Information Criterion (AIC; Akaike 1974) and the Bayesian Information
362 Criterion (BIC; Schwarz 1978), Bayes factors take into account the full posterior
363 densities of the model parameters and do not rely on point estimates. Furthermore
364 AIC and BIC ignore the priors assigned to parameters, whereas Bayes factors
365 penalizes parameters based on the informativeness of the prior. If the prior is
366 informative but overlaps little with the likelihood it is penalized more than a
367 diffuse uninformative prior that allows the parameter to take on whatever value is
368 informed by the data (Xie et al. 2011).

369

Implementation

370 The model and MCMC analyses described here are implemented in C++ in
371 the software RevBayes (Höhna et al. 2016). Rev scripts that specify the
372 chromosome number evolution model (ChromoSSE) described here as a
373 probabilistic graphical model (Höhna et al. 2014) and run the empirical analyses in
374 RevBayes are available at <http://github.com/wf8/ChromoSSE>. The RevGadgets
375 R package (available at <https://github.com/revbayes/RevGadgets>) contains
376 functions to summarize results and generate plots of inferred ancestral chromosome
377 numbers over a phylogeny.

378 The MCMC proposals used are outlined in Table 2. Aside from the
379 reversible jump MCMC proposals described above, all other proposals are standard
380 except for the ElementSwapSimplex move operated on the Dirichlet distributed root
381 frequencies parameter. This move randomly selects two elements r_1 and r_2 from the
382 root frequencies vector and swaps their values. The reverse move, swapping the
383 original values of r_1 and r_2 back, will have the same probability as the initial move
384 since r_1 and r_2 were drawn from a uniform distribution. Thus, the Hasting ratio is
385 1 and the ElementSwapSimplex move is a symmetric Metropolis move.

386 *Simulations*

387 We conducted a series of simulations to: 1) test the effect of unobserved
388 speciation events on chromosome number estimates when using a model that does
389 not account for unobserved speciation, 2) compare the accuracy of models of
390 chromosome evolution that account for unobserved speciation versus those that do
391 not, 3) test the effect of jointly estimating speciation and extinction rates with

392 chromosome number evolution, and 4) test for identifiability of cladogenetic
393 parameters. We will refer to each of the 4 simulations above as experiment 1,
394 experiment 2, experiment 3, and experiment 4.

395 For all 4 experiments the same set of simulated trees and chromosome
396 counts were used. 100 trees were simulated under the birth-death process with
397 $\lambda = 0.25$ and $\eta = 0.15$ (Figure 2) using the R package diversitree (FitzJohn 2012).
398 The trees were conditioned on an age of 25.0 time units and a minimum of 10
399 extant lineages. To test the effect of unobserved speciation events due to lineages
400 going extinct on cladogenetic estimates, chromosome number evolution was
401 simulated along the trees including their extinct lineages (unpruned) and the same
402 100 trees but with the extinct lineages pruned. All chromosome number
403 simulations were performed using RevBayes (Höhna et al. 2016).

404 Three models were used to generate simulated chromosome counts: a model
405 where all chromosome evolution was anagenetic, a model where all chromosome
406 evolution was cladogenetic, and a model that mixed both anagenetic and
407 cladogenetic changes (Table 3). Parameter values were roughly informed by the
408 mean values estimated from the empirical datasets. The mean length of the
409 simulated trees was 253.5 (Figure 2). Hence, the anagenetic rates were set to
410 $2/253.5 \approx 0.008$ which corresponds to an expected value of 2 events over the tree.
411 The root chromosome number was fixed to be 8. Simulating data for all 3 models
412 over both the pruned and unpruned tree resulted in 600 simulated datasets. To
413 reproduce the effect of using reconstructed phylogenies all inferences were
414 performed using the trees with extinct lineages pruned and with chromosome

415 counts from extinct lineages removed.

416 For all 4 experiments, MCMC analyses were run for 5000 iterations, where
417 each iteration consisted of 28 different moves in a random move schedule with 79
418 moves per iteration (Table 2). Samples were drawn with each iteration, and the
419 first 1000 samples were discarded as burn in. Effective sample sizes were
420 consistently over 200. To perform all 4 experiments 1300 MCMC analyses were run
421 requiring a total of 60927.8 CPU hours on the Savio computational cluster at the
422 University of California, Berkeley.

423 *Experiment 1.*—

424 In experiment 1 we tested the effect of unobserved speciation events on
425 chromosome number estimates when using a model that does not account for
426 unobserved speciation. Is the additional model complexity required to account for
427 unobserved speciation necessary, or are the effects of unobserved speciation
428 negligible and safe to ignore? Using the model described above that does not
429 account for unobserved speciation, ancestral chromosome numbers and chromosome
430 evolution model parameters were estimated for each of the 600 datasets.

431 *Experiment 2.*—

432 Here we compared the accuracy of models of chromosome evolution that
433 account for unobserved speciation versus those that do not. Since extinction can
434 safely be assumed to be present to some extent in all clades, it is likely all empirical
435 datasets contain some unobserved speciation. Do we see an increase in accuracy
436 when we account for unobserved speciation events, or conversely do we see an

Table 2: **MCMC moves used for chromosome number evolution analyses.** See the main text for further explanations of the moves used. Samples were drawn from the MCMC each iteration, where each iteration consisted of 28 different moves in a random move schedule with 79 moves per iteration.

	Parameter	X	Move	Weight
Anagenetic	Chromosome gain rate	γ_a	Scale($\lambda = 1$)	2
	Chromosome gain rate	γ_a	Reduce/Augment	2
	Chromosome loss rate	δ_a	Scale($\lambda = 1$)	2
	Chromosome loss rate	δ_a	Reduce/Augment	2
	Polyploidization rate	ρ_a	Scale($\lambda = 1$)	2
	Polyploidization rate	ρ_a	Reduce/Augment	2
	Demi-polyploidization rate	η_a	Scale($\lambda = 1$)	2
	Demi-polyploidization rate	η_a	Reduce/Augment	2
	Linear component of gain rate	γ_m	Slide($\delta = 0.1$)	1
	Linear component of gain rate	γ_m	Slide($\delta = 0.001$)	1
	Linear component of gain rate	γ_m	Reduce/Augment	2
	Linear component of loss rate	δ_m	Slide($\delta = 0.1$)	1
	Linear component of loss rate	δ_m	Slide($\delta = 0.001$)	1
	Linear component of loss rate	δ_m	Reduce/Augment	2
	Cladogenetic	No change	ϕ_c	Scale($\lambda = 5$)
Chromosome gain		γ_c	Scale($\lambda = 5$)	2
Chromosome gain		γ_c	Reduce/Augment	2
Chromosome loss		δ_c	Scale($\lambda = 5$)	2
Chromosome loss		δ_c	Reduce/Augment	2
Polyploidization		ρ_c	Scale($\lambda = 5$)	2
Polyploidization		ρ_c	Reduce/Augment	2
Demi-polyploidization		η_c	Scale($\lambda = 5$)	2
Demi-polyploidization		η_c	Reduce/Augment	2
All cladogenetic rates		$\phi_c, \gamma_c, \delta_c, \rho_c, \eta_c$	Joint Up-Down	2
Other	Root frequencies	π	BetaSimplex($\alpha = 0.5$)	10
	Root frequencies	π	ElementSwapSimplex	20
	Relative-extinction	r	Scale($\lambda = 5$)	3
	Relative-extinction and all clado rates	$r, \phi_c, \gamma_c, \delta_c, \rho_c, \eta_c$	Joint Up-Down	2
			Scale($\lambda = 0.5$)	
Total			28	79

437 increase in the variance of our estimates that perhaps describes true uncertainty
438 due to extinction? To test this, we estimated ancestral chromosome numbers and
439 chromosome evolution model parameters over the simulated datasets that included
440 unobserved speciation using both the chromosome model that accounts for
441 unobserved speciation as well as the model that does not.

442 *Experiment 3.*—

443 In experiment 3 we tested the effect of jointly estimating speciation and
444 extinction rates with chromosome number evolution. Estimating speciation and
445 extinction rates accurately is notoriously challenging (Nee et al. 1994a; Rabosky
446 2010; Beaulieu and O’Meara 2015; May et al. 2016), so how much of the variance in
447 chromosome evolution estimates made with models that jointly estimate speciation
448 and extinction are due to uncertainty in diversification rates? Here we compared
449 our estimates of ancestral chromosome numbers and chromosome evolution model
450 parameters using the model that accounts for unobserved speciation (and in which
451 speciation and extinction rates are jointly estimated) with estimates made from the
452 same model but where the true rates of speciation and extinction used to simulate
453 the data were fixed. The latter analyses were given the true rates of total
454 speciation and extinction, but still had to estimate the proportion of speciation
455 events for each type of cladogenetic event.

456 *Experiment 4.*—

457 Since we model the same chromosome number transitions as both
458 cladogenetic and anagenetic processes, it is possible that the two processes could be

459 confounded and our models may not be fully identifiable. Furthermore, preliminary
460 results suggested our models overestimate anagenetic changes and underestimate
461 cladogenetic changes when the true generating process includes cladogenetic
462 evolution. Here we compared cladogenetic and anagenetic estimates under
463 simulation scenarios that only included cladogenetic changes. Do we see an increase
464 in accuracy of cladogenetic parameter estimates when anagenetic changes are
465 disallowed (fixed to 0)?

466 *Summarizing Simulation Results.*—

467 To summarize the results of our simulations, we measured the accuracy of
468 ancestral state estimates as the percent of simulation in which the true root
469 chromosome number 8 was found to be the maximum a posteriori (MAP) estimate.
470 To evaluate the uncertainty of the simulations, we calculated the mean posterior
471 probability of root chromosome number for the simulation replicates that correctly
472 found 8 to be the MAP estimate. We also calculated the percentage of simulation
473 replicates for which the true model of chromosome number evolution used to
474 simulate the data (as given by Table 3) was estimated to be the MAP model, and
475 calculated the mean posterior probabilities of the true model. To compare the
476 accuracy of model averaged parameter value estimates we calculated coverage
477 probabilities. Coverage probabilities are the percentage of simulation replicates for
478 which the true parameter value falls within the 95% highest posterior density
479 (HPD). High accuracy is shown when coverage probabilities approach 1.0.

480

Empirical Data

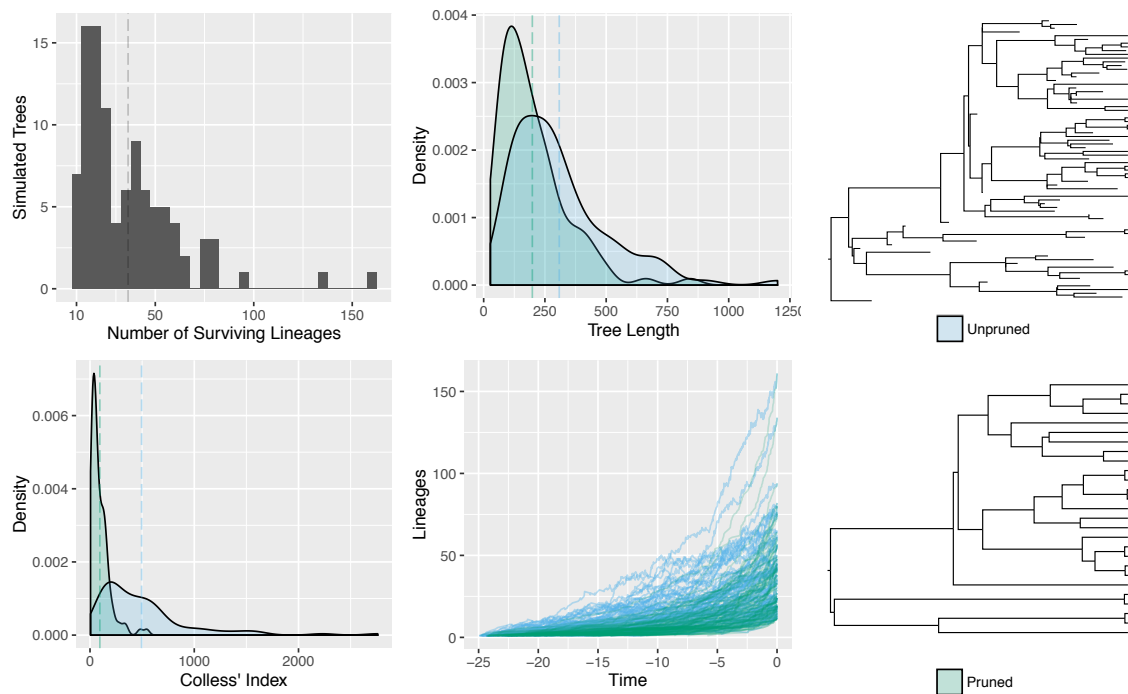


Figure 2: Tree simulations. 100 trees were simulated under the birth-death process as described in the main text. Chromosome number evolution was simulated over the unpruned trees that included all extinct lineages, as well as over the same trees but with extinct lineages pruned. This resulted in two simulated datasets: one simulated under a process that did have unobserved speciation events, and one simulated with no unobserved speciation events. Shown above is a histogram of the number of lineages that survived to the present, the tree lengths, Colless' Index (a measure of tree imbalance; Colless 1982), and lineage through time plots of the 100 pruned and unpruned trees.

Table 3: **Simulation parameter values.** Parameter values used to simulate datasets under 3 modes of chromosome number evolution: anagenetic only, cladogenetic only, and mixed. The total speciation rate $\lambda_t = 0.25$ and the extinction rate $\mu = 0.15$. The root state was fixed to 8.

Simulation mode	γ_a	δ_a	ρ_a	η_a	γ_m	δ_m	ϕ_c	γ_c	δ_c	ρ_c	η_c
Anagenetic	0.008	0.008	0.008	-	-	-	λ_t	-	-	-	-
Cladogenetic	-	-	-	-	-	-	$0.85\lambda_t$	$0.05\lambda_t$	$0.05\lambda_t$	$0.05\lambda_t$	-
Mixed	0.008	0.008	0.008	-	-	-	$0.85\lambda_t$	$0.05\lambda_t$	$0.05\lambda_t$	$0.05\lambda_t$	-

481 Phylogenetic data and chromosomes counts from five plant genera were
482 analyzed (see Table 4). Like in Mayrose et al. (2010) we assumed each species had
483 a single cytotype, however polymorphism could be accounted for by a vector of
484 probabilities for each chromosome count. Sequence data for *Aristolochia* was
485 downloaded from TreeBASE (Vos et al. 2010) study ID 1586. Sequences for
486 *Helianthus*, *Mimulus* sensu lato, and *Primula* were downloaded directly from
487 GenBank (Benson et al. 2005), reconstructing the sequence matrices from Timme
488 et al. (2007), Beardsley et al. (2004), and Guggisberg et al. (2009). For each of
489 these four datasets phylogenetic analyses were performed with all gene regions
490 concatenated and assuming the general time-reversible (GTR) nucleotide
491 substitution model (Tavaré 1986; Rodriguez et al. 1990) with among-site rate
492 variation modeled using a discretized gamma distribution (Yang 1994) with four
493 rate categories. Since divergence time estimation in years is not the objective of
494 this study, and only relative branching times are needed for our models of
495 chromosome number evolution, a birth-death tree prior was used with a fixed root
496 age of 10.0 time units. The MCMC analyses were sampled every 100 iterations and

497 run for a total of 400000 iterations, with samples from the first 100000 iterations
498 discarded as burnin. Convergence was assessed by ensuring that the effective
499 sample size for all parameters was over 200. For *Carex* section *Spirostachyae* the
500 time calibrated tree from Escudero et al. (2010) was used.

501 Ancestral chromosome numbers and chromosome evolution model
502 parameters were then estimated for each of the five clades. Since testing the effect
503 of incomplete taxon sampling on chromosome evolution inference was not a goal of
504 this work, we used a taxon sampling fraction of 1.0 for all empirical datasets
505 (though see the Discussion section for more on this). MCMC analyses were run for
506 11000 iterations, where each iteration consisted of 28 different moves in a random
507 move schedule with 79 moves per iteration (Table 2). Samples were drawn each
508 iteration, and the first 1000 samples were discarded as burn in. Effective sample
509 sizes for all parameters were over 200. For all datasets except *Primula* we used
510 priors as outlined in Table 1. To demonstrate the flexibility of our Bayesian
511 implementation and its capacity to incorporate prior information we used an
512 informative prior for the root chromosome number in the *Primula* section *Aleuritia*
513 analysis. Our dataset for *Primula* section *Aleuritia* also included samples from
514 *Primula* sections *Armerina* and *Sikkimensis*. Since we were most interested in
515 estimating chromosome evolution within section *Aleuritia*, we used an informative
516 Dirichlet prior $\{1, \dots, 1, 100, 1, \dots, 1\}$ (with 100 on the 11th element) to bias the root
517 state towards the reported base number of *Primula* $x = 11$ (Conti et al. 2000).
518 Note all priors can be easily modified in our implementation, thus the impact of
519 priors can be efficiently tested.

Table 4: **Empirical data sets analysed.**

Clade	Study	Gene region	Alignment length (bp)	Number of OTUs	Haploid chromosome numbers range
<i>Aristolochia</i>	Ohi-Toma et al. (2006)	matK	1268	34	3 - 16
<i>Carex</i> section <i>Spirostachyae</i>	Escudero et al. (2010)	ITS, trnK intron	see Escudero et al. (2010)	24	30 - 42
<i>Helianthus</i>	Timme et al. (2007)	ETS	3085	102	17 - 51
<i>Mimulus</i> sensu lato	Beardsley et al. (2004)	trnL intron, ETS, ITS	2210	115	8 - 46
<i>Primula</i> section <i>Aleuritia</i>	Guggisberg et al. (2009)	rpl16 intron, rps16 intron, trnL intron, trnL-trnF spacer, trnT-trnL spacer, trnD-trnT region	5705	56	9 - 36

520

RESULTS

521

Simulations

522 *General Results.*—

523 In all simulations, the true model of chromosome number evolution was
524 infrequently estimated to be the MAP model ($< 36\%$ of replicates), and when it
525 was the posterior probability of the MAP model was very low (< 0.12 ; Table 5).
526 We found that the accuracy of root chromosome number estimation was similar
527 whether the process that generated the simulated data was cladogenetic-only or
528 anagenetic-only (Tables 5 and 6). However, when the data was simulated under a
529 process that included both cladogenetic and anagenetic evolution we found a
530 decrease in accuracy in the root chromosome number estimates in all cases.

531 *Experiment 1 Results.*—

532 The presence of unobserved speciation in the process that generated the
533 simulated data decreased the accuracy of ancestral state estimates (Figure 3, Table
534 5). Similarly, uncertainty in root chromosome number estimates increased with
535 unobserved speciation (lower mean posterior probabilities; Table 5). The accuracy
536 of parameter value estimates (as measured by coverage probabilities) were similar
537 (results not shown).

538 *Experiment 2 Results.*—

539 When comparing estimates from models that did account for unobserved
540 speciation to estimates from models that did not, we found that the accuracy in
541 estimating model parameter values were mostly similar, though for some
542 cladogenetic parameters there was higher accuracy with the models that did
543 account for unobserved speciation (Figure 4). Estimates of anagenetic parameters
544 were more accurate than estimates of cladogenetic parameters when the true
545 generating model included cladogenetic changes.

546 We found that the models that accounted for unobserved speciation had
547 more uncertainty in their root chromosome number estimates (lower mean posterior
548 probabilities) compared to models that did not account for unobserved speciation.
549 Similarly, the root chromosome number was estimated with slightly lower accuracy
550 (Table 6).

551 *Experiment 3 Results.—*

552 We found that jointly estimating speciation and extinction rates with
553 chromosome number evolution slightly decreased the accuracy in estimating the
554 root chromosome number, and further it increased the uncertainty of root
555 chromosome number (as reflected in lower mean posterior probabilities; Table 6).
556 Fixing the speciation and extinction rates to their true value removed much of the
557 increased uncertainty associated with using a model that accounts for unobserved
558 speciation (Table 6).

559 *Experiment 4 Results.—*

560 Under simulation scenarios that had cladogenetic changes but no anagenetic

561 changes, we found that anagenetic parameters were overestimated and cladogenetic
562 parameters were underestimated (Figure 5 A), which explains the lower coverage
563 probabilities of cladogenetic parameters reported above for experiment 2 (Figure
564 4). When anagenetic parameters were fixed to 0.0 cladogenetic parameters were no
565 longer underestimated (Figure 5 A), and the coverage probabilities of cladogenetic
566 parameters increased slightly (Figure 5 B).

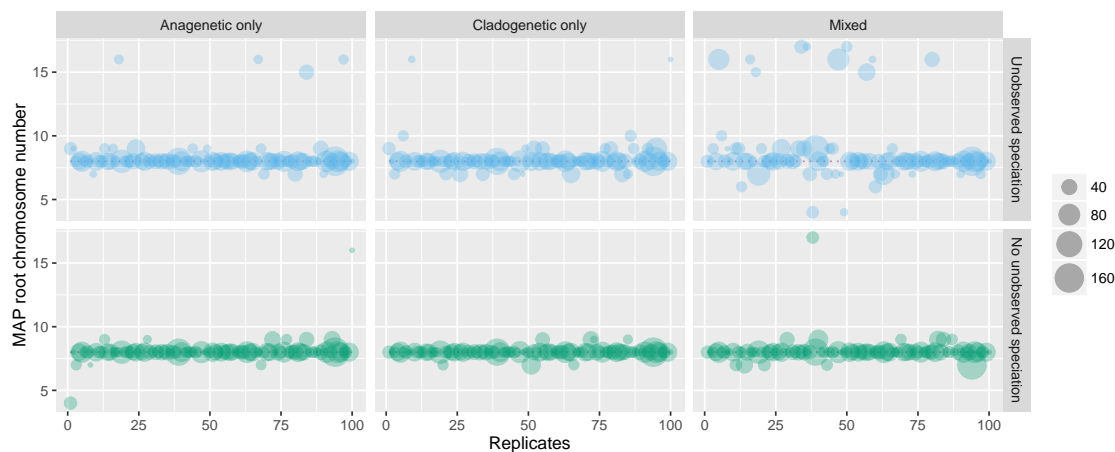


Figure 3: **Experiment 1 results: the effect of unobserved speciation events on the maximum a posteriori (MAP) estimates of root chromosome number.** Model averaged MAP estimates of the root chromosome number for 100 replicates of each simulation type on datasets that included unobserved speciation and datasets that did not include unobserved speciation. Each circle represents a simulation replicate, where the size of the circle is proportional to the number of lineages that survived to the present (the number of extant tips in the tree). The true root chromosome number used to simulate the data was 8 and is marked with a pink dotted line.

Table 5: **Experiment 1 results: the effect of ignoring unobserved speciation events on chromosome evolution estimates.** Regardless of the true mode of chromosome evolution, the presence of unobserved speciation decreases accuracy in estimating the true root state. The columns from left to right are: 1) an indication of whether or not the data was simulated with a process that included unobserved speciation, 2) the true mode of chromosome evolution used to simulate the data, (for description see main text and Table 3), 3) the percent of simulation replicates in which the true chromosome number at the root used to simulate the data was found to be the maximum a posteriori (MAP) estimate, 4) the mean posterior probability of the MAP estimate of the true root chromosome number, 5) the percent of simulation replicates in which the true model used to simulate the data was also found to be the MAP model, and 6) the mean posterior probability of the MAP estimate of the true model.

Simulated Data Included Unobserved Speciation?	Mode of Evolution Used to Simulate Data	True Root State Estimated (%)	Mean Posterior of True Root State	True Model Estimated (%)	Mean Posterior of True Model
No	Cladogenetic	93	0.92	13	0.10
No	Anagenetic	89	0.91	31	0.12
No	Mixed	88	0.84	0	0.0
Yes	Cladogenetic	78	0.87	15	0.09
Yes	Anagenetic	83	0.91	36	0.12
Yes	Mixed	62	0.80	2	0.10

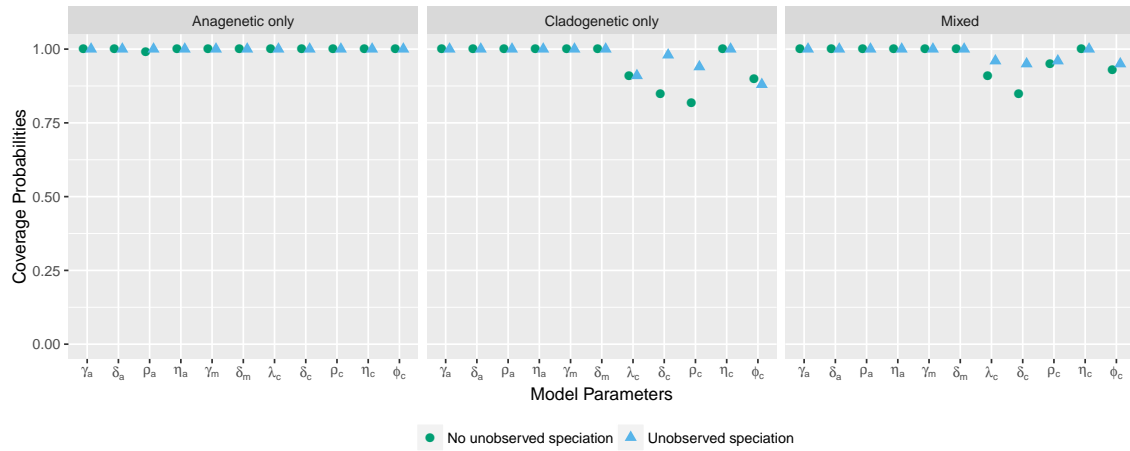


Figure 4: **Experiment 2 results: the effect of using a model that accounts for unobserved speciation on coverage probabilities of chromosome model parameters.** Each point represents the proportion of simulation replicates for which the 95% HPD interval contains the true value of the model parameter. Coverage probabilities of 1.00 mean perfect coverage. The circles represent coverage probabilities for estimates made using the model that does not account for unobserved speciation, and the triangles represent coverage probabilities for estimates made using the model that does account for unobserved speciation.

Table 6: Experiments 2 and 3 results: the effects of using a model that accounts for unobserved speciation and of jointly estimating diversification rates on ancestral chromosome number estimates. This table compares estimates of chromosome evolution using a model that does not account for unobserved speciation events with a model that does (Experiment 2), and compares estimates of chromosome evolution when jointly estimated with speciation and extinction rates versus when the true speciation and extinction rates are given (Experiment 3). Regardless of the true mode of chromosome evolution, the use of a model that accounts for unobserved speciation increases uncertainty in root state estimates. The columns from left to right are: 1) an indication of which experiment the results pertain to, 2) an indication of whether or not the estimates were made with a model that accounted for unobserved speciation, 3) whether diversification rates were jointly estimated with chromosome evolution, 4) the percent of simulation replicates in which the true chromosome number at the root used to simulate the data was found to be the MAP estimate, 5) the mean posterior probability of the MAP estimate of the true root chromosome number.

Experiment #	Estimates Made w/ Model That Accounted for Unobserved Speciation?	Speciation and Extinction Rates Jointly Estimated?	Mode of Evolution Used to Simulate Data	True Root State Estimated (%)	Mean Posterior of True Root State
2	No	No	Cladogenetic	78	0.87
2	No	No	Anagenetic	83	0.91
2	No	No	Mixed	62	0.80
2 & 3	Yes	Yes	Cladogenetic	78	0.81
2 & 3	Yes	Yes	Anagenetic	80	0.86
2 & 3	Yes	Yes	Mixed	61	0.72
3	Yes	No	Cladogenetic	78	0.84
3	Yes	No	Anagenetic	83	0.90
3	Yes	No	Mixed	62	0.76

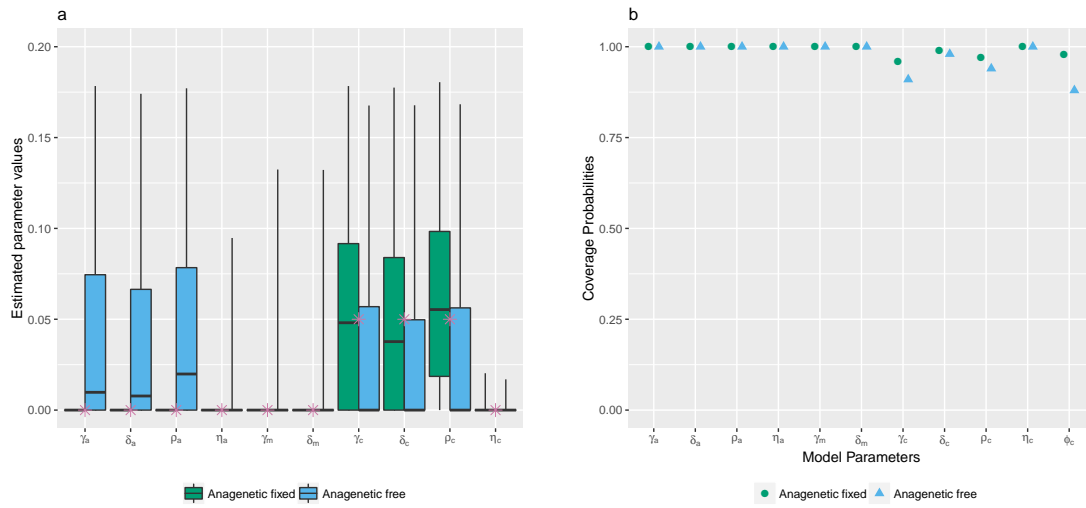


Figure 5: **Experiment 4 results: testing identifiability of cladogenetic parameters.** a) Chromosome parameter value estimates from 100 simulation replicates under a simulation scenario with no anagenetic changes (cladogenetic only). The stars represent true values. The box plots compare parameter estimates made when anagenetic parameters were fixed to 0 to estimates made when all parameters were free. When all parameters were free the anagenetic parameters were overestimated and cladogenetic parameters were underestimated. When the anagenetic parameters were fixed to 0 the estimates for the cladogenetic parameters were more accurate. b) Coverage probabilities of chromosome evolution parameters under the cladogenetic only model of chromosome evolution. The accuracy of cladogenetic parameter estimates increased when anagenetic parameters were fixed to 0.

567

Empirical Data

568 Model averaged MAP estimates of ancestral chromosome numbers for each
 569 of the five empirical datasets are show in Figures 6, 7, 8, 9, and 10. The mean
 570 model-averaged chromosome number evolution parameter value estimates for the
 571 empirical datasets are reported in Table 7. Posterior probabilities for the MAP

572 model of chromosome number evolution were low for all datasets, varying between
573 0.04 for *Carex* section *Spirostachyae* and 0.21 for *Helianthus* (Table 8). Bayes
574 factors supported unique, clade-specific combinations of anagenetic and
575 cladogenetic parameters for all five datasets (Table 8). None of the clades had
576 support for purely anagenetic or purely cladogenetic models of chromosome
577 evolution.

578 The ancestral state reconstructions for *Aristolochia* were highly similar to
579 those found by Mayrose et al. (2010). We found a moderately supported root
580 chromosome number of 8 (posterior probability 0.45), and a polyploidization event
581 on the branch leading to the *Isotrema* clade which has a base chromosome number
582 of 16 with high posterior probability (0.88; Figure 6). On the branch leading to the
583 main *Aristolochia* clade we found a dysploid loss of a single chromosome. Overall,
584 we estimated moderate rates of anagenetic dysploid and polyploid changes, and the
585 rates of cladogenetic change were 0 except for a moderate rate of cladogenetic
586 dysploid loss (Tables 7). There was only one cladogenetic change inferred in the
587 MAP ancestral state reconstruction, which was a recent possible dysploid
588 speciation event that split the sympatric west-central Mexican species *Aristolochia*
589 *tentaculata* and *A. taliscana*.

590 In *Helianthus*, on the other hand, we found high rates of cladogenetic
591 polyploidization, and low rates of anagenetic change (Tables 7). 12 separate
592 possible polyploid speciation events were identified over the phylogeny (Figure 7),
593 and cladogenetic polyploidization made up 16% of all observed and unobserved
594 speciation events. Bayes factors gave very strong support for models that included

595 cladogenetic polyploidization as well as anagenetic demi-polyploidization (Table 8),
596 the latter explaining the frequent anagenetic transitions from 34 to 51 chromosomes
597 found in the MAP ancestral state reconstruction. The well supported root
598 chromosome number of 17 (posterior probability 0.91) corresponded with the
599 findings of Mayrose et al. (2010).

600 As opposed to the *Helianthus* results, the *Carex* section *Spirostachyae*
601 estimates had very low rates of polyploidization and instead had high rates of
602 cladogenetic dysploid change (Tables 7). An estimated 36.9% of all observed and
603 unobserved speciation events included a cladogenetic gain or loss of a single
604 chromosome. Overall, the rates of anagenetic changes were estimated to be much
605 lower than the rates of cladogenetic changes. Bayes factors did not support either
606 anagenetic or cladogenetic polyploidization (Table 8). The MAP root chromosome
607 number of 37, despite being very weakly supported (0.08), corresponds with the
608 findings of Escudero et al. (2014), where it was also poorly supported (Figure 8).

609 In *Primula*, we found a base chromosome number for section *Aleuritia* of 9
610 with high posterior probability (0.82; Figure 9), which agrees with estimates from
611 Glick and Mayrose (2014). We estimated moderate rates of anagenetic and
612 cladogenetic changes, including both cladogenetic polyploidization and
613 demi-polyploidization (Table 7). The MAP ancestral state estimates include an
614 inferred history of possible polyploid and demi-polyploid speciation events in the
615 clade containing the tetraploid *Primula halleri* and the hexaploid *P. scotica*.
616 *Primula* is the only dataset out of the five analysed here for which Bayes factors
617 supported the inclusion of cladogenetic demi-polyploidization (Table 8).

Table 7: **Mean model-averaged parameter value estimates for empirical datasets.** Rates for all parameters are given in units of chromosome changes per branch length unit except for μ which is given in extinction events per time units.

Clade	γ_a	δ_a	ρ_a	η_a	γ_m	δ_m	ϕ_c	γ_c	δ_c	ρ_c	η_c	μ
<i>Aristolochia</i>	0.02	0.05	0.01	0.0	-0.01	-0.01	0.43	0.0	0.04	0.0	0.0	0.19
<i>Carex</i> section <i>Spirostachyae</i>	0.19	0.79	0.16	0.13	0.0	0.04	2.49	2.15	0.15	0.95	0.5	2.26
<i>Helianthus</i>	0.0	0.02	0.0	0.03	-0.0	-0.0	0.68	0.0	0.0	0.13	0.0	0.09
<i>Mimulus</i> s.l.	0.03	0.02	0.01	0.0	0.02	0.02	0.65	0.0	0.0	0.05	0.0	0.16
<i>Primula</i> section <i>Aleuritia</i>	0.01	0.05	0.01	0.01	-0.0	-0.0	2.39	0.01	0.03	0.15	0.09	2.47

618 The well supported root chromosome number of 8 (posterior probability
619 0.90) found for *Mimulus* s.l. corresponds with the inferences reported in Beardsley
620 et al. (2004). We estimated moderate rates of anagenetic dysploid gains and losses,
621 as well as a moderate rate of cladogenetic polyploidization (Table 7). Bayes factors
622 also supported models that included anagenetic dysploid gain and loss, as well as
623 cladogenetic polyploidization (Table 8). The MAP ancestral state reconstruction
624 revealed that most of the possible polyploid speciation events took place in the
625 *Diplacus* clade, particularly in the clade containing the tetraploids *Mimulus*
626 *cupreus*, *M. glabratus*, *M. luteus*, and *M. yecorensis* (Figure 10). Additionally, an
627 ancient cladogenetic polyploidization event is inferred for the split between the two
628 main *Diplacus* clades at about 5 million time units ago.

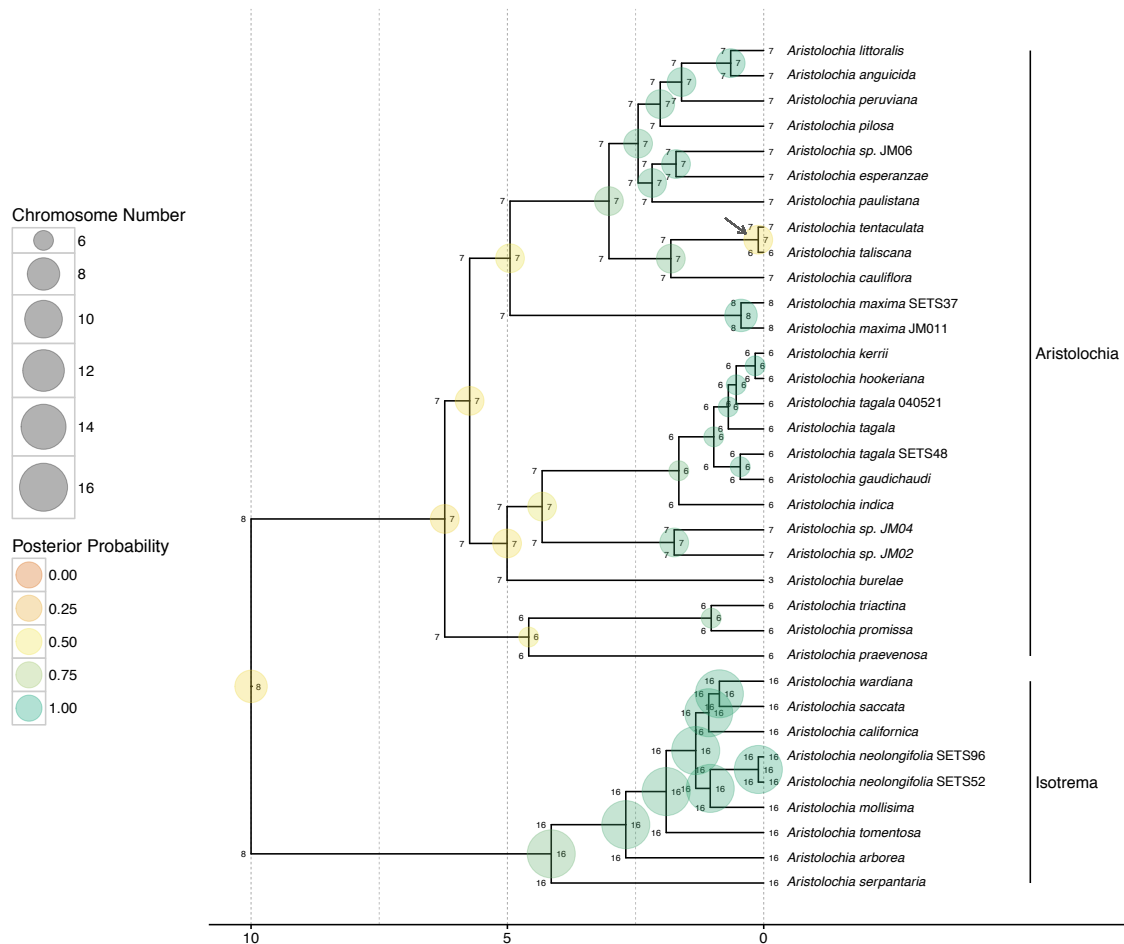


Figure 6: **Ancestral chromosome number estimates of *Aristolochia*.** The model averaged MAP estimate of ancestral chromosome numbers are shown at each branch node. The states of each daughter lineage immediately after cladogenesis are shown at the “shoulders” of each node. The size of each circle is proportional to the chromosome number and the color represents the posterior probability. The MAP root chromosome number is 8 with a posterior probability of 0.45. The grey arrow highlights the possible dysploid speciation event leading to the west-central Mexican species *Aristolochia tentaculata* and *A. taliscana*. Clades corresponding to subgenera are indicated at right.

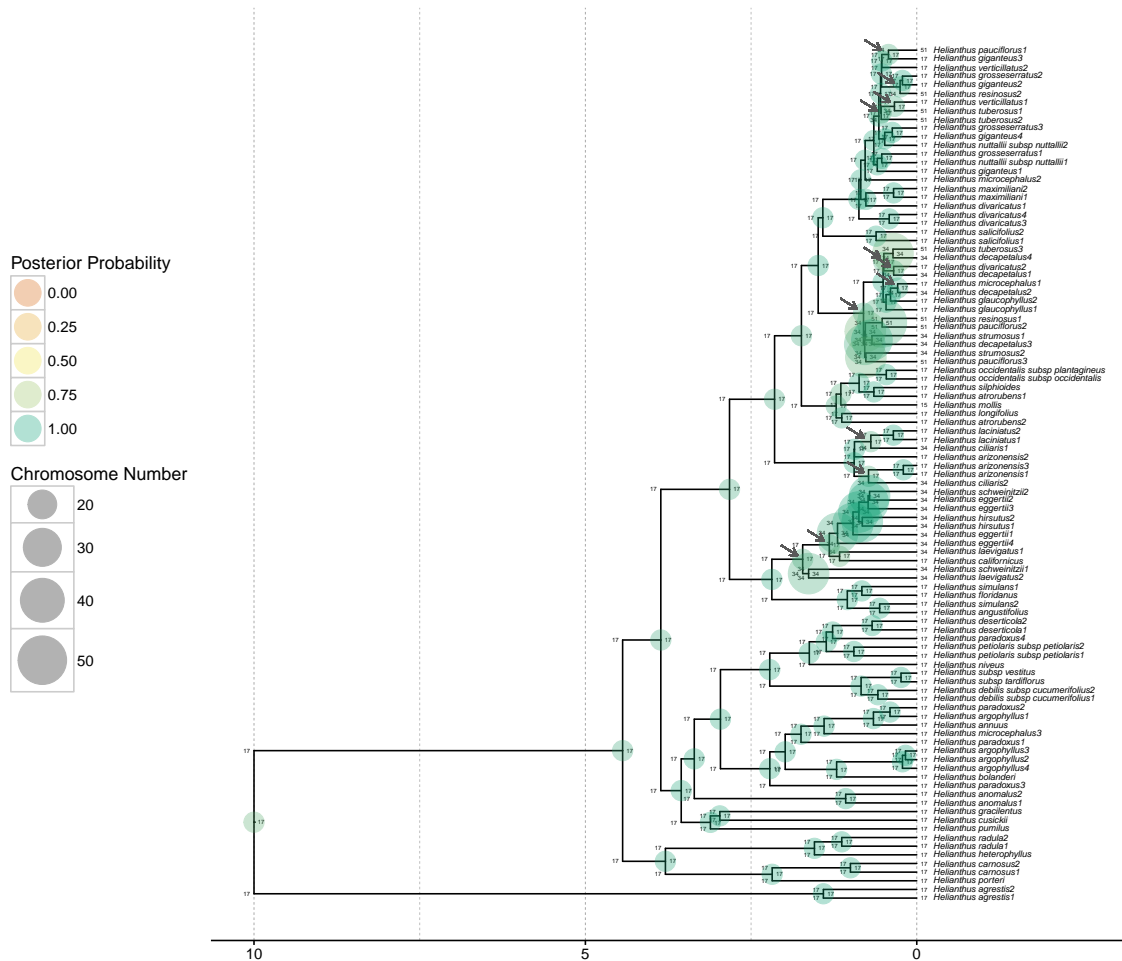


Figure 7: **Ancestral chromosome number estimates of *Helianthus*.** The model averaged MAP estimate of ancestral chromosome numbers are shown at each branch node. The states of each daughter lineage immediately after cladogenesis are shown at the “shoulders” of each node. The size of each circle is proportional to the chromosome number and the color represents the posterior probability. The MAP root chromosome number is 17 with a posterior probability of 0.91. The grey arrows show the locations of 12 inferred polyploid speciation events.

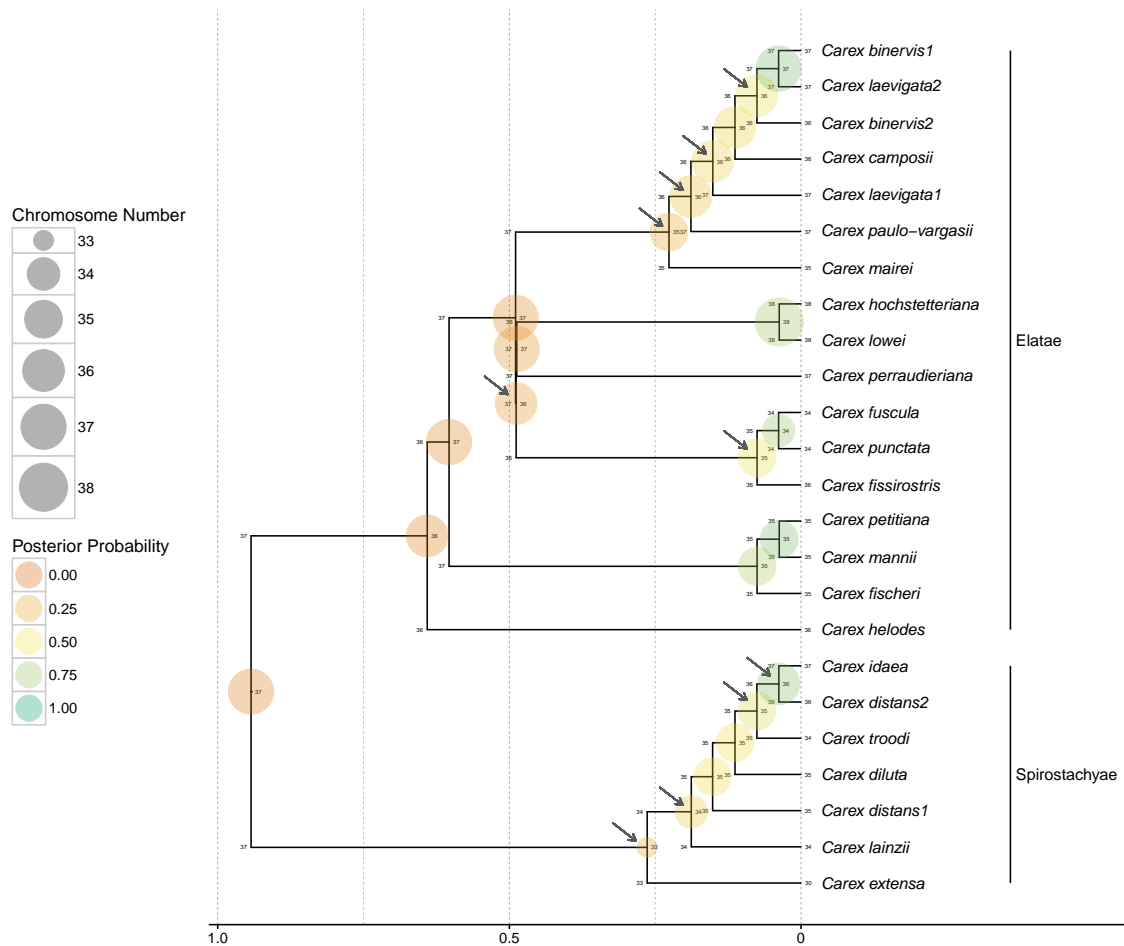


Figure 8: **Ancestral chromosome number estimates of *Carex* section *Spirostachyae*.** The model averaged MAP estimate of ancestral chromosome numbers are shown at each branch node. The states of each daughter lineage immediately after cladogenesis are shown at the “shoulders” of each node. The size of each circle is proportional to the chromosome number and the color represents the posterior probability. The MAP root chromosome number is 37 with a posterior probability of 0.08. Grey arrows indicate the location of possible dysploid speciation events. 36.9% of all speciation events include a cladogenetic gain or loss of a single chromosome. Clades corresponding to subsections are indicated at right.

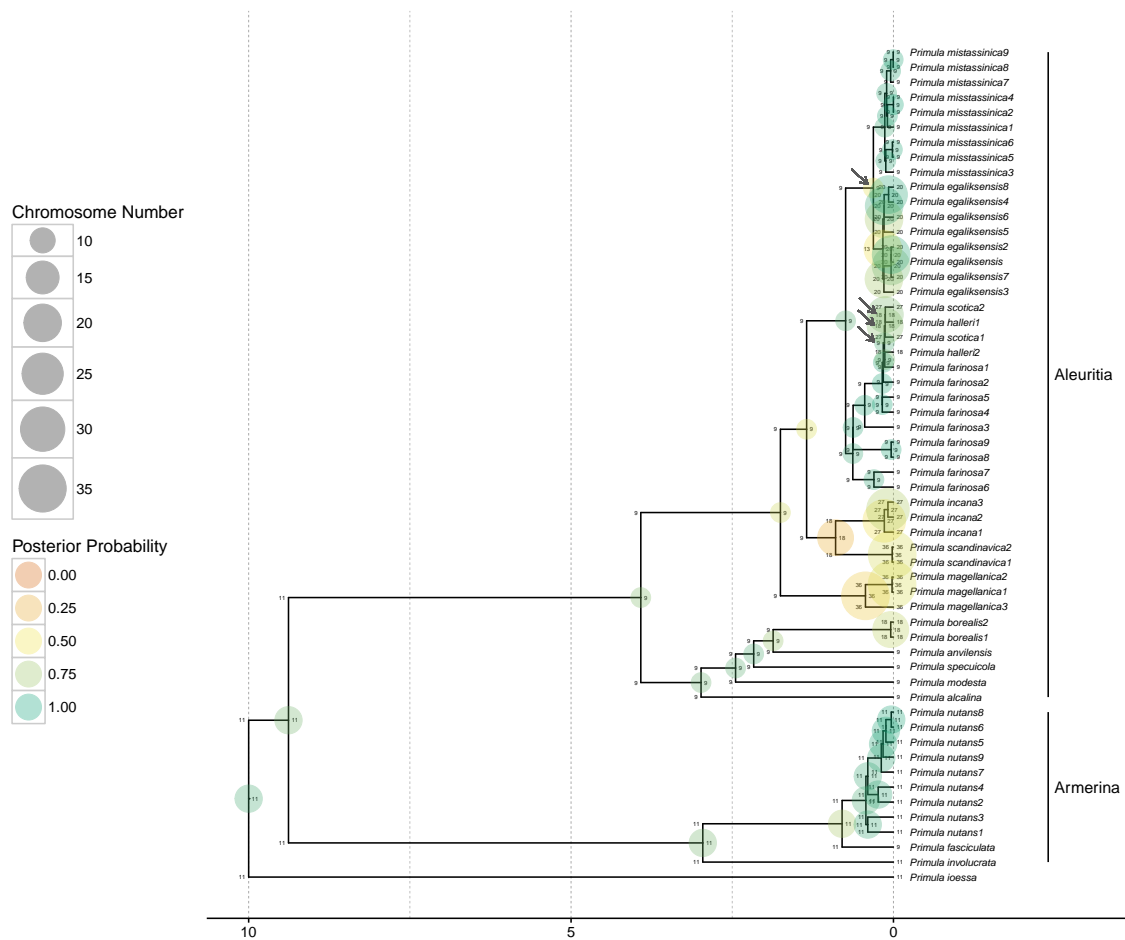


Figure 9: **Ancestral chromosome number estimates of *Primula* section *Aleuritia*.** The model averaged MAP estimate of ancestral chromosome numbers are shown at each branch node. The states of each daughter lineage immediately after cladogenesis are shown at the “shoulders” of each node. The size of each circle is proportional to the chromosome number and the color represents the posterior probability. The MAP root chromosome number of section *Aleuritia* is 9 with a posterior probability of 0.82. The arrows show the inferred history of possible polyploid and demi-polyploid speciation events in the clade containing the tetraploids *Primula egaliksensis* and *P. halleri* and the hexaploid *P. scotica*. Clades corresponding to sections are indicated at right.

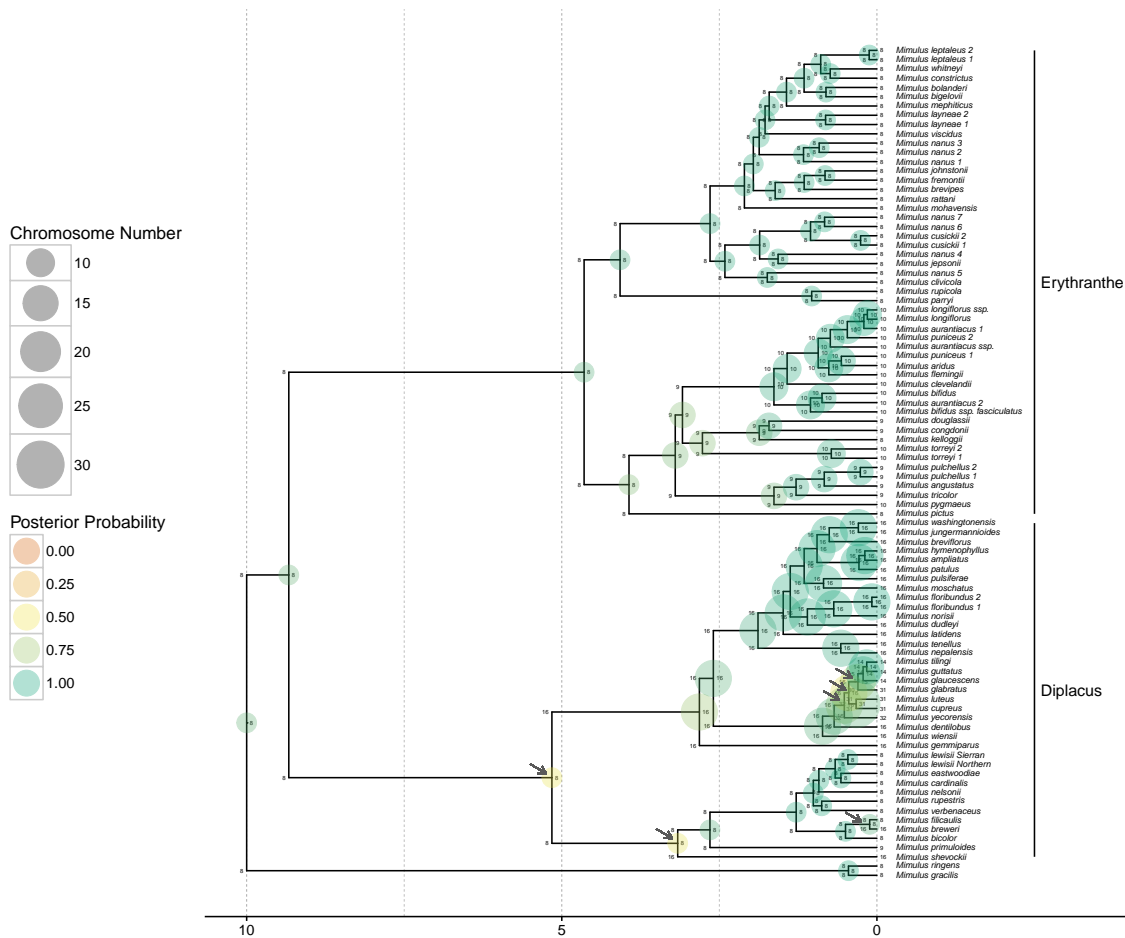


Figure 10: **Ancestral chromosome number estimates of *Mimulus sensu lato*.** The model averaged MAP estimate of ancestral chromosome numbers are shown at each branch node. The states of each daughter lineage immediately after cladogenesis are shown at the “shoulders” of each node. The size of each circle is proportional to the chromosome number and the color represents the posterior probability. The MAP root chromosome number is 8 with a posterior probability of 0.90. The arrows highlight the inferred history of repeated polyploid speciation events in the *Diplacus* clade, which contains the tetraploids *Mimulus cupreus*, *M. glabratus*, *M. luteus*, and *M. yecorensis*. Clades corresponding to segregate genera are indicated at right.

DISCUSSION

629

630 The results from the empirical analyses show that the ChromoSSE models
631 detect strikingly different modes of chromosome evolution with clade-specific
632 combinations of anagenetic and cladogenetic processes. Anagenetic dysploid gains
633 and losses were supported in nearly all clades; however, cladogenetic dysploid
634 changes were supported only in *Aristolochia* and *Carex*. The occurrence of
635 anagenetic dysploid changes in all clades suggest that small chromosome number
636 changes due to gains and losses may frequently have a minimal effect on the
637 formation of reproductive isolation, though our results suggest that *Carex* may be a
638 notable exception. Anagenetic polyploidization was only supported in *Aristolochia*,
639 while cladogenetic polyploidization was supported in *Helianthus*, *Mimulus* s.l., and
640 *Primula*. These findings confirm the evidence presented by Zhan et al. (2016) that
641 polyploidization events could play a significant role during plant speciation.

642 Our models shed new light on the importance of whole genome duplications
643 as a key driver in evolutionary diversification processes. *Helianthus* has long been
644 understood to have a complex history of polyploid speciation (Timme et al. 2007),
645 but our results here are the first to statistically show the prevalence of cladogenetic
646 polyploidization in *Helianthus* (occurring at 16% of all speciation events) and how
647 few of the chromosome changes are estimated to be anagenetic. Polyploid
648 speciation has also been suspected to be common in *Mimulus* s.l. (Vickery 1995),
649 and indeed we estimated that 7% of speciation events were cladogenetic
650 polyploidization events. We also estimated that the rates of cladogenetic
651 dysploidization in *Mimulus* s.l. were 0, which is in contrast to the parsimony based

652 inferences presented in Beardsley et al. (2004), which estimated 11.5% of all
653 speciation events included polyploidization and 13.3% included dysploidization.
654 Their estimates, however, did not distinguish cladogenetic from anagenetic
655 processes, and so they likely underestimated anagenetic changes. Our ancestral
656 state reconstructions of chromosome number evolution for *Helianthus*, *Mimulus* s.l.,
657 and *Primula* show that polyploidization events generally occurred in the relatively
658 recent past; few ancient polyploidization events were reconstructed (one exception
659 being the ancient cladogenetic polyploidization event in *Mimulus* clade *Diplacus*).
660 This pattern appears to be consistent with recent studies that show polyploid
661 lineages may undergo decreased net diversification (Mayrose et al. 2011; Scarpino
662 et al. 2014), leading some to suggest that polyploidization may be an evolutionary
663 dead-end (Arrigo and Barker 2012). While in the analyses presented here we fixed
664 rates of speciation and extinction through time and across lineages, an obvious
665 extension of our models would be to allow these rates to vary across the tree and
666 statistically test for rate changes in polyploid lineages.

667 Our findings also suggest dysploid changes may play a significant role in the
668 speciation process of some lineages. The genus *Carex* is distinguished by
669 holocentric chromosomes that undergo common fusion and fission events but rarely
670 polyploidization (Hipp 2007). This concurs with our findings from *Carex* section
671 *Spirostachyae*, where we saw no support for models including either anagenetic or
672 cladogenetic polyploidization. Instead we found high rates of cladogenetic dysploid
673 change, which is congruent with earlier results that show that *Carex* diversification
674 is driven by processes of fission and fusion occurring with cladogenetic shifts in

675 chromosome number (Hipp 2007; Hipp et al. 2007). Hipp (2007) proposed a
676 speciation scenario for *Carex* in which the gradual accumulation of chromosome
677 fusions, fissions, and rearrangements in recently diverged populations increasingly
678 reduce the fertility of hybrids between populations, resulting in high species
679 richness. More recently, Escudero et al. (2016) found that chromosome number
680 differences in *Carex scoparia* led to reduced germination rates, suggesting hybrid
681 dysfunction could spur chromosome speciation in *Carex*. Holocentricity has arisen
682 at least 13 times independently in plants and animals (Melters et al. 2012), thus
683 future work could examine chromosome number evolution in other holocentric
684 clades and test for similar patterns of cladogenetic fission and fusion events.

685 The models presented here could also be used to further study the role of
686 divergence in genomic architecture during sympatric speciation. Chromosome
687 structural differences have been proposed to perform a central role in sympatric
688 speciation, both in plants (Gottlieb 1973) and animals (Feder et al. 2005; Michel
689 et al. 2010). In *Aristolochia* we found most changes in chromosome number were
690 estimated to be anagenetic, with the only cladogenetic change occurring among a
691 pair of recently diverged sympatric species. By coupling our chromosome evolution
692 models with models of geographic range evolution it would be possible to
693 statistically test whether the frequency of cladogenetic chromosome changes
694 increase in sympatric speciation events compared to allopatric speciation events,
695 thereby testing for interaction between these two different processes of reproductive
696 isolation and evolutionary divergence.

697 The simulation results from Experiment 1 demonstrate that extinction

698 reduces the accuracy of inferences made by models of chromosome evolution that
699 do not take into account unobserved speciation events. Furthermore, the
700 simulations performed in Experiments 2 and 3 show that the substantial
701 uncertainty introduced in our analyses by jointly estimating diversification rates
702 and chromosome evolution resulted in lower posterior probabilities for ancestral
703 state reconstructions. We feel that this is a strength of our method; the lower
704 posterior probabilities incorporate true uncertainty due to extinction and so
705 represent more conservative estimates. Additionally, the simulation results from
706 Experiment 4 reveal that rates of anagenetic evolution were overestimated and
707 rates of cladogenetic change were underestimated when the generating process
708 consisted primarily of cladogenetic events. This suggests the possibility that our
709 models of chromosome number evolution are only partially identifiable, and that
710 the results of our empirical analyses may have a similar bias towards overestimating
711 anagenetic evolution and underestimating cladogenetic evolution. This bias may be
712 an issue for all ClaSSE type models, but the practical consequences here are
713 conservative estimates of cladogenetic chromosome evolution.

714 An important caveat for all phylogenetic methods is that estimates of model
715 parameters and ancestral states can be highly sensitive to taxon sampling (Heath
716 et al. 2008). All of the empirical datasets examined here included
717 non-monophyletic taxa that were treated as separate lineages. We made the
718 unrealistic assumptions that 1) each of the non-monophyletic lineages sharing a
719 taxon name have the same cytotype, and 2) the taxon sampling probability (ρ_s) for
720 the birth-death process was 1.0. The former assumption could drastically affect

721 ancestral state estimates, but its effect can only be confirmed by obtaining
722 chromosome counts for each lineage regardless of taxon name. While testing the
723 effect of incomplete taxon sampling on chromosome evolution inference was not a
724 goal of this work, analyses were performed with different values of ρ_s (results not
725 shown). The results indicated that speciation and extinction rates are sensitive to
726 ρ_s , but the relative speciation rates (e.g. between ϕ_c and γ_c) remained similar.
727 Thus, ancestral state estimates of cladogenetic and anagenetic chromosome changes
728 were robust to different values of ρ_s . This could vary among datasets and care
729 should be taken when considering which lineages to sample.

730 Bayesian model averaging is particularly appropriate for models of
731 chromosome number evolution since conditioning on a single model ignores the
732 considerable degree of model uncertainty found in both the simulations and the
733 empirical analyses. In the simulations the true model of chromosome evolution was
734 rarely inferred to be the MAP model (< 39% of replicates), and in the instances it
735 was correctly identified the posterior probability of the MAP model was < 0.13.
736 The posterior probabilities of the MAP models for the empirical datasets were
737 similarly low, varying between 0.04 and 0.22. Conditioning on a single poorly
738 fitting model of chromosome evolution, even when it is the best model available,
739 results in an underestimate of the uncertainty of ancestral chromosome numbers.
740 Furthermore, Bayesian model averaging enabled us to detect different modes of
741 chromosome number evolution without the limitation of traditional model testing
742 procedures in which multiple analyses are performed that each condition on a
743 different single model. This is a particularly useful approach when the space of all

744 possible models is large.

745 Our RevBayes implementation facilitates model modularity and easy
746 experimentation. Experimenting with different priors or MCMC moves is achieved
747 by simply editing the Rev scripts that describe the model. Though in our analyses
748 here we ignored phylogenetic uncertainty by assuming a fixed known tree, we could
749 easily incorporate this uncertainty by modifying a couple lines of the Rev script to
750 integrate over a previously estimated posterior distribution of trees. We could also
751 use molecular sequence data simultaneously with the chromosome models to jointly
752 infer phylogeny and chromosome evolution, allowing the chromosome data to help
753 inform tree topology and divergence times. In this paper we chose not to perform
754 joint inference so that we could isolate the behavior of the chromosome evolution
755 models; however, this is a promising direction for future research.

756 There are a number of challenging directions for future work on phylogenetic
757 chromosome evolution models. Models that incorporate multiple aspects of
758 chromosome morphology such as translocations, inversions, and other gene synteny
759 data as well as the presence of ring and/or B chromosomes have yet to be
760 developed. None of our models currently account for allopolyploidization; indeed
761 few phylogenetic comparative methods can handle reticulate evolutionary scenarios
762 that result from allopolyploidization and other forms of hybridization (Marcussen
763 et al. 2015). A more tractable problem is mapping chromosome number changes
764 along the branches of the phylogeny, as opposed to simply making estimates at the
765 nodes as we have done here. Since the approach described here models both
766 anagenetic and cladogenetic chromosome evolution processes while accounting for

767 unobserved speciation events, the rejection sampling procedure used in standard
768 stochastic character mapping (Nielsen 2002; Huelsenbeck et al. 2003) is not
769 sufficient. While data augmentation approaches such as those described by Bokma
770 (2008) could be utilized, they require complex MCMC algorithms that may have
771 difficulty mixing. Another option is to extend the method described in this paper
772 to draw joint ancestral states by numerically integrating root-to-tip over the tree
773 into a new procedure called joint conditional character mapping. This sort of
774 approach would infer the joint MAP history of chromosome changes both at the
775 nodes and along the branches of the tree, and provide an alternative to stochastic
776 character mapping that will work for all ClaSSE type models.

777 *Conclusions*

778 The analyses presented here show that the ChromoSSE models of
779 chromosome number evolution successfully infer different clade-specific modes of
780 chromosome evolution as well as the history of anagenetic and cladogenetic
781 chromosome number changes for a clade, including reconstructing the timing and
782 location of possible chromosome speciation events over the phylogeny. These
783 models will help investigators study the mode and history of chromosome evolution
784 within individual clades of interest as well as advance understanding of how
785 fundamental changes in the architecture of the genome such as whole genome
786 duplications affect macroevolutionary patterns and processes across the tree of life.

787 FUNDING

788 WAF was supported by a National Science Foundation Graduate Research
789 Fellowship under Grant DGE 1106400. SH was supported by the Miller Institute
790 for basic research in science. Analyses were computed using XSEDE, which is
791 supported by National Science Foundation grant number ACI-1053575, and the
792 Savio computational cluster provided by the Berkeley Research Computing
793 program at the University of California, Berkeley.

794 ACKNOWLEDGEMENTS

795 Thank you to Bruce Baldwin, Emma Goldberg, and Michael Landis for
796 valuable discussions that improved this work.

797 *

798 References

- 799 Akaike, H. 1974. A new look at the statistical model identification. IEEE
800 Transactions on Automatic Control 19:716–723.
- 801 Arrigo, N. and M. S. Barker. 2012. Rarely successful polyploids and their legacy in
802 plant genomes. Current Opinion in Plant Biology 15:140–146.
- 803 Ayala, F. J. and M. Coluzzi. 2005. Chromosome speciation: humans, *Drosophila*,
804 and mosquitoes. Proceedings of the National Academy of Sciences USA
805 102:6535–6542.

- 806 Beardsley, P. M., S. E. Schoenig, J. B. Whittall, and R. G. Olmstead. 2004.
807 Patterns of evolution in western North American *Mimulus* (Phrymaceae).
808 *American Journal of Botany* 91:474–489.
- 809 Beaulieu, J. M. and B. C. O’Meara. 2015. Extinction can be estimated from
810 moderately sized molecular phylogenies. *Evolution* 69:1036–1043.
- 811 Benson, D. A., I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and D. L. Wheeler.
812 2005. Genbank. *Nucleic Acids Research* 33:D34–D38.
- 813 Bokma, F. 2002. Detection of punctuated equilibrium from molecular phylogenies.
814 *Journal of Evolutionary Biology* 15:1048–1056.
- 815 Bokma, F. 2008. Detection of “punctuated equilibrium” by Bayesian estimation of
816 speciation and extinction rates, ancestral character states, and rates of anagenetic
817 and cladogenetic evolution on a molecular phylogeny. *Evolution* 62:2718–2726.
- 818 Colless, D. H. 1982. Review of phylogenetics: the theory and practice of
819 phylogenetic systematics. *Systematic Zoology* 31:100–104.
- 820 Conti, E., E. Suring, D. Boyd, J. Jorgensen, J. Grant, and S. Kelso. 2000.
821 Phylogenetic relationships and character evolution in *Primula* L.: the usefulness
822 of ITS sequence data. *Plant Biosystems* 134:385–392.
- 823 Coyne, J. A., H. A. Orr, et al. 2004. *Speciation*. Sinauer Associates Sunderland,
824 MA.

825 Dobzhansky, T. G. 1937. *Genetics and the Origin of Species*. Columbia University
826 Press.

827 Escudero, M., M. Hahn, B. H. Brown, K. Lueders, and A. L. Hipp. 2016.
828 Chromosomal rearrangements in holocentric organisms lead to reproductive
829 isolation by hybrid dysfunction: The correlation between karyotype
830 rearrangements and germination rates in sedges. *American Journal of Botany*
831 103:1529–1536.

832 Escudero, M., A. L. Hipp, and M. Luceño. 2010. Karyotype stability and predictors
833 of chromosome number variation in sedges: a study in *Carex* section
834 *Spirostachyae* (Cyperaceae). *Molecular Phylogenetics and Evolution* 57:353–363.

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

839 Feder, J. L., X. Xie, J. Rull, S. Velez, A. Forbes, B. Leung, H. Dambroski, K. E.
840 Filchak, and M. Aluja. 2005. Mayr, Dobzhansky, and Bush and the complexities
841 of sympatric speciation in *Rhagoletis*. *Proceedings of the National Academy of*
842 *Sciences USA* 102:6573–6580.

843 Felsenstein, J. 1981. Evolutionary trees from dna sequences: a maximum likelihood
844 approach. *Journal of Molecular Evolution* 17:368–376.

- 845 FitzJohn, R. G. 2012. Diversitree: comparative phylogenetic analyses of
846 diversification in R. *Methods in Ecology and Evolution* 3:1084–1092.
- 847 Glick, L. and I. Mayrose. 2014. Chromevol: assessing the pattern of chromosome
848 number evolution and the inference of polyploidy along a phylogeny. *Molecular*
849 *Biology and Evolution* 31:1914–1922.
- 850 Goldberg, E. E. and B. Igić. 2012. Tempo and mode in plant breeding system
851 evolution. *Evolution* 66:3701–3709.
- 852 Gottlieb, L. D. 1973. Genetic differentiation, sympatric speciation, and the origin of
853 a diploid species of *Stephanomeria*. *American Journal of Botany* Pages 545–553.
- 854 Green, P. J. 1995. Reversible jump Markov chain Monte Carlo computation and
855 Bayesian model determination. *Biometrika* 82:711–732.
- 856 Guggisberg, A., G. Mansion, and E. Conti. 2009. Disentangling reticulate evolution
857 in an arctic–alpine polyploid complex. *Systematic Biology* 58:55–73.
- 858 Hastings, W. K. 1970. Monte Carlo sampling methods using Markov chains and
859 their applications. *Biometrika* 57:97–109.
- 860 Heath, T. A., S. M. Hedtke, and D. M. Hillis. 2008. Taxon sampling and the
861 accuracy of phylogenetic analyses. *Journal of Systematics and Evolution*
862 46:239–257.
- 863 Hipp, A. L. 2007. Nonuniform processes of chromosome evolution in sedges (*Carex*:
864 *Cyperaceae*). *Evolution* 61:2175–2194.

- 865 Hipp, A. L., P. E. Rothrock, A. A. Reznicek, and P. E. Berry. 2007. Chromosome
866 number changes associated with speciation in sedges: a phylogenetic study in
867 *Carex* section *Ovales* (Cyperaceae) using AFLP data. *Aliso: A Journal of*
868 *Systematic and Evolutionary Botany* 23:193–203.
- 869 Hoeting, J. A., D. Madigan, A. E. Raftery, and C. T. Volinsky. 1999. Bayesian
870 model averaging: a tutorial. *Statistical Science* 14:382–401.
- 871 Höhna, S. 2015. The time-dependent reconstructed evolutionary process with a
872 key-role for mass-extinction events. *Journal of Theoretical Biology* 380:321–331.
- 873 Höhna, S., T. A. Heath, B. Boussau, M. J. Landis, F. Ronquist, and J. P.
874 Huelsenbeck. 2014. Probabilistic graphical model representation in phylogenetics.
875 *Systematic Biology* 63:753–771.
- 876 Höhna, S., M. J. Landis, T. A. Heath, B. Boussau, N. Lartillot, B. R. Moore, J. P.
877 Huelsenbeck, and F. Ronquist. 2016. RevBayes: Bayesian phylogenetic inference
878 using graphical models and an interactive model-specification language.
879 *Systematic Biology* 65:726–736.
- 880 Huelsenbeck, J. P. and J. P. Bollback. 2001. Empirical and hierarchical Bayesian
881 estimation of ancestral states. *Systematic Biology* 50:351–366.
- 882 Huelsenbeck, J. P., B. Larget, and D. L. Swofford. 2000. A compound Poisson
883 process for relaxing the molecular clock 154:1879–1892.
- 884 Huelsenbeck, J. P., R. Nielsen, and J. P. Bollback. 2003. Stochastic mapping of
885 morphological characters. *Systematic Biology* 52:131–158.

- 886 Kass, R. E. and A. E. Raftery. 1995. Bayes factors. *Journal of the American*
887 *Statistical Association* 90:773–795.
- 888 Landis, M. J. in press. Biogeographic dating of speciation times using
889 paleogeographically informed processes. *Systematic Biology* .
- 890 Landis, M. J., N. J. Matzke, B. R. Moore, and J. P. Huelsenbeck. 2013. Bayesian
891 analysis of biogeography when the number of areas is large. *Systematic Biology*
892 62:789–804.
- 893 Maddison, W. P., P. E. Midford, and S. P. Otto. 2007. Estimating a binary
894 character's effect on speciation and extinction. *Systematic Biology* 56:701–710.
- 895 Madigan, D. and A. E. Raftery. 1994. Model selection and accounting for model
896 uncertainty in graphical models using Occam's window. *Journal of the American*
897 *Statistical Association* 89:1535–1546.
- 898 Marcussen, T., L. Heier, A. K. Brysting, B. Oxelman, and K. S. Jakobsen. 2015.
899 From gene trees to a dated allopolyploid network: insights from the angiosperm
900 genus *Viola* (Violaceae). *Systematic Biology* 64:84–101.
- 901 May, M. R., S. Höhna, and B. R. Moore. 2016. A Bayesian approach for detecting
902 the impact of mass-extinction events on molecular phylogenies when rates of
903 lineage diversification may vary. *Methods in Ecology and Evolution* 7:947–959.
- 904 Mayrose, I., M. S. Barker, and S. P. Otto. 2010. Probabilistic models of
905 chromosome number evolution and the inference of polyploidy. *Systematic*
906 *Biology* 59:132–144.

- 907 Mayrose, I., S. H. Zhan, C. J. Rothfels, K. Magnuson-Ford, M. S. Barker, L. H.
908 Rieseberg, and S. P. Otto. 2011. Recently formed polyploid plants diversify at
909 lower rates. *Science* 333:1257–1257.
- 910 Melters, D. P., L. V. Paliulis, I. F. Korf, and S. W. Chan. 2012. Holocentric
911 chromosomes: convergent evolution, meiotic adaptations, and genomic analysis.
912 *Chromosome Research* 20:579–593.
- 913 Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller.
914 1953. Equation of state calculations by fast computing machines. *The Journal of*
915 *Chemical Physics* 21:1087–1092.
- 916 Michel, A. P., S. Sim, T. H. Powell, M. S. Taylor, P. Nosil, and J. L. Feder. 2010.
917 Widespread genomic divergence during sympatric speciation. *Proceedings of the*
918 *National Academy of Sciences USA* 107:9724–9729.
- 919 Nee, S., E. C. Holmes, R. M. May, and P. H. Harvey. 1994a. Extinction rates can
920 be estimated from molecular phylogenies. *Philosophical Transactions of the*
921 *Royal Society B: Biological Sciences* 344:77–82.
- 922 Nee, S., R. M. May, and P. H. Harvey. 1994b. The reconstructed evolutionary
923 process. *Philosophical Transactions of the Royal Society B: Biological Sciences*
924 344:305–311.
- 925 Nielsen, R. 2002. Mapping mutations on phylogenies. *Systematic Biology*
926 51:729–739.

- 927 Ohi-Toma, T., T. Sugawara, H. Murata, S. Wanke, C. Neinhuis, and J. Murata.
928 2006. Molecular phylogeny of *Aristolochia sensu lato* (Aristolochiaceae) based on
929 sequences of *rbcL*, *matK*, and *phyA* genes, with special reference to
930 differentiation of chromosome numbers. *Systematic Botany* 31:481–492.
- 931 Pagel, M. and A. Meade. 2006. Bayesian analysis of correlated evolution of discrete
932 characters by reversible-jump Markov chain Monte Carlo. *The American*
933 *Naturalist* 167:808–25.
- 934 Pagel, M., A. Meade, and D. Barker. 2004. Bayesian estimation of ancestral
935 character states on phylogenies. *Systematic Biology* 53:673–684.
- 936 Pires, J. C. and K. L. Hertweck. 2008. A renaissance of cytogenetics: Studies in
937 polyploidy and chromosomal evolution. *Annals of the Missouri Botanical Garden*
938 95:275–281.
- 939 Posada, D. and T. R. Buckley. 2004. Model selection and model averaging in
940 phylogenetics: advantages of Akaike information criterion and Bayesian
941 approaches over likelihood ratio tests. *Systematic Biology* 53:793–808.
- 942 Pupko, T., I. Pe, R. Shamir, and D. Graur. 2000. A fast algorithm for joint
943 reconstruction of ancestral amino acid sequences. *Molecular Biology and*
944 *Evolution* 17:890–896.
- 945 Rabosky, D. L. 2010. Extinction rates should not be estimated from molecular
946 phylogenies. *Evolution* 64:1816–1824.

- 947 Ree, R. H. and S. A. Smith. 2008. Maximum likelihood inference of geographic
948 range evolution by dispersal, local extinction, and cladogenesis. *Systematic*
949 *Biology* 57:4–14.
- 950 Rieseberg, L. H. and J. H. Willis. 2007. Plant speciation. *Science* 317:910–914.
- 951 Rodriguez, F., J. Oliver, A. Marin, and J. R. Medina. 1990. The general stochastic
952 model of nucleotide substitution. *Journal of theoretical biology* 142:485–501.
- 953 Scarpino, S. V., D. A. Levin, and L. A. Meyers. 2014. Polyploid formation shapes
954 flowering plant diversity. *The American Naturalist* 184:456–465.
- 955 Schwarz, G. 1978. Estimating the dimension of a model. *The Annals of Statistics*
956 6:461–464.
- 957 Stebbins, G. L. 1971. *Chromosomal evolution in higher plants*. Edward Arnold
958 Ltd., London.
- 959 Tank, D. C., J. M. Eastman, M. W. Pennell, P. S. Soltis, D. E. Soltis, C. E.
960 Hinchliff, J. W. Brown, E. B. Sessa, and L. J. Harmon. 2015. Nested radiations
961 and the pulse of angiosperm diversification: increased diversification rates often
962 follow whole genome duplications. *New Phytologist* 207:454–467.
- 963 Tavaré, S. 1986. Some probabilistic and statistical problems in the analysis of DNA
964 sequences. In: *Some Mathematical Questions in Biology—DNA Sequence*
965 *Analysis*, Miura RM (Ed.), American Mathematical Society, Providence (RI)
966 17:57–86.

- 967 Timme, R. E., B. B. Simpson, and C. R. Linder. 2007. High-resolution phylogeny
968 for *Helianthus* (Asteraceae) using the 18S-26S ribosomal DNA external
969 transcribed spacer. *American Journal of Botany* 94:1837–1852.
- 970 Vickery, R. K. 1995. Speciation by aneuploidy and polyploidy in *Mimulus*
971 (*Scrophulariaceae*). *The Great Basin Naturalist* 55:174–176.
- 972 Vos, R. A., H. Lapp, W. H. Piel, and V. Tannen. 2010. Treebase2: rise of the
973 machines .
- 974 White, M. J. D. 1978. Modes of speciation. San Francisco: WH Freeman
975 455p.-Illus., maps, chrom. nos.. General (KR, 197800185).
- 976 Xie, W., P. O. Lewis, Y. Fan, L. Kuo, and M.-H. Chen. 2011. Improving marginal
977 likelihood estimation for Bayesian phylogenetic model selection. *Systematic*
978 *Biology* 60:150–60.
- 979 Yang, Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences
980 with variable rates over sites: approximate methods. *Journal of Molecular*
981 *Evolution* 39:306–314.
- 982 Zhan, S. H., M. Drori, E. E. Goldberg, S. P. Otto, and I. Mayrose. 2016.
983 Phylogenetic evidence for cladogenetic polyploidization in land plants. *American*
984 *Journal of Botany* 103:1252–1258.