Statistical Inference of Allopolyploid Species Networks in the Presence of Incomplete Lineage Sorting

Graham Jones1,2,*, Sereb Sagitov2, and Bengt Oxelman1

1Department of Biological and Environmental Sciences, University of Gothenburg; and 2Mathematical Sciences, Chalmers University of Technology and the University of Gothenburg, Gothenburg, Sweden

*Correspondence to be sent to: 21e Balnakeil, Durness, Lairg, Sutherland IV27 4PT, UK; E-mail: art@gjones.name.

Received 23 November 2012; reviews returned 30 January 2013; accepted 12 February 2013

Associate Editor: Cédile Ant

Abstract.—Polyploidy is an important speciation mechanism, particularly in land plants. Allopolyploid species are formed after hybridization between otherwise intersterile parental species. Recent theoretical progress has led to successful implementation of species tree models that take population genetic parameters into account. However, these models have not included allopolyploid hybridization and the special problems imposed when species trees of allopolyploids are inferred. Here, 2 new models for the statistical inference of the evolutionary history of allopolyploids are evaluated using simulations and demonstrated on 2 empirical data sets. It is assumed that there has been a single hybridization event producing a species resulting in a genomic allotetraploid. The evolutionary history can be represented as a species network or as a multilabeled species tree in which some pairs of tips are labeled with the same species. In one of the models (AlloppMUL), the multilabeled species tree is inferred directly. This is the simplest model and the most widely applicable, since fewer assumptions are made. The second model (AlloppNET) incorporates the hybridization event explicitly which means that fewer parameters need to be estimated. Both models are implemented in the BEAST framework. Simulations show that both models are useful and that AlloppNET is more accurate if the assumptions it is based on are valid. The models are demonstrated on previously analyzed data from the genera Pachycladon (Brassicaceae) and Silene (Caryophyllaceae).

Polyploidy is an important mechanism for the emergence of new species, which is particularly prominent in plants. Allopolyploid species are formed after hybridization between two species and are considerably more common than autopolyploids, which are formed within species (Tate et al. 2005). Hybridization presents a challenge to phylogenetic analysis because the usual species tree is replaced by a species network. In addition, it becomes difficult to assign genome identities to allelic copies. This imposes a problem for the inference of species trees in a relevant multispecies coalescent framework (Rannala and Yang 2003) even if the hybridization event is ignored. This article explores the feasibility of making statistical inferences about the evolutionary history of allopolyploids using simulations of some simple scenarios and two novel models implemented in the BEAST software (Drummond and Rambaut 2007). The main restrictions made here are that there has been a single hybridization event between two diploids, and that the resulting hybrid is a genomic allopolyploid, in which the two diploid genomes (from the two parental diploid species) do not recombine with one another at meiosis because the chromosomes in the two parental species were too divergent by the time the hybrid formed. This leaves two major problems to deal with in the phylogenetic analysis. First, when the DNA from organisms is sequenced, it is not possible to assign sequences to their parental diploid species. Thus, although the sequences can be seen as the result of the evolution of diploid genomes, there is an ambiguity in the labeling of the sequences which is not normally present. Second, the issue of incomplete lineage sorting cannot be ignored. The latter problem is appropriately handled by approaches taking coalescent processes into account, and it has been shown both theoretically and empirically that these are superior to simple concatenation of gene alignments. In our case, concatenation is not even an option, because there is no way to tell a priori which sequences belong to the same homeologous genome (Huber and Moulton 2006; Lott et al. 2009b).

Figure 1 shows three ways of viewing the same evolutionary events. The three columns show three main scenarios labeled A, B, and C in which the allotetraploid could have arisen. In all scenarios, a speciation at the root produces two diploids a and b, and at some point later, a hybridization occurs between a and b or one or two of their extinct relatives. After hybridization, the tetraploid speciates to produce two species y and z. The sections from top to bottom show three different ways of viewing each of these scenarios. In the top section, hybridization and extinction events are explicitly represented. Note that more than one sequence of evolutionary events (speciation, extinction, and hybridization) can correspond to the same representations in the other two views. In the second section, the networks are represented as a collection of homoploid “trees with legs.” We use the term “leg” to mean a particular kind of edge in a species network involving diploids and allotetraploids. Each allotetraploid subtree has a pair of legs which lead from the past to the hybridization event which is the origin of the subtree. The term “hybridization edge” is sometimes used to denote such an edge, although that usually refers...
FIGURE 1. The 3 columns show three main Scenarios A, B, and C. In each one there are two diploid species a and b, and two tetraploid species y and z. See main text for more details.

The first view contains the most information, but in general it is difficult or impossible to infer the extra details represented in it. The methods of this article cannot be used to distinguish whether one or two extinct species hybridized to produce the allotetraploid. It might be possible in Scenario A to infer whether there was a direct hybridization between the two extant diploids, or whether extinct (or unsampled) species were involved, using estimates of node heights. It might also be possible to estimate the hybridization time from population sizes, if one assumes that the allotetraploid species arose from a single individual. However, it is difficult to obtain good estimates of ages and population sizes. In this article we focus on inferring the level of detail in the “trees with legs” or MUL-tree views.

The data available for inferring the species history consist of molecular sequences sampled from individuals belonging to species. One approach, pursued in Huber and Moulton (2006), Huber et al. (2006), and Lott et al. (2009b) and implemented in the software PADRE (Lott et al. 2009a), is to estimate the multilabeled gene trees first, and then search for the network that best accommodates them. Lott et al. (2009b) introduced an algorithm to make consensus MUL-trees of several input gene trees.

Various methods have been developed for detecting hybridization in the presence of lineage sorting, including Kubatko (2009), Gerard et al. (2011), and Yu et al. (2012). These are aimed at homoploid hybridization, where for example two diploid species form a third diploid species by hybridization. In this
situation, each allele is supposed to come from one or the other of the parental species, and a probability is assigned to each of these possibilities. Although it should be possible to use these methods on data from allopolyploids, they would not take into account the particular pattern that occurs in allopolyploids, where homeologs from all the parental species may be present in the hybrid species. The same limitation applies to hybridization networks (e.g., Chen and Wang 2010, 2012). All these approaches rely on gene trees as input data, thus uncertainty in their estimation is not taken into account. To overcome this, Cai et al. (2012) devised a strategy where a backbone species tree of diploids is first made using *BEAST, and then each polyploid allele is analyzed individually (with bootstrapping) to find its location on the backbone tree.

The approach here is the typically Bayesian one of “coestimating everything,” including DNA substitution parameters. The network node times and topology, the assignment of sequences obtained from tetraploid individuals to parental diploid species, and the node times and topologies of all the gene trees are all allowed to vary, and an MCMC algorithm is used to sample from the posterior distribution. The approach is similar to that of *BEAST as described in Heled and Drummond (2010) but the sequence assignment ambiguity is new.

### The Models

It is assumed that at some point in the past, a diploid species speciated to form two diploid species which both have survived to the present. The initial speciation forms the root of the network or MUL-tree. At some later time, a hybridization took place between these two diploid species or their extinct relatives, forming an allotetraploid in which the two parental diploid genomes continue to evolve without recombining with one another. After the hybridization, further speciation of the allotetraploid may have taken place. The models only allow for data sampled from two diploid species. Thus, data sets for these models have two diploid species, any number of tetraploid species, and the tetraploid species are all assumed to have descended from a single hybridization event.

Two models, denoted as AlloppMUL and AlloppNET, are considered. Generally, AlloppMUL models event at the level of the MUL-tree while AlloppNET models trees with legs. They are both based on the multispecies coalescent (Rannala and Yang 2003). It is assumed that there is free recombination between genes, but no recombination within genes. Although the assumption of no recombination within genes may well be unrealistic, simulations conducted in Lanier and Knowles (2012) suggest that this violation of the model does not pose a major problem. Also as in Heled and Drummond (2010), the term species “is not necessarily the same as a taxonomic rank, but designates any group of individuals that after some ‘divergence’ time, have no history of breeding with individuals outside that group.”

In the usual formulation of the multispecies coalescent, there is a species tree and a number of gene trees, and it is assumed that individuals can be assigned unambiguously to tips in the species tree, and that molecular sequences can also be assigned unambiguously to tips in the species tree. If the first assumption is relaxed, so that the assignment of individuals to species must be estimated, the problem is known as species delimitation and several approaches have been developed for this (Fujita et al. 2012). In particular, Yang and Rannala (2010) use a Bayesian approach to species delimitation. In the models considered here, the second assumption is relaxed, to cater for the ambiguity in assigning multiple sequences from the same individual to tips in the multilabeled species tree.

In the first model (AlloppMUL), the multilabeled species tree is inferred directly. The topology, the node times, and the population sizes along the branches are all allowed to vary freely, as if the diploid genomes within the allotetraploid(s) belonged to different species. This approach therefore throws away some information implicit in the assumptions. The two main advantages of this approach are that it may be more appropriate where the assumptions are dubious, especially when the number of hybridization events is not known, and the simplicity of implementation (since it is a relatively straightforward generalization of the implementation of the multispecies coalescent model in *BEAST).

The second model (AlloppNET) is more faithful to evolutionary events. The hybridization is modeled explicitly as a node in a species network, and from that time, the diploid genomes within the allotetraploid(s) must share population sizes and speciation events. Since this uses more information, it is expected to be more accurate. The network can be converted into a MUL-tree for calculations and for program output. The key point is that since the MUL-tree is derived from the network, the appropriate constraints on the topology, node times, and populations are enforced onto the MUL-tree.

#### The AlloppMUL Model

The posterior density for the AlloppMUL model is given by

\[
f(M, \theta, \tau, a, \gamma | d) \propto f_M (\lambda) f_{\theta}(\lambda) \times f_\tau (\tau_1) \times f_a(a) \times f_\gamma (\gamma_1) \times \prod_{i=1}^G \Pr(d_i | \tau_1, a_i). \tag{1}
\]

Here, the multilabeled species tree is denoted by \( M \), and the parameter(s) for the topology and node times in the prior for \( M \) are denoted by \( \lambda \). The population size parameters are denoted by the vector \( \theta \). The parameter \( \gamma \) is a scaling factor for the population sizes, appearing in
a hyperprior for $\theta$. The number of gene trees is denoted by $G$. The topology and set of node times for the $i$th gene tree is denoted by $t_i (1 \leq i \leq G)$. All the other parameters belonging to the $i$th gene tree are denoted by $a_i$; these are parameters for site rate heterogeneity, substitution model, branch rate model, and root model. Thus, $f(t_i, a_i)$ gives all the parameters for the $i$th gene tree. The permutations of sequences within polyploid individuals for the $i$th gene are denoted by $\gamma_i$. This parameter is the main addition to the usual formula for the multispecies coalescent. We only deal with tetraploids here, so $\gamma_i$ consists of transpositions (flips) of two sequences. The sequence data for the $i$th gene consists of transpositions (flips) of two sequences. The sequence data for the $i$th gene is denoted by $d_i$. We set $\tau = (t_1, \ldots , t_G)$, and similarly for $a$, $\gamma$, and $d$. The 5 terms in this expression will now be described in detail.

- The species tree prior $f_M(M|\lambda , \rho , \gamma)$ provides the probability of $M$ before seeing any molecular data. Little is known about what an appropriate prior should be. For the analyses in this article, $M$ is regarded as an ordinary tree, and a Yule prior is used for this. The single parameter $\lambda$ represents the birth rate. The hyperprior $f_\lambda$ for $\lambda$ is described later.

- The population size prior $f_\theta(\theta|\eta)f_\eta(\eta)$ is for the population size parameters $\theta$. There is one value at each tip, and one at the root-ward end of each branch in the MUL-tree. In the analyses done in this article, the priors for $\theta$ used were similar to those typically used by *BEAST*. An independent gamma distribution is assumed for each population size. The shape parameter is 4 for the populations at the tips and 2 for the rest. If it is assumed that the total population just before and just after a speciation is the same, then at tip-ward end of an internal branch, the population size is the sum of the two independent random variables each having a gamma distribution with shape parameter 2, and thus has a gamma distribution with shape parameter 4, like the tips. The scale parameter for all these gamma distributions is the hyperparameter $\eta$. Under this prior, and for a given value of $\eta$, the tip population sizes have an expectation of $4\eta$, and the rest have an expectation of $2\eta$. The population sizes are assumed to vary linearly along edges in the network, between the nodes where the population size parameters occur. The hyperprior $f_\eta$ for $\eta$ is described later.

- The permutation prior $f_\gamma(\gamma)$ is a discrete distribution on the set of sequence assignments. This is assumed to be uniform here, and thus could be omitted without affecting the inference.

- The term $f(t_i|M, \theta, \gamma)$ provides the distribution of $t_i$, when permuted by $\gamma_i$, fitting into the species tree $M$ with populations sizes determined by $\theta$. The value of $\gamma_i$ determines how the sequences for the $i$th gene are assigned to tips in the multilabeled tree species $M$. Note that this probability does not depend on $a_i$. Apart from this extra complexity due to the permutations, the value of $f(t_i|M, \theta, \gamma)$ is given by the multispecies coalescent, as used in Rannala and Yang (2003), Heled and Drummond (2010) and elsewhere. The formula is described in more detail in the Appendix.

- The term $\Pr(d_i|t_i, a_i)$ is the probability of the data for the $i$th gene given the $i$th gene tree and other parameters $a_i$. Regarded as a likelihood, it is the usual “Felsenstein likelihood.” The parameter $a_i$ is described in more detail in the subsection “Other parts of the models.” It may be helpful to think about this gene tree likelihood and the previous term $f(t_i|M, \theta, \gamma)$ in another way. One can think of the $\gamma_i$ as permuting the sequence data $d_i$, that is, swapping pairs of rows in a data matrix for the $i$th gene, where the pairs each consist of two homeologs from the same tetraploid individual. Then, for each $i$, the product $\Pr(d_i|t_i, a_i) \propto f_{fW}(t_i|M, \theta, \gamma)$ where $\gamma_i$ assigns sequences to tips in $M$ is replaced by $\Pr(d_i|t_i, a_i) \propto f_W(t_i|M, \theta)$ where $\gamma_i$ is now thought of as swapping rows in the data matrix. This is mathematically equivalent but does not work well in implementation.

The AlloppNET Model

The formula for the posterior density for the AlloppNET model is similar to that for AlloppMUL and is given by

$$f(W, \theta, \tau, \alpha, \gamma|d) \propto f_W(W|\lambda)f_\lambda(\lambda) \times f_\theta(\theta|\eta)f_\eta(\eta) \times f_\gamma(\gamma) \times \prod_{i=1}^{G} \left[ f(t_i|M_W, \theta, \gamma) \times \Pr(d_i|t_i, a_i) \right].$$

(2)

The network is denoted by $W$ and the multilabeled species tree derived from it is $M_W$. The other parameters are similar to those appearing in Equation (1), but the meanings of $\lambda$ and $\theta$ are somewhat different. The terms for the permutation prior and the gene tree likelihood are as before, but the models differ in the meaning of the other terms, as described next.

- The network prior is $f_W(W|\lambda)f_\lambda(\lambda)$, and again, little is known about what an appropriate prior should be. The prior used here was designed using the “trees with legs” representation. Thus, there is a diploid species tree with two tips and unknown age, a tetraploid species subtree with known age equal to the hybridization time, and the two legs. The priors for the diploid tree and the tetraploid subtree both use a birth–death model (Gernhard 2008) with the ratio of extinction rate to speciation rate fixed at 0.8, so that these
by Mau et al. (1999) was used to make changes within the tetraploid subtree. New operators were then designed to change the heights of the nodes at the bottom (most ancient) ends of the legs; to change the topology of the legs; to change the hybridization time; and to change the time of the diploid root. The move that changes the topology of the legs samples directly from the prior described in the subsection “The AlloppNET model.” The assumption of a single hybridization and just two diploids makes these operators quite simple, since they are only required to explore the topologies of what is essentially a tree with four tips. The same set of operators was used for all simulations and analyses of empirical data. The weights of the operators were chosen automatically using a formula which takes into account the number of genes, with the aim of spending a reasonable proportion of time exploring different gene trees versus different species networks.

Simulations and Empirical Data

Scenarios Used in Simulations

Six scenarios, as shown in Figure 2 were used to simulate DNA sequences. Each scenario represents a "true" MUL-tree. Heights are in units of expected substitutions per site. Population sizes are effective numbers of gene copies within diploid populations (twice the number of individuals), or numbers of gene copies with the same diploid parent, for allotetraploid populations. If the effective population size is $S$, the probability of coalescence between a pair of gene copies is $1/S$ per generation. Population sizes are 100,000 at tips, and at root-ward ends of branches, and 200,000 at tip-ward ends of internal branches and at the root. All genes have length 500.

These six scenarios were each tested with the number of genes $G$ equal to 1, 3, and 9, and the number of individuals $N$ per species equal to 1 and 3. The mutation rates $T$ were set to each of $4 \times 10^{-8}$, $8 \times 10^{-8}$, and $1.6 \times 10^{-7}$ for Scenarios A1, B1, C1, and to each of $4 \times 10^{-8}$ and $8 \times 10^{-8}$ for Scenarios A3, B3, and C3. The $T$ values are in expected substitutions per site per generation. Scenarios A1, B1, and C1 have a root height of 0.025. Scenarios A3, B3, and C3 have a root height of 0.03. Changing $T$ while keeping this height fixed changes the number of generations the tree represents. For example, $T = 4 \times 10^{-8}$ in Scenarios A3, B3, and C3 means $0.03/(4 \times 10^{-8}) = 750,000$ generations root to tip. In general, increasing $G$ and $N$ is expected to increase accuracy since there is more data, whereas increasing $T$ is expected to decrease accuracy since incomplete lineage sorting becomes more common.

Implementation of Simulations

The simulations and the analyses of results were implemented in R (R Development Core Team 2008). The
input scenarios, as shown in Figure 2 were converted to a MUL-tree, then gene tree topologies and coalescent times were simulated according to the coalescent model within branches. The sequences were then generated using Seq-Gen (Rambaut and Grassly 1997). One sequence was generated for each diploid individual, and two homeologous sequences were generated for each tetraploid individual. For one gene and a single individual per species, this amounts to one sequence for each tip of the MUL-tree.

Strict clock branch rates and no site rate heterogeneity were assumed in both the simulations and the inference. In the simulations, equal clock rates for all genes were assumed, whereas in the inference, the clock rate for one gene was fixed to 1.0, and the others were estimated. The HKY substitution model was assumed in simulations and the inference. In the simulations, the substitution model parameter $\theta$ was set to 3, and the frequencies were set to 0.3 for A and T, and 0.2 for C and G (Seq-Gen was called with parameters `-t3.0 -f0.3,0.2,0.2,0.3`). These were estimated in the inference. For the inference, the priors for the population scaling factor $\eta$ and the parameter $\lambda$ appearing in the network prior were the “OneOnX” distribution, which is improper, and the priors for relative clock rates were a very diffuse gamma distribution with mean 1 and shape parameter 0.1. These are the default priors in *BEAST. Note that improper priors are not recommended in general, since they can cause additional convergence problems, and may even result in the posterior becoming improper. However, our concern here is to assess the accuracy of the method, so using “unhelpful” priors seemed appropriate.

BEAST XML files were generated containing the simulated sequences for the AlloppNET and AlloppMUL models. There were 18 values for the triple (G, N, and T) for Scenarios A1, B1, and C1 and 12 for Scenarios A3, B3, and C3, making $54 + 36 = 90$ configurations in total. For each of these, 20 replicates were simulated and run for 1.5 million generations in BEAST using both models, making a total of $90 \times 2 \times 20 = 3600$ BEAST runs. MUL-trees were sampled every 1000 generations, and the first 501 samples (of 1501) discarded as burn-in. These numbers were chosen because they appeared sufficient to ensure convergence of the MCMC algorithm in most cases, but convergence was not tested beyond that. However, since the truth is known in simulations, the reported accuracy includes any failure to converge. For all six scenarios, 3 values for G, and 2 for N were used.

**Empirical Data**

The empirical data were analyzed using very similar assumptions to the simulated data. Strict clock branch rates and no site rate heterogeneity were assumed. The clock rate for one gene was fixed to 1.0, and for the others were estimated. The HKY substitution model was used. For the *Silene* data, the priors for $\eta$ and $\lambda$ and the priors for relative clock rates were the default priors used in *BEAST. For the *Pachycladon* data, these priors were changed, as described later.

Two sets of empirical data were analyzed. In both cases, at most one sequence from each diploid individual and two sequences from each tetraploid individual were available. In the latter case, these were assumed to be homeologs. Convergence of the MCMC algorithm was tested by visual comparison of the results of multiple independent runs with different random seeds.

The first data set comes from a study (Joly et al. 2009) of the genus *Pachycladon* (Brassicaceae) which consists of 8 polyploid species and a number of diploids. This study showed that the *Pachycladon* genus originated from an allopolyploidization which is estimated to have occurred...
FIGURE 3. Results for simulated Scenarios A1, B1, and C1. Results for the three scenarios are combined for each graph, so the maximum possible number of mismatches is 60. The number of generations root to tip is shown at the top of each graph. The x-axes show the number of genes \(G\) and the number of individuals per species \(N\); for example, ‘1,3’ means \(G=1\) and \(N=3\). The y-axes show the number of mismatched topologies; for details, see main text. The dotted line is for AlloppMUL and the solid line for AlloppNET.

between 1.6 and 0.8 Ma. For the present analysis, the eight polyploid species together with the two diploids Arabidopsis thaliana and Lepidium apelatum were used. There is one individual for each species, and five genes. There was a substantial amount of missing data: out of a possible 90 sequences (assuming that every diploid genome contributes with one allele each), 41 were unavailable.

The second data set comes from Silene (Caryophyllaceae). There is one allotetraploid Silene involucrata, here labeled “Si”. The species delimitations for the diploids are currently under investigation (Petri A., Oxelman B., unpublished data). For this analysis, the taxa Silene ajanensis, Silene linnaeana, Silene samojedona, and Silene villlosula were grouped together, and labeled “Salsv,” and Silene unalensis and Silene violascens were merged and labeled “Suw.” The sequence data come from the four low-copy nuclear genes NRP A2, NRPB2, NRPD2a, and NRPD2b (Popp et al. 2005; Petri A., Oxelman B., unpublished data). There were 12 individuals from Salsv, eight from Suw, and four from Si. Out of a possible 112 sequences, 34 were missing.

RESULTS

Simulations

Figures 3 and 4 show the results as number of times the correct topology of the MUL-tree was recovered as the maximum clade credibility tree produced by TreeAnnotator which is a tool supplied with BEAST. Suppose the 2 legs of the tetraploid subtree are labeled \(h_1\) and \(h_2\). For Scenarios A1 and A3, the network before hybridization is regarded as matching the true topology if it has topology (((\(h_1,a\)),\((h_2,b)\))) or (((\(h_2,a\)),\((h_1,b)\))). In Scenarios B1 and B3, the network before hybridization is regarded as matching the true topology if it has topology (((\(h_1,h_2))\),\(b\)) or (((\(h_2,h_1))\),\(b\)) (i.e., the order of the 2 joins is not considered). In Scenarios C1 and C3, the network before hybridization is regarded as matching the true topology if it has topology (((\(h_1,h_2))\),\(b\)). For Scenarios A3, B3, and C3, the “relaxed” match means that the network before hybridization matches the true topology in this manner. The “strict match” also requires that the topology of the tetraploid subtree also matches the true topology. The Supplementary Material (doi:10.5061/dryad.nn3j4) contains tables that show the results separately for each scenario.

The value of \(T\) has a major impact on the difficulty of the problem, as expected. \(T\) determines the lengths of the branches in generations. The key quantity is the ratio of the length of the branch measured in generations to the population size. Increasing \(G\) and \(N\) improves the accuracy, also as expected. In general, it appears that increasing \(N\) is most useful when the branches
T is most useful when the branches that need to be resolved are recent, whereas increasing time (going back in time) to be useful.

Each boxplot shows the results from 20 replicates for a model and particular numbers of genes and individuals. For example ‘NET 3,1’ means the AlloppNET model with three genes and one individual per species. The boxes show interquartile ranges, and the whiskers show the extremes of the ranges. The horizontal line is at the true value of log(0.008). Bottom: Estimates of root heights for Scenario B1, with $T = 8 \times 10^{-8}$. The horizontal line is at the true value, 0.025. Other details as for the population sizes.

that need to be resolved are recent, whereas increasing $G$ is most useful when the branches that need to be resolved are more ancient. This was also observed in similar scenarios with the same topology but different node times (results not shown). If the important branches are deep in the network, the sequences from different individuals usually coalesce too soon (going back in time) to be useful.

For A1, B1, and C1 together, AlloppNET is better in 28 cases, AlloppMUL is better in seven cases, with 19 cases where they give the same result. Thus, AlloppNET seems slightly better on the scenarios with one tetraploid species but the difference is of little practical importance. For the scenarios with three tetraploid species, AlloppNET is clearly a lot better: AlloppMUL appears to need around twice the amount of data to achieve similar accuracy. AlloppNET took around 1.5 times as much computational time per generation as AlloppMUL.

The sampled values of the population size parameters have high variance and are highly skewed, and so it seems preferable to work with the logarithms of these values. Estimates of the natural logarithms of population size parameters for Scenario B1 are shown at the top of Figure 5. Note that in the prior the mean of the tip population size parameters is twice that of those elsewhere in the tree, whereas in the simulations, all the parameters are equal. In most of the replicates, the influence of the prior was clear, even with $G=9$ and $N=3$, with the tip populations overestimated and the rest underestimated (data not shown). This indicates that there is little information in the data about the population sizes in individual branches.

Estimates of the root heights of the MUL-tree for Scenario B1 are shown at the bottom of Figure 5. Note that in three cases where there are little data, the estimates from AlloppMUL are occasionally extremely small. This may be due to a failure of the MCMC process to converge.

Empirical Data

Pachycladon Data.—There were some convergence problems with the AlloppMUL model when using the improper priors for the parameters $\eta$, $\lambda$, and the very diffuse prior for the relative clock rates. Convergence often failed to occur after 10 million generations, and the results were dubious even after 100 million generations. The lack of convergence may be due to a weak signal in the data combined with the improper priors—it does not necessarily imply poor mixing. Changing the priors to more realistic ones appeared to improve this behavior considerably. Log-normal distributions were used, and their parameters were as follows: for $\eta$, a mean on the log scale of $-6.0$ and a standard deviation on the log scale of $1.5$; for $\lambda$, a mean on the log scale of 4.6 and a standard deviation on the log scale of 2.0; and for the relative clock rates, a mean of 0.0 and a standard deviation of 1.0. No convergence problems were observed when using the AlloppNET model, and convergence to the expected topology usually appeared to occur within 2 million generations. It is not surprising that the AlloppMUL model has more difficulty than AlloppNET with this data set, since with 8 tetraploid species, there are far more possible topologies allowed by the model, as well as more node times and population size parameters to estimate.

The results reported here use the log-normal priors and a run length of 100 million generations, with the first half discarded as burn-in. They are shown in Figure 6. Both models infer that the Pachycladon genomes are both closer to Arabidopsis thaliana than Lepidium apelatum and that they emerged from the Arabidopsis lineage separately. This is the result expected from the previous analysis by Joly et al. (2009) and corresponds to Scenario B in the simulations.

There is no strong disagreement between the topology of the Pachycladon subtree estimated using AlloppMUL,
FIGURE 6. Pachycladon MUL-tree estimated using AlloppMUL (top) and network estimated using AllopNET (bottom). Heights are in expected substitutions per site based on the CHS gene. The gray bars at nodes indicate 95% HPD intervals for node height. Posterior clade probabilities are shown as numbers for internal branches. The position of the hybridization node in the network does not represent an estimate of the hybridization time.

AlloppNET, or the CHS gene tree from Joly et al. (2009). However, the AlloppNET subtree is fully resolved, whereas the CHS tree is not. The CHS gene is the only gene that was sequenced for all eight species. Although the other genes appear to contribute with little topological information, AlloppNET is capable of taking this information into account.

The mean $\zeta$ of the logarithms of the population sizes along branches was calculated for each MCMC sample. For AlloppMUL, the 95% HPD interval for $\zeta$ was $(-8.5, -6.2)$ with a median and mean of $-7.3$. For AlloppNET, the 95% HPD interval for $\zeta$ was $(-8.8, -6.2)$ with a median and mean of $-7.5$. A very approximate calculation can be made for the average population size. Taking $\exp(-7.4) \approx 6 \times 10^{-4}$ as a typical value for a population size along a branch, and an estimate of the mutation rate as approximately $3 \times 10^{-8}$ per site per generation from Joly et al. (2009), the effective number of gene copies in a typical population size is approximately 20,000, and so the effective number of individuals is estimated at 10,000. The 95% HPD interval for the number of individuals is approximately (2500, 40000).

Silene Data.—No convergence problems were observed here and the performance of the two models was very similar. The results are shown in Figure 7. For
AlloppMUL, the 95% HPD interval for $\tau$ was $(-7.3, -6.2)$ with a median and mean of $-6.7$. For AlloppNET, the 95% HPD interval for $\tau$ was $(-7.0, -5.7)$ with a median and mean of $-6.3$.

**DISCUSSION**

This article represents a first step toward the statistical inference of allopolyploid networks. The two models are complementary in that AlloppMUL is applicable to more data sets, while AlloppNET is more powerful if one can restrict to two diploids and a single hybridization. Under the assumption that the constituent genomes of an allopolyploid does not recombine, AlloppNET will reconstruct species trees under the multispecies coalescent, even if there is a substantial amount of missing data, as can be seen from the *Pachycladon* example. AlloppMUL could be useful for situations where the number of hybridizations are unknown. The number of hybridization could be inferred using methods such as PADRE (Huber et al. 2006; Lott et al. 2009; Marcussen et al. 2012). Both models are available as part of BEAST 1.7 (Drummond et al. 2012) and can take advantage of the numerous models for sequence evolution within gene trees. There is currently no support for the models in BEAUTi, but R scripts are available to aid the construction of suitable XML files (See Supplementary Material).

The methods here inherit some limitations and problems from the model and implementation of *BEAST*. If there are species delimitation errors, then any inferences made by the methods will be dubious at best. If there are gene duplications other than those produced by whole genome doubling, this will also violate the model assumptions. It is also to be expected that the methods here will have similar issues with poor mixing to those possessed by *BEAST*, and add some new ones of its own.

In both the simulations and the empirical data, there was at most one sequence from each diploid individual and two homeologous sequences from each tetraploid individual. In other data sets, there may be more alleles at the same locus due to heterozygosity and there may be cases where it is unclear whether the difference between alleles is because they are homeologs or due to heterozygosity. For diploids, this is unproblematic, and one can define the alleles as different “individuals,” because the multispecies coalescent model treats alleles as belonging to species, not individuals. Polyploids present more complications. For example, if 4 sequences are obtained from a single tetraploid individual, this provides more information than merely obtaining 4 sequences from a population. In the current version of our software, tetraploids with heterozygous loci can be divided into “individuals” if an assumption can be made about which sequences are homeologous. It is also possible in the current version to avoid making any assumptions about which sequences are homeologous by treating each sequence as belonging to a different “individual,” although the effects on accuracy of this type of data have not yet been assessed. Future versions of the software could provide more options than these “all or nothing” assumptions.

Note that in the current article, we do not attempt to detect hybridization; a single allopolyploid hybridization is assumed. The questions that our method is aimed at answering are how and when the hybridization occurred, and how the hybrid speciated afterward. Work is underway to extend the AlloppNET model to deal with arbitrary numbers of diploids, and to deal with an unknown number of hybridization events between diploids. This would enable quite general data sets from diploid and tetraploid species to be analyzed. It would also mean that the method could be used for detecting hybridization events in the sense of distinguishing between one and more than one such hybridization. Dealing with higher ploidy levels, for example, where a diploid and a tetraploid produce a hexaploid (e.g., Popp et al. 2005), would be a further step beyond this. The methods presented in this article assume that ploidy level is known. Sequencing a number of low-copy nuclear genes often gives a good idea about ploidy levels (e.g., Smedmark et al. 2005; Brysting et al. 2007; Petri and Oxelman 2011; Marcussen et al. 2012), in cases where chromosome counts are unavailable.

Both empirical data sets included large amounts of missing data (i.e., often the number of homeologs retrieved from one individual were less than “expected”). This is a situation expected to be typical. In the absence of whole genome sequence information, it will in principle be impossible to tell whether alleles are lost or just unsampled (e.g., due to poor PCR primer or hybridization probe match). The Bayesian approach mitigates this problem since missing data can be handled in a way that is statistically principled.
If one was sure that missing sequences were due to absence, the information could potentially be used as presence/absence data, and thus be informative.

When the number of hybridization events changes, the number of nodes in the network and therefore the number of node heights change. The number of population size parameters is also affected. Reversible-jump MCMC algorithms (Green 1995) have been successfully applied to problems in many areas, including, for example, species delimitation as in Yang and Rannala (2010). However, as pointed out by Green (1995), it is necessary that the prior densities be comparable across spaces of different dimensions. This is not a major problem for the population size parameters, but it is for the node heights in a network. For the species tree in Yang and Rannala (2010), a birth–death model can be used for the prior, and it is important that this prior is a normalized density, since that means the appropriate Hastings ratios can be calculated when jumping between different dimensionalities. But to the best of our knowledge, no similar densities are known for networks. It will probably therefore be necessary to sample from the network prior to estimate its properties when the model is extended in this direction.

SUPPLEMENTARY MATERIAL

Supplementary material, including data files and R scripts can be found in the Dryad data repository at http://datadryad.org, doi:10.5061/dryad.nn3j4.

FUNDING

This work was supported by the Swedish Research Council (621-2010-5623 to S.S. and 2009-5202 to B.O.) and the Centre for Theoretical Biology at the University of Gothenburg (G.J.).

ACKNOWLEDGMENTS

We thank Anna Petri for sharing some unpublished Silene sequences. We gratefully thank Cécile Ané, Simon Joly, and one anonymous reviewer for constructive criticisms to an earlier draft of this article.

APPENDIX

Here, we provide the formula for the multispecies coalescent likelihood. Consider a time interval \([t_0, t_{k+1}]\) within a single branch of the MUL-tree, where time is measured backwards in time from the present, and there are \(k \geq 0\) coalescences at times \(t_1 < \cdots < t_k\) within this interval. This time interval may represent a complete branch, or one of the two parts of a branch before and after hybridization. A particular gene tree which is compatible with the MUL-tree will have some number \(n\) lineages associated to this branch at \(t_0\) and \(n-k\) lineages at \(t_k\). The association of gene tree lineages to MUL-tree branches is calculated recursively from the tips, taking into account the sequence assignments, that is, the current value of the permutation \(\gamma\). The contribution to the multispecies coalescent likelihood from this time interval within the branch and this gene tree is given by Equation (3) of Heled and Drummond (2010), namely

\[
\prod_{j=0}^{k-1} N(t_{j+1})^{-1} \prod_{j=0}^{k} \exp \left( -\left( \frac{N-1}{2} \right) \int_{t_j}^{t_{j+1}} N(t)^{-1} \, dt \right),
\]

where \(N(t)\) is the population size at time \(t\). In our case, the population size varies linearly with time within each interval, so the integral can be calculated explicitly. The total multispecies coalescent likelihood is then a double integral, including, for example, species delimitation as in Yang and Rannala (2010). However, as pointed out by Green (1995), it is necessary that the prior densities be comparable across spaces of different dimensions. This is not a major problem for the population size parameters, but it is for the node heights in a network. For the species tree in Yang and Rannala (2010), a birth–death model can be used for the prior, and it is important that this prior is a normalized density, since that means the appropriate Hastings ratios can be calculated when jumping between different dimensionalities. But to the best of our knowledge, no similar densities are known for networks. It will probably therefore be necessary to sample from the network prior to estimate its properties when the model is extended in this direction.

REFERENCES

Brysting A.K., Oxelman B., Huber K.T., Moulton V., Brochmann C. 2007. Untangling complex histories of genome mergings in high polyploids. Syst. Biol. 56:467–476.
Cai D., Rodriguez F., Teng Y., Ané C., Bonierbale M., Mueller L.A., Spooner D.M. 2012. Single copy nuclear gene analysis of polyploidy in wild potatoes (Solanum section Petrole). BMC Evol. Biol. 12:70.
Chen Z.-Z., Wang L. 2010. Hybridnet: a tool for constructing hybridization networks. Bioinformatics 26:2912–2913.
Chen Z.-Z., Wang L. 2012. Algorithms for reconstructing networks of multiple phylogenetic trees. IEEE/ACM Trans. Comput. Biol. Bioinform. 9:372–384.
Cui L., Wall P.K., Leebens-Mack J.H., Lindsay B.G., Soltis D.E., Doyle J.J., Soltis P.S., Carlson J.E., Arutumyanathan K., Barakat A., Albert V.A., Ma H., dePamphilis C.W. 2006. Widespread genome duplications throughout the history of flowering plants. Genome Res. 16:736–749.
Drummond A.J., Rambaut A. 2007. BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol. Biol. 7:214.
Drummond A.J., Suchard M.A., Xie D., Rambaut A. 2012. Bayesian phylogenetics with BEAUti and the BEAST 1.7. Mol. Biol. Evol. 29:1969–1973.
Fujita M., Leaché A., Burbrik F., McGuire J., Moritz C. 2012. Coalescent-based species delimitation using an integrative taxonomy approach. Trends Ecol. Evol. 27:480–488.
Gerard D., Gibbs H., Kubatko L. 2011. Estimating hybridization in the presence of coalescence using phylogenetic intraspecific sampling. BMC Evol. Biol. 11:291.
Gernhard T. 2008. The conditioned reconstructed process. J. Theo. Biol. 253:769–778.
Green P.J. 1995. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika 82:771–782.
Heled J., Drummond A. 2010. Bayesian inference of species trees from multilocus data. Mol. Biol. Evol. 27:570–580.
Huber K.T., Moulton V. 2006. Phylogenetic networks from multilabeled trees. J. Math. Biol. 52:623–632.
Huber K.T., Oxelman B., Lott M., Moulton V. 2006. Reconstructing the evolutionary history of polyploids from multilabeled trees. Mol. Biol. Evol. 23:1794–1791.
Joly S., Heenan P.B., Lockhart P.J. 2009. A Pleistocene inter-tribal allopolyploidization event precedes the species radiation of Pachypanas (Brassicaceae) in New Zealand. Mol. Phylogenet. Evol. 51:365–372.
Kubatko L. 2009. Identifying hybridization events in the presence of coalescence using phylogenetic intraspecific sampling. Syst. Biol. 58:476–488.
Lanier H.C., Knowles L.L. 2012. Is recombination a problem for species-tree analyses? Syst. Biol. 61:691–701.
Lott M., Spillner A., Huber K.T., Moulton V. 2009a. Padre: a package for analyzing and displaying reticulate evolution. Bioinformatics 25:1199–1200.
Lott M., Spillner A., Huber K.T., Petri A., Oselman B., Moulton V. 2009b. Inferring polyploid phylogenies from multiply-labeled gene trees. BMC Evol. Biol. 9:226.
Marcussen T., Jakobsen K.S., Danihelka J., Ballard H.E., Blaxland K., Brysting A.K., Oselman B. 2012. Inferring species networks from gene trees in high-polyploid North American and Hawaiian violets (Viola, Violaceae). Syst. Biol. 61:107–26.
Mau B., Newton M.A., Larget B. 1999. Bayesian phylogenetic inference via Markov chain Monte Carlo methods. Biometrics 55:1–12.
Petri A., Oselman B. 2011. Phylogenetic relationships within silene (caryophyllaceae) section physolochius. Taxon 60:953–968.
Popp M., Erixon P., Eggens F., Oselman B. 2005. Origin and evolution of a circumpolar polyploid species complex in Silene (Caryophyllaceae) inferred from low copy nuclear RNA polymerase introns, rDNA, and chloroplast DNA. Syst. Bot. 30:302–313.
R Development Core Team. 2008. R: a language and environment for statistical computing, Vienna (Austria): R Foundation for Statistical Computing Vienna.
Rambaut A., Grassly N.C. 1997 Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. Comput. Appl. Biosci. 13:235–238.
Rannala B., Yang Z. 2003. Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. Genetics 164:1645–1656.
Smedmark J.E.E., Eriksson T., Evans R.C., Campbell C.S. 2003. Ancient allopolyploid speciation in geinae (rosaceae): evidence from nuclear granule-bound starch synthase (gbssi) gene sequences. Syst. Biol. 52:374–385.
Tate J.A., Soltis D.E., Soltis P.S. 2005. Polyploidy in plants. In: Gregory T.R., editor. The evolution of the genome. London: Elsevier Academic Press. p. 371–426.
Wendel J.F., Doyle J.J. 2005. Polyploidy and evolution in plants. In: Henry R.J., editor. Plant diversity and evolution: genotypic and phenotypic variation in higher plants. Cambridge (MA): CABI Publishing. p. 97–117.
Wood T.E., Takebayashi N., Barker M.S., Mayrose I., Greenspoon P.B., Rieseberg L.H. 2009. The frequency of polyploid speciation in vascular plants. Proc. Natl Acad. Sci. U S A 106: 13075–13079.
Yang Z., Rannala B. 2010. Bayesian species delimitation using multilocus sequence data. Proc. Natl Acad. Sci. U S A 107: 9264–9269.
Yu Y., Degnan J., Nakhleh L. 2012. The probability of a gene tree topology within a phylogenetic network with applications to hybridization detection. PLoS Genet. 8:e1002660.