Annual Review of Ecology, Evolution, and Systematics

Phylogenetics of Allopolyploids

Bengt Oxelman,[1] Anne Krag Brysting,[2] Graham R. Jones,[1] Thomas Marcussen,[2] Christoph Oberprieler,[3] and Bernard E. Pfeil[1]

[1]Gothenburg Global Biodiversity Centre, Department of Biology and Environmental Sciences, University of Gothenburg, SE405 30 Göteborg, Sweden; email: bengt.oxelman@gu.se

[2]Centre for Ecological and Evolutionary Synthesis, Department of Biosciences, University of Oslo, NO-0316 Oslo, Norway

[3]Evolutionary and Systematic Botany Group, Institute of Plant Sciences, University of Regensburg, D-93053 Regensburg, Germany

Annu. Rev. Ecol. Evol. Syst. 2017. 48:543–57

The *Annual Review of Ecology, Evolution, and Systematics* is online at ecolsys.annualreviews.org

https://doi.org/10.1146/annurev-ecolsys-110316-022729

Keywords

multispecies coalescent networks, allopolyploidy, phylogenetics

Abstract

We give an overview of recently developed methods to reconstruct phylogenies of taxa that include allopolyploids that have originated in relatively recent times—in other words, taxa for which at least some of the parental lineages of lower ploidy levels are not extinct and for which ploidy information is clearly shown by variation in chromosome counts. We review how these methods have been applied to empirical data, discuss challenges, and outline prospects for future research. In the absence of recombination between parental subgenomes, the allopolyploid phylogenetic histories can in principle be treated as genome tree inference. However, without whole genome or whole chromosome data, sequences must be assigned from genes sampled to parental subgenomes. The new version of the AlloppNET method, which now can handle any number of species at the diploid and tetraploid level and any number of hybridizations, is a promising attempt that can also treat gene tree discordance due to the coalescent process. The ongoing development of models that take migration, paralogy, and uncertainties in species delimitations into account offers exciting opportunities for the future of inference of species networks.

INTRODUCTION

Polyploidy, the presence of more than two nuclear genomes in the nongametic generation of eukaryotes, imposes certain problems on phylogenetic inference. Allopolyploidy, defined here as the merger of the nuclear genomes of two separate lineages into a new lineage in which the parental chromosomes continue to form bivalents at meiosis, violates the tree model usually used in phylogenetics and necessitates a more complicated network approach. A considerable amount of work has been done on hybridization and is often focused on homoploid hybridization (e.g., Kubatko 2009; Chen & Wang 2010, 2012; Gerard et al. 2011; Wen et al. 2016). Homoploid hybridization results in offspring of the same ploidy as the parents. In this review, we focus on the reconstruction of phylogenetic networks in which the reticulations stem from allopolyploidy.

Approximately 15% of speciation events in angiosperms and approximately 30% in ferns have been estimated to be associated with polyploidy (Wood et al. 2009), and polyploidy has been increasingly also found in other eukaryote taxa (Otto & Whitton 2000, Albertin & Marullo 2012). Allopolyploidy has long been considered to be the more common mode, but even if allo- and autopolyploidy are considered to be approximately equal in frequency (Barker et al. 2015, see also Doyle & Sherman-Broyles 2017), failure to account for allopolyploidy when reconstructing the past evolution of groups in which it has occurred will inevitably lead to inaccurate phylogenetic hypotheses.

In the pre-phylogenetic era, the ancestry of allopolyploids was typically inferred by artificial crossings of potential parental taxa, often coupled with cytological studies of meiotic behavior, and then by comparison of the obtained hybrid with the natural polyploid. Potential parental taxa were found by looking for traits from diploids that appeared to having been additive (or intermediate) in the polyploid. This approach can be viewed as purely inductive; few of these cases have been tested in a deductive phylogenetic framework, in which consideration is also taken for the possibility that the parental lineages may have undergone large changes since the hybridization event or may have even become extinct. It also fails to take into account that some traits may be transgressive (extreme in relation to the parents; see, e.g., Rieseberg et al. 1999), thus being poor predictors of the traits in the parental lineages.

Although the fundamental differences between gene and species trees were acknowledged almost three decades ago (Pamilo & Nei 1988, Doyle 1992), only recently has this notion become an integral part of phylogenetics. In particular, coalescent theory (Kingman 1982) has been merged with phylogenetics in the multispecies coalescent (MSC) model (Rannala & Yang 2003), which forms a framework in which gene trees constitute data for species tree inference. Thus, gene trees are expected to differ from species trees, as they usually coalesce deeper than the speciation events. In cases of ancestral polymorphisms, which are common if population sizes are large, gene tree topologies include stochastic components with a distribution dictated by coalescent theory (see **Figure 1**).

In this review, we focus on methodologies used for inferring the evolutionary histories (phylogenies) forming allopolyploids. We pay particular attention to explicit networks (i.e., those in which the nodes are thought to represent branching events) or merging events (hybridization), and we review empirical studies in which such methodology has been used.

ASSUMPTIONS AND DEFINITIONS

The definitions of auto- and allopolyploidy are somewhat ambiguous in the literature. In the taxonomic definition, autopolyploids arise within species, whereas allopolyploids result from hybridizations between species. Cytogenetic definitions emphasize whether bivalents or multivalents are formed at meiosis, with allopolyploids characterized by the former. In reality, auto- and

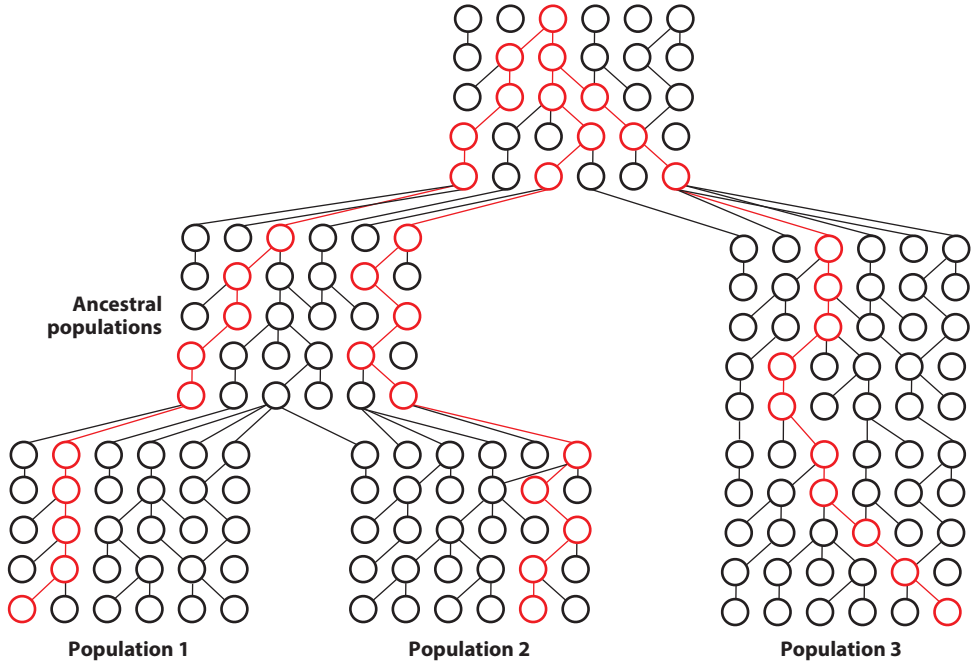

Figure 1

Illustration of the impact of incomplete lineage sorting on topologies of gene trees. A tree with three extant and two ancestral Wright-Fisher populations is shown. The circles represent allelic individuals, and the lines represent parent–offspring relationships. Following the red alleles in the extant populations as an example, we see that neither of them coalesce with another red lineage until the root population. Given that alleles choose their parent randomly in the previous population, each of the three possible topologies of these population trees will occur with probability 1/3.

allopolyploids represent the end points of a continuum, and many intermediate conditions exist (Soltis et al. 2003). In the methodology presented below, it is generally assumed that recombination between the parental subgenomes has not occurred, or at least not between the studied homoeologues (i.e., genes representing each parental subgenome). Moreover, in the MSC model, "species" form the branches of the tree and are assumed to be populations with no selection, random mating, nonoverlapping generations, and no migration after divergence, which is abrupt. Thus, these "species" are generally much more narrowly conceived than most taxonomic species (Toprak et al. 2016, Sukumarana & Knowles 2017).

Although the subgenomes of allopolyploids in due course are expected to be subject to recombination, they can remain distinguishable for long periods of time (Renny-Byfield et al. 2014). Therefore, sampling all homoeologous copies of nuclear genes should potentially enable the reconstruction of the genome trees. However, one obstacle is that when collecting data from more than one gene, it is usually not possible to a priori assign which gene copies belong to the same subgenome (but might be if whole genomes or large chunks of chromosomes are available; see Zwickl et al. 2014).

DATA ACQUISITION

In this review, we focus on methods aimed at the analysis of molecular sequence data. Several studies have utilized markers such as allozymes (e.g., Roose & Gottlieb 1976, Gastony 1986,

Hedrén 1996), microsatellites (e.g., Crespo-López et al. 2007), and random amplified polymorphic DNA (RAPD) or amplified fragment length polymorphism (AFLP) fragments (e.g., Brochmann et al. 1996, Oxelman 1996, Hedrén et al. 2001), sometimes in concert with gene phylogenies inferred from sequences. However, taken alone, they employ a similar approach to those of classical morphological studies. That is, additive patterns from putative parental species are used to confirm the ancestry hypothesis. We focus on methods that use DNA sequences from several species to infer gene and/or species trees or networks.

Analyses using nuclear ribosomal regions, plant plastid, and animal or fungal mitochondrial regions have dominated phylogenetic research for at least a couple of decades. However, the nuclear ribosomal DNA (nrDNA) cistrons typically undergo concerted evolution in an unpredictable direction, so although the plastid tree would be expected to usually reflect the maternal lineage of an allopolyploid, the nrDNA tree could reflect one or both parents, or it could be misleading because it is inferred from chimeric sequences (Álvarez & Wendel 2003).

By using sequences from putatively single-copy nuclear regions, both parental lineages can potentially be recovered. However, although the general phylogenetic utility of such sequences was suggested 20 years ago (reviewed by Sang 2002), the development remained slow until the recent replacement of polymerase chain reaction (PCR)/Sanger sequencing protocols with Targeted Sequence Enrichment (TSE) and Next Generation Sequencing (NGS) techniques (e.g., Lemmon & Lemmon 2013, Jones & Good 2016). The slow progress was also due to the difficulty in developing universal PCR primers, which often necessitated either knowledge of genomic sequences from the taxonomic group in focus or modification of primers developed for other taxa to successfully amplify the target regions of the organisms under study. Even if suitable primers are at hand, the problem of disentangling the different homoeologues (with possible allelic variation) into clean sequences remains. Direct Sanger sequencing of PCR products is unproblematic for haploid cytoplasmatic loci, but when employed for nrDNA, it results in a majority rule consensus sequence from the concerted repeats. If the amplified product contains homoeologues and their possible allelic variants, direct sequencing of PCR products often results in uninterpretable data as a consequence of polymorphic sites and sequence length variation among the amplified copies. The most common approach to circumvent this result involves bacterial subcloning of PCR fragments. However, to obtain all sequence variants, many sequencing reactions are needed. For example, to obtain all sequence variants with 95% probability for two heterozygous homoeologous loci in a tetraploid, at least 14 colonies must be sequenced (Rautenberg et al. 2008). Alternative, but still rather laborious, strategies involve designing homoeologue-specific PCR primers (e.g., Popp & Oxelman 2004, Marcussen et al. 2012), designing allele-specific primers for secondary sequencing reactions (Scheen et al. 2012) or single-molecule PCR (Kraytsberg & Khrapko 2005, Marcussen et al. 2012).

The recent development of TSE and NGS techniques has greatly increased the availability of sequences from many unlinked loci and holds great promise for the future. In the absence of genomic sequence information from the study group, probe design is a problem similar to PCR primer design. However, the scalability of TSE, the greedier nature of DNA–RNA hybridization (Gasc et al. 2016), and the possibility to multiplex barcoded samples with NGS make it much more efficient. Probe design can be facilitated by using transcriptome sequencing (RNA-seq) (e.g., Petri et al. 2013) or anonymous NGS sequencing such as restriction-site-associated DNA sequencing (RAD-seq) (Davey & Blaxter 2010), genome skimming (Dodsworth 2015), or genotyping by sequencing (Davey et al. 2011). The short read lengths of some sequencing protocols may introduce problems for accurate phasing of alleles, homoeologues, or paralogs (Lemmon & Lemmon 2013). Single-molecule sequencing techniques hold great promise to mitigate this problem (Rothfels et al. 2017).

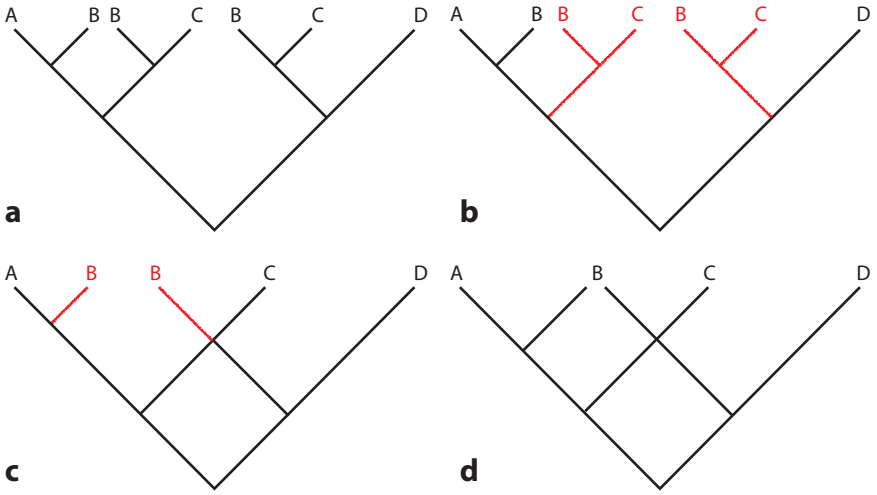

Figure 2

Schematic presentation of the PADRE algorithm of Huber et al. (2006). (*a*) The multilabeled tree is shown, (*b*) with the most inclusive duplicate subtrees identified in red. (*c*) These subtrees are merged into a network, with the remaining most inclusive duplicate subtrees identified in red. (*d*) The resulting network is shown. This algorithm ensures that a network with minimal hybridization events is found (see Huber et al. 2006 for details).

AD HOC PHYLOGENETIC METHODS

The first attempts to reconstruct plant allopolyploid origins used gene trees from the plastid genome and nuclear ribosomal regions (often together with other data, e.g., Oxelman 1996, Brochmann et al. 1996). The approach has been used to infer the phylogenetic origin of the polyploid from its divergent positions in the trees from the two regions, assuming maternal ancestry reflected by the chloroplast DNA (cpDNA) tree and paternal from the nrDNA tree, but as noted above this approach leaves many other possibilities.

By using sequences from a putatively single-copy nuclear region, both parental lineages can potentially be recovered, for example, as shown by Small et al. (1998), Sang & Zhang (1999), and Doyle et al. (2000). The gene trees that arise from sampling all homoeologues of a polyploid are multilabeled trees (MUL-trees), meaning that the same label occurs more than once in the tree (see **Figure 2*a***). Under the assumption that a single multilabeled gene tree accurately portrays the phylogenetic history of the sampled taxa, Smedmark et al. (2005) and Huber et al. (2006) developed an algorithm for converting such a tree into a phylogenetic network. The approach of the latter authors is an exact parsimony method (minimizing the number of reticulation events) and is illustrated in **Figure 2**. They also discuss an algorithm that is implemented in the PADRE software (Lott et al. 2009a).

Example: *Galeopsis*

The work of Müntzing (1932) represents a milestone in the study of polyploidy, as it was the first report of an experimental synthesis of a naturally occurring allopolyploid plant species, *Galeopsis tetrahit* (Lamiaceae). Müntzing succeeded, via a two-step process involving two diploid *Galeopsis* species and a triploid bridge, in recreating an allopolyploid speciation event producing an allotetraploid plant, which in chromosome number and morphology was very similar to and

cross-compatible with the naturally occurring tetraploid. The *Galeopsis* system provides an example of allopolyploid speciation of a relatively recent age, making it relatively easy to use molecular markers to detect both parental contributions and reconstruct the phylogenetic relationships. Even with a single nuclear marker (a noncoding region of the second largest subunit of RNA polymerase II, *NRPA2*), Bendiksby et al. (2011) confirmed Müntzing's conclusion regarding the allopolyploid origin of the naturally occurring *G. tetrahit* (as well as another naturally occurring tetraploid, *Galeopsis bifida*).

NRPA2 appeared to be a single-copy nuclear gene in *Galeopsis*, and parental-homoeologue specific primers were developed for direct sequencing on the basis of an initial nested PCR procedure, degenerate primers (Popp & Oxelman 2004), and subcloning of the products. Two diverged *NRPA2* copies were found in each of the two tetraploids, with the resulting MUL-tree clearly demonstrating that both tetraploids originated by allopolyploid speciation from the diploid *Galeopsis speciosa* and *Galeopsis pubescens* lineages. However, the data show that the parental genomes involved in the two tetraploids differed genetically and that the two tetraploids most likely originated by independent polyploidization events. The addition of cpDNA markers further allowed determination of the maternal parent of *G. tetrahit*. Bendiksby et al. (2011) also analyzed a larger population sampling of the two tetraploids and their parental species using AFLPs. These results, in combination with the DNA sequence data, suggest that both tetraploids appear to have originated only once, as opposed to recurrent origins usually reported for natural polyploids (e.g., Soltis et al. 2003), and that frequent hybridization and introgression occurs, especially within ploidy levels.

A problem that arises when sampling homoeologues from several unlinked loci is that it is generally not possible to know a priori which subgenome the sequences belong to, or even whether two alleles from different genes belong to the same subgenome. This can create further problems for some methods of analysis that require associating the alleles from each locus that belong to the same subgenome a priori (e.g., a concatenation approach in which subgenomes are handled as terminal taxa).

Example: *Cerastium*

As a typical example of recent and rapid speciation during the Pleistocene, the high-ploid *Cerastium alpinum* group (Caryophyllaceae) represents a much more complex history. The complex consists of six high-ploid species (octoploids, $2n = 8x = 72$, and dodecaploids, $2n = 12x = 108$) mainly distributed in arctic or alpine regions and for which no diploid progenitors are known. Brysting et al. (2007, 2011) sequenced noncoding regions of three single-copy nuclear RNA polymerase genes. The sequences were merged into consensus sequences representing monophyletic groups in initial phylogenetic analyses and used to produce MUL-trees, which were transformed into networks using the PADRE software (Lott et al. 2009a). The closest living relatives of the *C. alpinum* group are tetraploid species ($2n = 36$). Despite this, only one functional copy of each of the three genes was detected, and these tetraploid taxa are most likely the result of ancient polyploidization events. Conversely, the high-ploid species of the *C. alpinum* group likely result from much more recent polyploidization events related to recurrent episodes of range expansions and contractions during the Quaternary glaciations, and in most cases the copy number corresponds well with ploidy level. Overall, Brysting et al. (2007, 2011) were successful in disentangling the tetraploid progenitor lineages of the high-ploid species of the *C. alpinum* group. However, the three networks based on different RNA polymerase genes differed in several aspects and had small deviations from the general pattern, which could be better explained by gene duplication, lineage sorting events, or lack of information arising from incomplete sampling. The fact that gene loss,

pseudogenization, and possible lineage sorting are working independently in different parts of the polyploid genome may hamper the interpretation of reticulate evolution even in relatively young plant groups such as the *C. alpinum* complex and emphasizes the importance of approaches in which several unlinked regions of the genome are examined.

Examples: *Silene*, *Viola*, and *Fumaria*

Popp et al. (2005) sampled four RNA polymerase genes, nrDNA of the internal transcribed spacer, and cpDNA sequences from Arctic di-, tetra-, and hexaploid members of *Silene* section *Physolychnis*. Using an ad hoc procedure, they presented a consensus MUL-tree, which clearly showed that a diploid lineage that had never previously been suggested to have anything to do with the parentage of the polyploids has actually been involved in the origin of both the tetraploids and the hexaploids (resulting in a PADRE network like the one shown in **Figure 2**). Lott et al. (2009b) showed that finding consensus MUL-trees is computationally hard but presented a heuristic approach implemented in PADRE (Lott et al. 2009a). Marcussen et al. (2012) used PADRE to reconstruct an allopolyploid network from a MUL-tree of the glucose-6-phosphate isomerase (*GPI*) gene from *Viola* taxa with ploidy levels spanning from 2*x* to 18*x*. For some taxa, the expected number of homoeologues was not recovered. To assess whether this absence was primary (i.e., a result of the allopolyploid origin itself) or due to secondary gene loss (or detection failure), the different possible MUL-trees were analyzed separately in PADRE and the scenarios were compared to obtain the most parsimonious solution—that is, the one requiring the fewest polyploidization events and gene losses to explain the observed data.

In a study of *Viola*, Marcussen et al. (2014) identified the most parsimonious network topology from a set of five competing scenarios differing in the interpretation of homoeologue extinctions and lineage sorting, based on the (*a*) fewest possible ghost subgenome lineages, (*b*) fewest possible polyploidization events, and (*c*) least possible deviations from the expected ploidy as inferred from available chromosome counts of the involved polyploid taxa. They also estimated the polyploid speciation times by comparing branch lengths and speciation rates of lineages with and without ploidy shifts.

The studies of Marcussen et al. (2012, 2014) show that even high complexity networks can be reasonably well understood phylogenetically. However, a problem with the PADRE approach is that it does not take coalescent stochasticity into account. Using sequence data from the angiosperm genus *Fumaria* (ploidy levels ranging between 2x and 14x), Bertrand et al. (2015) proposed a workflow that assigns homoeologues to hypothetical diploid subgenomes prior to genome tree construction. Conflicting assignment hypotheses were evaluated against substitution model error and coalescent stochasticity. Incongruence that cannot be explained by stochastic mechanisms must be explained by other processes (e.g., homoploid hybridization or paralogy). The data can then be filtered to build multilabeled genome trees using inference methods that can recover species trees (e.g., under MSC) in the face of substitution model error and coalescent stochasticity. The network can then be obtained from PADRE.

EXPLICIT SPECIES NETWORK INFERENCE METHODS

Permutation Approach Using PhyloNet

A relatively simple and fast workflow pipeline for the reconstruction of species networks in polyploid complexes based on multilocus gene trees was described by Oberprieler et al. (2017). It uses a permutation strategy and a parsimony-based principle in species-tree reconstruction

(minimizing deep coalescences, or MDC) for the assignment of homoeologues to parental genomes in allopolyploids and eventually constructs a species MUL-tree in which polyploid taxa are represented by their diploid subgenomes. The method utilizes the PhyloNet software program (Than et al. 2008, Than & Nakhleh 2009), the first program in which an exact algorithm for inferring species trees from gene trees by minimizing the number of extra lineages (Maddison 1997, Maddison & Knowles 2006) had been implemented (Than & Nakhleh 2010).

In the first step, individually for each polyploid accession and each sequenced locus, the procedure forms all possible parental (diploid) allele pairs for the polyploid accession analyzed and runs an MDC analysis together with all diploid accessions, which tries to find a species tree that minimizes the number of deep coalescences for the given gene tree. For the subsequent permutation steps, the combination of parental alleles in the polyploid that has the lowest number of deep coalescences in the species tree inference procedure is kept.

After having determined the optimal allele combinations at all sequenced loci for the polyploid accession under study, a further MDC analysis is then run using all diploid accessions and the single polyploid accession concerned. At each turn, combinations of allele pairs across loci are submitted to a species tree reconstruction based on all gene trees, and this step is repeated for all possible combinations of allele groupings across loci. As in the first step of the procedure, the allele pair combination across loci that results in a species tree with the minimum number of deep coalescences is kept.

By repeating the two steps described for all polyploid accessions individually, this procedure produces a data set consisting of all diploid accessions (allele pairs) and all pseudo-diploid parental combinations of alleles and allele pairs across loci that could be subjected to a final MDC analysis to reconstruct the overall MUL-species tree. The inference of a species network is then easily accomplished by joining branches representing parental diploids to the polyploids into reticulations using the PADRE (Huber et al. 2006, Lott et al. 2009a) or Dendroscope (Huson & Scornavacca 2012) programs.

The permutations necessary for the two steps of the described procedure, along with the handover of retailored (pruned) gene tree topologies and further arguments for the MDC search to PhyloNet, is presently implemented in a Matlab v8.0.0.783 script utilizing the Matlab Bioinformatics Toolbox (Henson & Cetto 2005). Simulations carried out by Oberprieler et al. (2017) with varying effective population sizes, different temporal scenarios for diploid or polyploid formations, and different types of allopolyploid speciation (sister-species diploids, non-sister-species diploids, or polyploidization involving the participation of ancestral diploids) showed that the proposed workflow pipeline leads to reliable reconstructions if effective populations sizes are not large and divergences among ancestral taxa are not shallow.

Simultaneous Gene Tree and Species Network Inference (AlloppNET)

AlloppNET (Jones et al. 2013) is based on a fully parameterized stochastic model of the relevant evolutionary processes (i.e., the MSC). It uses a Bayesian approach in which parameters are coestimated by sampling from the posterior distribution using the Markov chain Monte Carlo (MCMC) algorithm. AlloppNET is an extension of the MSC model implemented in *BEAST (Heled & Drummond 2010). Here, we present an extension of the version presented in Jones et al. (2013) that caters to more than one hybridization and more than two diploid species (for details, see Jones 2017). It is implemented in BEAST 1.8.3 (Drummond et al. 2012). There is no support in BEAUTi, but R scripts are available (**http://www.indriid.com/2013/2013-05-15-manual.zip**) to aid the generation of suitable XML files. AlloppNET can be used with DNA sequence data from diploid species and allotetraploid species, and it can handle multiple individuals per species

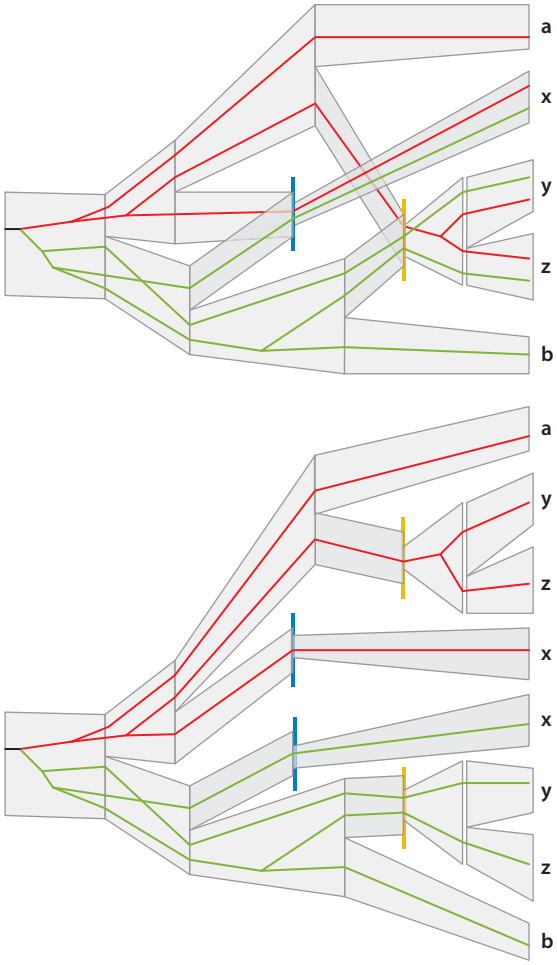

Figure 3

(*Top*) A species network for two diploid species (a and b) and three allotetraploid species (x, y, and z). The widths of the gray tubes indicate population sizes. The network contains a gene tree, with red and green branches indicating different ancestral species. The blue line indicates one hybridization, the yellow line another. (*Bottom*) The same scenario represented as a multilabeled genome tree. Note that the gene tree does not match the genome tree. Figure adapted from Jones (2017).

and multiple unlinked loci. The sequences from the allotetraploid species are assumed to belong to homoeologue pairs for each locus. The identity of the homoeologues (i.e., which sequence came from which diploid species) is estimated along with the other parameters.

Allotetraploid hybridization events in the species network may be followed by ordinary speciation events in the allotetraploid species (**Figure 3**). No assumption is made about the fate of the diploid species that hybridized: They may or may not leave descendants in the sample. There is no limit on the number of allotetraploid species or diploid species, and the number of hybridizations is estimated (Jones 2017).

The model includes an underlying species network W, which incorporates a topology, node heights, hybridization times, and population size parameters along each edge. From W, a multilabeled tree, M_W, can be derived. This is not an arbitrary multilabeled tree, because for each

hybridization there is a pair of identical allotetraploid subtrees. Given M_W, the homoeologue identities, and the population size parameters, a prior probability density for the gene trees is calculated using the multispecies coalescent model. Given the gene trees, the likelihood of the sequence data is calculated in the usual way (e.g., Felsenstein 2004). The posterior distribution is then a product of these terms and the prior probability densities for all the parameters. Jones (2017) contains a formal description of the model and further discussion.

To sample from the posterior distribution, proposal methods (operators) are needed for all the parameters. The operator that changes the number of hybridizations is particularly complicated, because it also changes the number of node heights and the number of population size parameters. This fact necessitates using a reversible jump move during the MCMC algorithm (Green 1995). Jones (2017) contains details of the operators and results for simulated data. Rothfels et al. (2017) used AlloppNET to derive a phylogenetic network in the fern family Cystopteridaceae.

Example: *Medicago*

Despite recent developments, some phylogenetic problems are still extremely difficult to solve because of complications arising from specific properties of the biological system under study. Polyploids arising from within (or among close relatives of) the *Medicago sativa* (Fabaceae) complex are among such cases because of the complexity of historical interactions among this group of species.

Small (2011) recognizes eight subordinate taxa within *M. sativa*, including the world's most important forage crop, alfalfa (*M. sativa* subsp. *sativa*). Hybridization has resulted in the successful introgression of alleles between subspecies of the complex at the same ploidy level (Sakiroglu et al. 2010, Kaljund & Leht 2013). Tetraploids within the complex are presumed to have arisen via autotetraploidy (Havananda et al. 2011), with alfalfa known to display tetrasomic inheritance (McCoy & Bingham 1988). Unreduced gamete production in diploids, and crosses between $2x$ and $4x$ plants to produce viable $4n$ plants, has been demonstrated experimentally (Bingham 1968, Veronesi et al. 1986) and probably allows gene flow from $2x$ to $4x$ populations to occur in the wild. Together, these lines of evidence have been used to propose a model of the *M. sativa* complex comprising four pillars (Small 2011), each pillar being a diploid–autotetraploid taxon pair with similar morphologies. Gene flow is readily achieved among diploid and among tetraploid members of each pillar and also occurs to some degree from diploid to tetraploid members (Small 2011).

Medicago arborea and *Medicago strasseri* are both tetraploids and morphologically very similar to one another and to *Medicago citrina*, a hexaploid; all three species are placed in the same section (Small 2011). *Medicago citrina* is thought to be an autohexaploid species (Quiros & Bauchan 1988), presumably on the basis of its morphological similarity, although this has been questioned on the basis of a lack of clear chromosomal similarities (Rosato et al. 2008). If the former case is correct, this implies that *M. arborea* and *M. strasseri* are also autotetraploids, a hypothesis explicitly made in an earlier study on the basis that no known diploid forms resembling *M. arborea* were known (Lesins & Lesins 1979). *Medicago arborea*, as a representative of these three closely related woody species, was also considered to be the oldest member of the genus and therefore not a likely candidate to be an immediate progenitor of other perennial species related to the *M. sativa* complex (Lesins & Lesins 1979, Quiros & Bauchan 1988). This reasoning was based at least in part on the presumption that woodiness is an ancestral trait in *Medicago* (Lesins & Lesins 1979), as it is in other groups of plants. However, both the assumption of *M. arboera*'s phylogenetic position and the mode of polyploid origin have been called into question by a recent molecular systematic study using ten nuclear genes (J. Eriksson, F. Sousa, Y. Bertrand, A. Antonelli, B. Oxelman & B. Pfeil, manuscript under review). In this work, phylogenetic evidence points strongly toward

an allopolyploid origin, because clades of alleles that presumably represent homoeologues are not sister in eight out of ten gene trees. The conclusion is also supported by AlloppNET, but homoploid hybridizations at the diploid level are likely violating the assumptions of the model and result in slow convergence of the MCMC. In this case at least, understanding of the polyploid origin is progressing, even though the relationships among diploids are contradictory across gene trees (e.g., Sousa et al. 2016, Eriksson et al. 2017).

FUTURE CHALLENGES AND PROSPECTS

Our understanding of polyploidy is rapidly improving, and the processes as well as the patterns produced are intricate (Barker et al. 2016 and references therein). Polyploidy has clearly played, and is playing, a central role in the evolution of some groups, obviously in plants but also in fungi (Albertin & Marullo 2012), invertebrates, vertebrates, and some other eukaryotes (Otto 2007). It is therefore of uttermost importance to have tools to reconstruct their phylogenetic history.

In this review, we have seen how early attempts based on single nuclear gene phylogenies have sometimes confirmed, and sometimes rejected, classical hypotheses of origins of polyploidy. As has long been appreciated, gene trees may differ from species trees for various reasons. For example, the *Medicago* data clearly demonstrates the need for several unlinked gene trees to disentangle the various sources of gene tree discordance (see, e.g., Zwickl et al. 2014). Such data are now realistically achievable, in principle, for any taxon.

Phylogenetics has developed to a high level of sophistication, and models that can take coalescent stochasticity into account are becoming standard (Degnan & Rosenberg 2009, Edwards & Rausher 2009). Recent advances have also made it possible to use these models without prior knowledge about species delimitations (e.g., Jones et al. 2014, Toprak et al. 2016), and promising attempts are being made to also account for migration (e.g., Tian & Kubatko 2016, Jones 2017, Wen & Nakhleh 2017), a probably very common violation to the MSC model at the homoploid level that can seriously affect species tree and network estimation (Leaché et al. 2014). However, homoploid and polyploid hybridization must also be taken into account. Another consideration is the need to further develop coalescent-aware methods of paralog detection (e.g., Sousa et al. 2017), because paralogy has been commonly overlooked and may especially confound phylogenetic inference of polyploid taxa. Finally, models that are able to take all relevant parameters into account are desirable.

We have reviewed some of the recent progress in developing methods to infer phylogenetic networks in the presence of allopolyploids, and although the advances are promising, more efforts are needed to exploit the information inherent in the enormous data sets now being generated with NGS. As in phylogenetics in general, we can use stepwise approaches that first estimate gene trees, which are then used as data for species tree inference, or make simultaneous coestimation of gene and species trees and their parameters. There is also a dichotomy between methods that first seek to infer a genome MUL-tree, which then may be converted to a network, and methods that infer the network directly. Fully parameterized, direct methods are so far restricted to two ploidy levels (i.e., AlloppNET; Jones et al. 2013, Jones 2017) but are perhaps the most promising. Conversely, they are sensitive to some model violations, which may compromise convergence of MCMC, so the use of simpler methods may be justified as well.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

The authors gratefully thank Michael Sanderson for many constructive comments and suggestions.

LITERATURE CITED

Albertin W, Marullo P. 2012. Polyploidy in fungi: evolution after whole-genome duplication. *Proc. R. Soc. B.* 279:2497–509

Álvarez I, Wendel JF. 2003. Ribosomal ITS sequences and plant phylogenetic inference. *Mol. Phylogenetics Evol.* 29:417–34

Barker MS, Arrigo N, Baniaga AE, Li Z, Levin DA. 2015. On the relative abundance of autopolyploids and allopolyploids. *New Phytol.* 210:391–98

Barker MS, Husband BC, Pires JC. 2016. Spreading Winge and flying high: the evolutionary importance of polyploidy after a century of study. *Am. J. Bot.* 103:1139–45

Bendiksby M, Tribsch A, Borgen L, Trávníček P, Brysting AK. 2011. Allopolyploid origins of the *Galeopsis* tetraploids—revisiting Müntzing's classical textbook example using molecular tools. *New Phytol.* 191:1150–67

Bertrand YJK, Scheen A-C, Marcussen T, Pfeil BE, Sousa F, Oxelman B. 2015. Assignment of homoeologs to parental genomes in allopolyploids for species tree inference, with an example from *Fumaria* (Papaveraceae). *Syst. Biol.* 64:448–71

Bingham ET. 1968. Transfer of diploid *Medicago* spp. germplasm to tetraploid *M. sativa* L. in 4x-2x crosses. *Crop Sci.* 8:760–62

Brochmann C, Nilsson T, Gabrielsen TM. 1996. A classic example of postglacial allopolyploid speciation re-examined using RAPD markers and nucleotide sequences: *Saxifraga osloensis* (Saxifragaceae). *Symb. Bot. Ups.* 31:75–89

Brysting AK, Mathiesen C, Marcussen T. 2011. Challenges in polyploid phylogenetic reconstruction: a case story from the arctic-alpine *Cerastium alpinum* complex. *Taxon* 60:333–47

Brysting AK, Oxelman B, Huber KT, Moulton V, Brochmann C. 2007. Untangling complex histories of genome merging in high polyploids. *Syst. Biol.* 56:467–76

Chen Z-Z, Wang L. 2010. HybridNET: a tool for constructing hybridization networks. *Bioinformatics* 26:2912–13

Chen Z-Z, Wang L. 2012. Algorithms for reticulate networks of multiple phylogenetic trees. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 9:372–84

Crespo-López ME, Pala I, Duarte TL, Dowling TE, Coelho MM. 2007. Genetic structure of the diploid-polyploid fish *Squalius alburnoides* in southern Iberian basins Tejo and Guadiana, based on microsatellites. *J. Fish Biol.* 71:423–36

Davey JW, Blaxter ML. 2010. RADSeq: next-generation population genetics. *Brief. Funct. Genom.* 9:416–23

Davey JW, Hohenlohe PA, Etter PD, Boone JQ, Catchen JM, Blaxter ML. 2011. Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nat. Rev. Genet.* 12:499–510

Degnan JH, Rosenberg NA. 2009. Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends Ecol. Evol.* 24:332–40

Dodsworth S. 2015. Genome skimming for next-generation biodiversity analysis. *Trends Plant Sci.* 20:525–27

Doyle JJ. 1992. Gene trees and species trees—molecular systematics as one-character taxonomy. *Syst. Bot.* 17:144–63

Doyle JJ, Doyle JL, Brown AHD, Pfeil BE. 2000. Confirmation of shared and divergent genomes in the *Glycine tabacina* polyploid complex (Leguminosae) using histone H3-D sequences. *Syst. Bot.* 25:437–48

Doyle JJ, Sherman-Broyles S. 2017. Double trouble: taxonomy and definitions of polyploidy. *New Phytol.* 213:487–93

Drummond AJ, Suchard MA, Xie D, Rambaut A. 2012. Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Mol. Biol. Evol.* 29:1969–73

Edwards SV, Rausher M. 2009. Is a new and general theory of molecular systematics emerging? *Evolution* 63:1–19

Eriksson JS, Blanco Pastor JL, Sousa F, Bertrand YJK, Pfeil BE. 2017. A cryptic species produced by autopolyploidy and subsequent introgression involving *Medicago prostrata* (Fabaceae). *Mol. Phylogenetics Evol.* 107:367–81

Felsenstein J. 2004. *Inferring Phylogenies*. Sunderland, MA: Sinauer

Gasc C, Peyretaillade E, Peyret P. 2016. Sequence capture by hybridization to explore modern and ancient genomic diversity in model and nonmodel organisms. *Nucleic Acids Res*. 44:4504–18

Gastony GJ. 1986. Electrophoretic evidence for the origin of fern species by unreduced spores. *Am. J. Bot.* 73:1563–69

Gerard D, Gibbs HL, Kubatko LS. 2011. Estimating hybridization in the presence of coalescence using phylogenetic intraspecific sampling. *BMC Evol. Biol.* 11:291

Green PJ. 1995. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* 82:711–32

Havananda T, Brummer EC, Doyle JJ. 2011. Complex patterns of autopolyploid evolution in alfalfa and allies (*Medicago sativa*; Leguminosae). *Am. J. Bot.* 98:1633–46

Hedrén M. 1996. Genetic differentiation, polyploidization and hybridization in northern European *Dactylorhiza* (*Orchidaceae*): evidence from allozyme markers. *Plant Syst. Evol.* 201:31–55

Hedrén M, Fay MF, Chase MW. 2001. Amplified fragment length polymorphisms (AFLP) reveal details of polyploid evolution in *Dactylorhiza* (Orchidaceae). *Am. J. Bot.* 88:1868–80

Heled J, Drummond AJ. 2010. Bayesian inference of species trees from multilocus data. *Mol. Biol. Evol.* 27:570–80

Henson R, Cetto L. 2005. The MATLAB bioinformatics toolbox. In *Online Encyclopedia of Genetics*, *Genomics*, *Proteomics*, *and Bioinformatics*, ed. LB Jorde, PFR Little, MJ Dunn, S Subramaniam, Part 4, 4.8:105. Wiley. **https://doi.org/10.1002/047001153X.g409308**

Huber KT, Oxelman B, Lott M, Moulton V. 2006. Reconstructing the evolutionary history of polyploids from multilabeled trees. *Mol. Biol. Evol.* 23:1784–91

Huson DH, Scornavacca C. 2012. Dendroscope 3—an interactive viewer for rooted phylogenetic trees and networks. *Syst. Biol.* 61:1061–67

Jones G. 2017. Bayesian phylogenetic analysis for diploid and allotetraploid species networks. bioRxiv 129361. **http://dx.doi.org/10.1101/129361**

Jones G, Aydin Z, Oxelman B. 2014. DISSECT: an assignment-free Bayesian discovery method for species delimitation under the multispecies coalescent. *Bioinformatics* 31:991–98

Jones G, Sagitov S, Oxelman B. 2013. Statistical inference of allopolyploid species networks in the presence of incomplete lineage sorting. *Syst. Biol.* 62:467–78

Jones MR, Good JM. 2016. Targeted capture in evolutionary and ecological genomics. *Mol. Ecol.* 25:185–202

Kaljund K, Leht M. 2013. Extensive introgressive hybridization between cultivated lucerne and the native sickle medic (*Medicago sativa* ssp. *falcata*) in Estonia. *Ann. Bot. Fenn.* 50:23–31

Kingman JFC. 1982. On the genealogy of large populations. *J. Appl. Probab.* 19A:27–43

Kraytsberg Y, Khrapko K. 2005. Single-molecule PCR: an artifact-free PCR approach for the analysis of somatic mutations. *Expert Rev. Mol. Diagn.* 5:809–15

Kubatko LS. 2009. Identifying hybridization events in the presence of coalescence via model selection. *Syst. Biol.* 58:478–88

Leaché AD, Harris RB, Rannala B, Yang Z. 2014. The influence of gene flow on species tree estimation: a simulation study. *Syst. Biol.* 63:17–30

Lemmon EM, Lemmon AR. 2013. High-throughput genomic data in systematics and phylogenetics. *Annu. Rev. Ecol. Evol. Syst.* 44:99–121

Lesins KA, Lesins I. 1979. Genus *Medicago* (Leguminosae). *A Taxogenetic Study*. The Hague, Neth.: Dr. W. Junk

Lott M, Spillner A, Huber KT, Moulton V. 2009a. PADRE: a package for analyzing and displaying reticulate evolution. *Bioinformatics* 25:1199–200

Lott M, Spillner A, Huber KT, Petri A, Oxelman B, Moulton V. 2009b. Inferring polyploid phylogenies from multiply-labeled gene trees. *BMC Evol. Biol.* 9:216

Maddison WP. 1997. Gene trees in species trees. *Syst. Biol.* 46:523–36

Maddison WP, Knowles LL. 2006. Inferring phylogeny despite incomplete lineage sorting. *Syst. Biol.* 55:21–30
Marcussen T, Heier L, Brysting AK, Oxelman B, Jakobsen K. 2014. From gene trees to a dated allopolyploid network: insights from the angiosperm genus *Viola* (Violaceae) *Syst. Biol.* 64:84–101
Marcussen T, Jakobsen K, Danihelka J, Ballard H, Blaxland K, et al. 2012. Inferring species networks from gene trees in high-polyploid North American and Hawaiian violets (*Viola*, Violaceae). *Syst. Biol.* 61:107–26
McCoy TJ, Bingham ET. 1988. Cytology and cytogenetics of alfalfa. *Agron. Monogr.* 29:737–76
Müntzing A. 1932. Cytogenetic investigations on synthetic *Galeopsis tetrahit*. *Hereditas* 16:105–54
Oberprieler C, Wagner F, Tomasello S, Konowalik K. 2017. A permutation approach for inferring species networks from gene trees in polyploid complexes by minimising deep coalescences. *Methods Ecol. Evol.* 8:835–49
Otto SP. 2007. The evolutionary consequences of polyploidy. *Cell* 131:452–62
Otto SP, Whitton J. 2000. Polyploid incidence and evolution. *Annu. Rev. Genet.* 34:401–37
Oxelman B. 1996. RAPD patterns, nrDNA ITS sequences, and morphological patterns in the *Silene sedoides* group (Caryophyllaceae). *Plant Syst. Evol.* 201:93–116
Pamilo P, Nei M. 1988. Relationships between gene trees and species trees. *Mol. Biol. Evol.* 5:568–83
Petri A, Pfeil BE, Oxelman B. 2013. Introgressive hybridization between anciently diverged lineages of *Silene* (Caryophyllaceae). *PLOS ONE* 8:e67729
Popp M, Erixon P, Eggens F, Oxelman B. 2005. Origin and evolution of a circumpolar polyploid species complex in *Silene* (Caryophyllaceae). *Syst. Bot.* 30:302–13
Popp M, Oxelman B. 2004. Evolution of a RNA polymerase gene family in *Silene* (Caryophyllaceae)—incomplete concerted evolution and topological congruence among paralogues. *Syst. Biol.* 53:914–932
Quiros CF, Bauchan GR. 1988. The genus *Medicago* and the origin of the *Medicago sativa* complex. *Agron. Monogr.* 29:737–76
Rannala B, Yang Z. 2003. Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. *Genetics* 164:1645–56
Rautenberg A, Filatov D, Svennblad B, Heidari N, Oxelman B. 2008. Conflicting phylogenetic signals in the SlX1/Y1 gene in *Silene*. *BMC Evol. Biol.* 8:299
Renny-Byfield S, Gallagher JP, Grover CE, Szadkowski E, Page JT, et al. 2014. Ancient gene duplicates in *Gossypium* (cotton) exhibit near-complete expression divergence. *Genome Biol. Evol.* 6:559–71
Rieseberg LH, Archer MA, Wayne K. 1999. Transgressive segregation, adaptation and speciation. *Heredity* 83:363–72. **https://doi.org/10.1046/j.1365-2540.1999.00617.x**
Roose ML, Gottlieb LD. 1976. Genetic and biochemical consequences of polyploidy in *Tragopogon*. *Evolution* 30:818–30
Rosato M, Castro M, Rosselló JA. 2008. Relationships of the woody *Medicago* species (section *Dendrotelis*) assessed by molecular cytogenetic analyses. *Ann. Bot.* 102:15–22
Rothfels CJ, Pryer KM, Li FW. 2017. Next generation polyploid phylogenetics: rapid resolution of hybrid polyploid complexes using PacBio single-molecule sequencing. *New Phytol.* 213:413–29
Sakiroglu M, Doyle JJ, Brummer EC. 2010. Inferring population structure and genetic diversity of broad range of wild diploid alfalfa (*Medicago sativa* L.) accessions using SSR markers. *Theor. Appl. Genet.* 121:403–15
Sang T. 2002. Utility of low-copy nuclear gene sequences in plant phylogenetics. *Crit. Rev. Biochem. Mol. Biol.* 37:121–47
Sang T, Zhang D. 1999. Reconstructing hybrid speciation using sequences of low-copy nuclear genes: hybrid origins of five *Paeonia* species based on Adh gene phylogenies. *Syst. Bot.* 24:148–63
Scheen A-C, Pfeil BE, Petri A, Heidari N, Nylinder S, Oxelman B. 2012. Use of allele-specific sequencing primers is an efficient alternative to PCR subcloning of low-copy nuclear genes. *Mol. Ecol. Res.* 12:128–35
Small E. 2011. Alfalfa and relatives: evolution and classification of *Medicago*. Ottawa, Can.: NRC Research
Small RL, Ryburn JA, Cronn RC, Seelanan T, Wendel JF. 1998. The tortoise and the hare: choosing between noncoding plastome and nuclear Adh sequences for phylogeny reconstruction in a recently diverged plant group. *Am. J. Bot.* 85:1301–15
Smedmark JEE, Eriksson T, Bremer B. 2005. Allopolyploid evolution in *Geinae* (Colurieae: Rosaceae)—building reticulate species trees from bifurcating gene trees. *Org. Divers. Evol.* 5:275–83
Soltis DE, Soltis PS, Tate JA. 2003. Advances in the study of polyploidy since *Plant Speciation*. *New Phytol.* 161:173–91

Sousa F, Bertrand YJK, Doyle JJ, Oxelman B, Pfeil BE. 2017. Using genomic location and coalescent simulation to investigate gene tree discordance in *Medicago*. L *Syst. Biol*. **https://doi.org/10.1093/sysbio/syx035**

Sousa F, Bertrand YJK, Pfeil BE. 2016. Patterns of phylogenetic incongruence in *Medicago* found among six loci. *Plant Syst. Evol.* 302:493–513

Sukumarana J, Knowles LL. 2017. Multispecies coalescent delimits structure, not species. *PNAS* 114:1607–12

Than C, Nakhleh L. 2009. Species tree inference by minimizing deep coalescences. *PLOS Comput. Biol.* 5:e1000501

Than C, Nakhleh L. 2010. Inference of parsimonious species tree from multilocus data by minimizing deep coalescences. In *Estimating Species Trees: Practical and Theoretical Aspects*, ed. LL Knowles, LS Kubatko, pp. 79–97. Hoboken, NJ: Wiley

Than C, Ruths D, Nakhleh L. 2008. PhyloNet: a software package for analyzing and reconstructing reticulate evolutionary relationships. *BMC Bioinform*. 9:322

Tian Y, Kubatko L. 2016. Distribution of coalescent histories under the coalescent model with gene flow. *Mol. Phylogenetics Evol.* 105:177–92

Toprak Z, Pfeil BE, Jones G, Marcussen T, Ertekin AS, Oxelman B. 2016. Species delimitation without prior knowledge: DISSECT reveals extensive cryptic speciation in the *Silene aegyptiaca* complex (Caryophyllaceae). *Mol. Phylogenetics Evol.* 102:1–8

Veronesi F, Mariani A, Bingham ET. 1986. Unreduced gametes in diploid *Medicago* and their importance in alfalfa breeding. *Theor. Appl. Genet.* 72:37–41

Wen D, Yu Y, Nakhleh L. 2016. Bayesian inference of reticulate phylogenies under the multispecies network coalescent. *PLOS Genet*. 12:e1006006

Wen D, Nakhleh L. 2017. Co-estimating reticulate phylogenies and gene trees from multi-locus sequence data. bioRxiv 095539. **http://dx.doi.org/10.1101/095539**

Wood TE, Takebayashi N, Barker MS, Mayrose I, Greenspoon PB, Rieseberg LH. 2009. The frequency of polyploid speciation in vascular plants. *PNAS* 106:13875–79

Zwickl DJ, Stein JC, Wing RA, Ware D, Sanderson MJ. 2014. Disentangling methodological and biological sources of gene tree discordance on *Oryza* (Poaceae) chromosome 3. *Syst. Biol.* 63:645–59