Running Head: Phased Target Enrichment for Polyploids

Phasing Alleles Improves Network Inference with Allopolyploids

George P. Tiley[1,†,*], Andrew A. Crowl[1,†], Paul S. Manos[1], Emily B. Sessa[2], Claudia Solís-Lemus[3], Anne D. Yoder[1], J. Gordon Burleigh[2]

[1]Department of Biology, Duke University, Durham NC, 27708, USA

[2]Department of Biology, University of Florida, Gainesville FL, 32611, USA

[3]Wisconsin Institute for Discovery and Department of Plant Pathology, University of Wisconsin – Madison, Madison WI, 53706, USA

[†]These authors contributed equally

[*]Author for correspondence: george.tiley@duke.edu

Abstract

Accurately reconstructing the reticulate histories of polyploids remains a central challenge for understanding plant evolution. Although phylogenetic networks can provide insights into relationships among polyploid lineages, inferring networks may be hampered by the complexities of homology determination in polyploid taxa. We use simulations to show that phasing alleles from allopolyploid individuals can improve inference of phylogenetic networks under the multispecies coalescent. Phased allelic data can also improve divergence time estimates for networks, which is helpful for evaluating allopolyploid speciation hypotheses and proposing mechanisms of speciation. To achieve these outcomes, we present a novel pipeline that leverages a recently developed phasing algorithm to reliably phase alleles from polyploids. This pipeline is especially appropriate for target enrichment data, where depth of coverage is typically high enough to phase entire loci. We provide an empirical example in the North American *Dryopteris* fern complex that demonstrates how phasing can help reveal the mode of polyploidization and improve network inference. We establish that our pipeline (PATÉ: Phased Alleles from Target Enrichment data) is capable of recovering a high proportion of phased loci from both diploids and polyploids, and that these data improve network estimates compared to using haplotype consensus assemblies. This approach is shown to be especially effective in reticulate complexes where there are multiple hybridization events. The pipeline is available at: https://github.com/gtiley/Phasing.

Key words: Introgression; Hybridization; Reticulate Evolution; Multispecies Coalescent; Divergence Time Estimation; Polyploidy; Target Enrichment; *Dryopteris*

INTRODUCTION

The phenomenon of polyploidy, or whole-genome duplication, occurs throughout the tree of life. Nowhere, perhaps, is its evolutionary significance more evident than in plants, with recent estimates suggesting up to 35% of vascular plant species are of recent polyploid origin (Wood et al. 2009; Barker et al. 2016). Despite advances in genomic data generation and a long-term interest in understanding the role of whole-genome duplication in driving plant speciation and local adaptation (reviewed in Soltis et al. 2014), polyploids remain a central challenge for the field of phylogenetics. One persistent problem when analyzing sequence data from polyploid taxa, and especially allopolyploids, is identifying the alleles and divergent homeolog copies from parental lineages. Most bioinformatic tools for processing next generation sequence data were developed with diploids, or specifically humans, in mind. These approaches often collapse variable homeolog sequences into a single consensus sequence for *de novo* assemblies or assume the organism is diploid when performing genotyping and phasing for reference-based assembly. For polyploids, this creates chimeric sequences that may interfere with phylogenetic reconstruction and obscure signals of polyploidy and polyploid mode-of-origin. Using allelic data that more accurately capture the complex genomic histories of polyploids should enable the incorporation of divergent signals from polyploid loci into phylogenomic inference, distinguish allopolyploidy from autopolyploidy, and identify parental taxa. However, few studies have examined the potential benefits of using phased versus unphased data to reconstruct polyploid histories, and there are few formal methods and little guidance for phasing alleles from polyploid taxa. Here we explore the value of using phased data to reconstruct polyploid networks leverage recent algorithmic advances in polyploid phasing (Xie et al. 2016) to develop a bioinformatic pipeline that can phase alleles from polyploids using target enrichment sequence data.

Previous studies of reticulate complexes have suggested phasing alleles is crucial for accurate evolutionary reconstruction, at least when sampling relatively few loci (e.g., 4 to 10;

Rothfels et al. 2017; Eriksson et al. 2018). The applications of phased sequencing data for phylogenomic studies of polyploid complexes are thus enticing; however, it remains challenging to genotype and phase next generation sequencing data from polyploids. Methods exist to genotype consensus loci from target enrichment data, but these have been either limited to diploids (Kates et al. 2018; Andermann et al. 2019) or manual curation of variants with polyploids where both parental populations are available (Eriksson et al. 2018). Otherwise, obtaining phased sequence data for polyploids has largely depended on costly long-read sequencing to recover complete haplotype sequences (e.g., Rothfels et al. 2017), or cloning PCR products (e.g., Sessa et al. 2012; Oberprieler et al. 2017).

Target enrichment (or HybSeq), where specific regions of the genome are isolated and sequenced (Faircloth et al. 2012; Lemmon et al. 2012), is an increasingly common method for collecting large-scale phylogenomic datasets, and these data can provide insights into the evolutionary history of reticulate complexes (e.g., Crowl et al. 2017; Karimi et al. 2020) and sources of gene tree discordance (e.g., Morales-Briones et al. 2018; Stull et al. 2020). Probe kits for target enrichment have been developed in many land plant lineages (Wolf et al. 2018; Johnson et al. 2019; Liu et al. 2019; Breinholt et al. 2021), and there are bioinformatic pipelines available for custom probe design (e.g., Jantzen et al. 2020). The most common approach to assemble phylogenetic datasets from target enrichment data has been to use *de novo* assembly pipelines (Faircloth 2016; Johnson et al. 2016; Andermann et al. 2018; Breinholt et al. 2018). The assembly algorithms within these pipelines typically treat variable base calls as sequencing errors and consider only the most frequent nucleotide sequence while discarding the alternatives (Bankevich et al. 2012; Iqbal et al. 2012; Luo et al. 2012). This results in loci in which variable positions are collapsed to a single base call (haplotype consensus loci), losing information related to heterozygosity. While this may be appropriate, or at least benign, for phylogenetic analyses of diploid taxa (Kates et al. 2018), it may pose substantial problems when attempting to investigate the evolutionary history of polyploid taxa or reticulate lineages.

While phylogenetic studies in plants often infer strictly bifurcating trees, the complexities of allopolyploid evolution can be represented more accurately using networks. Simulations and empirical analyses have suggested that phylogenetic networks can recover the reticulate histories of polyploid lineages with few loci, at least when gene tree discordance due to incomplete lineage sorting (ILS; Hudson et al. 1983; Pamilo and Nei 1988) is low (Oberprieler et al. 2017) using parsimony methods (Huber et al. 2006; Lott et al. 2009), or even with moderate ILS when explicitly modeled (Jones et al. 2013). Contemporary phylogenetic network models and software packages jointly consider gene tree variation due to allele sampling error as described by the multispecies coalescent (MSC; Rannala and Yang 2003) and gene flow modeled as episodic introgression events (Solis-Lemus and Ané 2016; Wen et al. 2016; Zhang, Ogilvie et al. 2018, Flouri et al. 2020). Depending on the complexity and goals of the research question, these methods can search for networks with a constrained number of reticulation events using quartet-based maximum pseudolikelihood (Solis-Lemus and Ané 2016; Wen et al. 2018) or a full-likelihood Bayesian model where the number of reticulations is a parameter (Wen et al. 2018; Zhang et al. 2018). Also, it is possible to estimate model parameters (divergence times, population sizes, and the fraction of introgressed genes) on a fixed species network using a full-likelihood Bayesian model that allows efficient computation with large numbers of loci (Flouri et al. 2020). We refer to these network models from here on as the multispecies coalescent with introgression (MSci), consistent with Flouri et al. (2020); although, other names have been used, such as the network multispecies coalescent (NMSC; e.g Zhu and Degnan 2017) and multispecies network coalescent (MSNC; e.g., Wen et al. 2016). We emphasize that network approaches to investigate polyploid complexes are not novel (e.g., Huber et al. 2006; Lott et al. 2009; Jones et al. 2013), but the difficulty of collecting appropriate genomic data from polyploids for such analyses has limited their use.

To address the issues outlined above, we have developed a pipeline, PATÉ (Phased Alleles from Target Enrichment data), that can phase genotyping data for individuals of a known ploidy without the need for sampling their parental lineages. PATÉ was designed with scalability and population-level sampling in mind for target enrichment projects where deep coverage from paired-end Illumina data allow calling of high-quality variants. In this study, we first use simulations to explore the ability of network approaches to reconstruct the history of allopolyploidy in the presence of ILS, and whether phasing the data affects the accuracy of the reconstruction. We show that using phased allelic data can improve network estimation and divergence time estimation compared to using haplotype consensus sequences, but also highlight scenarios where phasing may not be necessary or beneficial. Next, we describe the individual steps used by PATÉ to phase target enrichment data. For an empirical example of the benefits of PATÉ, we compare phased and unphased (haplotype consensus) data to infer the evolutionary history of the North American *Dryopteris* complex, a model system for reticulate polyploid evolution (Sessa et al. 2012a; Sessa et al. 2012b), using new targeted enrichment data. The system includes four diploid species, as well as one extinct diploid, that have formed five allopolyploids in which there is high confidence in the parent-progeny relationships (Fig. 1), although numerous sterile allopolyploid species have also been reported within the complex (Montgomery and Paulton 1981). The allopolyploid species have relatively ancient origins, with the best estimates placing hybridization events between six and 13 Ma (Sessa et al. 2012b). PATÉ is largely successful in recovering phased haploid sequences from polyploid individuals, and networks inferred from phased data more accurately represent the North American *Dryopteris* complex than those inferred from unphased data.

MATERIALS & METHODS

Testing the Effects of Phasing on Network Inference through Simulation

Simulating phased and unphased sequence data for an allopolyploid — We simulated gene trees and their nucleotide sequence data using the MSC model with BPP v1.4.1 (Flouri et al. 2018) under a five-species network (Fig. 2). All simulations used the Jukes-Cantor (Jukes and Cantor 1969) model of sequence evolution with no rate heterogeneity. The allopolyploid species *E* was treated as two lineages (*E* sister to *B* and *F* sister to *C*) whose alleles were pooled to form the hybrid species *E* at time τ_h. This makes species *E* a tetraploid hybrid with the parents *B* and *C*. θ was constant among lineages and set at 0.01. Assuming a per-generation mutation rate (μ) of 1×10^{-8} and one year per generation yields an effective population size (N_e) of 250,000 and root age of 10 Ma for the simulation network (Fig. 2). We also simulated data under a shallow divergence scenario, in which all node ages were divided by 10, and a deep divergence scenario, in which all node ages were multiplied by 10. This changes the root age (τ_r) to 0.01 (1 Ma) and 1.0 (100 Ma), respectively.

Because the distance between speciation nodes is 0.025 and $\theta = 0.01$ for the simulation network (Fig. 2), there are five coalescent units between nodes, $T = \frac{\tau}{\left(\frac{\theta}{2}\right)}$ (Yang 2006), which implies a near-zero probability of gene tree discordance due to ILS (Hudson 1983). We incorporated ILS into the simulation by reducing the node heights τ_u and τ_s to 0.0375 and 0.05 (for a low level of ILS) and 0.03 and 0.035 (for a moderate level of ILS). This increases the probability of ILS at node *u* and *s* to 0.05 and 0.25, respectively. Thus, we used a total of nine simulation conditions that combined three levels of evolutionary distance and three levels of ILS. While our simulations do not explore extreme levels of gene tree discordance, they allow us to learn about some general features of increasing ILS on network inference with different data types. We simulated 1000 gene trees and their sequences, 500bp in length, under the MSC for each of the nine conditions. We also explored the effects of sampling fewer genes on downstream analyses. For each replicate of 1000 gene trees and their sequences, we randomly sampled 400, 40, and four genes without replacement.

All simulations sampled haploid data. Two haploid sequences were sampled for each diploid species and four haploid sequences were sampled for the allopolyploid species *E,* where two sequences came from each parental lineage. We then investigated unphased data in three ways. First, to generate unphased genotype sequences, the simulated haploid sequences for each species were collapsed into a single sequence in which heterozygous sites were represented by IUPAC ambiguity codes (genotype). The allopolyploid species was not restricted to only biallelic sites. Second, we generated haploid consensus sequences, where for each variable site, only one base was randomly retained (consensus). This could represent a case in which read coverage across a locus is highly uneven such that a haploid sequence is actually a chimera of two or more alleles. Finally, we simply picked one phased haploid sequence, which is possible when one parental haplotype has a majority of reads for a locus (pick one). This scenario where only one parent's sequence would be recovered in the offspring could be anticipated in real data due to subgenome dominance (e.g., Buggs et al. 2014; Bird et al. 2018). In practice, we expect most *de novo* assemblers to generate output in between the haploid consensus data and pick one data.

Inferring species networks with phased and unphased simulated data — We estimated species networks with PhyloNetworks v0.12.0 (Solis-Lemus et al. 2017) using Julia v.1.4.1 (Bezanson et al. 2015) from either the true gene trees used to simulate the data, or gene trees estimated from the phased or unphased sequence data. For estimated trees, we used IQTREE v1.6.10 (Nguyen et al. 2015) with the same model used for simulation. Each PhyloNetworks analysis used the species tree (A,((B,E),(C,D))) as the starting tree and allowed zero, one, or two reticulation events. Each analysis included ten independent optimizations of the pseudolikelihood score. We considered larger numbers of reticulations an improvement if they were two or more pseudolikelihood units lower than the best model. We compared the estimated networks with one reticulation to the true network with the *hardwiredClusterDistance* function (Huson et al. 2010) in PhyloNetworks. This allowed us to score the number of replicates that 1) recovered the correct number of reticulations and 2) matched the true network when the number of reticulations was set to one. We estimated networks for samples of four, 40, 400, and 1000 gene trees for each of the 30 replicates for each of the nine simulation divergence and ILS conditions.

Effect of phasing on divergence time estimation — We also used our simulated phased and unphased data to estimate divergence times under the MSci model (Flouri et al. 2020) using BPP v4.2.9. Here, we estimate divergence times (τs), population sizes (θs), and the proportion of introgressed loci (φ) on the correct fixed species network. MSCi analyses used diffuse priors on τ_r and θ with a mean on their simulated values with $\varphi \sim \beta(1{,}1)$. Phased, consensus, and pick one sequences were treated as haploids while genotype sequences were treated as unphased diploids and used the analytical phasing (Gronau et al. 2011) implemented in BPP. Although this is not correct for the tetraploid, it is arguably more appropriate than treating all of the genotype sequences as haploid. Each Markov chain Monte Carlo (MCMC) analysis collected 10,000 posterior samples, saving every 100 generations, while discarding the first 100,000 generations (i.e., 10% of the total run) as burnin. All scripts for simulation and subsequent analyses of simulated data are available in Dryad (X).

A Phasing Pipeline for Polyploids

Target enrichment data — We were motivated by the general premise of using phased data to infer reticulate evolutionary histories of polyploids based on the success of empirical studies where phasing was informative about hybridization or introgression events (e.g., Oberprieler et al. 2017; Eriksson et al. 2018). We were aware of few instances of phasing genomic or phylogenomic data in polyploids, except in cases where chromosome-level whole-genome assemblies have characterized subgenomes in allopolyploid crops (Yang et al. 2017; Colle et al. 2019) or emerging results that are dependent on the sampling of parental lineages (Freyman et al. 2020; Nauheimer et al. 2020). We designed PATÉ for target enrichment data because of the availability of such data for many of taxa, but it is applicable to other types of data with paired-end Illumina reads.

The end product of a *de novo* target enrichment assembly pipeline (such as HybPiper; Johnson et al. 2016) generally is a single consensus sequence for each locus for each individual. Allelic variation may be represented by ambiguous nucleotide codes within the single consensus sequence or lost when the pipeline outputs the haplotype consensus sequence where the majority vote from a collection of reads is used. We use these existing *de novo* assembly pipelines as a starting point to provide the reference sequence for each locus for each individual and leverage a recent phasing algorithm with high-quality variant calls to recover phased haplotype sequences for taxa with known ploidy levels. Ploidy levels were well-characterized for individuals in our *Dryopteris* analyses, but for unknown systems there are existing methods to estimate ploidy directly from target enrichment data (Weiss et al. 2018; Viruel et al. 2019) in the absence of other sources, such as flow cytometry (Farhat et al. 2019).

Phasing alleles within loci — PATÉ (Fig. 3) starts with assembled target enrichment loci, such as the supercontig files output from HybPiper (Johnson et al. 2016) that contain a single haplotype consensus sequence from each individual per locus. Reads for each individual are then realigned to their consensus locus using BWA v0.7.17 (Li and Durbin 2009). PCR duplicates are flagged with MarkDuplicates in Picard v2.9.2 (http://broadinstitute.github.io/picard), and variant calls are computed with HaplotypeCaller in GATK v.4.1.4 (McKenna et al. 2010). We applied the following hard filters with VariantFiltration in GATK: (1) QD < 2.0, (2) FS > 60.0, (3) MQ < 40.0, (4) ReadPosRankSum < –8.0, (5) AF < 0.05 || AF > 0.95. These loosely follow community recommendations on filters for germline variant discovery (DePristo et al. 2011). Notably, we do not perform quality score recalibration or filter on the mapping quality rank sum, as we anticipate allopolyploids could have a lower mapping quality associated with an alternate allele due to sequence divergence or structural variation among homeologous chromosomes. We also consider a very narrow window for filtering on allele frequency. Because increasing ploidy levels will generate smaller anticipated ratios of alternate to reference alleles, coupled with sequencing error and read stochasticity, we only aim to remove the most extreme cases. For example, if almost all reads support the alternate allele at a site, it is difficult to diagnose if the error lies in the consensus assembly or the read alignment. In these cases, only the reference site is retained, and the variant does not pass the allele frequency filter. However, the allele frequency filter could be removed if investigators are focused on organisms with extremely high ploidy levels.

Biallelic SNPs that pass filters are then phased with H-PoPG v.0.2.0 (Xie et al. 2016). H-PopG solves a heuristic phasing problem efficiently using dynamic programming. Although not guaranteed to be an optimal solution, H-PoPG has been shown to have high accuracy while also being fast (Xie et al. 2016; He et al. 2018; Moeinzadeh et al. 2020). Phasing variants in polyploids is difficult because for *n* variants and *k* ploidy, there are $2^{n-1}(k-1)^n$ possible ways to link the sites together. H-PoPG evaluates possible solutions efficiently by grouping reads into *k* groups in such a way that differences within the groups are minimized. Focusing on target enrichment data also constrains the complexity of the phasing problem compared to whole-genome alignments (i.e., haplotype blocks are constrained to about 1000 bp). We then used the phased variants to create individual allele sequences, where invariable sites are filled in based on the reference sequence. In cases where there is no linkage information to phase across the entire locus, we retain the phasing only for the longest block. The variants for the shorter haplotype blocks can be collapsed into IUPAC ambiguity codes or treated as missing data based on the investigator's preferences. PATÉ outputs analysis-ready fasta files with multiple alleles per species. Variants are only phased within loci; we do not attempt to assign loci to parental subgenomes. While this may complicate analyses of concatenated multi-locus datasets, it is ideal for the MSC that assumes free recombination between loci and can leverage multiple alleles per species for estimating θs. Those interested in concatenated analyses can use other recent approaches that assign gene copies to parental subgenomes (Freyman et al. 2020; Nauheimer et al. 2020).

Analyses of a Species Complex with Allopolyploidy

The North American wood fern complex (Dryopteris) — We tested PATÉ using new target enrichment data from nine North American *Dryopteris* species, including both allotetraploid and allohexaploid taxa, with well-studied reticulate relationships (Sessa et al. 2012a, b) as well as two outgroup taxa from the sister genus *Polystichum*. All putative parental lineages are represented in our dataset, with the exception of a hypothesized extinct lineage (*D. semicristata*; Sessa et al. 2012b). We sampled two or three individuals for each *Dryopteris* taxon (Table S1). The target enrichment data were generated from the GoFlag 408 flagellate land plant probe set (Breinholt et al. 2021) at RAPID Genomics (Gainesville, FL). The target regions for this probe set are 408 exons found in 229 single or low-copy nuclear genes. We generated haplotype consensus assemblies for each with HybPiper (Johnson et al. 2016). The resulting supercontig sequences became our reference sequences for genotyping and phasing with PATÉ. We aligned both phased and unphased (i.e., the reference supercontig sequences) with MUSCLE with default settings (Edgar 2004).

Three species tests — We first explored the value of phasing data when estimating reticulate relationships among three species, including two diploid parental lineages (*D. expansa* and *D. intermedia*) and their putative allotetraploid descendent (*D. campyloptera*). The two diploid parents last shared a common ancestor during the Late Eocene and Early Miocene, approximately 23 Ma (Sessa et al. 2012b). We used both a full-likelihood Bayesian approach and a topology-based pseudolikelihood approach to estimate the correct species relationships from phased and unphased data. First, using BPP v.4.1.4 (Flouri et al. 2020), we estimated log-marginal likelihoods (ln *mL*) with stepping-stone sampling (Xie et al. 2011) for the three possible rooted three taxon trees and twelve possible network models that imply differences for the timing and direction of allopolyploidy and the presence of unsampled ancestral lineages (Supplementary Fig. S1). Each ln *mL* estimate used 24 steps, and each step had a posterior sample of 10,000, saving every 100 generations after a 100,000 generation burnin (10% of the total run). The ln *mL* values were then used to calculate the fifteen model probabilities following equation 1 (e.g., Beerli et al. 2019).

$$P(model) = \frac{exp(ln\, mL_{model}\ - ln\, mL_{max}\ )}{\Sigma_i[exp(ln\, mL_i\ - ln\, mL_{max}\ )]} \qquad Equation\ 1$$

We repeated analyses for 30 random subsets of 40 and then four loci to explore the effects of the number of loci on inferring allopolyploidy. Next, we used PhyloNetworks to test the presence and placement of gene flow between the three species. For that analysis, we also included the sequences from the two *Polystichum* outgroups. IQ-TREE v1.6.10 (Nguyen et al. 2015) and model selection by ModelFinder (Kalyaanamoorthy et al. 2017) was used to estimate gene trees for the phased and unphased data. The starting tree was obtained with ASTRAL III v5.6.3 (Zhang, Rabiee et al. 2018). We tested the presence of zero, one, or two reticulations with slope heuristics (Solis-Lemus and Ané 2016). Each analysis used ten independent optimizations. In addition to the dataset of all loci, we analyzed the same 30 random subsets of 40 and four loci from the marginal likelihood analyses.

Nine species tests — We also investigated the differences in results from phased versus consensus sequences when estimating a network for the nine-species complex, which involves multiple reticulation events on an edge and thus should be difficult for network estimation (Solis-Lemus and Ané 2016). Because the increased number of species and complexity of reticulation in *Dryopteris,* evaluation with marginal likelihoods was not computationally feasible. Instead, we performed analyses of the nine-species complex and two outgroups with phased and unphased haplotype consensus data with PhyloNetworks as described above, but we allowed up to six reticulation events. We used ASTRAL III v5.6.3 (Zhang, Rabiee et al. 2018) to generate the starting species tree for network estimation using gene trees inferred from IQ-TREE v1.6.10 (Nguyen et al. 2015) with the best model selected by ModelFinder (Kalyaanamoorthy et al. 2017).

Results

Simulation Shows Benefits of Phased Data

Network inference — In our simulation results, both phased (i.e., the haplotypic allele sequences) and unphased data (i.e., genotype, consensus, and pick one) performed well when the goal is only to detect the correct number of reticulations (Supplementary Fig. S2). The only case where analyses did not converge to the true number of reticulations was in the presence of moderate ILS and high nucleotide divergence; however, this appears largely due to gene tree estimation error, as analyses using the true simulated gene trees greatly outperformed those using gene trees estimated from the simulated sequence data. However, using phased data provides more accurate estimates of the placement of the reticulation edge in comparison to using genotype data, and to a lesser extent, consensus sequences (Fig. 4). When the true gene trees were used, which have information about the allopolyploid’s hybridization event (i.e., the allele sequences are sister to their respective parents in every tree), the correct network can be inferred with 40 or fewer loci when nucleotide divergence is moderate. The gene trees estimated from phased data perform equally well, although they require a few more gene trees when nucleotide divergence is low and ILS moderate. Analyses using the genotype data almost never recover the true network for these medium and low divergence scenarios, regardless of the amount of ILS. Analyses based on pick one data perform similarly well to the phased and true data when sampling 400 loci, but they are less accurate for four or 40 loci at low and medium divergences. Analyses using the consensus data perform poorly for low numbers of loci under a low divergence and no ILS scenario, but they are capable of recovering the true reticulation when sampling 400 or more loci for the other low and medium divergence cases. In the high divergence simulations, all data types could infer the true network if enough genes were sampled, but analyses with the phased data required fewer genes than others.

Divergence times — Using phased data also improves divergence time estimates when nucleotide divergence is low, but not when divergence is moderate or high (Fig. 5). When divergence was low, the timing of reticulate events was accurately estimated when using phased haplotypic data, but was overestimated when using genotype and especially consensus data. For analyses with genotype and consensus data, as the number of loci increased and uncertainty in the posterior was reduced, the posterior mean did not converge to the true estimate and the simulated value was not within the highest posterior density (HPD) interval. Additionally, all other nodes in the species network were overestimated with genotype or consensus data, while the phased data were capable of recovering the simulated divergence times (Supplementary Figs. S3-S5). Aside from some cases with the consensus sequences, all four data types performed similarly with four loci; however, this is likely due to the posterior being dominated by the prior in the absence of enough data. There was little improvement in divergence time estimates when going from 40 to 400 loci, aside from further reduction in the HPD intervals.

For medium sequence divergence, there was little difference between the phased and unphased data. Phased sequences slightly underestimated the timing of hybridization while unphased data slightly overestimated the timing of these events. However, phased data accurately estimated the age of older speciation nodes that were again systematically overestimated with unphased data (Supplementary Figs. S6-S8). At a high level of nucleotide divergence, genotype data were capable of accurately estimating all divergence time parameters while the phased data underestimated the timing of hybrid events (Fig. 5; Supplementary Figs. S9-S11). Notably, the pick one data performed very well for all divergence time estimation scenarios. Divergence times were not strongly affected by increasing levels of ILS, likely because estimates were performed with the MSci model and we did not explore very high ILS scenarios, but age estimates improved slightly for the genotypes, consensus, and pick one data with increasing ILS.

Analyses of target enrichment data from Dryopteris

Recovery of Phased Loci — On average, 62% of loci sequenced for an individual were phased with eight variants passing filters (Table S2). The ploidy level appears to be strongly associated with the number of phased loci. Among diploids, only 30% of loci were phased; the other 70% of diploid loci were either homozygous or had too few linked variants for phasing. For tetraploids and hexaploids, 87% and 94% of loci, respectively, were phased such that two or more phased haplotype sequences could be recovered. Among loci where phasing was possible, variants were almost always resolved as a single haplotype block, as opposed to being split into two or more blocks because not enough reads were available to physically link variants. For polyploids, the number of phased haplotype sequences most frequently matched the ploidy level except in the case of a single *D. campyloptera* individual (B087-D08), which also had few recovered loci. Phasing data only extended sequence alignment length by about four base pairs, but it more than doubled the number of parsimony informative sites (Table S3).

Placing a single reticulation event — For our three-species full-likelihood analyses with the MSci, both phased and unphased data recovered the anticipated reticulation hypothesis, identifying *D. campyloptera* as an allotetraploid with the two diploid parental lineages *D. expansa* and *D. intermedia*, when using all loci (Table 1). Model probabilities show decisive support for a scenario where there are two unsampled ancestral populations that were the progenitors of *D. campyloptera*, as opposed to *D. campyloptera* being a hybrid species with extant *D. expansa* and *D. intermedia* as parents. All model parameters converged for both phased (Supplementary Figs. S12 and S13) and unphased data (Supplementary Figs. S14 and S15). Divergence time estimates were older in the analysis of phased data, although the relative order of divergence events was the same using phased and unphased data (Fig. 6). There was more uncertainty in the θ estimates, especially for the alloplyploid species *D. campyloptera* and the two ancestral populations of the parental lineages that formed the polyploid ancestor (Supplementary Figure S16). Repeating the analyses with fewer loci did not always produce the same result, but there was either decisive support for a model or multiple plausible models that all had the correct species relationships and direction of introgression for both phased and unphased data when using 40 loci (Fig. 7). The only difference between models was the presence or absence of ancestral θs for unsampled lineages. Analyses with four loci produced less reliable results for both phased and unphased data. The four-locus analyses of phased and unphased data found some non-negligible model probability for trees without hybridization or networks where hybridization was incorrect in nine and twelve out of 30 replicates, respectively (Fig. 7).

When performing a network search based on gene tree distributions, both the phased and unphased data were able to recover the hypothesized allopolyploidy event (Fig. 8; Supplementary Table S5). Both analyses inferred the major branch to be *D. intermedia* with the minor branch from *D. expansa*. Phased data estimated a slightly smaller inheritance probability compared to the unphased data. These findings from PhyloNetworks are consistent with parameter estimation under the MSci model, where phased data estimated a slightly smaller mean φ_h compared to unphased data (Table S4). When sampling 100 replicates of 40 loci, 98% of replicates for phased data and 100% for unphased data were capable of detecting a single hybrid event in the data. When sampling four loci, this drops to 66% and 85%, respectively (Fig. 8). Phased data more frequently recovered the correct network (58%) compared to unphased data (38%) with 40 loci; however, the converse is true for four loci, with 34% correct for phased and 41% correct for unphased (Fig. 8). Unphased data also got the network wrong more frequently than phased data, such that phased data had the direction of introgression incorrect in 36% and 12% of replicates while unphased data were incorrect in 60% and 38% replicates for 40 and four loci, respectively (Fig. 8).

Inferring relationships among a complex with multiple reticulation events — The network recovered for phased data identified three out of five hypothesized reticulation events among the nine *Dryopteris* species (Fig. 9; Supplementary Table S5). The allotetraploid *D. celsa* was correctly identified with diploid *D. ludoviciana* and *D. goldiana* as parents. Analysis of the phased data detected a low level of gene flow from the common ancestor of this clade into tetraploid *D. cristata,* which has *D. ludoviciana* as one hypothesized parent while the other parent lineage (*D. semicristata*) is assumed to be extinct (Sessa et al. 2012b). *Dryopteris carthusiana* is another tetraploid that is assumed to share the extinct common ancestor with *D. cristata*, but has experienced introgression from *D. intermedia*, with a high inheritance probability of 0.43 (Fig. 9). However, the phased data missed the putative allotetraploid case of *D. campyloptera*, despite the strong evidence for this reticulation event in our earlier three-taxon analyses. Our network with phased data also failed to identify the putative reticulate evolutionary history of *D. clintoniana*, an allohexaploid where allotetraploid *D. cristata* and diploid *D. goldiana* are assumed to be the parents (Sessa et al. 2012b). The spectra of quartet concordance factors were overall similar between the phased and unphased data (Supplementary Fig. S17), but the phased data were arguably more accurate.

Although the phased data were not successful in recovering all hypothesized reticulate relationships, they performed better than the unphased data. Analyses of unphased data were capable of finding the allotetraploid history of *D. celsa* with an inheritance probability similar to the phased data (Fig. 9). Hybridization between *D. intermedia* and *D. carthusiana* was also detected; however, the directionality was reversed, with gene flow going from the allotetraploid into the diploid. A third reticulation edge was found in the unphased analysis, from the common ancestor of *Dryopteris* into the common ancestor of *D. clintoniana* and its sister clade. This hybrid edge is difficult to reconcile because of the hypothesized extinct common ancestor that contributed to both *D. cristata* and *D. carthusiana*. *Dryopteris clintoniana* is correctly placed in the major species tree topology, as a grade between *D. cristata* and *D. goldiana*. This reticulation edge from the *Dryopteris* common ancestor may reflect the extinct lineage and a high degree of gene tree variation.

Discussion

New phylogenetic network methods offer the promise of elucidating the often complex reticulate histories of polyploid lineages, even in the presence of ILS. Our results demonstrate that phasing polyploid target enrichment data can improve the accuracy of such network inferences as well as divergence time estimates for the networks, and we describe a novel pipeline (PATÉ) to address the difficult problem of phasing polyploid data. Although PATÉ could handle different types of genomic data, such as transcriptomes and whole genomes, target enrichment data are ideal for investigation because they often yield high and even coverage across loci. Because MSC methods assume treat loci independently and assume free recombination between loci, it is not necessary to assign individual loci to parental subgenomes. However, the allele sequences output by PATÉ can also be used as input for the recently developed *Homologizer* (Freyman et al. 2020) or *HybPhaser* (Nauheimer et al. 2020), which attempt to phase across loci and recover parental subgenomes. Our methods also enable population genomic studies of polyploids where accurate estimation of site frequency spectra can be used for demographic analyses otherwise complicated by polyploidy (e.g., Excoffier et al. 2013; Liu and Fu 2020) or SNP-based network inference in the absence of variation suitable for gene tree estimation (Blischak et al 2018; Olave and Meyer 2020).

Promises and Pitfalls of Phasing

The prospect of using alleles from phased genomic data presents an exciting step towards revealing the evolutionary history of polyploids, which remains a critical impediment within the plant evolution community (McKain et al. 2018). Strategies for explicitly addressing this challenge are only now emerging (Freyman et al. 2020; Nauheimer et al. 2020), and PATÉ can be a useful tool by phasing variants for many individuals while leveraging genotyping information. Our simulations demonstrate that phasing can improve estimates of reticulate evolutionary relationships using network methods. Phased data can more accurately recover the placement and directionality of hybrid edges than various types of unphased data in simulations (Fig. 4) and empirical analyses with limited numbers of loci (Fig. 7). The advantages of using phased versus unphased data for network estimation decrease when a large number of loci are sampled (Table 1; Fig. 8).

Perhaps an underappreciated aspect of phased data is their ability to improve divergence times estimates (Fig. 5; Supplementary Figs. S3-S11; Anderman et al. 2019). Our empirical analyses also demonstrated how the timing of introgression τ_h can be greatly affected by whether phased or consensus data are used. In our *Dryopteris* analyses, the estimate from phased data was nearly four times older than the estimate from unphased data (Supplementary Table S4). The direction of this difference was unanticipated, because our simulations suggested the consensus data should overestimate age compared to phased data. This highlights the difficulties of simulating data that capture real complexities and makes deciding which estimate is more reliable somewhat difficult; however, the uncertainty of τ_h for haplotype consensus data reflected in its posteriors (Supplementary Fig. S14) gives us more confidence in the phased estimates (Supplementary Fig. S12). Because we used the MSci model for divergence time estimation, we did not observe any effect of ILS on age estimates in our simulations, but if we were using concatenation methods to date the divergence times for the allopolyploids two subgenomes, nodes affected by ILS should be overestimated (Tiley et al. 2020).

In most cases phasing appears to be beneficial, but it may be problematic when the parental lineages are deeply diverged. Although the phased data were able to accurately estimate the age of older speciation nodes (Supplementary Figs. S3-S11), as phylogenetic information is lost from multiple hits, the influence of the prior becomes more substantial. These deep divergence simulation scenarios may border on being biologically unrealistic, as identifying an allopolyploid and its two parental lineages becomes more difficult over time due to extinction and population genetic processes, but it provides some expectations for the performance of phased and unphased data in the presence of high nucleotide divergence. Our simulations showed that when alleles are phased but only one is sampled, as in our pick one simulations, network and divergence time estimation is similar to having all phased alleles present. We also showed where consensus data can perform poorly through simulation (Figs. 3 and 4) as well as empirical analyses where the direction of introgression was more frequently reversed in a simple example of two parental lineages and an allopolyploid (Fig. 8). When enough reads are available to call high-confidence variants, we suggest that phasing with PATÉ can improve network and divergence time estimation for species complexes with low to moderate sequence divergence. However, when investigating very ancient hybrid events, unphased genotype data may be preferential, and using analytical approaches that integrate over phases (Gronau et al. 2011; Flouri et al. 2020) may outperform analyses with phased sequences because allele-specific information no longer captures shared ancestry with parental lineages and haplotype blocks become smaller due to recombination.

Challenges of Network Estimation

Our analyses highlight the difficulty of estimating the evolutionary histories of reticulate complexes using phylogenetic methods, regardless of data type. The full-likelihood implementation of the MSci model appears to be useful for limited cases, but these methods are computationally demanding. Thus, they may not be practical for generating hypotheses and exploring unknown relationships for large numbers of taxa (e.g., Zhang et al. 2018). Quartet-based methods are fast and accurate when there are limited numbers of reticulations (Solis-Lemus and Ané 2016) and show similar accuracy to full-likelihood methods for estimating introgression probabilities (Flouri et al. 2020). However, there are scenarios where true network topologies become non-identifiable (Solis-Lemus and Ané 2016). For example, when multiple introgression events affect the same lineage, the expected quartet distribution under the MSci model becomes a poor fit for the empirical data (Cai and Ané 2020) and the network estimated may be incorrect. These effects were evident in our empirical analyses where the relationship between *D. campyloptera*, *D. expansa*, and *D. intermedia* was missing in the nine-species analysis (Fig. 9), which we expect is due to a reticulation edge present between *D. intermedia* and *D. carthusiana*. Phasing sequence data adds information that can improve estimates (Supplementary Table S3), but unsampled or extinct lineages, such as the hypothesized *D. semicristata,* can create significant barriers to recovering the true evolutionary history of reticulate complexes, regardless of how many loci or individuals are available.

Insights into Dryopteris Evolution

The North American *Dryopteris* complex has been well-characterized through the study of multi-locus nuclear and chloroplast phylogenies, morphology, cytological observations, chromatography, and isozyme analyses (reviewed in Sessa et al. 2012b). This makes it a useful system for testing our phasing pipeline, and our analyses add nuance to our understanding of some of the relationships among *Dryopteris* species. For example, our results indicate *D. campyloptera* has received slightly more loci from *D. intermedia* than *D. expansa* (Fig. 8), and *D. intermedia* is thought to be the likely maternal progenitor (Sessa et al. 2012b). Although our analysis of marginal likelihoods for all target enrichment loci suggested the presence of unsampled ancestral populations (Table 1), the age of introgression and divergence of the *D. campyloptera* ancestral population from the *D. intermedia* and *D. expansa* parental lineages is consistent with hybrid speciation rather than a lineage that was isolated from *D. intermedia* and received more recent gene flow from *D. expansa* (Fig. 6). Following the allopolyploidy event, the *D. intermedia* genome was likely dominant, providing some selective advantage for *D. campyloptera* in its distribution at the time (Bird et al. 2018). Similar insights can be gained from the nine-taxon analyses, which suggests *D. ludoviciana* is the dominant genome in *D. celsa*. *Dryopteris ludoviciana* is the maternal parent of *D. celsa*, again suggesting some bias in retaining homeologous alleles from the maternal lineage, which provides the chloroplast genome in ferns (Sessa et al. 2012b).

The nine-taxon analyses also support the hypothesis of the unsampled diploid lineage *D. semicristata*, based on the placement of *D. carthusiana* as sister to the rest of *Dryopteris* in the phased nine-taxon analyses, but with *D. carthusiana* having received over 40% of its genes from *D. intermedia* more recently. Our analyses suggest that *D. cristata* did not have *D. ludoviciana* as a progenitor, but rather an unsampled common ancestor of *D. goldiana* and *D. ludoviciana*. In the case of *D. cristata*, both parental diploid lineages, including *D. semicristata,* may have gone extinct.

Conclusions

Combining phased data with recent network methods holds much promise for confronting a major challenge of plant phylogenetics: resolving the complex histories of polyploids. The PATÉ pipeline can enhance systematic, speciation genomic, and population genomic analyses of groups containing polyploids. While haplotype consensus sequences may be adequate for resolving single reticulation events where both parents are sampled, using phased sequences can improve inferences of more complicated allopolyploid events, demonstrating how allelic variation can be leveraged for MSC methods that account for reticulation. Still, some reticulate complexes can be difficult to disentangle with any data when there are multiple hybrid events involving the same branch. PATÉ is available through GitHub (https://github.com/gtiley/Phasing) and can be run on any UNIX environment after installing basic genotyping software and H-PoPG.

Data Availability

PATÉ is freely available through GitHub at https://github.com/gtiley/Phasing. Simulated and empirical data supporting findings and files for replicating analyses are available from the Dryad Digital Repository: (X). Raw Fastq reads for *Dryopteris* individuals are available through the NCBI SRA database and are associated with BioProject PRJNA725004. Individual SRA Identifiers are available in Supplementary Table S1.

Supplementary Material

Data available from the Dryad Digital Repository: (X).

Acknowledgements

The authors thank A.M. Duffy, K. Imwattana, M. Nieto-Lugilde, B.T. Piatkowski, and A.J. Shaw for helpful discussions and providing feedback on the manuscript. We also thank M.G Johnson, J. Mendez Reneau, and L. Nauheimer for discussions and sharing their strategies for phasing sequence data. We are grateful to the people who made data collection possible; we used *Dryopteris* samples collected by C.J. Rothfels, M.A. Sundue, and W. Testo, and DNA extractions were performed by S.B. Carey and E. Lockwood.

Funding

Funding was provided by National Science Foundation awards DEB-2038213 to AAC and PSM, and from DEB-1541506 to JGB and EBS. This work was also partially supported by the Department of Energy award DE-SC0021016 to CSL. ADY and GPT gratefully acknowledge support from Duke University.

LITERATURE CITED

Andermann T., Cano A., Zizka A., Bacon C., Antonelli A. 2018. SECAPR-a bioinformatics pipeline for the rapid and user-friendly processing of targeted enriched Illumina sequences, from raw reads to alignments. *PeerJ* 6:e5175.

Andermann T., Fernandes A.M., Olsson U., Topel M., Pfeil B., Oxelman B., Aleixo A., Faircloth B.C., Antonelli A. 2019. Allele Phasing Greatly Improves the Phylogenetic Utility of Ultraconserved Elements. *Syst. Biol.* 68:32-46.

Baaijens J.A., Schonhuth A. 2019. Overlap graph-based generation of haplotigs for diploids and polyploids. *Bioinformatics* 35:4281-4289.

Bankevich A., Nurk S., Antipov D., Gurevich A.A., Dvorkin M., Kulikov A.S., Lesin V.M., Nikolenko S.I., Pham S., Prjibelski A.D., Pyshkin A.V., Sirotkin A.V., Vyahhi N., Tesler G., Alekseyev M.A., Pevzner P.A. 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* 19:455-477.

Barker M.S., Arrigo N., Baniaga A.E., Li Z., Levin D.A. 2016. On the relative abundance of autopolyploids and allopolyploids. *New Phytol.* 210:391-398.

Beerli P., Mashayekhi S., Sadeghi M., Khodaei M., Shaw K. 2019. Population Genetic Inference With MIGRATE. *Curr. Protoc. Bioinformatics* 68:e87.

Berger E., Yorukoglu D., Peng J., Berger B. 2014. HapTree: a novel Bayesian framework for single individual polyplotyping using NGS data. *PLoS Comput. Biol.* 10:e1003502.

Bezanson J., Edelman A., Karpinski S., Shah V.B. 2015. Julia: A Fresh Approach to Numerical Computing. *arXiv*1411.1607v4.

Bird K.A., VanBuren R., Puzey J.R., Edger P.P. 2018. The causes and consequences of subgenome dominance in hybrids and recent polyploids. *New Phytol.* 220:87-93.

Blischak P.D., Chifman J., Wolfe A.D., Kubatko, L.S. 2018. HyDe: a Python package for genome-scale hybridization detection. *Syst. Biol.*, 67:821-829.

Breinholt J.W., Carey S.B., Tiley G.P., Davis E.C., Endara L., McDaniel S.F., Neves L.G., Sessa E.B., von Konrat M., Chantanaorrapint S., Fawcett S., Ickert-Bond S.M., Labiak P.H., Larraín J., Lehnert M., Lewis L.R., Nagalingum N.S., Patel N., Rensing S.A., Testo W., Vasco A., Villarreal J.C., Williams E.W., Burleigh J.G. 2021. A target enrichment probe set for resolving the flagellate plant tree of life. *Appl. Plant Sci.* 9:e11406.

Breinholt J.W., Earl C., Lemmon A.R., Lemmon E.M., Xiao L., Kawahara A.Y. 2018. Resolving relationships among the megadiverse butterflies and moths with a novel pipeline for anchored phylogenomics. *Syst. Biol.* 67:78-93.

Buggs J.A., Wendel J.F., Doyle J.J., Soltis D.E., Soltis P.S., Coate J.E. 2014. The legacy of diploid progenitors in allopolyploid gene expression patterns. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 368:20130354.

Cai R., Ané C. 2020. Assessing the fit of the multi-species network coalescent to multi-locus data. *Bioinformatics* btaa863. doi: https://doi.org/10.1093/bioinformatics/btaa863

Colle M., Leisner C.P., Wai C.M., Ou S., Bird K.A., Wang J., Wisecaver J.H., Yocca A.E., Alger E.I., Tang H., Xiong Z., Callow P., Ben-Zvi G., Brodt A., Baruch K., Swale T., Shiue L., Song G.-Q., Childs K.L., Schilmiller A., Vorsa N., Buell C.R., VanBuren R., Jiang N., Edger P.P. 2019. Haplotype-phased genome and evolution of phytonutrient pathways of tetraploid blueberry. *GigaScience* 8:giz012.

Crowl A.A., Myers C., Cellinese N. 2017. Embracing discordance: Phylogenomic analyses provide evidence for allopolyploidy leading to cryptic diversity in a Mediterranean Campanula (Campanulaceae) clade. *Evolution* 71:913-922.

DePristo M., Banks E., Poplin R., Garimella K., Maguire J., Hartl C., Philippakis A., del Angel G., Rivas M.A., Hanna M., McKenna A., Fennell T., Kernytsky A., Sivachenko A., Cibulskis K., Gabriel S., Altschuler S., Daly M. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* 43:491-498.

Edgar R.C. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32:1792-1797.

Eriksson J.S., de Sousa F., Bertrand Y.J.K., Antonelli A., Oxelman B., Pfeil B.E. 2018. Allele phasing is critical to revealing a shared allopolyploid origin of Medicago arborea and M. strasseri (Fabaceae). *BMC Evol. Biol.* 18:9.

Excoffier L., Dupanloup I., Huerta-Sanchez E., Sousa V.C., Foll M. 2013. Robust demographic inference from genomic and SNP data. *PLoS Genet.* 9:e1003905.

Faircloth B.C. 2016. PHYLUCE is a software package for the analysis of conserved genomic loci. *Bioinformatics* 32:786-788.

Faircloth B.C., McCormack J.E., Crawford N.G., Harvey M.G., Brumfield R.T., Glenn T.C. 2012. Ultraconserved elements anchor thousands of genetic markers spanning multiple evolutionary timescales. *Syst. Biol.* 61:717-726.

Farhat P., Hidalgo O., Robert T., Siljak-Yakovlev S., Leitch I.J., Adams R.P., Bou Dagher-Kharrat M. 2019. Polyploidy in the Conifer Genus *Juniperus*: An Unexpectedly High Rate. *Front. Plant Sci.* 10:676.

Flouri T., Jiao X., Rannala B., Yang Z. 2020. A Bayesian Implementation of the Multispecies Coalescent Model with Introgression for Phylogenomic Analysis. *Mol. Biol. Evol.* 37:1211-1223.

Flouri T., Jiao X., Rannala B., Yang Z.. 2018. Species Tree Inference with BPP Using Genomic Sequences and the Multispecies Coalescent. *Mol. Biol. Evol.* 35:2585-2593.

Freyman W.A,. Johnson M.G., Rothfels C.J. 2020. Homologizer: Phylogenetic phasing of gene copies into polyploid subgenomes. *bioRxiv* doi: https://doi.org/10.1101/2020.10.22.351486.

Gronau I., Hubisz M.J., Gulko B., Danko C.G., Siepel A. 2011. Bayesian inference of ancient human demography from individual genome sequences. *Nat. Genet.* 43:1031-1034.

He D., Saha S., Finkers R., Parida L. 2018. Efficient algorithms for polyploid haplotype phasing. *BMC Genomics* 19:110.

Huang J., Flouri T., Yang Z. 2020. A simulation study to examine the information content in phylogenomic datasets under the multispecies coalescent model. *Mol. Biol. Evol.* 37:3211-3224.

Huber K.T., Oxelman B., Lott M., Moulton V. 2006. Reconstructing the evolutionary history of polyploids from multi-labelled trees. *Mol. Biol. Evol.* 23:1784-1791.

Hudson R.R. 1983. Testing the Constant-Rate Neutral Allele Model with Protein Sequence Data. *Evolution* 37:203-217.

Huson D.H., Rupp R., Scornavacca C. 2010. Phylogenetic networks: concepts, algorithms and applications. Cambridge University Press.

Iqbal Z., Caccamo M., Turner I., Flicek P., McVean G. 2012. De novo assembly and genotyping of variants using colored de Bruijn graphs. *Nat. Genet.* 44:226-232.

Jantzen J.R., Amarasinghe P., Folk R.A., Reginato M., Michelangeli F.A., Soltis D.E., Cellinese N., Soltis P.S. 2020. A two-tier bioinformatic pipeline to develop probes for target capture of nuclear loci with applications in Melastomataceae. *Appl. Plant Sci.* 8:e11345.

Johnson M.G., Gardner E.M., Liu Y., Medina R., Goffinet B., Shaw A.J., Zerega N.J., Wickett N.J. 2016. HybPiper: Extracting coding sequence and introns for phylogenetics from high-throughput sequencing reads using target enrichment. *Appl. Plant Sci.* 4:1600016.

Johnson M.G., Pokorny L., Dodsworth S., Botigue L.R., Cowan R.S., Devault A., Eiserhardt W.L., Epitawalage N., Forest F., Kim J.T., Leebens-Mack J.H., Leitch I.J., Maurin O., Soltis D.E., Soltis P.E., Wong G.K.-S., Baker W.J., Wickett N.J. 2019. A Universal Probe Set for Targeted Sequencing of 353 Nuclear Genes from Any Flowering Plant Designed Using k-Medoids Clustering. *Syst. Biol.* 68:594-606.

Jones G., Sagitov S., Oxelman B. 2013. Statistical inference of allopolyploid species networks in the presence of incomplete lineage sorting. *Syst. Biol.* 62:467-478.

Jukes T.H., Cantor C.R. 1969. Evolution of protein molecules. In: Munro H.N., editor. Mammalian Protein Metabolism. New York, NY: Acedemic Press. p. 21-132.

Kalyaanamoorthy S., Minh B.Q., Wong T.K.F., von Haeseler A., Jermiin L.S. 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* 14:587-589.

Karimi N., Grover C.E., Gallagher J.P., Wendel J.F., Ané C., Baum D.A. 2020. Reticulate Evolution Helps Explain Apparent Homoplasy in Floral Biology and Pollination in Baobabs (Adansonia; Bombacoideae; Malvaceae). *Syst. Biol.* 69:462-478.

Kates H.R., Johnson M.G., Gardner E.M., Zerega N.J.C., Wickett N.J. 2018. Allele phasing has minimal impact on phylogenetic reconstruction from targeted nuclear gene sequences in a case study of Artocarpus. *Am. J. Bot.* 105:404-416.

Lemmon A.R., Emme S.A., Lemmon E.M. 2012. Anchored hybrid enrichment for massively high-throughput phylogenomics. *Syst. Biol.* 61:727-744.

Li H., Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25:1754-1760.

Liu X., Fu Y.X. 2020. Stairway Plot 2: demographic history inference with folded SNP frequency spectra. *Genome Biol.* 21:280.

Liu Y., Johnson M.G., Cox C.J., Medina R., Devos N., Vanderpoorten A., Hedenas L., Bell N.E., Shevock J.R., Aguero B., Quandt D., Wickett N.J., Shaw A.J., Goffinet B. 2019. Resolution of the ordinal phylogeny of mosses using targeted exons from organellar and nuclear genomes. *Nat. Commun.* 10:1485.

Lott M., Spillner A., Huber K.T., Moulton V. 2009. PADRE: a package for analyzing and displaying reticulate evolution. *Bioinformatics* 25:1199-1200.

Luo R., Liu B., Xie Y., Li Z., Huang W., Yuan J., He G., Chen Y., Pan Q., Liu Y., Tang J., Wu G., Zhang H., Shi Y., Liu Y., Yu C., Wang B., Lu Y., Han C., Cheung D.W., Yiu S.-M., Peng S., Xiaoqian Z., Liu G., Liao X., Li Y., Yang H., Wang J., Lam T-W., Wang J. 2012. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *GigaScience* 1:18.

McKain M.R., Johnson M.G., Uribe-Convers S., Eaton D., Yang Y. 2018. Practical considerations for plant phylogenomics. *Appl. Plant Sci.* 6:e1038.

McKenna A., Hanna M., Banks E., Sivachenko A., Cibulskis K., Kernytsky A., Garimella K., Altshuler D., Gabriel S., Daly M., DePristo M.A. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 20:1297-1303.

Moeinzadeh M.H., Yang J., Muzychenko E., Gallone G., Heller D., Reinert K., Haas S., Vingron M. 2020. Ranbow: A fast and accurate method for polyploid haplotype reconstruction. *PLoS Comput*. *Biol.* 16:e1007843.

Montgomery J.D., Paulton E.M. 1981. *Dryopteris* in North America. *Fiddlehead Forum* 8:25-31.

Morales-Briones D.F., Liston A., Tank D.C. 2018. Phylogenomic analyses reveal a deep history of hybridization and polyploidy in the Neotropical genus Lachemilla (Rosaceae). *New Phytol.* 218:1668-1684.

Nauheimer L., Weigner N., Joyce E., Crayn D., Clarke C., Nargar K. 2020. HybPhaser: a workflow for the detection and phasing of hybrids in target capture datasets. bioRxiv. doi: https://doi.org/10.1101/2020.10.27.354589.

Nguyen L.-T., Schmidt H.A., von Haeseler A., Minh B.Q. 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* 32:268-274.

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 in Ecology and Evolution* 8:835-849.

Olave M., Meyer A. 2020. Implementing Large Genomic Single Nucleotide Polymorphism Data Sets in Phylogenetic Network Reconstructions: A Case Study of Particularly Rapid Radiations of Cichlid Fish. *Syst. Biol.* 69:848-862.

Pamilo P., Nei M. 1988. Relationships between gene trees and species trees. *Mol. Biol. Evol.* 5:568-583.

Rothfels C.J., Pryer K.M., Li F.-W. 2017. Next-generation polyploid phylogenetics: rapid resolution of hybrid polyploid complexes using PacBio single-molecule sequencing. *New Phytol.* 213:413-429.

Saada O.A., Tsouris A., Freidrich A., Schachrer J. 2020. nPhase: An accurate and contiguous phasing method for polyploids. *bioRxiv* doi:https://doi.org/10.1101/2020.07.24.219105.

Sessa E.B., Zimmer E.A., Givnish T.J. 2012a. Reticulate evolution on a global scale: a nuclear phylogeny for New World Dryopteris (Dryopteridaceae). *Mol. Phylogenet. Evol.* 64:563-581.

Sessa E.B., Zimmer E.A., Givnish T.J. 2012b. Unraveling reticulate evolution in North American Dryopteris (Dryopteridaceae). *BMC. Evol. Biol.* 12:104.

Solis-Lemus C., Ané C. 2016. Inferring Phylogenetic Networks with Maximum Pseudolikelihood under Incomplete Lineage Sorting. *PLoS Genet.* 12:e1005896.

Solis-Lemus C., Bastide P., Ané C. 2017. PhyloNetworks: A Package for Phylogenetic Networks. *Mol. Biol. Evol.* 34:3292-3298.

Soltis, D.E., Visger C.J., Soltis P.S. 2014. The polyploidy revolution then...and now: Stebbins revisited. *American Journal of Botany* 101: 1057–1078.

Stull G.W., Soltis P.S., Soltis D.E., Gitzendanner M.A., Smith S.A. 2020. Nuclear phylogenomic analyses of asterids conflict with plastome trees and support novel relationships among major lineages. *Am. J. Bot.* 107:790-805.

Tiley G.P., Poelstra J.W., dos Reis M., Yang Z., Yoder A.D. 2020. Molecular Clocks without Rocks: New Solutions for Old Problems. *Trends Genet.* 36:845-856.

Viruel J., Conejero M., Hidalgo O., Pokorny L., Powell R.F., Forest F., Kantar M.B., Soto Gomez M., Graham S.W., Gravendeel B., Wilkin P., Leitch I.J. 2019. A Target Capture-Based Method to Estimate Ploidy from Herbarium Specimens. *Front. Plant Sci.* 10:937.

Weiss C.L., Pais M., Cano L.M., Kamoun S., Burbano H.A. 2018. nQuire: a statistical framework for ploidy estimation using next generation sequencing. *BMC Bioinformatics* 19:122.

Wen D., Yu Y., Nakhleh L. 2016. Bayesian Inference of Reticulate Phylogenies under the Multispecies Network Coalescent. *PLoS Genet.* 12:e1006006.

Wen D., Yu Y., Zhu J., Nakhleh L. 2018. Inferring Phylogenetic Networks Using PhyloNet. *Syst. Biol.* 67:735-740.

Wolf P.G., Robison T.A., Johnson M.G., Sundue M.A., Testo W.L., Rothfels C.J. 2018. Target sequence capture of nuclear-encoded genes for phylogenetic analysis in ferns. *Appl. Plant Sci.* 6:e01148.

Wood T.E., Takebayashi N., Barker M.S., Mayrose I., Greenspoon P.B., Rieseberg L.H. 2009. The frequency of polyplid speciation in vascular plants. *Proc. Natl. Acad. Sci. U. S. A.* 106:13875-13879.

Xie M., Wu Q., Wang J., Jiang T. 2016. H-PoP and H-PoPG: heuristic partitioning algorithms for single individual haplotyping of polyploids. *Bioinformatics* 32:3735-3744.

Xie W., Lewis P.O., Fan Y., Kuo L., Chen M.H. 2011. Improving marginal likelihood estimation for Bayesian phylogenetic model selection. *Syst. Biol.* 60:150-160.

Yang J., Moeinzadeh M-H., Kuhl H., Helmuth J., Xiao P., Haas S., Liu G., Zheng J., Sun Z., Fan W., Deng G., Wang H., Hu F., Zhao S., Fernie A.R., Boerno S., Timmermann B., Zhang P., Vingron M. 2017. Haplotype-resolved sweet potato genome traces back its hexaploidization history. *Nat Plants* 3:696-703.

Yang Z. 2006. Computational molecular evolution: Oxford University Press.

Zhang C., Ogilvie H.A., Drummond A.J., Stadler T. 2018. Bayesian Inference of Species Networks from Multilocus Sequence Data. *Mol. Biol. Evol.* 35:504-517.

Zhang C., Rabiee M., Sayyari E., Mirarab S. 2018. ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. *BMC Bioinformatics* 19:153.

Zhu S., Degnan J.H. 2017. Displayed trees do not determine distinguishability under the network multispecies coalescent. *Syst. Biol.* 66:283-298.

Tables

Topology	Phased Marginal lnL	Unphased Marginal lnL	Phased Model Probability	Unphased Model Probability
1	-613405.3154	-552720.6776	1.006E-278	3.2574E-214
2	-613223.0755	-552534.6540	1.4073E-199	2.0039E-133
3	-613366.4128	-552644.7603	7.9026E-262	3.0431E-181
4	-612866.6368	-552245.2218	8.86633E-45	1.00164E-07
5	-612828.7348	-552243.7091	2.56082E-28	4.54611E-07
6	-612822.4509	-552238.6462	1.37227E-25	7.18535E-05
7†	-612765.2027	-552229.1054	1	0.999927592
8	-613211.8583	-552484.8320	1.047E-194	8.6956E-112
9	-613168.7562	-552479.3732	5.4825E-176	2.0419E-109
10	-613102.1509	-552454.2002	4.6266E-147	1.74796E-98
11	-613079.9316	-552426.8867	2.0654E-137	1.27244E-86
12	-613379.6350	-552538.6937	1.4303E-267	3.5276E-135
13	-613390.9585	-552533.6798	1.7287E-272	5.3085E-133
14	-613272.7561	-552561.3165	3.7359E-221	5.2787E-145
15	-613247.7503	-552566.6741	2.7054E-210	2.4873E-147

Table 1 — Marginal Likelihoods for Possible Topological Hypotheses. Topologies of the 15 models are displayed in Figure S1.

†*a priori* allopolyploid hypothesis

FIGURES

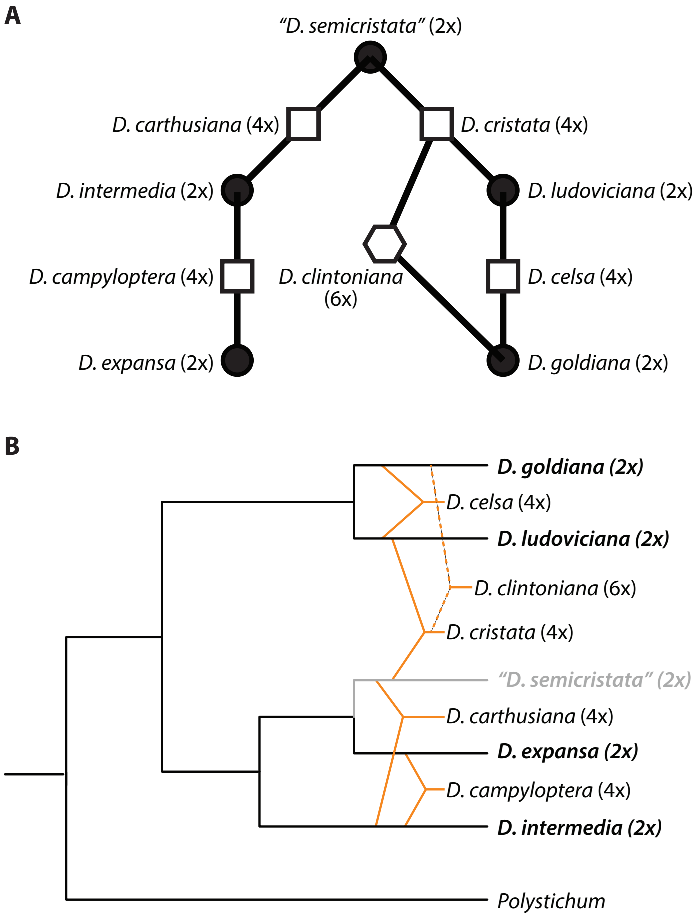


Figure 1 — Hypothesized Relationships among North American *Dryopteris*. Synthesis of results from Sessa et al. 2012a and Sessa et al. 2012b. A) Links between shapes show the putative parents and their allopolyploid derivatives. Black circles are diploids, squares are tetraploids, and the hexagon is the one hexaploid species in the group. *Dryopteris semicristata* is presumed extinct. B) Placement of allopolyploids in the context of the backbone relationships among diploids. Tetraploids are indicated with solid orange lines and the hexaploid with dotted lines. The grey line denoting a sister relationship between *D. semicristata* and *D. expansa* reflects one possible placement for the extinct taxon based on previous analyses (Sessa et al., 2012b).

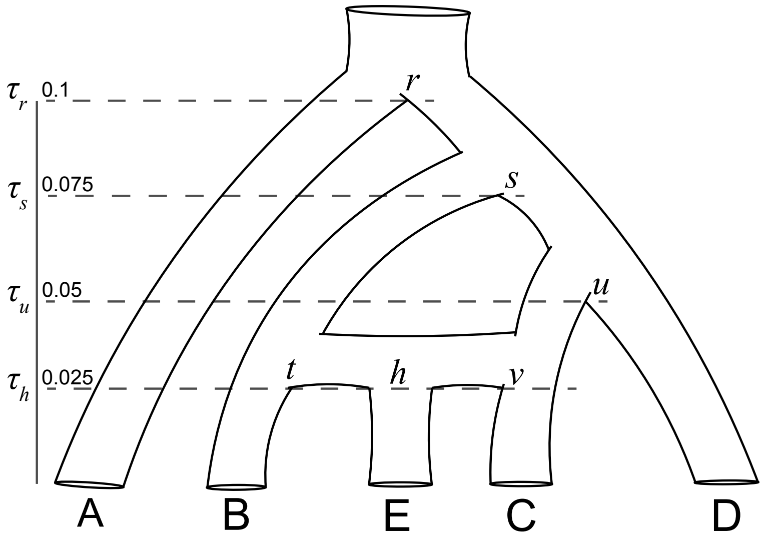


Figure 2 — Species network used for simulation. The divergence times in expected substitutions per site are given for each node, and *h* is the hybrid node where two alleles enter from both *t* and *v*. E is an allotetraploid while other species are diploid. Nucleotide divergence was reduced by dividing all τ by 10 or increased by multiplying all τ by 10. ILS was increased by halving the distance between τ_h and τ_u and between τ_u and τ_s either once (for the medium ILS condition) or twice (for the high ILS condition).

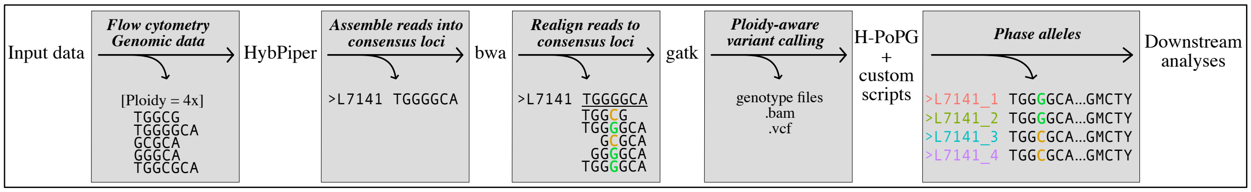


Figure 3 — PATÉ Phasing Pipeline. Overview of data input, output, and steps taken to phase alleles. Input data are assumed to be paired-end Illumina reads and reference sequences for each individual are required (consensus loci from HybPiper can be used). The ploidy of each individual must be specified. Only biallelic sites are used for phasing.

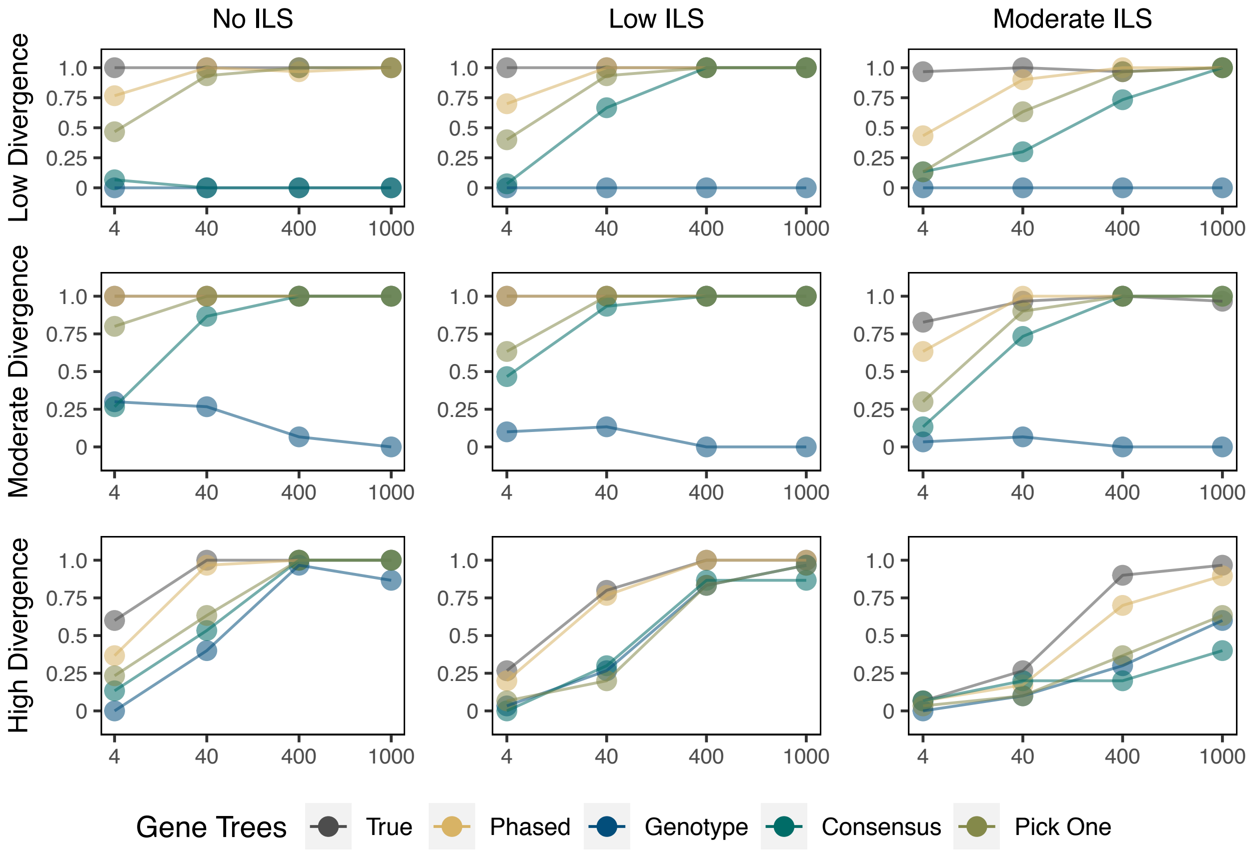
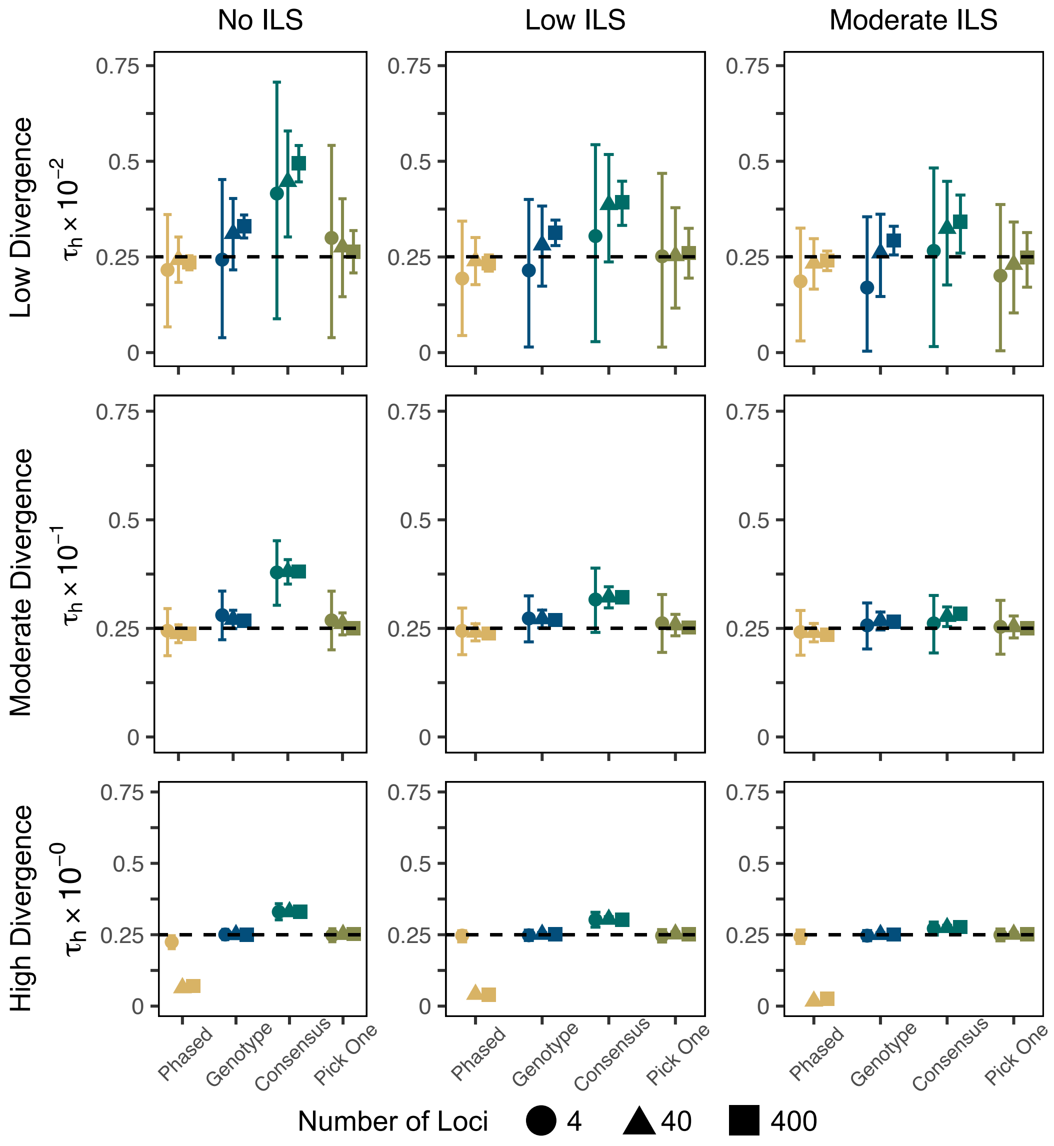


Figure 4 — Proportion of simulations that correctly identify the allopolyploid lineage. The x-axes are the number of loci sampled for each simulation. Th y-axes are the proportions of correct networks. Results are based on networks estimated with a single reticulation, even if that network was considered less optimal than networks with zero or two reticulations. We saved the true gene trees from the simulations, while we estimated gene trees with the phased and unphased (Genotype, Consensus, and Pick One) data.



929

930 **Figure 5 — Estimating the Timing of Introgression.** Divergence times are for node *h* in 931 Figure 2. The low divergence simulation corresponds to the y-axis units of $\tau_h \times 10^{-2}$ while the 932 high divergence case is represented by τ_h.Divergence times are measured as the expected 933 number of substitutions per site. The dashed line represents the true simulated values. Points 934 are posterior means, and error bars are 95% HPD intervals, averaged across 30 replicates.

935

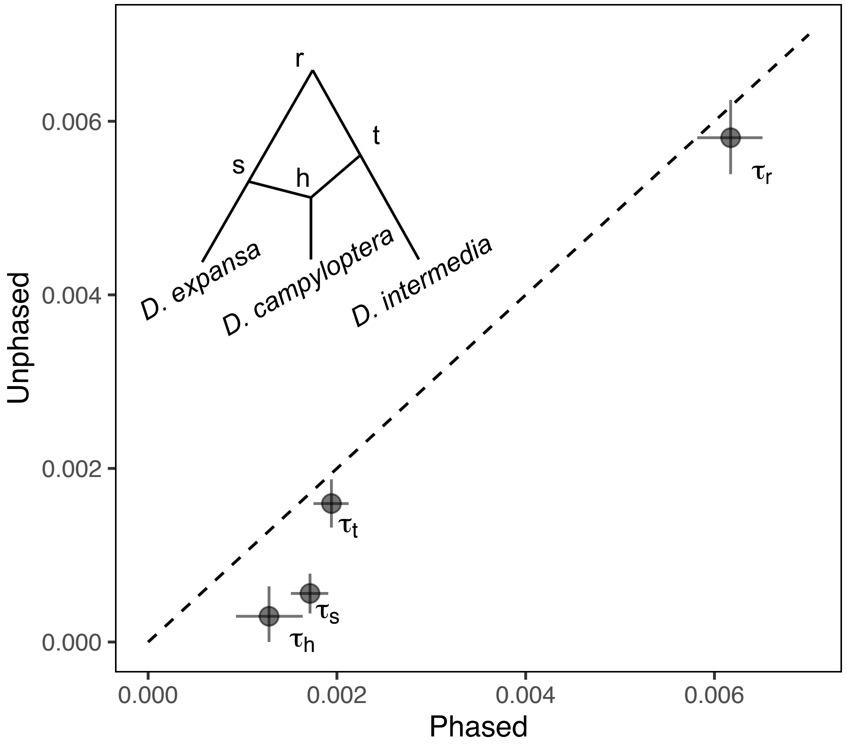


Figure 6— *Dryopteris* Divergence Times under the MSci Model. Divergence times are measured as the expected number of substitutions per site. The x-axis shows estimates from phased data and the y-axis shows estimates from unphased (haplotype consensus) data for the inset network (equivalent to model 7 from Supplementary Fig. S1). The dashed one-to-one line shows where older age estimates are consistently obtained from phased data. Error bars on points show the 95% HPD intervals for phased and unphased data.

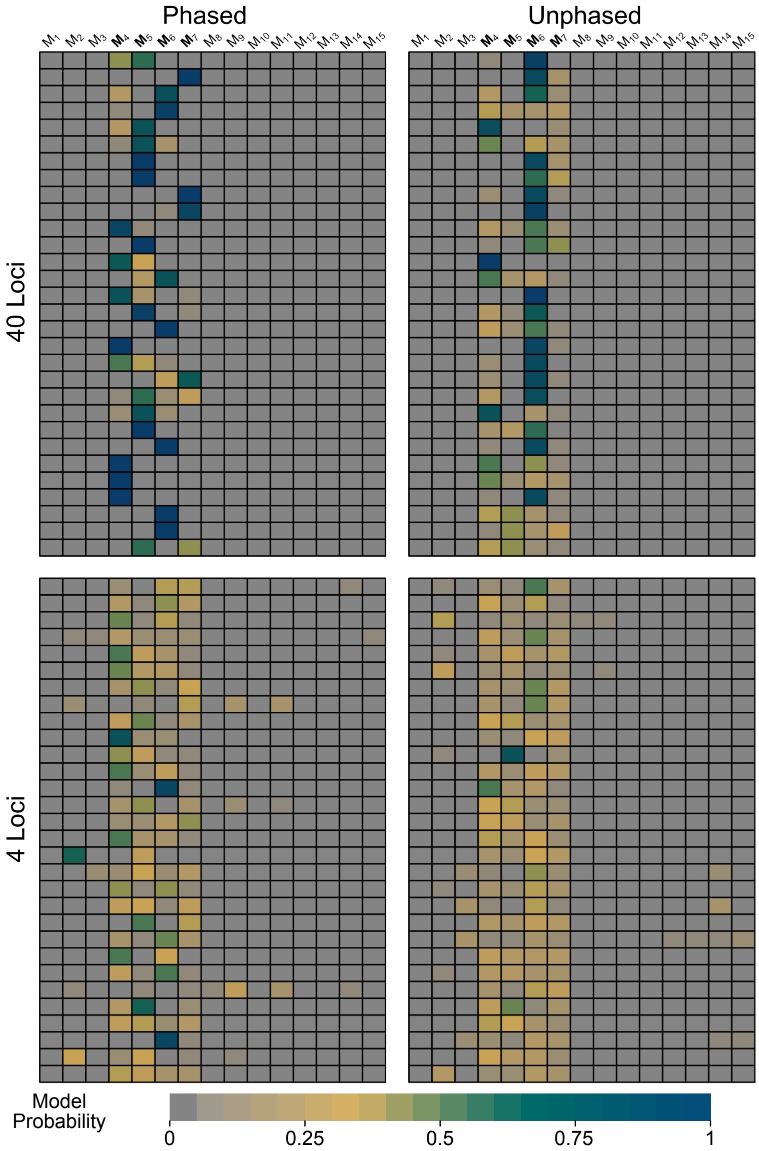


Figure 7— Probabilities for three-taxon *Dryopteris* MSci Models with fewer Loci. Marginal likelihoods were estimated for the 15 MSci models (Supplementary Fig. S1). Model weights were used to obtain probabilities. We consider model probability greater than 0.95 as decisive evidence in favor of a model, and a probability less than 0.05 is evidence against a model. We considered probabilities between 0.05 and 0.95 to be ambiguous. Models four through seven (in bold) all have the correct reticulate relationships between the parental diploid lineages and the allopolyploid. Models one through three do not have introgression, and models eight through 15 have incorrect introgression events. Each row represents one of 30 sampling replicates.

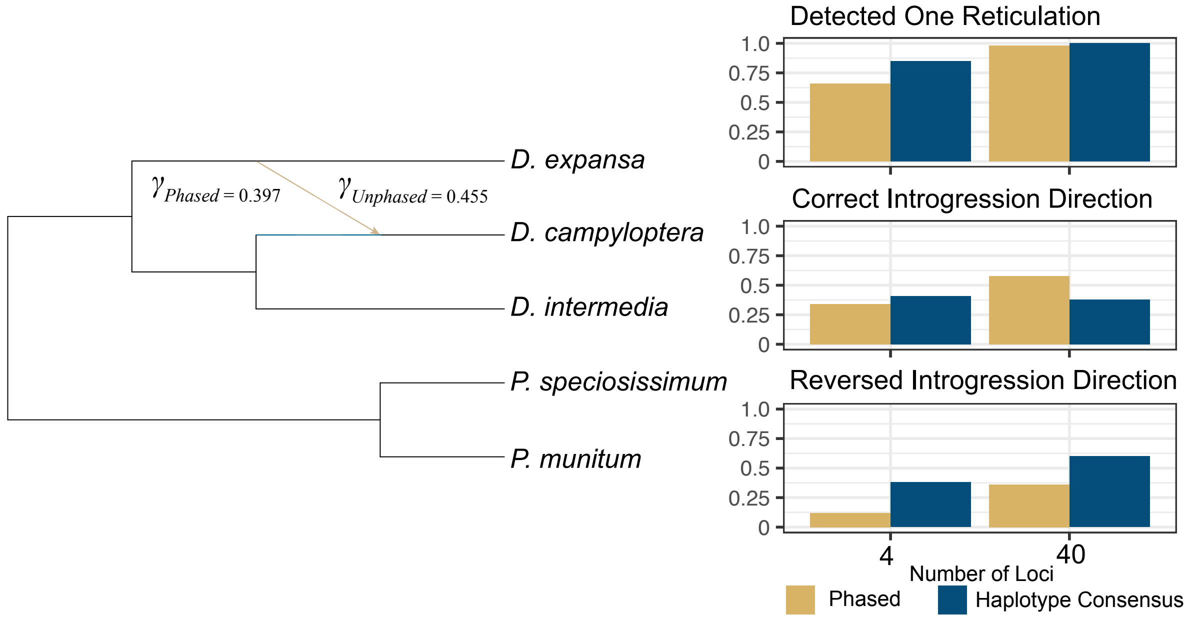


Figure 8 — Network search results for three-taxon *Dryopteris* example. Both phased and unphased haplotype consensus data recovered the same network topology with introgression occurring in the expected direction. The major topology is indicated by the blue edge and the minor edge (direction of introgression) is shown in tan. The inheritance probability γ was slightly higher in the haplotype consensus data ($\gamma_{\text{unphased}} = 0.455$). Bar plots show the proportion of 100 replicates when sampling four or 40 loci that correctly detect one reticulation based on the pseudolikelihood scores (top), correctly estimate the network with introgression going from one of the diploids into *D. campyloptera* (middle), and estimate a network where the direction of introgression is from *D. campyloptera* into one of the diploids (bottom).

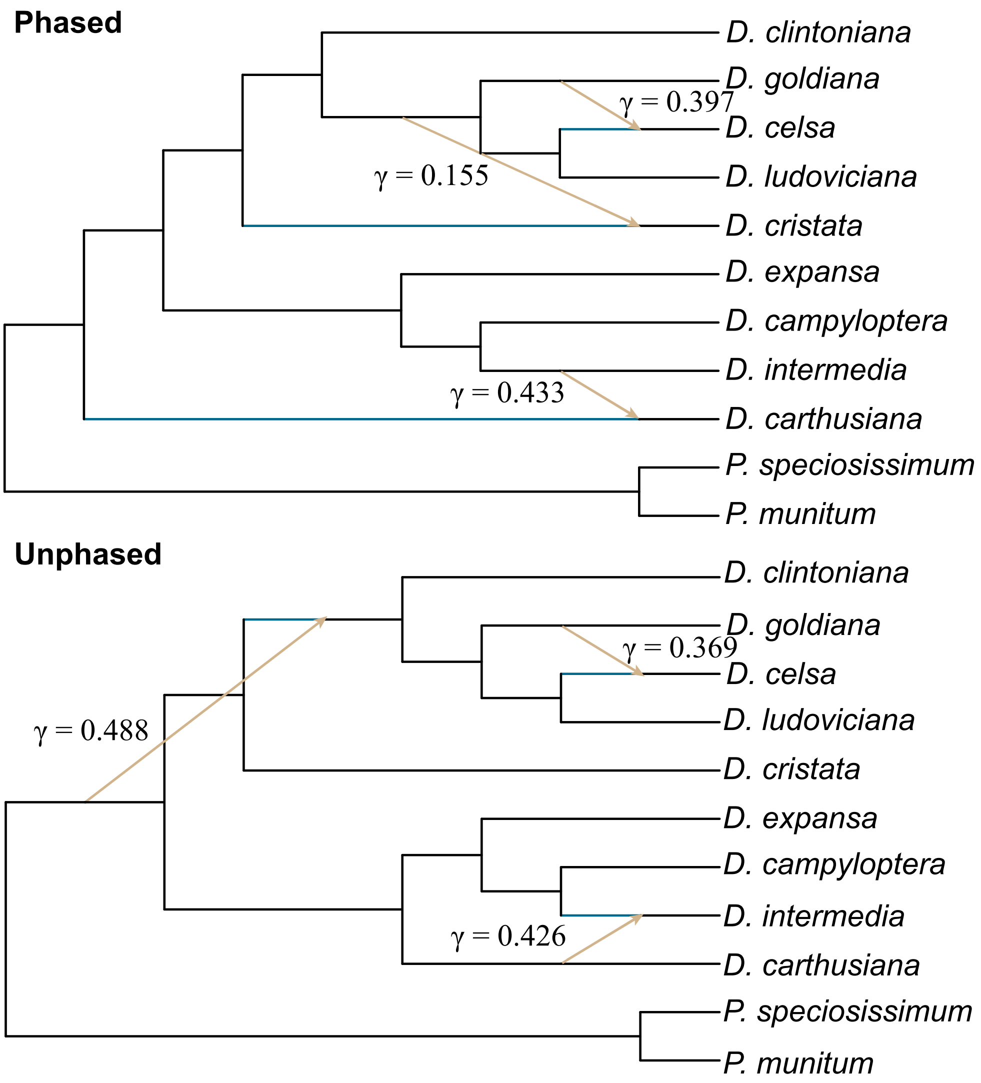


Figure 9 — Networks for Nine-Taxon *Dryopteris* example. Both data types recovered optimal networks with three reticulation events. The major topology edge is blue and the minor (reticulation) edges shown in tan, with the direction of introgression flowing into the major edge. The position of *D. carthusiana* changes in the major topology between phased and unphased data but the relationships are otherwise the same. All three reticulation events in the phased data are plausible, but in the unphased data, the direction of introgression from *D. carthusiana* into *D. intermedia* is incorrect and the reticulate edge from the common ancestor of *Dryopteris* is difficult to reconcile. Inheritance probabilities for each introgression event are shown next to reticulation edges.