Running Head: Allopolyploid networks with phased sequences

Benefits and Limits of Phasing Alleles for Network Inference of Allopolyploid Complexes

George P. Tiley[1,*], Andrew A. Crowl[2], Paul S. Manos[2], Emily B. Sessa[3], Claudia Solís-Lemus[4], Anne D. Yoder[2], J. Gordon Burleigh[3]

[1]Royal Botanic Gardens Kew, Richmond, TW9 3AE, UK

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

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

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

[*]Author for correspondence: g.tiley@kew.org

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 hindered 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 in empirical data, 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 insights from phased data as well as the challenges of 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. These data may improve network estimates compared to using haplotype consensus assemblies by accurately inferring the direction of gene flow, but statistical identifiability of phylogenetic networks poses a barrier to inferring the evolutionary history of reticulate complexes.

Key words: Introgression; Hybridization; Phylogenetic Networks; Statistical Identifiability; Multispecies Coalescent; Divergence Time Estimation; Polyploidy; Target Enrichment

INTRODUCTION

Polyploidy, or whole-genome duplication, occurs throughout the eukaryotic tree of life. However, its evolutionary significance is especially evident 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 numerous studies investigating the role of whole-genome duplication in plant speciation and local adaptation (reviewed in Soltis et al. 2014), polyploidy 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 evidence of polyploidy and its origins. Using allelic data that more accurately capture the complex genomic histories of polyploids should enable analyses to incorporate 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, leveraging fast polyploid phasing algorithms (Xie et al. 2016) to develop a bioinformatic pipeline that can phase alleles from polyploids using target enrichment sequence data.

Previous studies have suggested phasing alleles is crucial for accurate evolutionary reconstruction of reticulate complexes, at least when sampling relatively few loci (e.g., 4 to 10; Rothfels et al. 2017; Eriksson et al. 2018). However, it remains challenging to genotype and

63 phase next generation sequencing data from polyploids. Methods exist to genotype consensus
64 loci from target enrichment data, but these have been either limited to diploids (Kates et al.
65 2018; Andermann et al. 2019) or used manual curation of variants with polyploids where both
66 parental populations are available (Eriksson et al. 2018). Otherwise, obtaining phased sequence
67 data for polyploids has largely depended on long-read sequencing to recover complete
68 haplotype sequences (e.g., Rothfels et al. 2017) or cloning PCR products (e.g., Sessa et al.
69 2012; Oberprieler et al. 2017).

70 Target enrichment (or HybSeq), where specific regions of the genome are isolated and
71 sequenced (Albert et al. 2007; Gnirke et al. 2009; Faircloth et al. 2012; Lemmon et al. 2012), is
72 an increasingly common method for collecting large-scale phylogenomic datasets, and these
73 data can provide insights into the evolutionary history of reticulate complexes (e.g., Crowl et al.
74 2017; Karimi et al. 2020) and sources of gene tree discordance (e.g., Morales-Briones et al.
75 2018; Stull et al. 2020). Probe kits for target enrichment have been developed in many land
76 plant lineages (Wolf et al. 2018; Johnson et al. 2019; Liu et al. 2019; Breinholt et al. 2021), and
77 there are bioinformatic pipelines available for custom probe design (e.g., Jantzen et al. 2020).
78 The most common approach to assemble phylogenetic datasets from target enrichment data
79 has been to use haplotype consensus assemblies from various pipelines (e.g. Faircloth 2016;
80 Johnson et al. 2016; Andermann et al. 2018; Breinholt et al. 2018). The *de novo* assembly
81 algorithms within these pipelines typically consider only the most frequent nucleotide sequence
82 and discard the alternatives (Bankevich et al. 2012; Iqbal et al. 2012; Luo et al. 2012). This
83 results in loci in which variable positions are collapsed to a single base call (haplotype
84 consensus loci), losing information related to heterozygosity. While this may be appropriate, or
85 at least benign, for phylogenetic analyses of diploid taxa (Kates et al. 2018), it may pose
86 substantial problems when attempting to investigate the evolutionary history of polyploid taxa or
87 reticulate lineages. Some of the before mentioned pipelines can recover the heterozygosity

information and phase sequences by mapping reads back to their haplotype consensus sequences (Andermann et al. 2018), but this utility is, at the time of writing, limited to diploids.

While phylogenetic studies in plants usually 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 with moderate ILS when it is 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; Zhu and Degnan 2017) and multispecies network coalescent (MSNC; Wen et al. 2016). We emphasize that using network approaches to investigate polyploid complexes is 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; although, the switch errors among these haplotypes sequences is unknown. Networks inferred from phased data more accurately represent some relationships within the North American *Dryopteris* complex than unphased strategies, but statistical identifiability problems confound all data types.

MATERIALS & METHODS

140 ***Testing the Effects of Phasing on Network Inference through Simulation***

141 *Simulating phased and unphased sequence data for an allopolyploid* — We simulated
142 gene trees and their nucleotide sequence data using the MSC model (Rannala and Yang 2003)
143 with BPP v4.4.0 (Flouri et al. 2020) under a network with five tips, such that one taxon was a
144 hybrid of the ancestors of two other tip taxa (Fig. 2). All simulations used the HKY85 (Hasegawa
145 et al 1985) model of sequence evolution with a transition-transversion rate ratio of 3 and
146 gamma-distributed among-site rate heterogeneity with a shape parameter of 0.6. Equilibrium
147 frequencies were 0.3, 0.2, 0.3, and 0.2 for nucleotides A, C, G, and T. The allopolyploid species
148 *E* was treated as two lineages (*E* sister to *B* and *F* sister to *C*) whose alleles were pooled to
149 form the hybrid species *E* at time τ_h. This makes species *E* a tetraploid hybrid with the parents
150 the ancestors of *B* and *C*. The mutation-scaled effective population size (θ) was constant
151 among lineages and set at 0.01. Assuming a per-generation mutation rate (μ) of 1×10^{-8} and
152 one year per generation yields an effective population size (N_e) of 250,000 and hybridization
153 age of 10 Ma (Fig. 2). We also simulated data such that θ was 0.001 (hybridization at 1 Ma) or
154 0.1 (hybridization at 100 Ma). Investigating such extreme values can generate general
155 expectations for downstream estimation under a wide range of parameters, but has some
156 biological salience as there is increasing evidence for reticulate evolution early in the
157 diversification of plant orders and families (Sun et al. 2015; Stull et al. 2020; Liu et al. 2022)

158 Because the distance between speciation nodes is 0.01 substitutions per site and $\theta =$
159 0.01 (Fig. 2), there are two coalescent units between nodes, $T = \frac{\tau}{\left(\frac{\theta}{2}\right)}$ (Yang 2006), which yields a
160 low probability (0.09) of gene tree discordance due to ILS (Hudson 1983). We incorporated
161 more ILS into the simulation by keeping the node heights fixed but increasing θ to 0.02, 0.03,
162 and 0.04. This increased the probability of a non-speciodentric split among all nodes to 0.25,
163 0.34, and 0.4, respectively. With three different starting values of θ and τ, this gave a total of 12
164 simulation conditions. While our simulations do not explore the limits 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, 500 base pairs (bp) in length, under the MSC 100 times for each of the 12 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 chosen and 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 the SNaQ function (Solís-Lemus and Ané 2016) from PhyloNetworks v0.12.0 (Solís-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, pick one, genotype, and consensus sequence data. For estimated trees, we used IQTREE v1.6.10

191 (Nguyen et al. 2015) with the same substitution model used to simulate the data. Each SNaQ
192 analysis used the species tree (A,((B,E),(C,D))) as the starting tree and allowed zero, one, or
193 two reticulation events. Each analysis included ten independent optimizations of the
194 pseudolikelihood score. We considered increasing numbers of reticulations if there was a
195 pseudolikelihood score improvement of two or greater compared to the previous model with one
196 fewer allowed hybrid edges. Note that pseudolikelihoods violate standard likelihood-based
197 model selection criteria such as likelihood ratio tests or using AIC weights, and slope heuristics
198 (Baudry et al. 2012) are often used instead. The choice of a cutoff here was necessary to
199 streamline processing across thousands of simulations. We compared the estimated networks
200 with one reticulation to the true network with the *hardwiredClusterDistance* function (Huson et
201 al. 2010) in PhyloNetworks. This allowed us to score the number of replicates that 1) recovered
202 the correct number of reticulations and 2) matched the true network when the number of
203 reticulations was set to one. We estimated networks for samples of four, 40, 400, and 1000
204 gene trees for each of the 100 replicates from the 12 simulation conditions.
205
206 *Effect of phasing on divergence time estimation* — We also used our simulated phased
207 and unphased data to estimate divergence times under the MSci model (Flouri et al. 2020)
208 using BPP v4.4.0. Here, we estimate divergence times (τ), population sizes (θ), and the
209 introgression probability, or proportion of immigrant loci in the receiving population, (φ) on the
210 correct fixed species network. φ was originally represented by γ (Meng and Kubatko 2009) and
211 referred to as an inheritance probability (Yu et al. 2011), but we use φ and the term
212 introgression probability to be consistent with BPP (Flouri et al. 2020). Our MSci analyses used
213 diffuse priors on τ_r and θ, with a mean on their simulated values and $\varphi \sim \beta(1,1)$. Phased, pick
214 one, and consensus sequences were treated as haploids while genotype sequences were
215 treated as unphased diploids and used for analytical integration over phases (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 haploids with ambiguous states. 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. Scripts for simulation and subsequent analyses of simulated data are available in Dryad (http://dx.doi.org/10.5061/dryad.[NNNN]).

A Phasing Pipeline for Polyploids

Target enrichment data — The success of empirical studies in which phasing was informative about hybridization or introgression events (e.g., Kamneva et al. 2017; Oberprieler et al. 2017; Eriksson et al. 2018) motivated us to use phased data to infer reticulate evolutionary histories of polyploids. 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 tree-based methods 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 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

241 phased haplotype sequences for taxa with known ploidy levels. Ploidy levels were characterized
242 for individuals in our *Dryopteris* analyses, but for unknown systems there are existing methods
243 to estimate ploidy directly from target enrichment data (Weiss et al. 2018; Viruel et al. 2019) in
244 the absence of other sources, such as flow cytometry (Farhat et al. 2019). PATÉ implements a
245 test for ploidy based on allele balance distributions for biallelic sites (Tiley et al. 2023a), but we
246 implore heavy caution with these tests.
247
248 *Phasing alleles within loci* — PATÉ (Supplementary Fig. S1) starts with assembled
249 target enrichment loci, such as the supercontig files output from HybPiper (Johnson et al. 2016),
250 that contain a single haplotype consensus sequence from each individual per locus. Reads for
251 each individual are then realigned to their consensus locus using BWA v0.7.17 (Li and Durbin
252 2009). PCR duplicates are flagged with MarkDuplicates in Picard v2.9.2
253 (http://broadinstitute.github.io/picard), and variant calls are computed with HaplotypeCaller in
254 GATK v.4.1.4 (McKenna et al. 2010). We applied the following hard filters with VariantFiltration
255 in GATK: (1) QD < 2.0, (2) FS > 60.0, (3) MQ < 40.0, (4) ReadPosRankSum < –8.0, (5) AF <
256 0.05 || AF > 0.95. These loosely follow community recommendations on filters for germline
257 variant discovery (DePristo et al. 2011). Notably, we do not perform quality score recalibration
258 or filter on the mapping quality rank sum, as we anticipate allopolyploids could have a lower
259 mapping quality associated with an alternate allele due to sequence divergence or structural
260 variation among homeologous chromosomes. We also consider a very narrow window for
261 filtering on allele frequency. Because increasing ploidy levels will generate smaller anticipated
262 ratios of alternate to reference alleles, coupled with sequencing error and read stochasticity, we
263 only aim to remove the most extreme cases. For example, if almost all reads support the
264 alternate allele at a site, it is difficult to diagnose if the error lies in the consensus assembly or
265 the read alignment. In these cases, only the reference site is retained, and the variant does not
266 pass the allele frequency filter. However, the allele frequency filter could be removed if

267 investigators are focused on organisms with higher ploidy levels. Additional options for pipeline
268 configuration are available on GitHub (https://github.com/gtiley/Phasing).
269 Biallelic SNPs that pass filters are then phased with H-PoPG v.0.2.0 (Xie et al. 2016). H-
270 PopG solves a heuristic phasing problem efficiently using dynamic programming. Although not
271 guaranteed to be an optimal solution, H-PoPG has been shown to have high accuracy while
272 also being fast (Xie et al. 2016; He et al. 2018; Moeinzadeh et al. 2020). Phasing variants in
273 polyploids is difficult because for *n* variants and *k* ploidy, there are $2^{n-1}(k-1)^n$ possible ways
274 to link the sites together. H-PoPG evaluates possible solutions efficiently by grouping reads into
275 *k* groups in such a way that differences within the groups are minimized. Focusing on target
276 enrichment data also constrains the complexity of the phasing problem compared to whole-
277 genome alignments (i.e., haplotype blocks are constrained to about 1000 bp). We then used the
278 phased variants to create individual allele sequences, where invariable sites are filled in based
279 on the reference sequence. In cases where there is no linkage information to phase across the
280 entire locus, we retain the phasing only for the longest block. The variants for the shorter
281 haplotype blocks can be collapsed into IUPAC ambiguity codes or treated as missing data
282 based on the investigator's preferences. PATÉ outputs analysis-ready fasta files with multiple
283 alleles per species. Variants are only phased within loci; we do not attempt to assign loci to
284 parental subgenomes. While this may complicate analyses of concatenated multi-locus
285 datasets, it is ideal for the MSC that assumes free recombination between loci and can leverage
286 multiple alleles per species for estimating θ. Assigning alleles to parental subgenomes should
287 be possible with PATÉ outputs where both parental lineages are sampled and there is sufficient
288 divergence for gene tree estimation.
289
290 ***Analyses of a Species Complex with Allopolyploidy***
291 *The North American wood fern complex (Dryopteris)* — We tested PATÉ using new
292 target enrichment data from nine North American *Dryopteris* species, including both

293 allotetraploid and allohexaploid taxa, with well-studied reticulate relationships (Sessa et al.
294 2012a, b), as well as two outgroup taxa from the sister genus *Polystichum*. All putative parental
295 lineages are represented in our dataset, with the exception of a hypothesized extinct lineage (*D.*
296 *semicristata*; Sessa et al. 2012b). We sampled two or three individuals for each *Dryopteris*
297 taxon (Table S1). The target enrichment data were generated from the GoFlag 408 flagellate
298 land plant probe set (Breinholt et al. 2021) at RAPID Genomics (Gainesville, FL). The target
299 regions for this probe set are 408 exons found in 229 single or low-copy nuclear genes. We
300 generated haplotype consensus assemblies for each with HybPiper (Johnson et al. 2016). The
301 resulting supercontig sequences, which includes the exon region targeted by probes for
302 enrichment and flanking intron regions, became our reference sequences for genotyping and
303 phasing with PATÉ. We aligned both phased and haplotype consensus (i.e., the reference
304 supercontig) sequences with MUSCLE v.3.8.31 with default settings (Edgar 2004). For
305 comparison with simulation results, we generated a "pick one" dataset where only one phased
306 haplotype per locus was randomly selected from each individual. A "genotype" dataset was
307 generated by collapsing allelic variation into IUPAC codes for each individual as a locus.
308
309 *Three species tests* — We first explored the value of phasing data when estimating
310 reticulate relationships among three species, where we sampled two extant parental lineages
311 and their putative hybrid. We examined the three such extant relationships in the North
312 American *Dryopteris* complex:
313 1) Two diploid parental lineages (*D. ludoviciana* and *D. goldiana*) and their putative
314 allotetraploid descendent (*D. celsa*).
315 2) Two diploid parental lineages (*D. expansa* and *D. intermedia*) and their putative
316 allotetraploid descendent (*D. campyloptera*).
317 3) One tetraploid parent (*D. cristata*), one diploid parent (*D. goldiana*), and their putative
318 allohexaploid descendent (*D. clintoniana*).

These three scenarios represent different ages of parental divergence, with a common ancestor near 10 Ma for *D. ludoviciana* and *D. goldiana*, 15 Ma for *D. expansa* and *D. intermedia*, and 30 Ma for *D. cristata* and *D. goldiana* (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 the different data types. First, using BPP v.4.4.0 (Flouri et al. 2020), we estimated log-marginal likelihoods (ln *mL*) with stepping-stone sampling (Xie et al. 2011) for the three possible rooted 3-taxon trees and network models that imply differences in the timing and direction of gene flow among species (Supplementary Fig. S2). Each ln *mL* estimate used 24 steps, and each step had a posterior sample of 10,000, saving every 30 generations after a 30,000 generation burnin (10% of the total run). The ln *mL* values were then used to calculate the six model probabilities following equation 1 using the bppr v.0.6.3 R package (https://github.com/dosreislab/bppr). A percentile-based 95% confidence interval was generated for each model probability by bootstrapping posterior disributions 1,000 times to investigate uncertainty in model selection.

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

We performed four independent runs for each model for each data type prior to marginal likelihood estimation to ensure convergence and the appropriateness of the MCMC settings. The combined posteriors from the correct model were also used estimate all τ and θ for comparison across data types. Notably, BPP can analytically integrate over possible phases of sequence data when given IUPAC codes representing heterozygous sites (Gronea et al. 2011), which is as accurate as perfectly phased data for diploids (Huang et al. 2022). In comparison, standard gene tree estimation under the Felsenstein likelihood (Felsenstein 1981) treat IUPAC codes as uncertainty and integrate over the possible tip states at sites independently.

343 For our topology-based method, we used the SNaQ function (Solís-Lemus and Ané 2016) from
344 PhyloNetworks v0.12.0 (Solís-Lemus et al. 2017) to test the presence and placement of gene
345 flow between the three species, while also including the two *Polystichum* outgroups. Gene trees
346 were estimated for each data type across the three cases with IQ-TREE v1.6.10 (Nguyen et al.
347 2015), using ModelFinder (Kalyaanamoorthy et al. 2017) for model selection on the
348 unpartitioned alignments. The starting tree was obtained with ASTRAL III v5.6.3 (Zhang et al.
349 2018) from the collection of ML gene trees. We tested the presence of zero, one, or two
350 reticulations based on a two pseudolikelihood point cutoff for accepting an additional
351 reticulation. Each analysis used ten independent optimizations.
352
353 *Nine species tests* — We investigated the differences in networks from SNaQ across
354 data types for the nine-species complex. This complex involves multiple reticulation events on
355 an edge, and thus the true network is not identifiable with topology-based methods (Solis-
356 Lemus and Ané 2016). Although full-likelihood methods should not suffer from the some of the
357 identifiability issues (Yu and Nakhleh 2015), they are not computationally feasible for the North
358 American *Dryopteris* complex. Despite the inherent identifiability issues, it is worthwhile to
359 examine the performance of scalable topology-based methods in such contexts to see if they
360 will identify some of the correct reticulate relationships or if they also will identify erroneous
361 relationships. Thus, we performed analyses of the nine-species complex and two outgroups
362 across the four data types with SNaQ as described above, but we allowed up to six reticulation
363 events. We used ASTRAL III v5.6.3 (Zhang et al. 2018) to generate the starting species tree for
364 network estimation using gene trees inferred from IQ-TREE v1.6.10 (Nguyen et al. 2015) with
365 the best model selected by ModelFinder (Kalyaanamoorthy et al. 2017). 100 standard bootstrap
366 replicates were performed for each gene tree, which were used to evaluate gene tree
367 incongruence and informativeness with respect to the ASTRAL species tree from phased data
368 for the consensus, genotype, and pick one gene trees with phyparts (Smith et al. 2015). To

assess node support for SNaQ, we performed 100 non-parametric bootstrap replicates, in which both loci and bootstrap trees per locus were sampled with replacement. Each bootstrap replicate was distributed independently to hasten the computation using a Perl script that accommodates multiple alleles available on GitHub (https://github.com/gtiley/bootSNaQ-MA).

RESULTS

Simulation Shows Benefits of Phased Data

Network inference — In our simulation results, all data types performed well estimating the correct number of reticulations (Supplementary Fig. S3). The analyses typically converged to the true number of reticulations with 40 or more gene trees, except when using gene trees estimated from genotype data with a 25% ($\theta = 2 \times \tau_h$) chance of a discordance gene tree. However, using phased data provided more accurate estimates of the placement of the reticulation edge in comparison to the genotype, consensus, and pick one data (Fig. 3). 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 loci regardless of the age of hybridization or degree of ILS. The gene trees estimated from phased data perform equally well, aside from requiring a few more gene trees for recent hybridization events with high ILS ($\theta = 4 \times \tau_h$). Analyses based on genotype, consensus, and pick one data almost always recover the true network when sampling 400 or more loci, but can miss the true network over half the time with 40 or fewer loci. However, the genotype data failed to converge to the true network with increasing loci for a hybridization age of 10 Ma with low to moderate amounts of ILS. Otherwise, varying degrees of ILS had a negligible effect on network estimation among data types, and estimated gene trees, aside from those from phased data, were less efficient when the hybridization age was more recent.

394 *Divergence times* — Using phased data, whether using all alleles or only one, typically
395 resulted in accurate estimation of the hybridization age (Fig. 4). When divergence was low,
396 there was a small amount of estimation error, such that the phased data underestimated and
397 the pick one data overestimated the hybridization age, but the pattern suggests estimates could
398 converge to the true values with more loci. Interestingly, convergence towards the true value
399 appeared more efficient with higher ILS (Fig. 4). The phased and pick one data could otherwise
400 accurately estimate the hybridization time with 40 or fewer loci across all other divergence and
401 ILS scenarios. For analyses with genotype and consensus data, as the number of loci increased
402 and uncertainty in the posterior was reduced, the posterior mean did not converge to the true
403 estimate, and the simulated value was not within the highest posterior density (HPD) interval.
404 There was always a small degree of estimation error with the genotype data despite using
405 analytical phasing, and the consensus data overestimated the hybridization age between 1.5x
406 and 2x when using 400 loci across all simulation scenarios. These biases from genotype or
407 consensus data were also propagated for the parent nodes of the hybridization event across all
408 levels of ILS for $\tau_h = 0.001$ (Supplementary Figs. S4-S7), $\tau_h = 0.01$ (Supplementary Figs. S8-
409 S11), and $\tau_h = 0.1$ (Supplementary Figs. S12-S15). Nodes from speciation events could always
410 be estimated accurately with high precision by the phased and pick one data types provided
411 enough loci were sampled. The phased data was more efficient, always converging to the true
412 value by 40 loci, while the pick one data required 400 to recover τ_u when the sequence
413 divergence was low at $\tau_h = 0.001$. There was a small degree of estimation error among
414 speciation nodes for the genotype and consensus data, and the patterns suggested they might
415 not converge to the true values with more loci. However, speciation times were overestimated
416 by genotype and consensus data for $\tau_h = 0.001$ (Supplementary Figs. S4-S7) and $\tau_h = 0.01$
417 (Supplementary Figs. S8-S11), but underestimated by the genotype data and overestimated by
418 the consensus data for $\tau_h = 0.1$ (Supplementary Figs. S12-S15). Across all simulations, there

419 was little effect of increasing ILS aside from increasing HPD intervals, but the uncertainty was
420 negligible when 400 loci were sampled.
421
422 ***Analyses of target enrichment data from Dryopteris***
423 *Recovery of Phased Loci* — On average, 62% of loci sequenced for an individual were
424 phased with eight variants passing filters per locus per individual (Supplementary Table S2).
425 The ploidy level appears to be strongly associated with the number of phased loci. Among
426 diploids, only 32% of loci were phased; the other 68% of diploid loci were either homozygous or
427 had too few linked variants for phasing. For tetraploids and hexaploids, 87% and 94% of loci,
428 respectively, were phased such that two or more phased haplotype sequences could be
429 recovered. Among loci where phasing was possible, variants were almost always resolved as a
430 single haplotype block, as opposed to being split into two or more blocks because not enough
431 reads were available to physically link variants. For polyploids, the number of unique phased
432 haplotype sequences most frequently matched the ploidy level, except in the case of a single *D.*
433 *campyloptera* individual (B087-D08), which also had relatively few recovered loci. Phasing data
434 only extended sequence alignment length by about two base pairs relative to the haplotype
435 consensus sequences (Supplementary Table S3). However, phased data more than tripled the
436 number of parsimony informative sites per alignment, similar to the increase in number of tips
437 (Supplementary Table S3).
438
439 *Placing a single reticulation event* — For our three-species full-likelihood analyses with
440 the MSci model, all data types decisively supported the putative reticulation hypothesis in the
441 cases of the allotetraploids *D. celsa* or *D. campyloptera* (Table 1). However, only the genotyped
442 data, coupled with the analytical phasing technique of BPP, chose the putatively correct model
443 for the allohexaploid *D. clintoniana* (Table 1). The phased and pick one data types selected the
444 network with the diploid *D. goldiana* as the hybrid descendent of the allohexaploid *D. clintoniana*

445 and allotetraploid *D. cristata*, while the best model for consensus data was a binary tree with *D.*
446 *clintoniana* sister to *D. cristata* (Table 1). However, model probability confidence intervals
447 suggest a great deal of uncertainty underlies model selection results for the allohexaploid case,
448 with the upper end of the interval for the putatively correct model at 0.603, 0.582, and 0.996 for
449 the phased, consensus, and pick one data, respectively (Table 1). Model parameters typically
450 converged across all analyses with the exception of the phased data for the *D. clintoniana* triplet
451 (Supplementary Fig. S16), which was due to one MCMC run despite convergence of the lnL
452 values.

453 The simulation experiments established some expectations that the phased and pick
454 one data would provide unbiased estimates of divergence times for nodes involved in the
455 hybridization events, while the genotyped and consensus data would provide under- and
456 overestimates, respectively. However, phased data generally produced younger mean τ
457 estimates for the *D. celsa* triplet with reduced credible intervals relative to other data types, but
458 much older mean τ estimates with more uncertainty for the *D. campyloptera* and *D. clintoniana*
459 triplets (Fig. 5; Supplementary Table S4). Similarly, there was inconsistency in θ estimates
460 observed among data types across triplets, with notably larger credible intervals for phased data
461 in the *D. celsa* and *D. clintoniana* triplets. Estimates of φ were sensitive to data type across
462 triplets, such that the mean φ suggested the major contribution to the focal allopolyploid from
463 phased data was *D. goldiana* into *D. celsa*, *D. intermedia* into *D. campyloptera*, and *D. goldiana*
464 into *D. clintoniana*, with other data types sometimes disagreeing (Fig. 5). Credible intervals
465 around φ always spanned $\varphi = 0.5$, but estimates for genotype, consensus, and phased data
466 suggested some biased gene retention for the *D. celsa* and *D. clintoniana* triplets
467 (Supplementary Table S4).

468 When performing a network search based on gene tree distributions with SNaQ, all data
469 types recovered the putatively correct relationship among *D. celsa* and its diploid parental

470 lineages (Fig. 6a). For *D. campyloptera*, all data types correctly detected one hybrid edge
471 (Supplementary Table S5), but they got the direction of gene flow wrong (Fig. 6b). All data types
472 also detected one hybrid edge for the allohexaploid *D. clintoniana* (Supplementary Table S5),
473 but there was disagreement over the underlying network. Both the phased and consensus data
474 invoked a ghost lineage sister to the three sampled *Dryoptersis,* with gene flow into the
475 allotetraploid *D. cristata*. While not unreasonable given the putatively extinct *D. semiscristata*
476 and missing *D. ludoviciana* from the sample, the relationship between *D. clintoniana* and its
477 parental lineages goes undetected (Fig. 6c). The genotype and pick one data recovered the
478 correct network, but reversed the direction of gene flow from allohexaploid *D. clintoniana* into
479 allotetraploid *D. cristata* (Fig 6c).
480
481 *Inferring relationships among a complex with multiple reticulation events* — The nine
482 *Dryopteris* species SNaQ network analysis with phased data identified two out of five
483 hypothesized reticulation events correctly along with one incorrect reticulation edge (Fig. 7a;
484 Supplementary Table S6). The analysis correctly identified the diploid *D. ludoviciana* and *D.*
485 *goldiana* as parents to the allotetraploid *D. celsa*, as well as the diploids *D. expansa* and *D.*
486 *intermedia* as parents to the allotetraploid *D. campyloptera*. It also detected a moderate level of
487 gene flow ,with an introgression probability of 0.335 from the common ancestor of *D. celsa*, *D.*
488 *ludoviciana*, and *D. goldiana* into the common ancestor of the outgroup *Polystichum*, albeit with
489 low bootstrap support. The parents of *D. celsa* were also detected by the other data types, but
490 all had some degree of error. The genotype data found a reticulation edge between the
491 allotetraploid *D. carthusiana* and diploid parent *D. intermedia*, but reversed the direction of gene
492 flow (Fig. 7b). A reticulation edge was also detected between *D. cristata* and the common
493 ancestor of *D. clintoniana*, *D. goldiana*, *D. celsa*, and *D. ludoviciana*, which is not completely
494 unreasonable given the allotetraploid *D. cristata* is believed to be a parent of *D. clintoniana*. The
495 consensus data only detected two reticulation edges (Supplementary Table S6), with the

496 second edge correctly showing gene flow from the diploid parent *D. intermedia* into the
497 allotetraploid *D. carthusiana*, provided the other parent *D. semicristata* is presumed extinct
498 obscuring the network interpretation (Fig. 7c). The pick one data, similar to the genotype data,
499 inferred a reticulation edge between *D. carthusiana* and *D. intermedia* with the wrong direction
500 of gene flow (Fig. 7d). The pick one data also detected a reticulation edge from *D. cristata* into
501 the common ancestor of *D. carthusiana D. intermedia*, *D. campyloptera*, and *D. expansa*, which
502 lakcs a clear interpretation other than invoking missing lineages underlying complexity of the
503 backbone topology (Fig. 1). In the one case where all data types detected the correct
504 relationship regarding *D. celsa*, the introgression probability was approximately 0.5 with phased
505 data while between 0.37 and 0.39 with others (Fig. 7; Supplementary Fig. S17).
506 Differences in the network estimates among data types may be explained by their
507 underlying spectra of quartet concordance factors (Supplementary Fig. S18). While spectra
508 were overall similar among data types, some deviations from expectations under ILS alone were
509 stronger in the phased data versus others (Supplementary Fig. S18). Analyses of gene tree
510 incongruence among the consensus (Supplementary Fig. S19), genotype (Supplementary Fig.
511 S20), and pick one (Supplementary Fig. S21) data showed similar levels of conflict among the
512 different data types with respect to a species tree estimated from the phased data (such
513 analyses with multiple alleles per species is not possible). The frequency of conflicting splits
514 with high bootstrap support was similar across datatypes for all nodes (Supplementary Fig.
515 S22), and low-support conflicts represented on average 56% of splits within *Dryopteris*.
516
517 **DISCUSSION**
518 New phylogenetic network methods offer the promise of elucidating the often complex
519 reticulate histories of polyploid lineages, even in the presence of ILS. Our results demonstrate
520 that under some conditions phasing polyploid target enrichment data can improve the accuracy
521 of such network inferences as well as divergence time estimates for the networks, and we

522 describe a novel pipeline (PATÉ) to address the difficult problem of phasing polyploid data.
523 Although PATÉ could handle different types of genomic data, such as transcriptomes and whole
524 genomes, target enrichment data are ideal for computational phasing because they can yield
525 high (i.e. greater than 100x) and even coverage within a locus. This was demonstrated by
526 almost all loci under investigation here being recovered as single haplotype blocks
527 (Supplementary Table S2). Because MSC methods treat loci independently and assume free
528 recombination between loci, it is not necessary to assign individual loci to parental subgenomes
529 for network analyses. However, the allele sequences output by PATÉ can also be used as input
530 for other tools such as *Homologizer* (Freyman et al. 2023) or *HybPhaser* (Nauheimer et al.
531 2020), which attempt to phase across loci and assign parental subgenomes. Our methods also
532 enable population genomic studies of polyploids (Tiley et al. 2023) where accurate estimation of
533 site frequency spectra can be used for demographic analyses otherwise complicated by
534 polyploidy (e.g., Excoffier et al. 2013; Liu and Fu 2020) or SNP-based network inference in the
535 absence of variation suitable for gene tree estimation (Blischak et al. 2018; Olave and Meyer
536 2020).
537
538 ***Promises and Limits of Phasing***
539 The prospect of using alleles from phased genomic data presents an exciting step
540 towards revealing the evolutionary history of polyploids, which remains a critical impediment
541 within the plant evolution community (McKain et al. 2018). New strategies for explicitly
542 addressing this challenge are emerging (Freyman et al. 2023; Nauheimer et al. 2020; Mendez-
543 Reneau et al. 2023), and PATÉ can be a useful tool by phasing variants for many individuals
544 while leveraging genotype quality information. Our simulations demonstrated that phasing can
545 improve estimates of reticulate evolutionary relationships using network methods. In
546 simulations, phased data with all alleles represented in gene trees could more efficiently recover
547 the placement and directionality of hybrid edges than the genotyped, haplotype consensus, or

548 single haplotype sequences (Fig. 3). The advantages of using phased data for network
549 estimation diminished in simulations as the number of loci increased (Fig. 3), and this was
550 evident from our empirical analyses of *Dryopteris* triplets which produced similar results across
551 data types with 330 or 328 loci (Table 1; Fig. 6), except for the correct inference of allopolyploid
552 and parent relationships with genotyped data in the case of *D. clintoniana* (Table 1).

553 Both our simulations (Fig. 4; Supplementary Figs. S4-S15) and previous investigations
554 (Anderman et al. 2019). Our empirical analyses also demonstrated how the timing of
555 introgression τ_h and other parameters can be greatly affected by the data type used (Fig. 5). In
556 our *Dryopteris* analyses, τ_h for *D. celsa* was at least in the same order-of-magnitude among
557 data types, which had partially overlapping credible intervals, while phased τ_h estimate *for D.*
558 campyloptera was nearly four times older than the other data types (Fig. 5; Supplementary
559 Table S4). The *D. clintoniana* τ_h estimate from phased data was also much older than other
560 data types; albeit, with a 95% HPD interval. There were noticeable mixing problems in our *D.*
561 *clintoniana* full-likelihood analyses, which makes it difficult to evaluate whether the uncertainty
562 was due to extreme differences in θ among phased alleles within *D. clintoniana* and/or the
563 computational difficulty of the Bayesian methods with increasing sequence data. In all cases,
564 the differences between τ_h among data types is not consistent with results from the simulation
565 experiements, which suggested the consensus data should overestimate age compared to
566 phased or pick one data, with genotyped data having slightly younger estimates than phased
567 data. This highlights the difficulties of simulating data that capture the complexities of empirical
568 data and makes deciding which estimate is more reliable somewhat difficult. For example the
569 uncertainty of τ_h with phased data for *D. celsa* and *D. campyloptera* triplets was reduced
570 compared to the consensus and pick one estimates, but greater uncertainty was observed for τ
571 and θ at other nodes (Fig. 5). A potential compromise from simulations and empirical analyses
572 is the use of analytical phasing of polyploid data, as done with our genotyped data for BPP

analyses. Analytical phasing (Gronau et al. 2011; Flouri et al. 2020) has been shown to perform well in diploids (Huang et al. 2022), and any potential biases from ignoring ploidy seem outweighed by getting the *D. clintoniana* model correct (Table 1) as well as seemingly reasonable parameter estimates under the MSci model with tight credible intervals (Fig. 5).

Our analyses also suggest that phasing 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 in simulations (Supplementary Figs. S4-S15), as phylogenetic information is lost from multiple hits, the influence of the prior in full likelihood analyses should become stronger. Deeper divergences will likely generate more switch errors among phased haplotype sequences generated by PATÉ (Xie et al. 2016), as there will be more opportunities for paralogy, structural variants, and sequence divergence to confound the read mapping, genotyping, and phasing processes. When enough reads are available to call high-confidence variants, we suggest that phasing with PATÉ can improve network estimation based on the correct identification of the allotetraploids *D, celsa* and *D. campyloptera* and their parental lineages in the analyses of the North American *Dryopteris* complex (Fig. 7a); however, phasing alone cannot overcome some inherent challenges of network estimation.

Challenges of Network Estimation

Our analyses highlight the difficulty of estimating the evolutionary histories of reticulate complexes, regardless of data type. The full-likelihood implementation of the MSci model appears to be useful when the number of tips and plausible models for investigation are low, but these methods 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 much faster and accurate when model assumptions are met (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 the true networks are non-identifiable (Yu and

599 Nakhleh 2015; Solís-Lemus and Ané 2016). For example, when multiple introgression events
600 affect the same lineage, such that there are overlapping cycles in a network (Solís-Lemus and
601 Ané 2016), the expected quartet distribution under the MSci model becomes a poor fit for the
602 empirical data (Cai and Ané 2020) and the network estimated may be incorrect. These effects
603 were evident in our empirical analyses where the relationship between *D. campyloptera*, *D.*
604 *expansa*, and *D. intermedia* was missing in the nine-species analysis (Fig. 7), which we expect
605 is due to a reticulation edge present between *D. intermedia* and *D. carthusiana*. Phasing
606 sequence data adds information that can improve estimates (Supplementary Table S3), but
607 unsampled or extinct lineages, such as the hypothesized *D. semicristata,* can create significant
608 barriers to recovering the true evolutionary history of reticulate complexes, regardless of how
609 many loci or individuals are available. Another identifiability problem is getting the direction of
610 gene flow correct (Yu and Nakhleh 2015; Thawornwattana et al. 2023), and these scenarios
611 were evident in our quartet-based analyses of a triplet that reversed gene flow from the
612 allotetraploid *D. campyloptera* into the diploid *D. expansa* (Fig. 6d) or in the analyses of the full
613 complex that reversed gene flow from the allotetraploid *D. carthusiana* into the diploid *D.*
614 *intermedia* (Fig. 7b and 7d).
615
616 ***Insights into Dryopteris Evolution***
617 The North American *Dryopteris* complex has been well-characterized through the study
618 of multi-locus nuclear and chloroplast phylogenies, morphology, cytological observations,
619 chromatography, and isozyme analyses (reviewed in Sessa et al. 2012b). This makes it a useful
620 system for testing our phasing pipeline, and our analyses add nuance to our understanding of
621 some of the relationships among *Dryopteris* species. For example, our results with phased data
622 indicated *D. campyloptera* has received more loci from *D. intermedia* than *D. expansa* (Fig. 7a),
623 and *D. intermedia* is likely the maternal progenitor of *D. campyloptera* (Sessa et al. 2012b).
624 Following the allopolyploidy event, the *D. intermedia* subgenome may have been dominant,

625 providing a selective advantage for *D. campyloptera* in its distribution at the time (Bird et al.
626 2018). The phased nine-taxon analyses also supported the hypothesis of the unsampled diploid
627 lineage *D. semicristata*, based on the placement of *D. carthusiana* as sister to the rest of
628 *Dryopteris*, and the general sensitivity of *D. carthusiana* and *D. cristata* in the phylogenetic
629 backbone across data types (Fig. 7). Additional insights may be gained by a more complete
630 sampling of the genus, as the North American *Dryopteris* complex is not a clade (Sessa et al.
631 2012b), which should alleviate some of the identifiability problems. Analyses of the North
632 American *Dryopteris* complex alone highlight the need for caution when interpreting networks,
633 that are increasingly prominent plant evolutionary research.
634
635 ***Does Phasing Matter?***
636 Combining phased data with recent network methods can help confront a major
637 challenge of plant phylogenetics: resolving the complex histories of polyploids. While our
638 analyses showed haplotype consensus sequences can be adequate for resolving single
639 reticulation events where both parental lineages are sampled, phased sequences may provide
640 more information with few loci sequence divergence is not too high or improve divergence time
641 estimates (see also Huang et al. 2022). Sequencing hundreds to thousands of loci is becoming
642 easier, but large studies with degraded herbarium material can have low returns in target
643 enrichment studies. Analytical phasing approaches (Gronau et al. 2011; Flouri et al. 2020)
644 provide a way forward available for full-likelihood methods while avoiding the potential for switch
645 errors from read-based phasing as done in PATÉ, but these will become difficult with many tips.
646 This highlights the need for continued development of models that explicitly consider
647 polyploidization mechanisms (Jones et al. 2013) while addressing scalability and limiting
648 assumptions for contemporary genomic data (Yan et al. 2022). Some reticulate complexes will
649 be difficult to disentangle with any data when confronted with statistical identifiability problems,
650 such as networks with multiple reticulations on a single edge (Solís-Lemus et al. 2016) when

using quartet-based methods, and our *Dryopteris* analyses demonstrate that caution is needed when interpreting biological processes from phylogenetic networks in the absence of corroborating evidence. As haplotype-phased genomes from long-read data become increasingly available, the discussion of phase resolution will become trivial, but progress on models that leverage such data and improve interpretation of processes, especially in polyploids that maintain distinct subgenomes, from networks is still needed.

DATA AVAILABILITY

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. (https://github.com/gtiley/Phasing). Simulated and empirical data supporting findings and files for replicating analyses are available from the Dryad Digital Repository: http://dx.doi.org/10.5061/dryad.[NNNN]. 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: http://dx.doi.org/10.5061/dryad.[NNNN]

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, L. Nauheimer, and C.J. Rothfels 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. This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No. 101026923, awarded to GPT.

LITERATURE CITED

Albert T.J., Molla M.N., Muzny D.M., Nazareth L., Wheeler D., Song X., Richmond T.A., Middle C.M., Rodesch M.J., Packard C.J., Weinstock G.M., Gibbs R.A. 2007. Direct selection of human genomic loci by microarray hybridization, *Nature Methods*. 4:903-905.

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.

Baudry J.-P., Maugis C., Michel B. 2012. Slope heuristics: overview and implementation. *Statistics and Computing*. 22:455-470.

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.

Felsenstein J. 1981. Evolutionary trees from DNA sequences: A maximum likelihood approach. *Journal of Molecular Evolution*. 17:368-376.

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. 2023. Homologizer: Phylogenetic phasing of gene copies into polyploid subgenomes. *Methods in Ecology and Evolution*. 14:1230-1244.

Gnirke A., Melnikov A., Maguire J., Rogov P., LeProust E.M., Brockman W., Fennell T., Giannoukos G., Fisher S., Russ C., Gabriel S., Jaffe D.B., Lander E.S., Nusbaum C. 2009. Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nature Biotechnology*. 27:182-189.

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., Bennett J., Flouri T., Leaché A.D., Yang Z. 2022. Phase resolution of heterozygous sites in diploid genomes is important to phylogenomic analysis under the multispecies coalescent model. *Syst Biol.* 71:334-352.

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.

Ji J., Jackson D.J., Leaché A.D., Yang Z. 2022. Power of Bayesian and heuristic tests to detect cross-species introgression with reference to gene flow in the *Tamias quadrivittatus* group of North American chipunks. *Syst Biol.* syac077.

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.

Kamneva O.K., Syring J., Liston A., Rosenberg N.A. 2017. Evaluating allopolyploid origins in strawberries (*Fragaria*) using haplotypes generated from target capture sequencing. *BMC Evol Biol*. 17:180.

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.

Kass R.E., Raftery A.E. 1995. Bayes factors. *Journal of the American Statistical Association*. 90:773-795.

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.

Liu X.-Q., Xia X.-M., Chen L., Wang X-Q. 2022. Phylogeny and evolution of Cupressaceae: Updates on intergeneric realtionships and new insights on ancient intergeneric hybridization. *Mol. Phylogenet. Evol.* 177:107606.

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.

Mendez-Reneau J., Burleigh J.G., Sigel E.M. 2023. Target capture methods offer ingsight into the evolution of rapidly diverged taxa and resolve allopolyploid homeologs in the fern genus *Polypodium* s.s. *Systematic Botany* 48:96-109.

Meng C., Kubatko L.S. 2009. Detecting hybrid speciation in the presence of incomplete lineage sorting using gene tree incongruence: a model. *Theor. Popul. Biol.* 75:35-45.

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.

Sun M., Soltis D.E., Soltis P.S., Zhu X., Burleigh J.G., Chen Z. 2015. Deep phylogenetic incongruence in the angiosperm clade *Rosidae*. *Mol. Phylogenet. Evol.* 83:156-166.

Thawornwattana Y., Huang J., Flouri T., Mallet J., Yang Z. 2023. Inferring the direction of introgression using genomic sequence data. *Mol Biol Evol*. 40:msad178. https://doi.org/10.1093/molbev/msad178

Tiley G.P., Crowl A.A., Almary T.O.M., Luke W.R.Q., Solofondranohatra C.L., Besnard G., Lehmann C.E.R., Yoder A.D., Vorontsova M.S. 2023. Genetic variation in *Loudetia simplex* supports the presence of ancient grasslands in Madagascar. *Plants People Planet* in press. https://doi.org/10.1002/ppp3.10437

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 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.

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.

Yan Z., Cao Z., Liu Y., Ogilvie H.A., Nakhleh L. 2022. Phylogenetic networks in the presence of polyploid complexes. *Syst Biol.* 71:706-720.

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.

Yu Y., Nakhleh L. 2015. A maximum pseudo-likelihood approach for phylogenetic networks. *BMC Genomics*. 16:S10.

Yu Y., Than C., Degnan J.H., Nakhleh L. 2011. Coalescent histories on phylogenetic networks and detection of hybridization despite incomplete lineage sorting. *Syst. Biol.* 60:138-149.

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.

1024 **FIGURES**

1025

1026

1027

1028

1029

1030

1031

1032

1033

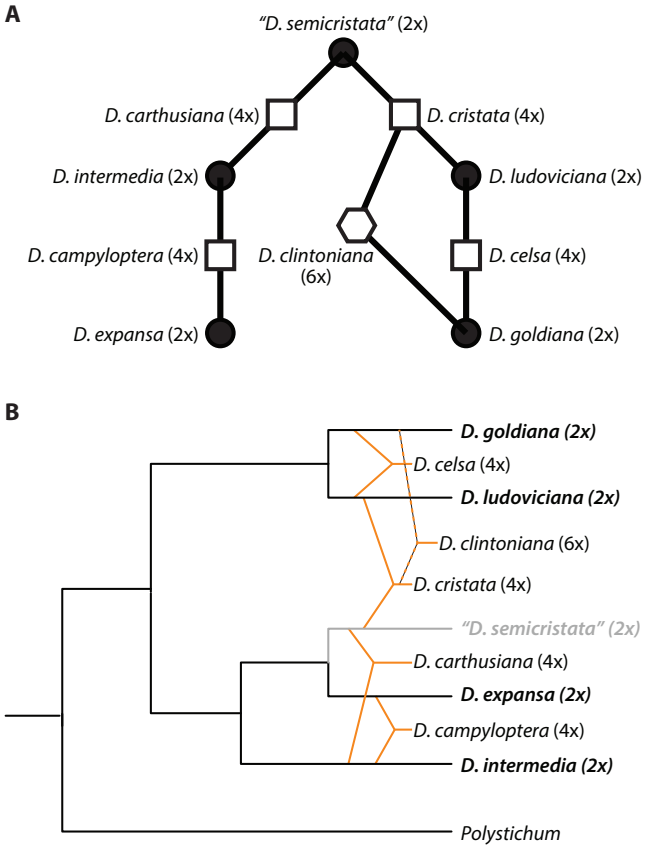
1034

1035

1036

1037

1038



1039 **Figure 1 — Hypothesized Relationships among North American *Dryopteris*.** Synthesis of

1040 results from Sessa et al. 2012a and Sessa et al. 2012b. A) Links between shapes show the

1041 putative parents and their allopolyploid derivatives. Black circles are diploids, squares are

1042 tetraploids, and the hexagon is the one hexaploid species in the group. *Dryopteris semicristata*

1043 is presumed extinct. B) Placement of allopolyploids in the context of the backbone relationships

1044 among diploids. Tetraploids are indicated with solid orange lines and the hexaploid with dotted

1045 lines. The grey line denoting a sister relationship between *D. semicristata* and *D. expansa*

1046 reflects one possible placement for the extinct taxon based on previous analyses (Sessa et al.,

1047 2012b).

1048

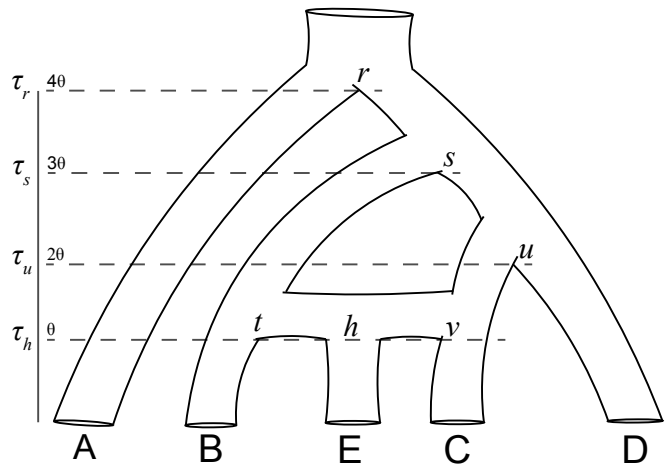


Figure 2 — Species network used for simulation. The divergence times in expected substitutions per site are given for each node, and *h* is the hybrid node where two alleles enter from both *t* and *v*. E is an allotetraploid while other species are diploid. Absolute divergence times were changed by using values of θ at 0.001, 0.01, and 0.1. ILS was increased by multiplying θ by a factor of 2, 3, or 4 while keeping all $\tau_u$fixed to their initial values.

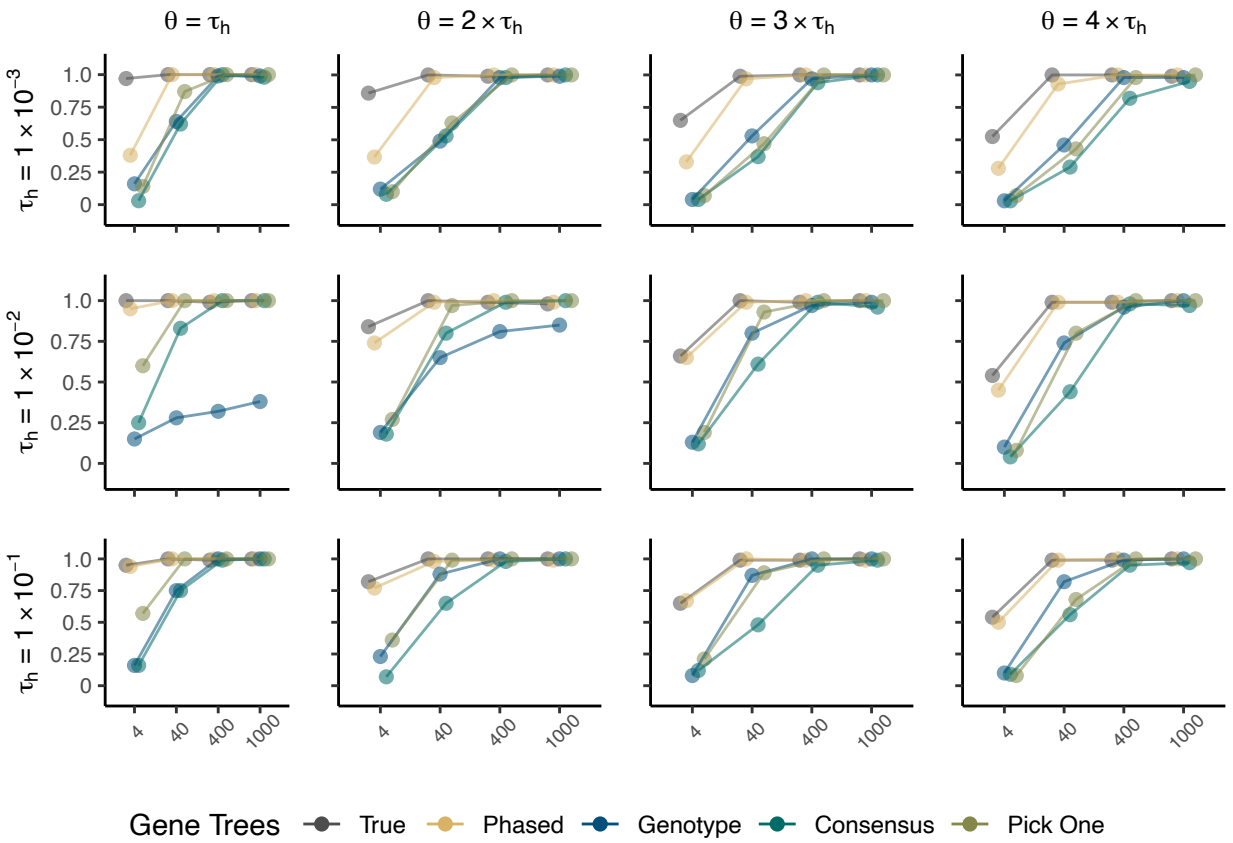


Figure 3 — Proportion of simulation replicates that correctly identify the allopolyploid lineage using the SNaQ function of Phylonetworks. The x-axes are the number of loci sampled for each simulation. The y-axes are the proportions of correct networks. Results are based on networks estimated with a 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, genotype, consensus, and pick one data types.

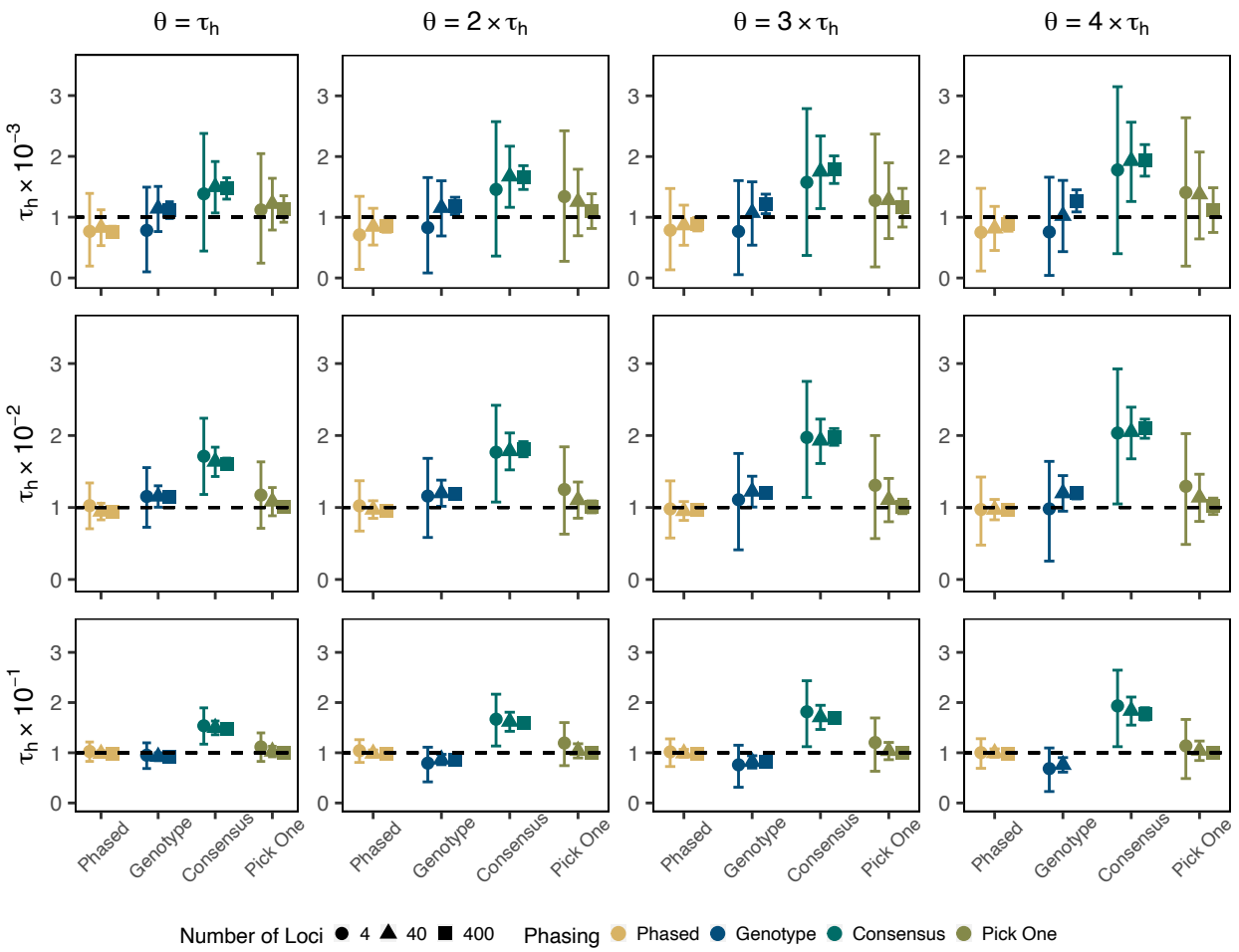


Figure 4 — Estimating the Timing of Hybridization with the MSci Model. Divergence times are for node *h* in Figure 2. Divergence times are measured as the expected number of substitutions per site. The dashed line represents the true simulated values. Points are posterior means, and error bars are 95% HPD intervals, averaged across 100 replicates. Results are not shown for the genotype data when $\tau_h = 0.1$ and $\theta = 4\tau_h$ because posterior means went towards infinity and numerically unstable values that cannot be displayed.

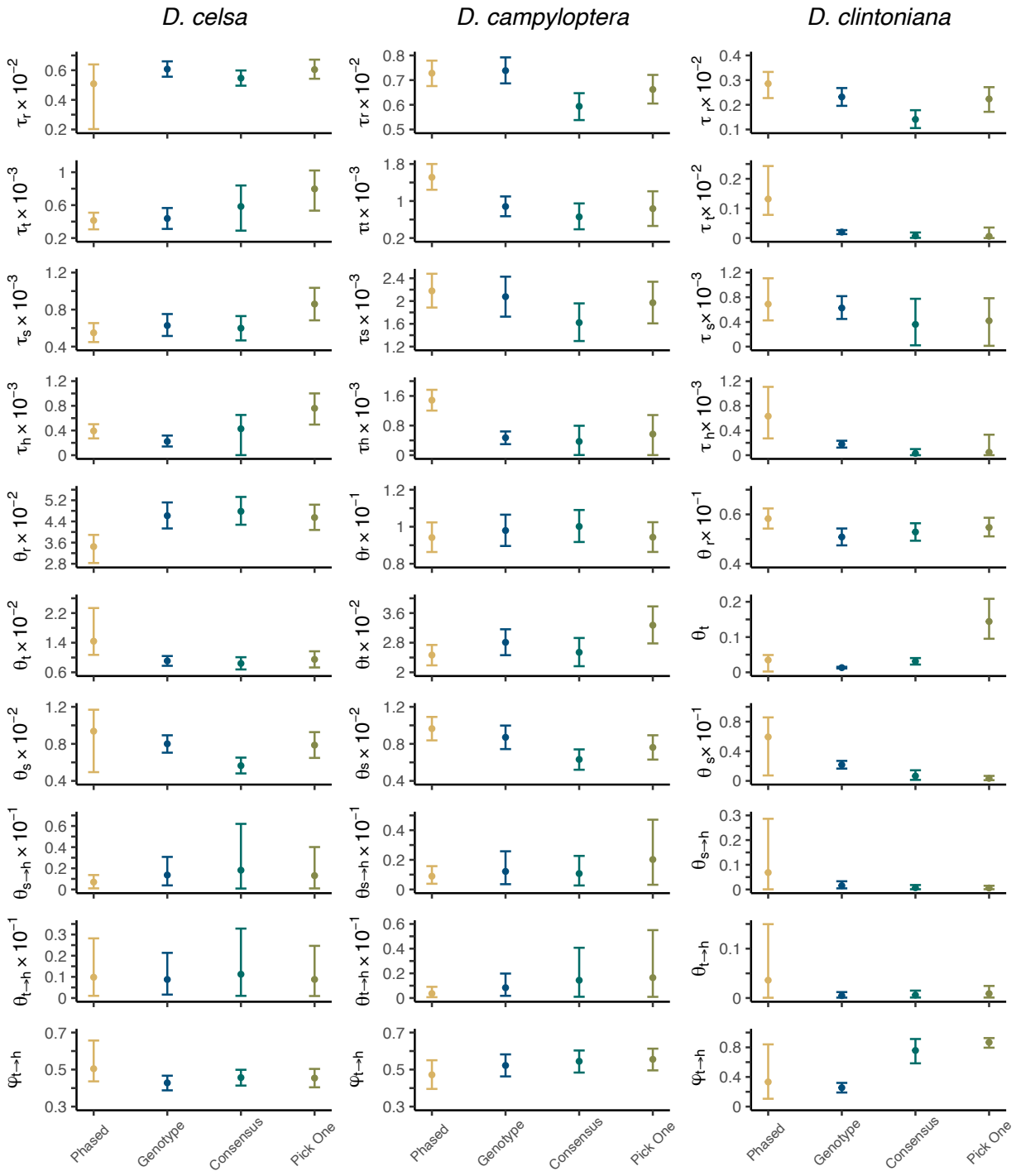
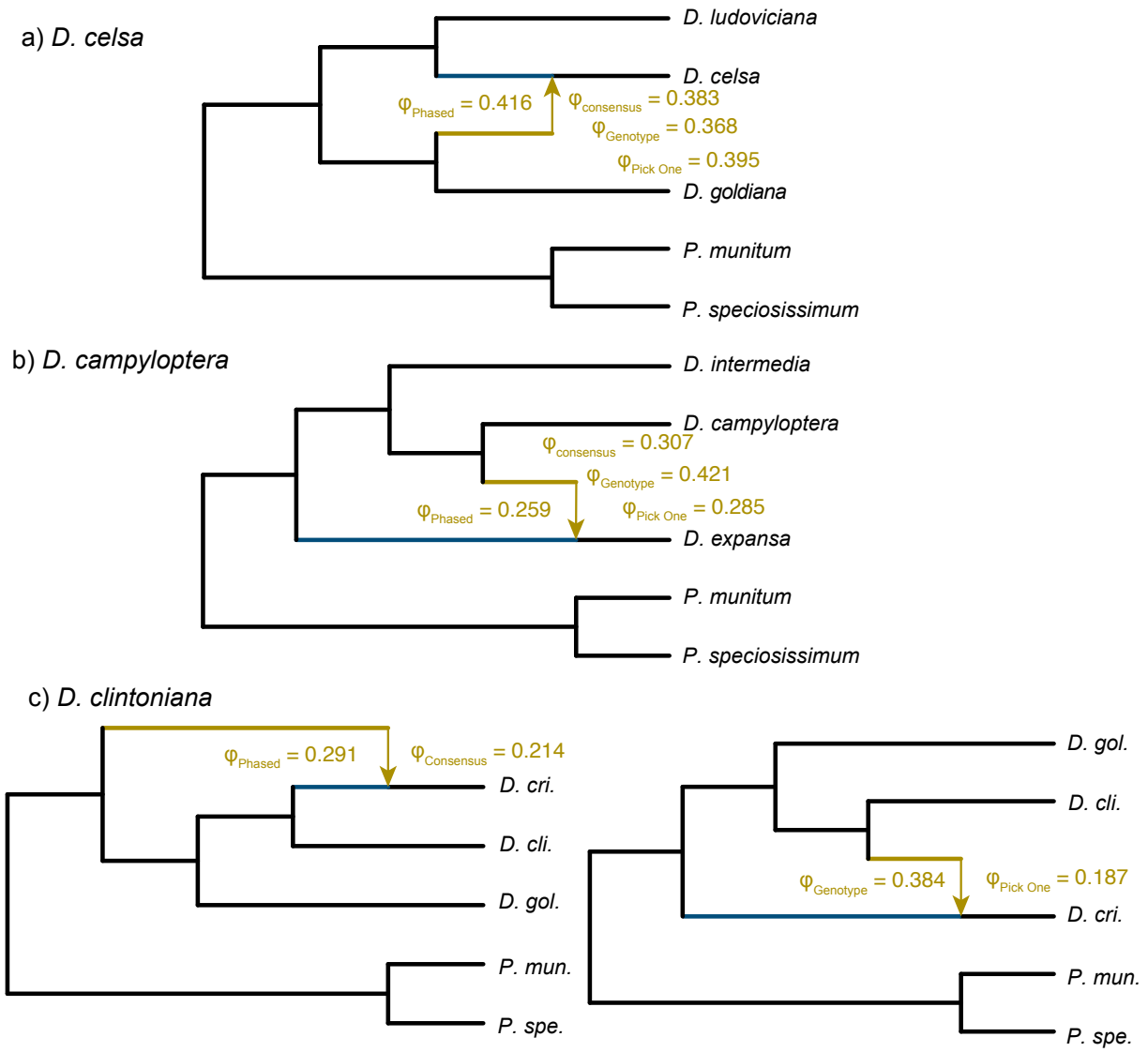


Figure 5 — *Dryopteris* Divergence Times, Population Sizes, and Introgression Probabilities under the MSci Model. Divergence times are measured as the expected number of substitutions per site. The x-axis shows estimates from phased data and the y-axis shows estimates from unphased (haplotype consensus) data for the inset network (equivalent to model

1085 7 from Supplementary Fig. S1). The dashed one-to-one line shows where older age estimates
1086 are consistently obtained from phased data. Error bars on points show the 95% HPD intervals
1087 for phased and unphased data.
1088

1089



1090 **Figure 6 — Network search results for three-taxon *Dryopteris* cases with SNaQ.** The major
1091 topology edge is indicated in blue and the minor edge (direction of introgression) is shown in
1092 gold. The introgression probability (φ) is shown for each data type. In cases where a) *D. celsa*
1093 or b) *D. campyloptera* were the allotetraploids under investigation, all data types recovered the
1094 same species network. When c) *D. clintoniana* was the allohexaploid lineage from the
1095 allotetraploid *D. cristata* and diploid *D. goldiana* progenitors, different networks were recovered
1096 by the phased or consensus versus genotype or pick one data types.

1097

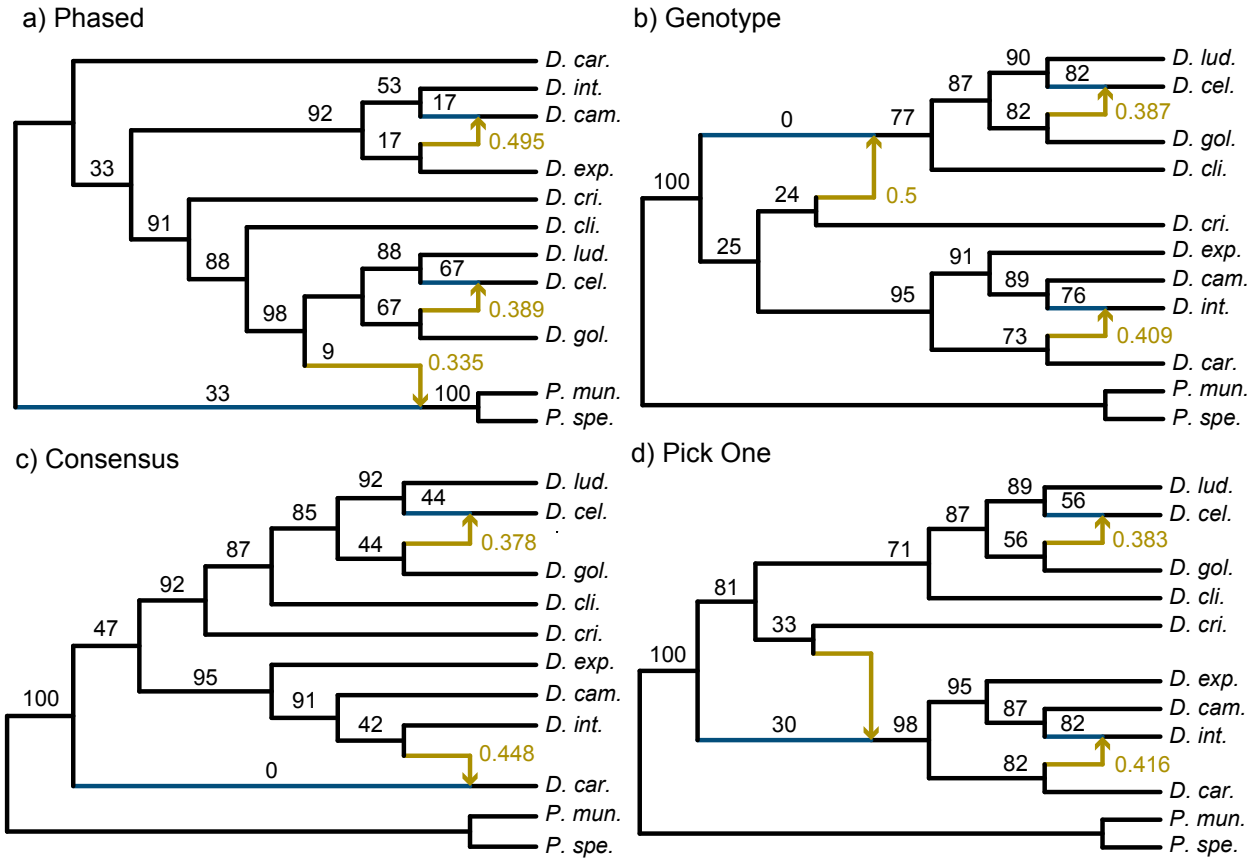


Figure 7 — Networks for all North American *Dryopteris* inferred from SNaQ. The major topology edge is blue and the minor (reticulation) edges are shown in gold, with the direction of introgression flowing into the major edge. Introgression probabilities for each reticulation edge are shown in gold. Numbers on branches subtending nodes are bootstrap support values. Results are shown separately for a) phased data where all haplotype sequences are included in gene trees, b) genotyped data with a single sequence for individual with IUPAC ambiguity codes, c) haplotype consensus sequences, and d) a single haplotype sequence representing each individual among gene trees.

TABLES

Hybrid	Topology	Phased Model Probability	Genotype Model Probability	Consensus Model Probability	Pick One Model Probability
D. celsa	1	8.02×10^{-142} (3.62×10^{-146}, 3.31×10^{-137})	1.95×10^{-173} (1.03×10^{-177}, 2.78×10^{-169})	5.79×10^{-168} (3.15×10^{-172}, 8.7×10^{-164})	4.32×10^{-105} (1.99×10^{-109}, 2.27×10^{-101})
	2	3.18×10^{-126} (4.89×10^{-131}, 1.34×10^{-121})	5.44×10^{-136} (2.17×10^{-140}, 7.99×10^{-132})	1.66×10^{-113} (9.18×10^{-118}, 2.83×10^{-109})	2.78×10^{-84} (1.24×10^{-88}, 2.14×10^{-80})
	3	3.33×10^{-124} (5.19×10^{-129}, 1.66×10^{-119})	5.77×10^{-149} (1.63×10^{-153}, 1.31×10^{-144})	5.66×10^{-162} (6.09×10^{-166}, 9.31×10^{-158})	1.44×10^{-109} (6.87×10^{-114}, 2.06×10^{-105})
	4	1.36×10^{-146} (5.71×10^{-151}, 5.14×10^{-142})	6.98×10^{-149} (2.24×10^{-153}, 1.62×10^{-144})	2.89×10^{-152} (1.19×10^{-156}, 4.33×10^{-148})	1.44×10^{-109} (8.54×10^{-114}, 8.43×10^{-106})
	5†	**1 (1, 1)**	**1 (1, 1)**	**1 (1, 1)**	**1 (1, 1)**
	6	3.69×10^{-113} (1.31×10^{-117}, 1.45×10^{-108})	1.59×10^{-143} (7.07×10^{-148}, 3.22×10^{-139})	1.97×10^{-118} (1.83×10^{-122}, 4.43×10^{-114})	1.46×10^{-92} (1.12×10^{-96}, 8.06×10^{-89})
D. campyloptera	1	1.42×10^{-251} (1.6×10^{-256}, 5.32×10^{-247})	3.77×10^{-251} (2.11×10^{-255}, 6.16×10^{-247})	3.15×10^{-154} (3.78×10^{-158}, 5.25×10^{-150})	1.52×10^{-130} (5.8×10^{-135}, 3.11×10^{-126})
	2	1.03×10^{-210} (1.12×10^{-215}, 8.01×10^{-206})	2.54×10^{-226} (1.33×10^{-230}, 2.41×10^{-222})	1.69×10^{-110} (9.98×10^{-115}, 1.72×10^{-106})	1.11×10^{-90} (5.02×10^{-95}, 2.23×10^{-86})
	3	2.16×10^{-242} (2.64×10^{-247}, 6.06×10^{-238})	1.53×10^{-184} (7.18×10^{-189}, 5.03×10^{-180})	1.24×10^{-129} (6.49×10^{-134}, 2.39×10^{-125})	5.72×10^{-123} (5.18×10^{-127}, 1.1×10^{-118})
	4	1.22×10^{-133} (7.33×10^{-139}, 7.63×10^{-129})	1.87×10^{-193} (6.47×10^{-198}, 2.12×10^{-189})	2.8×10^{-134} (1.98×10^{-138}, 7.92×10^{-130})	4.19×10^{-114} (2.78×10^{-118}, 6.67×10^{-110})
	5†	**1 (1, 1)**	**1 (1, 1)**	**1 (1, 1)**	**1 (1, 1)**
	6	4.96×10^{-173} (6.57×10^{-178}, 3.7×10^{-168})	1.88×10^{-189} (9.15×10^{-194}, 1.93×10^{-185})	5.11×10^{-103} (2.49×10^{-107}, 4.81×10^{-99})	5.44×10^{-86} (3.65×10^{-90}, 1.32×10^{-81})
D. clintoniana	1	3.44×10^{-110} (5.27×10^{-115}, 1.25×10^{-105})	6.01×10^{-65} (2.21×10^{-69}, 1.99×10^{-60})	2.62×10^{-12} (8.24×10^{-17}, 1.82×10^{-8})	1.52×10^{-13} (3.47×10^{-18}, 4.54×10^{-10})
	2	1.44×10^{-04} (1.46×10^{-9}, 0.826)	1.91×10^{-47} (4.69×10^{-52}, 3×10^{-43})	4.49×10^{-9} (1.83×10^{-13}, 4.18×10^{-5})	5.84×10^{-6} (9.84×10^{-11}, 7.03×10^{-2})
	3	4.6×10^{-134} (6.97×10^{-139}, 5.65×10^{-130})	2.31×10^{-55} (5.51×10^{-60}, 5.35×10^{-51})	**1 (4.84×10^{-2}, 1)**	2.23×10^{-12} (5.82×10^{-17}, 1.38×10^{-8})
	4	1.23×10^{-115} (3.09×10^{-120}, 2.69×10^{-111})	2.18×10^{-27} (3.04×10^{-32}, 3.68×10^{-23})	*1.88×10^{-4} (1.05×10^{-8}, 0.728)*	2.02×10^{-27} (3.5×10^{-32}, 1.34×10^{-23})
	5†	*4.23×10^{-5} (7.74×10^{-10}, 0.603)*	**1 (1, 1)**	*1.73×10^{-4} (6.85×10^{-09}, 0.582)*	*5.7×10^{-3} (1.31×10^{-7},0.996)*
	6	**1 (3.81×10^{-2}, 1)**	2.48×10^{-31} (6.11×10^{-36}, 5.76×10^{-27})	1.04×10^{-15} (6.71×10^{-20}, 6.87×10^{-12})	**0.994 (3.6×10^{-3}, 1)**

†The *a priori* allopolyploid hypothesis is always number five.

Table 1 — Model Probabilities for Possible Topological Hypotheses Using the MSci model. Topologies of six models that enumerate the possible binary and reticulate evolutionary histories are displayed in Figure S2. The best model for each data type and triplet is shown in bold. A percentile-based 95% confidence interval is given for each probability by bootstrapping the posteriors.