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: george.tiley@gmail.com

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 phylogenetic network inference under the multispecies coalescent by obtaining the true network with fewer loci compared to haplotype consensus sequences or sequences with heterozygous bases represented as ambiguity codes. 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 non-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 remains 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 can create 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 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 (Kamneva et al. 2017; Oberprieler et al. 2017; Eriksson et al. 2018), 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;

64 Rothfels et al. 2017; Eriksson et al. 2018). However, it remains challenging to genotype (e.g.
65 Blischak et al. 2018b) and phase next generation sequencing data from polyploids using short
66 reads, in which individual variants may lack physical linkage information (e.g. Schrinner et al.
67 2020). Methods exist to genotype consensus loci from target enrichment data, but these have
68 been either limited to diploids (Kates et al. 2018; Andermann et al. 2019) or used manual
69 curation of variants with polyploids where both parental populations are available (Eriksson et
70 al. 2018). Otherwise, obtaining phased sequence data for polyploids has largely depended on
71 long-read sequencing to recover complete haplotype sequences (e.g., Rothfels et al. 2017) or
72 cloning PCR products (e.g., Sessa et al. 2012; Oberprieler et al. 2017).
73 Target enrichment (or HybSeq), in which specific regions of the genome are isolated and
74 sequenced (Albert et al. 2007; Gnirke et al. 2009; Faircloth et al. 2012; Lemmon et al. 2012), is
75 an increasingly common method for collecting large-scale phylogenomic datasets, and these
76 data can provide insights into the evolutionary history of reticulate complexes (e.g., Crowl et al.
77 2020; Karimi et al. 2020) and sources of gene tree discordance (e.g., Crowl et al. 2017;
78 Morales-Briones et al. 2018; Stull et al. 2020). Probe kits for target enrichment have been
79 developed in many land plant lineages (Wolf et al. 2018; Johnson et al. 2019; Liu et al. 2019;
80 Breinholt et al. 2021), and there are bioinformatic pipelines available for custom probe design
81 (e.g., Jantzen et al. 2020). The most common approach to assemble phylogenetic datasets from
82 target enrichment data has been to use haplotype consensus assemblies from various pipelines
83 (e.g. Faircloth 2016; Johnson et al. 2016; Andermann et al. 2018; Breinholt et al. 2018). The *de*
84 *novo* assembly algorithms within these pipelines often consider only the most frequent
85 nucleotide sequence and discard the alternatives (Bankevich et al. 2012; Iqbal et al. 2012; Luo
86 et al. 2012). This results in loci in which variable positions are collapsed to a single base call
87 (haplotype consensus loci), losing information related to heterozygosity. While this may be
88 appropriate, or at least benign, for phylogenetic analyses of diploid taxa (Kates et al. 2018), it
89 may pose substantial problems when attempting to investigate the evolutionary history of

90 polyploid taxa or reticulate lineages. Some of the before-mentioned pipelines can recover
91 heterozygosity information and phase sequences by mapping reads back to their haplotype
92 consensus sequences (Andermann et al. 2018), but this ability is, at the time of writing, limited
93 to diploids.

94 Recovering phased haplotype sequences from polyploid individuals, especially
95 allopolyploids with distinct subgenomes, should be valuable when estimating phylogenetic
96 networks that likely are more accurate representations of reticulate evolution than a bifurcating
97 tree. Simulations and empirical analyses have suggested that phylogenetic networks can
98 recover the reticulate histories of polyploid lineages with few loci, at least when gene tree
99 discordance due to incomplete lineage sorting (ILS; Hudson et al. 1983; Pamilo and Nei 1988)
100 is low (Oberprieler et al. 2017) using parsimony methods (Huber et al. 2006; Lott et al. 2009), or
101 with moderate ILS when it is explicitly modeled (Jones et al. 2013). Contemporary phylogenetic
102 network models and software packages jointly consider gene tree variation due to allele
103 sampling error as described by the multispecies coalescent (MSC; Rannala and Yang 2003)
104 and gene flow modeled as episodic introgression events (Solis-Lemus and Ané 2016; Wen et al.
105 2016; Zhang, Ogilvie et al. 2018, Flouri et al. 2020). Depending on the complexity and goals of
106 the research question, these methods can search for networks with a constrained number of
107 reticulation events using quartet-based maximum pseudolikelihood (Solis-Lemus and Ané 2016;
108 Wen et al. 2018) or a full-likelihood Bayesian model where the number of reticulations is a
109 parameter (Wen et al. 2018; Zhang et al. 2018). Also, it is possible to estimate model
110 parameters (e.g., divergence times, population sizes, and the fraction of introgressed genes) on
111 a fixed species network using a full-likelihood Bayesian model that allows efficient computation
112 with large numbers of loci (Flouri et al. 2020). We refer to these network models from here on as
113 the multispecies coalescent with introgression (MSci), consistent with Flouri et al. (2020),
114 although other names have been used, such as the network multispecies coalescent (NMSC;
115 Zhu and Degnan 2017) and multispecies network coalescent (MSNC; Wen et al. 2016). We

116 emphasize that using network approaches to investigate polyploid complexes is not novel (e.g.,
117 Huber et al. 2006; Lott et al. 2009; Jones et al. 2013; Crowl et al. 2017), but the difficulty of
118 collecting appropriate genomic data from polyploids for such analyses has limited their use.
119 To address the issues outlined above, we have developed a pipeline, PATÉ (Phased
120 Alleles from Target Enrichment data), that can phase genotyping data for individuals of a known
121 ploidy without the need for sampling their parental lineages. PATÉ was designed with scalability
122 and population-level sampling in mind for target enrichment projects where deep coverage from
123 paired-end Illumina data allows calling of high-quality variants. In this study, we first use
124 simulations to explore the ability of network approaches to reconstruct the history of
125 allopolyploidy in the presence of ILS, and whether phasing the data affects the accuracy of the
126 reconstruction. We show that using phased allelic data can improve network estimation and
127 divergence time estimation compared to using haplotype consensus sequences, but also
128 highlight scenarios where phasing may not be necessary or beneficial. For an empirical
129 example of how PATÉ output can be used in studies of reticulate evolution, we compared
130 phased (either all alleles or only one) and unphased (either haplotype consensus or IUPAC
131 ambiguity codes) data to infer the evolutionary history of the North American *Dryopteris* fern
132 complex, a model system for reticulate polyploid evolution (Sessa et al. 2012a; Sessa et al.
133 2012b), using new targeted enrichment data. The system includes four diploid species, as well
134 as one extinct diploid, that have formed five allopolyploids in which there is high confidence in
135 the parent-progeny relationships (Fig. 1), although numerous sterile allopolyploid species have
136 also been reported within the complex (Montgomery and Paulton 1981). The allopolyploid
137 species have relatively ancient origins, with the best estimates placing hybridization events
138 between six and 13 Ma (Sessa et al. 2012b). PATÉ is capable of recovering phased haploid
139 sequences almost always the length of the original haplotype consensus assembly from
140 polyploid individuals, although the switch errors among these haplotype sequences is unknown.
141 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.

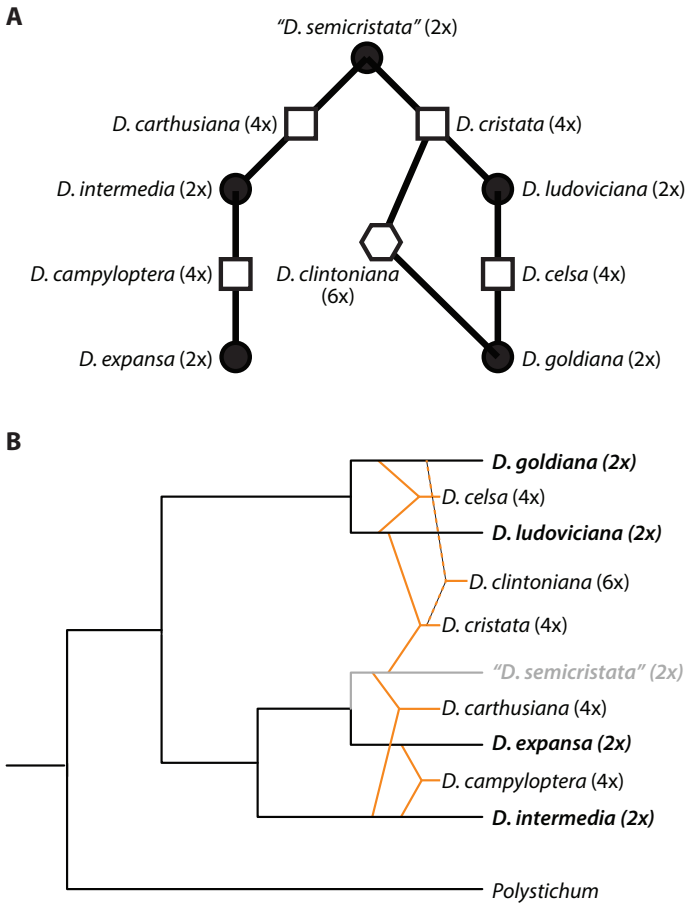

Figure 1 — Hypothesized Relationships among North American *Dryopteris*. Synthesis of results from Sessa et al. 2012a and Sessa et al. 2012b. A) Links between shapes show the putative parents and their allopolyploid derivatives. Black circles are diploids, squares are tetraploids, and the hexagon is 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).

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 (Rannala and Yang 2003) with BPP v4.4.0 (Flouri et al. 2020) under a network with five tips, such that one taxon was a hybrid of the ancestors of two other tip taxa (Fig. 2). All simulations used the HKY85 (Hasegawa et al 1985) model of sequence evolution with a transition-transversion rate ratio of 3 and

gamma-distributed among-site rate heterogeneity with a shape parameter of 0.6. Equilibrium frequencies were 0.3, 0.2, 0.3, and 0.2 for nucleotides A, C, G, and T. These parameters for the sequence evolution model were based on the observed means of parameter distributions across gene trees for our empirical *Dryopteris* data (see Materials and Methods: *Analyses of a Species Complex with Allopolyploidy*). 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 the ancestors of *B* and *C*. The mutation-scaled effective population size (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 hybridization age of 10 Ma (Fig. 2). We also simulated data such that θ was 0.001 (hybridization at 1 Ma) or 0.1 (hybridization at 100 Ma). Investigating such extreme values can generate general expectations for downstream estimation under a wide range of parameters, but it also has some biological salience as there is increasing evidence for reticulate evolution early in the diversification of plant orders and families (Sun et al. 2015; Stull et al. 2020; Liu et al. 2022).

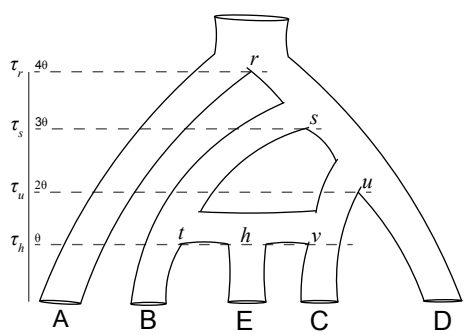

Figure 2 — Species network used for simulation. The divergence times in expected substitutions per site are given for each node, and *h* is the hybrid node where two alleles enter from both *t* and *v*. E is an allotetraploid while other species are diploid. Absolute divergence times were changed by 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 τ_u fixed to their initial values.

Because the distance between speciation nodes is 0.01 substitutions per site and $\theta = 0.01$ (Fig. 2), there are two coalescent units between nodes, $T = \frac{\tau}{\left(\frac{\theta}{2}\right)}$ (Yang 2006), which yields a low probability (0.09) of gene tree discordance due to ILS (Hudson 1983). We incorporated more ILS into the simulation by keeping the node heights fixed but increasing θ to 0.02, 0.03,

176 and 0.04. This increased the probability of a non-speciodentric split among all nodes to 0.25,
177 0.34, and 0.4, respectively. With three different starting values of θ and τ, this gave a total of 12
178 simulation conditions. While our simulations do not explore the limits of gene tree discordance,
179 they allow us to learn about some general features of increasing ILS on network inference with
180 different data types. We simulated 1000 gene trees and their sequences, 500 base pairs (bp) in
181 length, under the MSC 100 times for each of the 12 conditions. We also explored the effects of
182 sampling fewer genes on downstream analyses. For each replicate of 1000 gene trees and their
183 sequences, we randomly sampled 400, 40, and four genes without replacement.
184 All simulations sampled haploid data. Two haploid sequences were sampled for each
185 diploid species, and four haploid sequences were sampled for the allopolyploid species *E,*
186 where two sequences came from each parental lineage. We then investigated unphased data in
187 three ways. First, to generate unphased genotype sequences, the simulated haploid sequences
188 for each species were collapsed into a single sequence in which heterozygous sites were
189 represented by IUPAC ambiguity codes (genotype). The allopolyploid species was not restricted
190 to only biallelic sites. Second, we generated haploid consensus sequences where for each
191 variable site only one base was randomly chosen and retained (consensus). This could
192 represent a case in which read coverage across a locus is highly uneven such that a haploid
193 sequence is actually a chimera of two or more alleles. Finally, we simply picked one phased
194 haploid sequence, which is possible when one parental haplotype has a majority of reads for a
195 locus (pick one). This scenario where only one parent's sequence would be recovered in the
196 offspring could be anticipated in real data due to subgenome dominance (e.g., Buggs et al.
197 2014; Bird et al. 2018). In practice, we expect most *de novo* assemblers to generate output in
198 between the haploid consensus data and pick one data.
199
200 *Inferring species networks with phased and unphased simulated data* — We estimated
201 species networks with the SNaQ function (Solís-Lemus and Ané 2016) from PhyloNetworks

202 v0.12.0 (Solís-Lemus et al. 2017) using Julia v.1.4.1 (Bezanson et al. 2015) from either the true
203 gene trees used to simulate the data or gene trees estimated from the phased, pick one,
204 genotype, and consensus sequence data. For estimated trees, we used IQTREE v1.6.10
205 (Nguyen et al. 2015) with the same substitution model used to simulate the data. Each SNaQ
206 analysis used the species tree (A,((B,E),(C,D))) as the starting tree and allowed zero, one, or
207 two reticulation events. Each analysis included ten independent optimizations of the
208 pseudolikelihood score. We considered increasing numbers of reticulations if there was a
209 pseudolikelihood score improvement of two or greater compared to the previous model with one
210 fewer allowed hybrid edges. Note that pseudolikelihoods violate standard likelihood-based
211 model selection criteria such as likelihood ratio tests or using AIC weights, and slope heuristics
212 (Baudry et al. 2012) are often used instead. The choice of a cutoff here was necessary to
213 streamline processing across thousands of simulations. We compared the estimated networks
214 with one reticulation to the true network with the *hardwiredClusterDistance* function (Huson et
215 al. 2010) in PhyloNetworks. This allowed us to score the number of replicates that 1) recovered
216 the correct number of reticulations and 2) matched the true network when the number of
217 reticulations was set to one. We estimated networks for samples of four, 40, 400, and 1000
218 gene trees for each of the 100 replicates from the 12 simulation conditions.
219
220 *Effect of phasing on divergence time estimation* — We also used our simulated phased
221 and unphased data to estimate divergence times under the MSci model (Flouri et al. 2020)
222 using BPP v4.4.0. Here, we estimate divergence times (τ), population sizes (θ), and the
223 introgression probability, or proportion of immigrant loci in the receiving population, (φ) on the
224 correct fixed species network. φ was originally represented by γ (Meng and Kubatko 2009) and
225 referred to as an inheritance probability (Yu et al. 2011), but we use φ and the term
226 introgression probability to be consistent with BPP (Flouri et al. 2020). Our MSci analyses used

227 diffuse priors on τ_r and θ, with a mean on their simulated values and $\varphi \sim \beta(1,1)$. Phased, pick
228 one, and consensus sequences were treated as haploids while genotype sequences were
229 treated as unphased diploids and used for analytical integration over phases (Gronau et al.
230 2011) implemented in BPP. Although this is not correct for the tetraploid, it is arguably more
231 appropriate than treating all of the genotype sequences as haploids with ambiguous states.
232 Each Markov chain Monte Carlo (MCMC) analysis collected 10,000 posterior samples, saving
233 every 100 generations, while discarding the first 100,000 generations (i.e., 10% of the total run)
234 as burnin. Scripts for simulation and subsequent analyses of simulated data are available in
235 Dryad (https://doi.org/10.5061/dryad.5qfttdz53).
236

237 ***A Phasing Pipeline for Polyploids***

238 *Target enrichment data* — The success of empirical studies in which phasing was
239 informative about hybridization or introgression events (e.g., Kamneva et al. 2017; Oberprieler
240 et al. 2017; Eriksson et al. 2018) motivated us to use phased data to infer reticulate evolutionary
241 histories of polyploids. We were aware of few instances of phasing genomic or phylogenomic
242 data in polyploids, except in cases where chromosome-level whole-genome assemblies have
243 characterized subgenomes in allopolyploid crops (Yang et al. 2017; Colle et al. 2019) or tree-
244 based methods that are dependent on the sampling of parental lineages (Freyman et al. 2020;
245 Nauheimer et al. 2020). We designed PATÉ for target enrichment data because of the
246 availability of such data for many taxa, but it is applicable to other types of data with paired-end
247 Illumina reads.

248 The end product of many *de novo* target enrichment assembly pipelines (such as
249 HybPiper; Johnson et al. 2016) is a single consensus sequence for each locus for each
250 individual. Allelic variation may be represented by ambiguous nucleotide codes within the single
251 consensus sequence or lost when the pipeline outputs the haplotype consensus sequence
252 where the majority vote from a collection of reads is used. We use these existing *de novo*

253 assembly pipelines as a starting point to provide the reference sequence for each locus for each
254 individual and leverage a recent phasing algorithm with high-quality variant calls to recover
255 phased haplotype sequences for taxa with known ploidy levels. Ploidy levels are well
256 established and relationships have previously been characterized for individuals in our
257 *Dryopteris* analyses (Sessa 2012b).
258
259 *Phasing alleles within loci* — PATÉ (Supplementary Fig. S1) starts with assembled
260 target enrichment loci, such as the supercontig files output from HybPiper (Johnson et al. 2016),
261 that contain a single haplotype consensus sequence from each individual per locus. Reads for
262 each individual are then realigned to their consensus locus using BWA v0.7.17 (Li and Durbin
263 2009). PCR duplicates are flagged with MarkDuplicates in Picard v2.9.2
264 (http://broadinstitute.github.io/picard), and variant calls are computed with HaplotypeCaller in
265 GATK v.4.1.4 (McKenna et al. 2010). We applied the following hard filters with VariantFiltration
266 in GATK: (1) the variant quality divided by the depth of the alternate allele (QD) should be two or
267 greater; (2) a Fisher Exact test statistic based on a contingency table of read direction and
268 whether that read has the reference or alternate allele (FS) should be 60 or less; (3) the root
269 mean square of mapping quality scores (MQ) should be 40 or greater; (4) an approximation of
270 the Rank Sum Test for variant position in a read should be negative eight or greater to avoid
271 variants consistently towards the end of reads; (5) the fraction of reads with the alternate allele
272 (AF) should be between 5% and 95%, and the total read depth of the site (DP) should be at
273 least ten. The filter expressions are as follows: (1) QD < 2.0, (2) FS > 60.0, (3) MQ < 40.0, (4)
274 ReadPosRankSum < −8.0, (5) AF < 0.05, AF > 0.95, (6) DP < 10. These loosely follow
275 community recommendations on filters for germline variant discovery (DePristo et al. 2011).
276 Notably, we do not perform quality score recalibration or filter on the mapping quality rank sum,
277 as we anticipate allopolyploids could have a lower mapping quality associated with an alternate
278 allele due to sequence divergence or structural variation among homeologous chromosomes.

279 We also consider a very narrow window for filtering on allele frequency. Because increasing
280 ploidy levels will generate smaller anticipated ratios of alternate to reference alleles, coupled
281 with sequencing error and read stochasticity, we only aim to remove the most extreme cases.
282 For example, if almost all reads support the alternate allele at a site, it is difficult to diagnose if
283 the error lies in the consensus assembly or the read alignment. In these cases, only the
284 reference site is retained, and the variant does not pass the allele frequency filter. However, the
285 allele frequency filter could be removed if investigators are focused on organisms with higher
286 ploidy levels. Additional options for pipeline configuration are available on GitHub
287 (https://github.com/gtiley/Phasing).

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

305 multiple alleles per species for estimating θ. Assigning alleles to parental subgenomes should
306 be possible with PATÉ outputs where both parental lineages are sampled and there is sufficient
307 divergence for gene tree estimation.
308
309 ***Analyses of a Species Complex with Allopolyploidy***
310 *The North American wood fern complex (Dryopteris)* — We tested PATÉ using new
311 target enrichment data from nine North American *Dryopteris* species, including both
312 allotetraploid and allohexaploid taxa, with well-studied reticulate relationships (Sessa et al.
313 2012a, b), as well as two outgroup taxa from the sister genus *Polystichum*. All putative parental
314 lineages are represented in our dataset, with the exception of a hypothesized extinct lineage (*D.*
315 *semicristata*; Sessa et al. 2012b). We sampled two or three individuals for each *Dryopteris*
316 taxon (Table S1). The target enrichment data were generated from the GoFlag 408 flagellate
317 land plant probe set (Breinholt et al. 2021) at RAPID Genomics (Gainesville, FL). The target
318 regions for this probe set are 408 exons found in 229 single or low-copy nuclear genes. We
319 generated haplotype consensus assemblies for each with HybPiper (Johnson et al. 2016). The
320 resulting supercontig sequences, which includes the exon region targeted by probes for
321 enrichment and flanking intron regions, became our reference sequences for genotyping and
322 phasing with PATÉ. We aligned both phased and haplotype consensus (i.e., the reference
323 supercontig) sequences with MUSCLE v.3.8.31 with default settings (Edgar 2004). For
324 comparison with simulation results, we generated a “pick one” dataset where only one phased
325 haplotype per locus was randomly selected from each individual. A “genotype” dataset was
326 generated by collapsing allelic variation into IUPAC codes for each individual as a locus.
327
328 *Three species tests* — We first explored the value of phasing data when estimating
329 reticulate relationships among three species, where we sampled two extant parental lineages

330 and their putative hybrid. We examined the three such extant relationships in the North
331 American *Dryopteris* complex:

332 1) Two diploid parental lineages (*D. ludoviciana* and *D. goldiana*) and their putative
333 allotetraploid descendent (*D. celsa*).
334 2) Two diploid parental lineages (*D. expansa* and *D. intermedia*) and their putative
335 allotetraploid descendent (*D. campyloptera*).
336 3) One tetraploid parent (*D. cristata*), one diploid parent (*D. goldiana*), and their putative
337 allohexaploid descendent (*D. clintoniana*).

338 These three scenarios represent different ages of parental divergence, with a common ancestor
339 near 10 Ma for *D. ludoviciana* and *D. goldiana*, 15 Ma for *D. expansa* and *D. intermedia*, and 30
340 Ma for *D. cristata* and *D. goldiana* (Sessa et al. 2012b). We used both a full-likelihood Bayesian
341 approach and a topology-based pseudolikelihood approach to estimate the correct species
342 relationships from the different data types. First, using BPP v.4.4.0 (Flouri et al. 2020), we
343 estimated log-marginal likelihoods (ln *mL*) with stepping-stone sampling (Xie et al. 2011) for the
344 three possible rooted 3-taxon trees and network models that imply differences in the timing and
345 direction of gene flow among species (Supplementary Fig. S2). Each ln *mL* estimate used 24
346 steps, and each step had a posterior sample of 10,000, saving every 30 generations after a
347 30,000 generation burnin (10% of the total run). The ln *mL* values were then used to calculate
348 the six model probabilities following equation 1 using the bppr v.0.6.3 R package
349 (https://github.com/dosreislab/bppr). A percentile-based 95% confidence interval was generated
350 for each model probability by bootstrapping posterior distributions 1,000 times to investigate
351 uncertainty in model selection.

352 $$P(model) = \frac{exp(ln\ mL_{model}\ - ln\ mL_{max}\ )}{\Sigma_i[exp(ln\ mL_i\ - ln\ mL_{max}\ )]} \qquad Equation\ 1$$

353 We performed four independent runs for each model for each data type prior to marginal
354 likelihood estimation to evaluate mixing and diagnose potential convergence issues. The

355 combined posteriors from the models that represent *a priori* hypotheses of relationships
356 between allopolyploids and their parental lineages were used estimate all τ and θ for
357 comparisons across data types. Notably, BPP can analytically integrate over possible phases of
358 sequence data when given IUPAC codes representing heterozygous sites (Gronea et al. 2011),
359 which is as accurate as perfectly phased data for diploids (Huang et al. 2022). In comparison,
360 standard gene tree estimation under the Felsenstein likelihood (Felsenstein 1981) treat IUPAC
361 codes as uncertainty and integrate over the possible tip states at sites independently.
362
363 For our topology-based method, we used the SNaQ function (Solís-Lemus and Ané 2016) from
364 PhyloNetworks v0.12.0 (Solís-Lemus et al. 2017) to test the presence and placement of gene
365 flow between the three species, while also including the two *Polystichum* outgroups. Gene trees
366 were estimated for each data type across the three cases with IQ-TREE v1.6.10 (Nguyen et al.
367 2015), using ModelFinder (Kalyaanamoorthy et al. 2017) for model selection on the
368 unpartitioned alignments. The starting tree was obtained with ASTRAL III v5.6.3 (Zhang et al.
369 2018) from the collection of ML gene trees. We tested the presence of zero, one, or two
370 reticulations based on a two pseudolikelihood point cutoff for accepting an additional
371 reticulation. Each analysis used ten independent optimizations. For SNaQ and ASTRAL
372 analyses, all haplotypes were mapped to their respective species, as opposed to individuals.
373
374 *Nine species tests* — We investigated the differences in networks from SNaQ across
375 data types for the nine-species complex. This complex involves multiple reticulation events on
376 an edge, and thus the true network is not identifiable with topology-based methods (Solis-
377 Lemus and Ané 2016). Although full-likelihood methods should not suffer from the some of the
378 identifiability issues (Yu and Nakhleh 2015), they can impose a heavy computational burden
379 (Wen et al. 2018; Zhang et al. 2018). Despite the inherent identifiability issues, it is worthwhile to
380 examine the performance of scalable topology-based methods in such contexts to see if they

will identify some of the correct reticulate relationships and if they also recover erroneous ones. Thus, we performed analyses of the nine-species complex and two outgroups across the four data types with SNaQ as described above, but we allowed up to six reticulation events. We used ASTRAL III v5.6.3 (Zhang 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). 100 standard bootstrap replicates were performed for each gene tree, which were used to evaluate gene tree incongruence and informativeness with respect to the ASTRAL species tree from phased data 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 discordant 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.

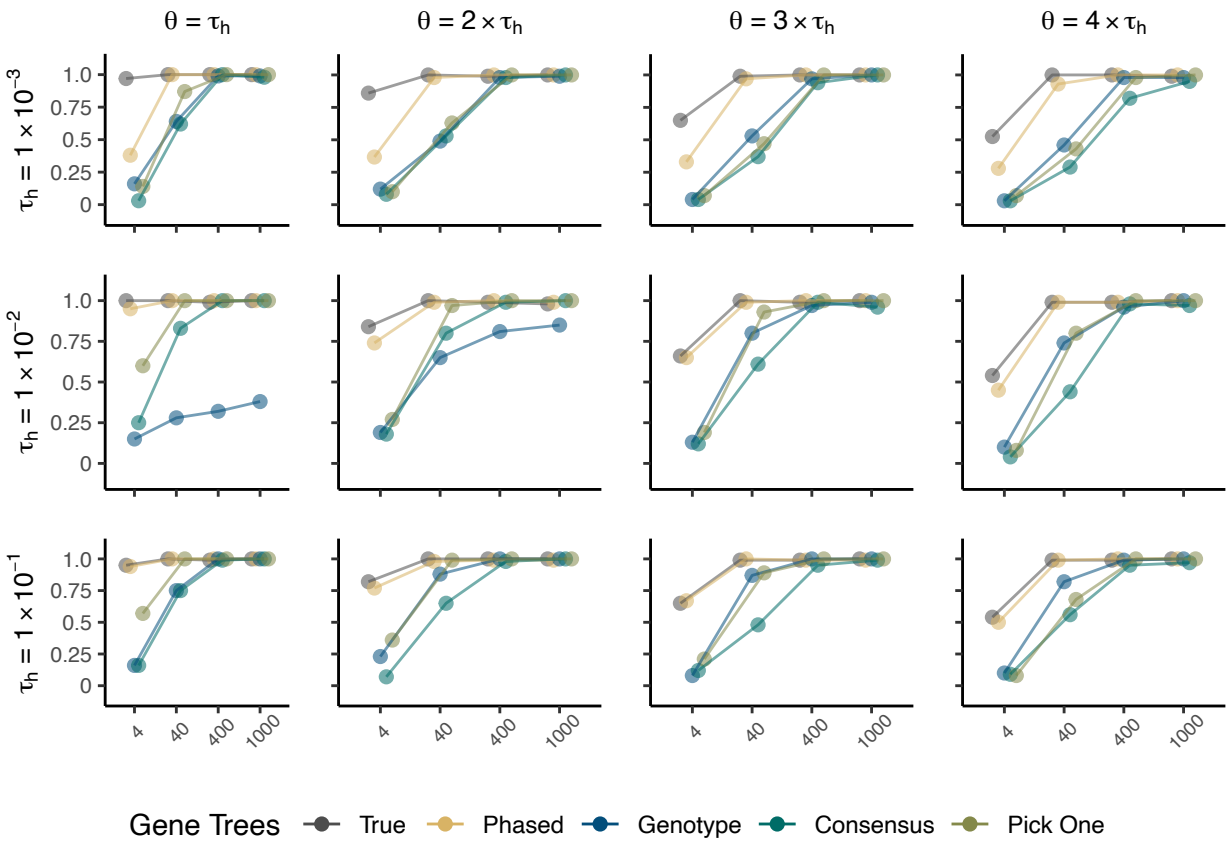

Figure 3 — Proportion of simulation replicates that correctly identify the allopolyploid lineage using the SNaQ function of PhyloNetworks. The x-axes are the number of loci sampled for each simulation. The y-axes are the proportions of correct networks. Results are based on networks estimated with a 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.

Divergence times — Using phased data, whether using all alleles or only one, typically resulted in accurate estimation of the hybridization age (Fig. 4). When divergence was low, there was a small amount of estimation error, such that the phased data underestimated and the pick one data overestimated the hybridization age, but the pattern suggests estimates could converge to the true values with more loci. Interestingly, convergence towards the true value appeared more efficient with higher ILS (Fig. 4). The phased and pick one data could otherwise accurately estimate the hybridization time with 40 or fewer loci across all other divergence and ILS scenarios. 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

429 estimate, and the simulated value was not within the highest posterior density (HPD) interval.
430 There was always a small degree of estimation error with the genotype data despite using
431 analytical phasing, and the consensus data overestimated the hybridization age between 1.5x
432 and 2x when using 400 loci across all simulation scenarios. These biases from genotype or
433 consensus data were also propagated for the parent nodes of the hybridization event across all
434 levels of ILS for $\tau_h = 0.001$ (Supplementary Figs. S4-S7), $\tau_h = 0.01$ (Supplementary Figs. S8-
435 S11), and $\tau_h = 0.1$ (Supplementary Figs. S12-S15). Nodes from speciation events could always
436 be estimated accurately with high precision by the phased and pick one data types provided
437 enough loci were sampled. The phased data was more efficient, always converging to the true
438 value by 40 loci, while the pick one data required 400 to recover τ_u when the sequence
439 divergence was low at $\tau_h = 0.001$. There was a small degree of estimation error among
440 speciation nodes for the genotype and consensus data, and the patterns suggested they might
441 not converge to the true values with more loci. However, speciation times were overestimated
442 by genotype and consensus data for $\tau_h = 0.001$ (Supplementary Figs. S4-S7) and $\tau_h = 0.01$
443 (Supplementary Figs. S8-S11), but underestimated by the genotype data and overestimated by
444 the consensus data for $\tau_h = 0.1$ (Supplementary Figs. S12-S15). Across all simulations, there
445 was little effect of increasing ILS aside from increasing HPD intervals, but the uncertainty was
446 negligible when 400 loci were sampled.
447

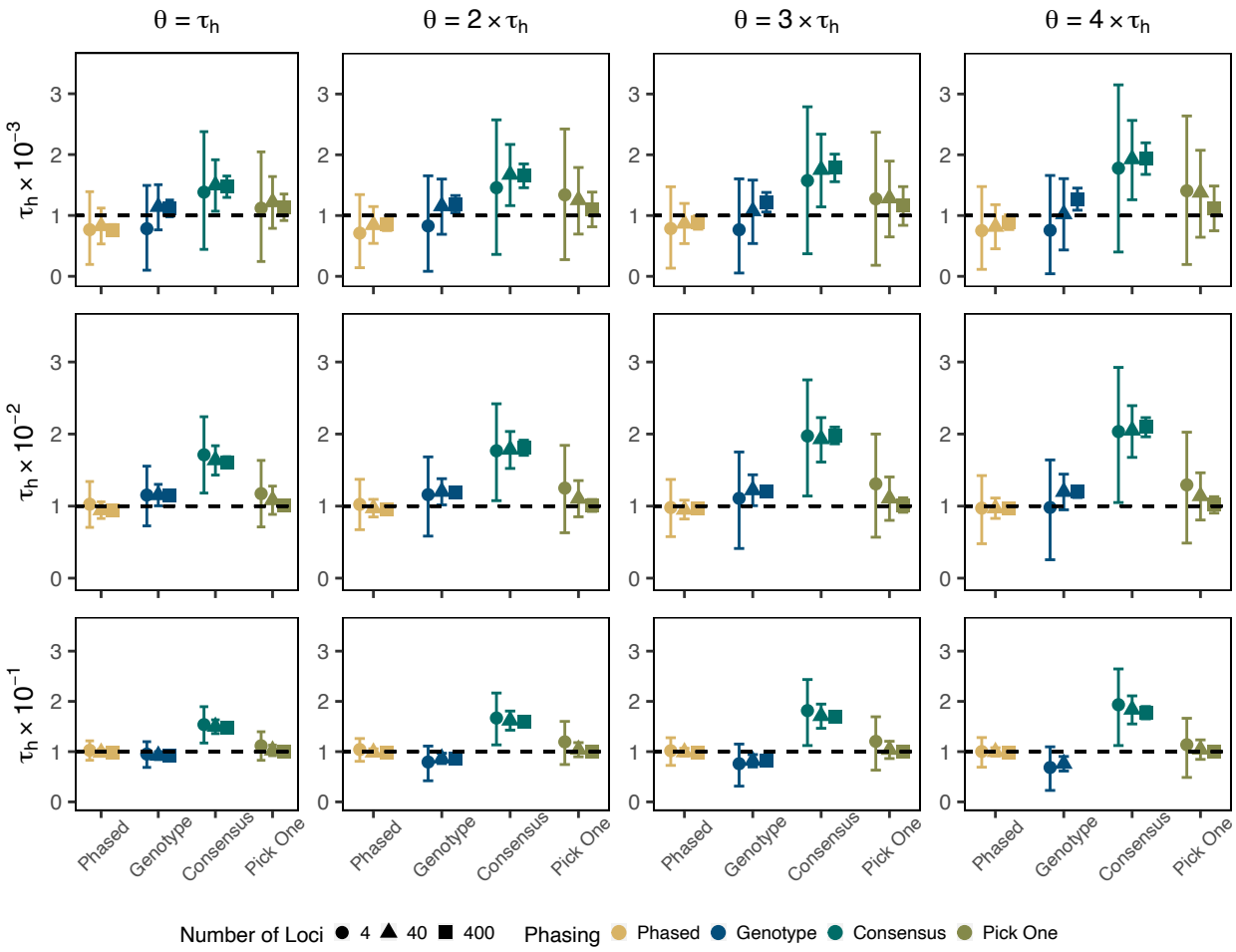

Figure 4 — Estimating the Timing of Hybridization with the MSci Model. Divergence times, measured as the expected number of substitutions per site, are for node *h* in Figure 2. The dashed line represents the true simulated values. Points are posterior means, and error bars are 95% HPD intervals averaged 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.

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 per locus per individual (Supplementary Table S2). The ploidy level appears to be strongly associated with the number of phased loci. Among diploids, only 32% of loci were phased; the other 68% 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

460 single haplotype block, as opposed to being split into two or more blocks because not enough
461 reads were available to physically link variants. For polyploids, the number of unique phased
462 haplotype sequences most frequently matched the ploidy level, except in the case of a single *D.*
463 *campyloptera* individual (B087-D08), which also had relatively few recovered loci. Phasing data
464 only extended sequence alignment length by about two base pairs relative to the haplotype
465 consensus sequences (Supplementary Table S3). However, phased data more than tripled the
466 number of parsimony informative sites per alignment, similar to the increase in number of tips
467 (Supplementary Table S3).
468
469 *Placing a single reticulation event* — For our three-species full-likelihood analyses with
470 the MSci model, all data types decisively supported the putative reticulation hypothesis in the
471 cases of the allotetraploids *D. celsa* or *D. campyloptera* (Table 1). However, only the genotyped
472 data, coupled with the analytical phasing technique of BPP, chose the putatively correct model
473 for the allohexaploid *D. clintoniana* (Table 1). The phased and pick one data types selected the
474 network with the diploid *D. goldiana* as the hybrid descendent of the allohexaploid *D. clintoniana*
475 and allotetraploid *D. cristata*, while the best model for consensus data was a binary tree with *D.*
476 *clintoniana* sister to *D. cristata* (Table 1). However, model probability confidence intervals
477 suggest a great deal of uncertainty underlies model selection results for the allohexaploid case,
478 with the upper end of the interval for the putatively correct model at 0.603, 0.582, and 0.996 for
479 the phased, consensus, and pick one data, respectively (Table 1). Model parameters typically
480 converged across all analyses with the exception of the phased data for the *D. clintoniana* triplet
481 (Supplementary Fig. S16), which was due to one MCMC run despite convergence of the lnL
482 values.
483 The simulation experiments established some expectations that the phased and pick
484 one data would provide unbiased estimates of divergence times for nodes involved in the
485 hybridization events, while the genotyped and consensus data would provide under- and

Table 1 — Model Probabilities for Possible Topological Hypotheses Using the MSci model.

Hybrid	Model	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)**

Note: Topologies of the 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.

overestimates, respectively. However, phased data generally produced younger mean τ estimates for the *D. celsa* triplet with reduced credible intervals relative to other data types, but much older mean τ estimates with more uncertainty for the *D. campyloptera* and *D. clintoniana* triplets (Fig. 5; Supplementary Table S4). Similarly, there was inconsistency in θ estimates observed among data types across triplets, with notably larger credible intervals for phased data in the *D. celsa* and *D. clintoniana* triplets. Estimates of φ were sensitive to data type across triplets, such that the mean φ suggested the major contribution to the focal allopolyploid from phased data was *D. goldiana* into *D. celsa*, *D. intermedia* into *D. campyloptera*, and *D. goldiana* into *D. clintoniana*, with other data types sometimes disagreeing (Fig. 5). Credible intervals around φ always spanned $\varphi = 0.5$, but estimates for genotype, consensus, and phased data suggested some biased gene retention for the *D. celsa* and *D. clintoniana* triplets (Supplementary Table S4).

When performing a network search based on gene tree distributions with SNaQ, all data types recovered the putatively correct relationship among *D. celsa* and its diploid parental lineages (Fig. 6a). For *D. campyloptera*, all data types correctly detected one hybrid edge (Supplementary Table S5), but they got the direction of gene flow wrong (Fig. 6b). All data types also detected one hybrid edge for the allohexaploid *D. clintoniana* (Supplementary Table S5), but there was disagreement over the underlying network. Both the phased and consensus data invoked a ghost lineage sister to the three sampled species, such that gene flow was from the ghost into the allotetraploid *D. cristata*. While not unreasonable given the putatively extinct *D. semiscristata* and missing *D. ludoviciana* from the sample, the relationship between *D. clintoniana* and its parental lineages was undetected (Fig. 6c). The genotype and pick one data recovered the correct network, but reversed the direction of gene flow from allohexaploid *D. clintoniana* into allotetraploid *D. cristata* (Fig 6c).

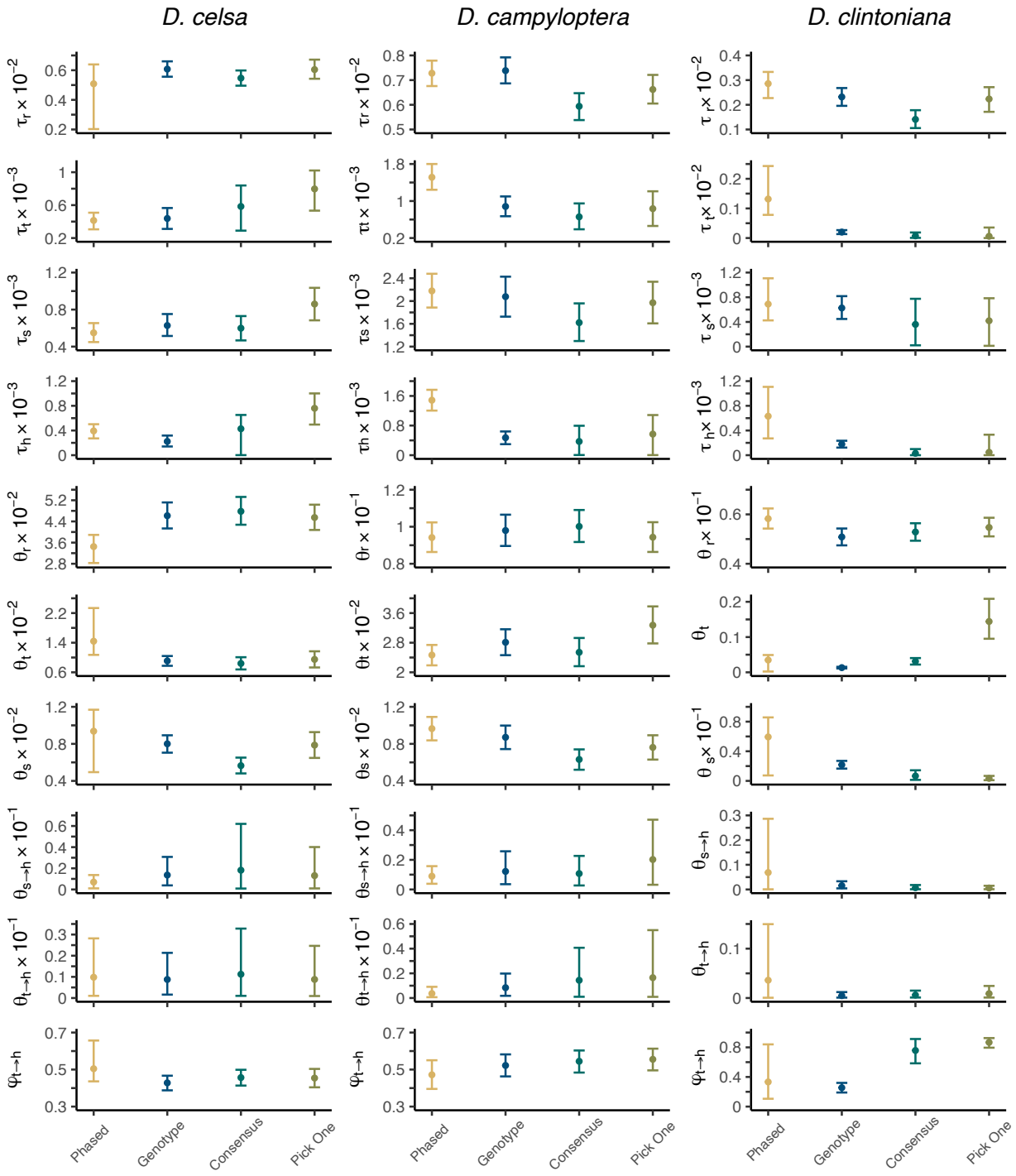

Figure 5 — *Dryopteris* Divergence Times, Population Sizes, and Introgression Probabilities under the MSci Model. Points are posterior means and error bars show the 95% HPD intervals. Divergence times are measured as the expected number of substitutions per site.

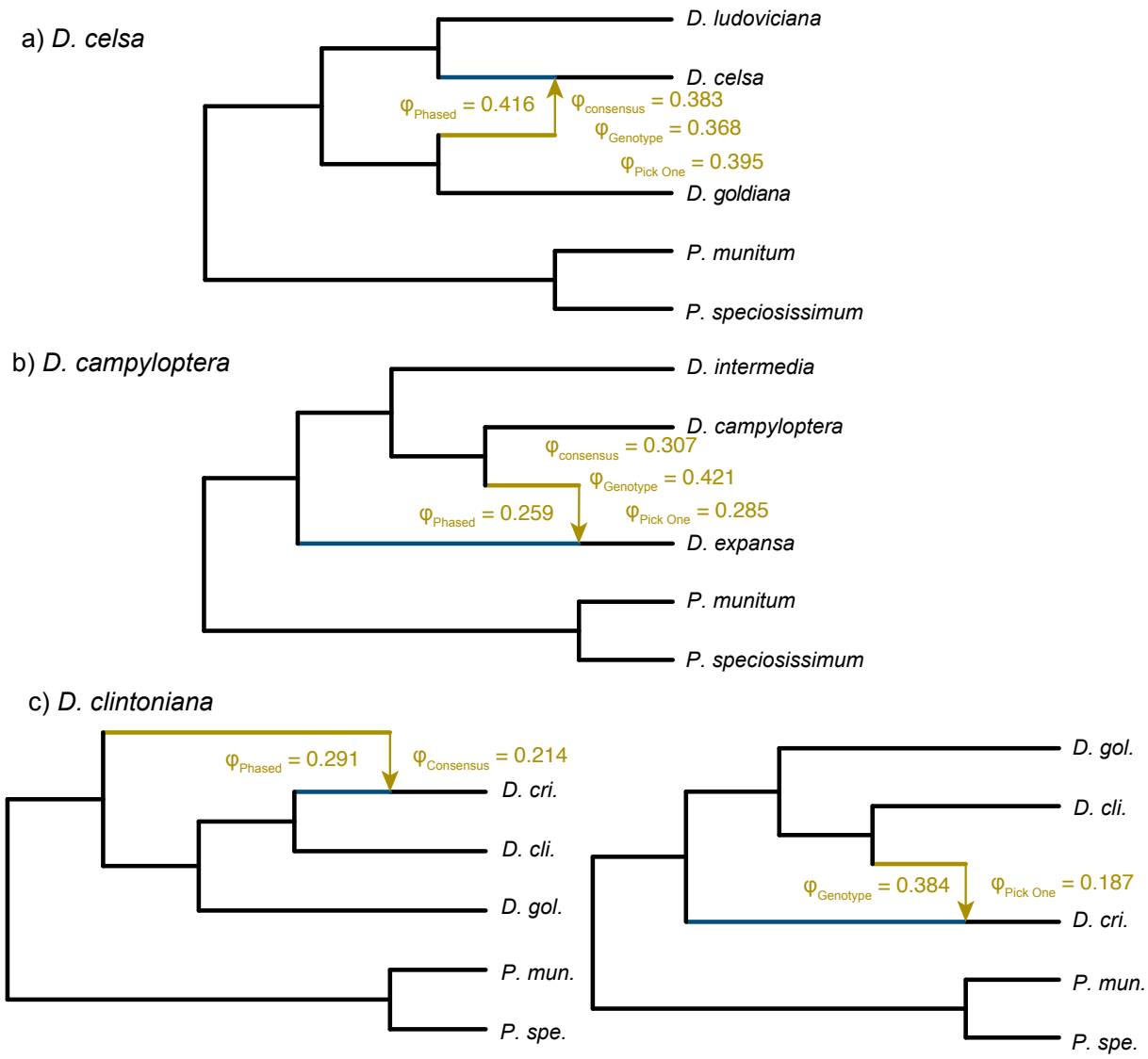

Figure 6 — Network search results for three-taxon *Dryopteris* cases with SNaQ. The major topology edge is indicated in blue and the minor edge (direction of introgression) is shown in gold. The introgression probability (φ) is shown for each data type. In cases where a) *D. celsa* or b) *D. campyloptera*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

525 *Inferring relationships among a complex with multiple reticulation events* — The nine
526 *Dryopteris* species SNaQ network analysis with phased data identified two out of five
527 hypothesized reticulation events correctly along with one incorrect reticulation edge (Fig. 7a;
528 Supplementary Table S6). The analysis correctly identified the diploids *D. ludoviciana* and *D.*
529 *goldiana* as parents to the allotetraploid *D. celsa*, as well as the diploids *D. expansa* and *D.*
530 *intermedia* as parents to the allotetraploid *D. campyloptera*. A hybrid edge from the common
531 ancestor of *D. celsa*, *D. ludoviciana*, and *D. goldiana* to the common ancestor of *Polystichum*
532 was also recovered with an introgression probability of 0.335, although with low bootstrap
533 support. The parents of *D. celsa* were also detected by the other data types, but all had some
534 degree of error. The genotype data found a reticulation edge between the allotetraploid *D.*
535 *carthusiana* and diploid parent *D. intermedia*, but reversed the direction of gene flow (Fig. 7b). A
536 reticulation edge was also detected between *D. cristata* and the common ancestor of *D.*
537 *clintoniana*, *D. goldiana*, *D. celsa*, and *D. ludoviciana*, which is not completely unreasonable
538 given the allotetraploid *D. cristata* is believed to be a parent of *D. clintoniana*. The consensus
539 data only detected two reticulation edges (Supplementary Table S6), with the second edge
540 correctly showing gene flow from the diploid parent *D. intermedia* into the allotetraploid *D.*
541 *carthusiana*, provided the other parent *D. semicristata* is presumed extinct obscuring the
542 network interpretation (Fig. 7c). The pick one data, similar to the genotype data, inferred a
543 reticulation edge between *D. carthusiana* and *D. intermedia* with the wrong direction of gene
544 flow (Fig. 7d). The pick one data also detected a reticulation edge from *D. cristata* into the
545 common ancestor of *D. carthusiana D. intermedia*, *D. campyloptera*, and *D. expansa*, which
546 lacks a clear interpretation other than invoking missing lineages underlying complexity of the
547 backbone topology (Fig. 1). In the one case where all data types detected the correct
548 relationship regarding *D. celsa*, the introgression probability was approximately 0.5 with phased
549 data while between 0.37 and 0.39 with others (Fig. 7; Supplementary Fig. S17).
550

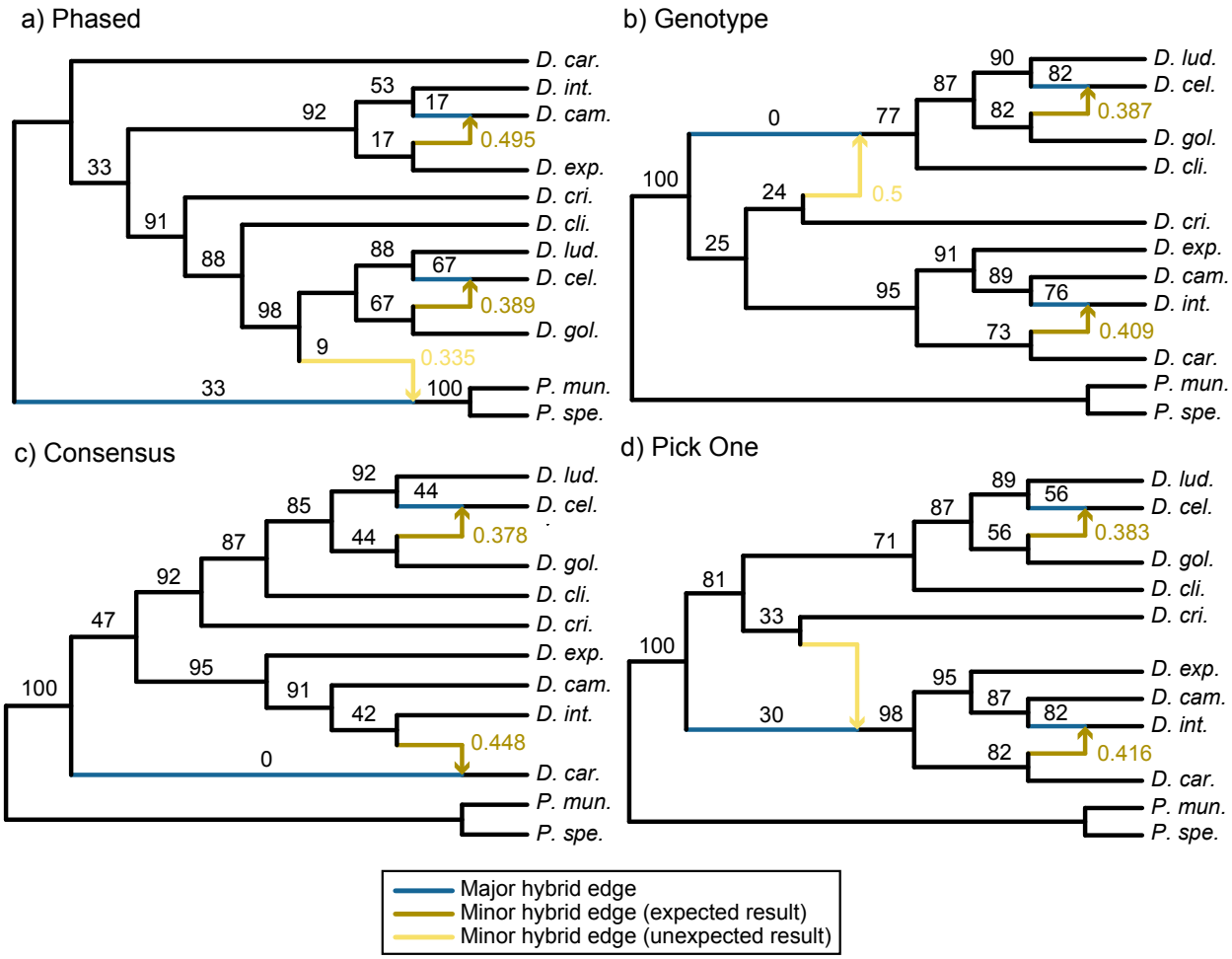

Figure 7 — Networks for all North American *Dryopteris* inferred from SNaQ. Colors indicate major and minor (reticulation) edges, and which minor edges were hypothesized or an unexpected result. Introgression probabilities for each minor edge are provided. Numbers on branches subtending nodes are bootstrap 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.

Differences in the network estimates among data types may be explained by their underlying spectra of quartet concordance factors (Supplementary Fig. S18). While spectra were overall similar among data types, some deviations from expectations under ILS alone were stronger in the phased data versus others (Supplementary Fig. S18). Analyses of gene tree incongruence among the consensus (Supplementary Fig. S19), genotype (Supplementary Fig. S20), and pick one (Supplementary Fig. S21) data showed similar levels of conflict among the

563 different data types with respect to a species tree estimated from the phased data (such
564 analyses with multiple alleles per species is not possible). The frequency of conflicting splits
565 with high bootstrap support was similar across datatypes for all nodes (Supplementary Fig.
566 S22), and low-support conflicts represented on average 56% of splits within *Dryopteris*.
567

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 under some conditions phasing polyploid target enrichment data can improve the accuracy of such network inferences as well as divergence time estimates for 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 computational phasing because they can yield high (i.e. greater than 100x) and even coverage within a locus. This was demonstrated in our *Dryopteris* analyses when almost all loci were recovered as single haplotype blocks (Supplementary Table S2). Because MSC methods treat loci independently and assume free recombination between loci, it is not necessary to assign individual loci to parental subgenomes for network analyses. However, the allele sequences output by PATÉ can also be used as input for other tools such as *Homologizer* (Freyman et al. 2023) or *HybPhaser* (Nauheimer et al. 2020), which attempt to phase across loci and assign parental subgenomes. Our methods also enable population genomic studies of polyploids (Tiley et al. 2024) 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. 2018a; Olave and Meyer 2020).

Promises and Limits 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). New strategies for explicitly addressing this challenge are emerging (Freyman et al. 2023; Nauheimer et al. 2020; Mendez-

Reneau et al. 2023), and PATÉ can be a useful tool by phasing variants for many individuals while leveraging genotype quality information. Our simulations demonstrated that phasing can improve estimates of reticulate evolutionary relationships using network methods. In simulations, phased data with all alleles represented in gene trees could more efficiently recover the placement and directionality of hybrid edges than the genotyped, haplotype consensus, or single haplotype sequences (Fig. 3). The advantages of using phased data for network estimation diminished in simulations as the number of loci increased (Fig. 3), and this was evident from our empirical analyses of *Dryopteris* triplets, which produced similar results across data types with 330 or 328 loci (Table 1; Fig. 6), except for the correct inference of allopolyploid and parent relationships with genotyped data in the case of *D. clintoniana* (Table 1).

Both our simulations (Fig. 4; Supplementary Figs. S4-S15) and previous investigations (Anderman et al. 2019) suggested phased data can be important for unbiased divergence time estimation, especially estimating the origin of a hybrid lineage (Fig. 4). Our empirical analyses also demonstrated how the timing of introgression τ_h and other parameters can be greatly affected by the data type used (Fig. 5). In our *Dryopteris* analyses, τ_h for *D. celsa* was at least in the same order-of-magnitude among data types, which had partially overlapping credible intervals, while phased τ_h estimate *for D. campyloptera* was nearly four times older than the other data types (Fig. 5; Supplementary Table S4). This was surprising given that simulations suggested haplotype consensus data should generate older estimates than phased data (Fig. 4). The *D. clintoniana* τ_h estimate from phased data was also much older than other data types, albeit with a wide 95% HPD interval. There were noticeable mixing problems in our *D. clintoniana* full-likelihood analyses, which makes it difficult to evaluate whether the uncertainty was due to extreme differences in θ among phased alleles within *D. clintoniana* and/or the computational difficulty of the Bayesian methods with increasing sequence data. In all cases, the differences between τ_h among data types is not consistent with results from the simulation

experiments, which suggested the consensus data should overestimate age compared to phased or pick one data, with genotyped data having slightly younger estimates than phased data. This highlights the difficulties of simulating data that capture the complexities of empirical data and makes deciding which estimate is more reliable difficult. For example, the uncertainty of τ_h with phased data for *D. celsa* and *D. campyloptera* triplets was reduced compared to the consensus and pick one estimates, but greater uncertainty was observed for τ and θ at other nodes (Fig. 5). A potential compromise from simulations and empirical analyses 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

645 Our analyses highlight the difficulty of estimating the evolutionary histories of reticulate
646 complexes, regardless of data type. The full-likelihood implementation of the MSci model
647 appears to be useful when the number of tips and plausible models for investigation are low, but
648 these methods may not be practical for generating hypotheses and exploring unknown
649 relationships for large numbers of taxa (e.g., Zhang et al. 2018). Quartet-based methods are
650 much faster and accurate when model assumptions are met (Solis-Lemus and Ané 2016) and
651 show similar accuracy to full-likelihood methods for estimating introgression probabilities (Flouri
652 et al. 2020). However, there are scenarios where the true networks are non-identifiable (Yu and
653 Nakhleh 2015; Solís-Lemus and Ané 2016). For example, when multiple introgression events
654 affect the same lineage, such that there are overlapping cycles in a network (Solís-Lemus and
655 Ané 2016), the expected quartet distribution under the MSci model becomes a poor fit for the
656 empirical data (Cai and Ané 2020), and the network estimated may be incorrect. These effects
657 were evident in our empirical analyses where the relationship between *D. campyloptera*, *D.*
658 *expansa*, and *D. intermedia* was missing in the nine-species analysis (Fig. 7), which we expect
659 is due to a reticulation edge present between *D. intermedia* and *D. carthusiana*. Phasing
660 sequence data adds information that can improve estimates (Supplementary Table S3), but
661 unsampled or extinct lineages, such as the hypothesized *D. semicristata,* can create significant
662 barriers to recovering the true evolutionary history of reticulate complexes, regardless of how
663 many loci or individuals are available. Another identifiability problem is getting the direction of
664 gene flow correct (Yu and Nakhleh 2015; Thawornwattana et al. 2023), and these scenarios
665 were evident in our quartet-based analyses of a triplet that reversed gene flow from the
666 allotetraploid *D. campyloptera* into the diploid *D. expansa* (Fig. 6d) or in the analyses of the full
667 complex that reversed gene flow from the allotetraploid *D. carthusiana* into the diploid *D.*
668 *intermedia* (Fig. 7b and 7d).
669
670 ***Insights into Dryopteris Evolution***

671 The North American *Dryopteris* complex has been well-characterized through the study
672 of multi-locus nuclear and chloroplast phylogenies, morphology, cytological observations,
673 chromatography, and isozyme analyses (reviewed in Sessa et al. 2012b). This makes it a useful
674 system for testing our phasing pipeline, and our analyses add nuance to our understanding of
675 some of the relationships among *Dryopteris* species. For example, our results with phased data
676 indicated *D. campyloptera* has received more loci from *D. intermedia* than *D. expansa* (Fig. 7a),
677 and *D. intermedia* is likely the maternal progenitor of *D. campyloptera* (Sessa et al. 2012b).
678 Following the allopolyploidy event, the *D. intermedia* subgenome may have been dominant,
679 providing a selective advantage for *D. campyloptera* in its distribution at the time (Bird et al.
680 2018). The phased nine-taxon analyses also supported the hypothesis of the unsampled diploid
681 lineage *D. semicristata*, based on the placement of *D. carthusiana* as sister to the rest of
682 *Dryopteris*, and the general sensitivity of *D. carthusiana* and *D. cristata* in the phylogenetic
683 backbone across data types (Fig. 7). Additional insights may be gained by a more complete
684 sampling of the genus, as the North American *Dryopteris* complex is not a clade (Sessa et al.
685 2012b), which should alleviate some of the identifiability problems. Analyses of the North
686 American *Dryopteris* complex alone highlight the need for caution when interpreting networks,
687 which are increasingly prominent plant evolutionary research.
688
689 ***Does Phasing Matter?***
690 Combining phased data with recent network methods can help confront a major
691 challenge of plant phylogenetics: resolving the complex histories of polyploids. While our
692 analyses showed haplotype consensus sequences can be adequate for resolving single
693 reticulation events where both parental lineages are sampled, phased sequences may provide
694 more information when only a few loci are available for cases where sequence divergence is not
695 too high, while more generally improving divergence time estimates (see also Huang et al.
696 2022). Analytical phasing approaches (Gronau et al. 2011; Flouri et al. 2020) provide a way

697 forward for full-likelihood methods while avoiding the potential for switch errors from read-based
698 phasing as done in PATÉ, but these will become difficult with many tips. This highlights the
699 need for continued development of models that explicitly consider polyploidization mechanisms
700 (Jones et al. 2013) while addressing scalability and limiting assumptions for contemporary
701 genomic data (Yan et al. 2022). Some reticulate complexes will be difficult to disentangle with
702 any data when confronted with statistical identifiability problems, such as networks with multiple
703 reticulations on a single edge (Solís-Lemus et al. 2016) when using quartet-based methods, and
704 our *Dryopteris* analyses demonstrate that caution is needed when interpreting biological
705 processes from phylogenetic networks in the absence of corroborating evidence. As haplotype-
706 phased genomes from long-read data become increasingly available, the discussion of phase
707 resolution will become trivial, but progress on models that leverage such data and improve
708 interpretation of processes, especially in polyploids that maintain distinct subgenomes, from
709 networks is still needed.
710

711 **DATA AVAILABILITY**

712 PATÉ is available through GitHub (https://github.com/gtiley/Phasing) and can be run on any
713 UNIX environment after installing basic genotyping software and H-PoPG. Simulated and
714 empirical data supporting findings and files for replicating some analyses are available from the
715 Dryad Digital Repository: https://doi.org/10.5061/dryad.5qfttdz53. Raw Fastq reads for
716 *Dryopteris* individuals are available through the NCBI SRA database and are associated with
717 BioProject PRJNA725004. Individual SRA Identifiers are available in Supplementary Table S1.
718

719 **SUPPLEMENTARY MATERIAL**

720 Data and supplementary material is available from the Dryad Digital Repository:
721 https://doi.org/10.5061/dryad.5qfttdz53
722

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. The manuscript was improved through helpful comments from the associate editor S.D. Smith and three anonymous reviewers.

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. 2018a. HyDe: a Python package for genome-scale hybridization detection. *Syst. Biol.*, 67:821-829.

Blischak PD, Kubatko LS, Wolfe AD. 2018b. SNP genotyping and parameter estimation in polyploids using low-coverage sequencing data. *Bioinformatics* 34:407-415.

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.

Crowl A.A., Manos P.S., McVay J.D., Lemmon A.R., Lemmon E.M., Hipp A.L. 2020. Uncovering the genomic signature of ancient introgression between white oak lineages (*Quercus*). *New Phytologist* 226: 1158–1170.

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.

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.

Schrinner S.D., Mari R.S., Ebler J., Rautianen M., Seillier L., Reimer J.J., Usadel B., Marschall T., Klau G.W. 2020. Haplotype threading: accurate polyploid phasing from long reads. *Genome Biology* 21:252.

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. 2024. Genetic variation in *Loudetia simplex* supports the presence of ancient grasslands in Madagascar. *Plants People Planet* 6:315-329.

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.

1078 **FIGURE CAPTIONS**

1079 **Figure 1 — Hypothesized Relationships among North American *Dryopteris*.** Synthesis of
1080 results from Sessa et al. 2012a and Sessa et al. 2012b. A) Links between shapes show the
1081 putative parents and their allopolyploid derivatives. Black circles are diploids, squares are
1082 tetraploids, and the hexagon is the one hexaploid species in the group. *Dryopteris semicristata*
1083 is presumed extinct. B) Placement of allopolyploids in the context of the backbone relationships
1084 among diploids. Tetraploids are indicated with solid orange lines and the hexaploid with dotted
1085 lines. The grey line denoting a sister relationship between *D. semicristata* and *D. expansa*
1086 reflects one possible placement for the extinct taxon based on previous analyses (Sessa et al.,
1087 2012b).

1088

1089 **Figure 2 — Species network used for simulation.** The divergence times in expected
1090 substitutions per site are given for each node, and *h* is the hybrid node where two alleles enter
1091 from both *t* and *v*. E is an allotetraploid while other species are diploid. Absolute divergence
1092 times were changed by using values of θ at 0.001, 0.01, and 0.1. ILS was increased by
1093 multiplying θ by a factor of 2, 3, or 4 while keeping all $\tau_u$fixed to their initial values.

1094

1095 **Figure 3 — Proportion of simulation replicates that correctly identify the allopolyploid**
1096 **lineage using the SNaQ function of PhyloNetworks.** The x-axes are the number of loci
1097 sampled for each simulation. The y-axes are the proportions of correct networks. Results are
1098 based on networks estimated with a single reticulation, even if that network was considered less
1099 optimal than networks with zero or two reticulations. We saved the true gene trees from the
1100 simulations, while we estimated gene trees with the phased, genotype, consensus, and pick one
1101 data types.

1102

1103 **Figure 4 — Estimating the Timing of Hybridization with the MSci Model.** Divergence times
1104 are for node *h* in Figure 2. Divergence times are measured as the expected number of
1105 substitutions per site. The dashed line represents the true simulated values. Points are posterior
1106 means, and error bars are 95% HPD intervals, averaged across 100 replicates. Results are not
1107 shown for the genotype data when $\tau_h = 0.1$ and $\theta = 4\tau_h$ because posterior means went towards
1108 infinity and numerically unstable values that cannot be displayed.
1109
1110 **Figure 5 — *Dryopteris* Divergence Times, Population Sizes, and Introgression**
1111 **Probabilities under the MSci Model.** Points are posterior means and error bars show the 95%
1112 HPD intervals. Divergence times are measured as the expected number of substitutions per
1113 site.
1114
1115 **Figure 6 — Network search results for three-taxon *Dryopteris* cases with SNaQ.** Colors
1116 indicate the major and minor edge. The arrow from the minor edge shows the direction of
1117 introgression. The introgression probability (φ) is shown for each data type. In cases where a)
1118 *D. celsa* or b) *D. campyloptera* were the allotetraploids under investigation, all data types
1119 recovered the same species network. When c) *D. clintoniana* was the allohexaploid lineage from
1120 the allotetraploid *D. cristata* and diploid *D. goldiana* progenitors, different networks were
1121 recovered by the phased or consensus versus genotype or pick one data types.
1122
1123 **Figure 7 — Networks for all North American *Dryopteris* inferred from SNaQ.** Colors
1124 indicate major and minor (reticulation) edges, and which minor edges were hypothesized or an
1125 unexpected result. Introgression probabilities for each minor edge are provided. Numbers on
1126 branches subtending nodes are bootstrap support values. Results are shown separately for a)
1127 phased data where all haplotype sequences are included in gene trees, b) genotyped data with

1128 a single sequence for individual with IUPAC ambiguity codes, c) haplotype consensus

1129 sequences, and d) a single haplotype sequence representing each individual among gene trees.

1130