homologizer: Phylogenetic phasing of gene copies into polyploid subgenomes

William A. Freyman[1,*], Matthew G. Johnson[2], and Carl J. Rothfels[3,*]

[1]*23andMe, Inc., Sunnyvale, CA, USA*
[2]*Department of Biological Sciences, Texas Tech University, Lubbock, Texas*
[3]*University Herbarium and Department of Integrative Biology, University of California, Berkeley*
[*]*Corresponding authors: William A. Freyman (willfreyman@gmail.com); Carl J. Rothfels (crothfels@berkeley.edu)*

Summary

1. Organisms such as allopolyploids and F1 hybrids contain multiple distinct subgenomes, each potentially with its own evolutionary history. These organisms present a challenge for multilocus phylogenetic inference and other analyses since it is not apparent which gene copies from different loci are from the same subgenome and thus share an evolutionary history.

2. Here we introduce `homologizer`, a flexible Bayesian approach that uses a phylogenetic framework to infer the phasing of gene copies across loci into their respective subgenomes.

3. Through the use of simulation tests we demonstrate that `homologizer` is robust to a wide range of factors, such as incomplete lineage sorting and the phylogenetic informativeness of loci. Furthermore, we establish the utility of `homologizer` on real data, by analyzing a multilocus dataset consisting of nine diploids and 19 tetraploids from the fern family Cystopteridaceae.

4. Finally, we describe how `homologizer` may potentially be used beyond its core phasing functionality to identify non-homologous sequences, such as hidden paralogs or contaminants.

Key-Words: allopolyploidy, homeologs, hybridization, haplotype assembly, incongruence, phasing, polyploidy, polyploid phylogenetics, reticulate evolution, `RevBayes`

Introduction

Some individual organisms—such as allopolyploids and F1 hybrids—contain multiple distinct subgenomes, each with its own evolutionary history (in the case of polyploids, the homology of these subgenomes is related to a genome duplication event and they are referred to as "homeologous"). Such organisms present a particular challenge for multilocus phylogenetic inference because a researcher must take care to avoid assuming that particular sequences from different loci share an evolutionary history, *i.e.*, that they are from the same subgenome. For example, a diploid F1 hybrid individual will have one copy of a given nuclear locus from its maternal parent, and another from its paternal parent. If the maternal copy from one locus is treated as sharing an evolutionary history with the paternal version of another locus (*e.g.*, by subjecting them to classic concatenated or multispecies coalescent analysis), the underlying bifurcating model will be violated and the inferences will be unreliable (McDade, 1990, 1992; Oxelman et al., 2017; Rothfels, 2021). This issue is not limited to multilocus phylogenies of single-copy nuclear loci—the same problem applies when attempting to include organellar and nuclear loci in a common analysis, or when analyzing ITS sequences with other loci.

Phasing copies across loci is related to the problem of haplotype assembly—the phasing of sequencing reads within a locus: in both cases, the goal is to avoid chimeric data that are a mix of multiple evolutionary histories. In the assembly problem, however, a researcher can rely on physical linkage to determine which reads belong to which haplotype (Kates et al., 2018; Schrinner et al., 2020; Majidian et al., 2020; Nauheimer et al., 2021; Tiley et al., 2021). This approach is not available in the locus-phasing case, where the loci are separated from each other by unsequenced regions; the only information available to determine whether two copies come from the same subgenome is in the phylogenetic history itself.

The need to phase loci is particularly prevalent in plants, since over one-third of extant species are inferred to be recent polyploids (Wood et al., 2009; Såstad, 2005).

However, the issue is not restricted to plants. For example, both insects and diatoms have been reported to have an extensive history of polyploidy (Li et al., 2018; Parks et al., 2018), all salmonid fish are ancestrally polyploid (Alexandrou et al., 2013), and squamates and amphibians have clades with frequent allopolyploid lineage formation (Bogart, 1980; Bogart and Licht, 1986; Hedges et al., 1992; Lowe and Wright, 1966).

Historically, groups with extensive polyploidy have been under-studied phylogenetically (Soltis et al., 2014). Often, polyploids are dropped from phylogenetic analyses, in a "diploids-first" (or diploids-only) approach (*e.g.*, Beck et al., 2010; Govindarajulu et al., 2011; Lee et al., 2002) or, if polyploids are included, authors tend to infer gene trees for each locus individually (*e.g.*, Rothfels et al., 2014; Sousa et al., 2016; Chrtek et al., 2019; Melichárková et al., 2019; Griffin et al., 2011; Sessa et al., 2012a; Fortune et al., 2008; Rousseau-Gueutin et al., 2009; Kao et al., 2020). By omitting polyploids from multilocus analyses, we are not just removing the information those accessions could provide about general evolutionary patterns. Rather, we are systematically biasing our studies against groups with lots of polyploids, and we are impairing our ability to investigate questions where polyploidy may itself be an important factor (Mayrose et al., 2015; Ramsey and Ramsey, 2014; Rothfels, 2021). For example, if we are unable to include polyploids in multilocus phylogenies, we are limited in our ability to investigate classic questions in evolutionary biology such as the impact of polyploidy on diversification rates (Stebbins Jr, 1940; Wagner Jr., 1970; Mayrose et al., 2011; Tank et al., 2015; Soltis et al., 2009; Zhan et al., 2014; Freyman and Höhna, 2018; Zenil-Ferguson et al., 2019), or the association of polyploidy with niche breadth or with dispersal and establishment ability (Stebbins Jr, 1985; Marchant et al., 2016; Ramsey and Schemske, 2002).

In theory, researchers could get around this locus-phasing problem by inferring individual gene trees, and then using the phylogenetic position of the copies to determine to which subgenome they belong (the "phase" of the copies). For example, if a polyploid always has two gene copies, one of which is closely related to diploid species A, and the other to diploid species B, and this is true across all loci sampled, then one could confidently conclude that the "A" copies share one evolutionary history, and the "B" copies another (*e.g.*, Sessa et al., 2012b; Dauphin et al., 2018). However, given variable amounts of missing data (such as failure to recover one of the copies for a subset of the loci, or failure to sequence the related diploids for all loci) and the phylogenetic uncertainty inherent in inferring single-gene phylogenies, this method can be exceedingly difficult to apply, especially in datasets with many polyploids, or with many loci (Bertrand et al., 2015).

Beyond this by-eye approach, and approaches that rely upon existing reference sequences, such as that of Hénocq et al. (2020) and Nauheimer et al. (2021), there are, to our knowledge, three currently available methods for phasing copies across loci. First, Bertrand et al. (2015) developed an approach to phase two loci by finding the largest set of sequence pairs such that any incongruence between the two resulting gene trees could be due to stochastic or coalescent error. A somewhat similar approach was developed and refined by Oberprieler and colleagues (Oberprieler et al., 2017; Lautenschlager et al., 2020). Instead of explicitly testing for sources of incongruence, this method looks for the subgenome assignments of copies within a locus, and then across loci, that minimizes the number of inferred deep coalescent events in the context of the phylogeny of the diploid species. Both methods are relatively fast, although the complexity of the analysis increases quickly with the number of polyploid accessions, and with the number of loci. Perhaps more fundamentally, these methods add in accessions sequentially, so the phylogenetic information in the full dataset is not available to inform the phasing of each polyploid copy.

In contrast to the two methods above, which require an iterative series of analyses and are based on particular test statistics, the third available method, `alloPPnet` (Jones et al., 2013; Jones, 2017) is a one-step parametric Bayesian model for the inference of polyploid species trees (more accurately, species networks) under the multi-species coalescent. As such it explicitly models polyploid formation events, takes into account all the data from the full sample, and, for example, requires that the post-polyploidization subgenome lineages have the same effective population size. The chief limitations of `alloPPnet` are that it cannot accommodate ploidy levels higher than tetraploid and is computationally demanding, so can only be used on relatively small datasets.

Here we introduce `homologizer`, a simple, flexible, tree-based method to infer the posterior probabilities of the phasing of gene copies across loci that can be performed on a fixed topology, or while simultaneously estimating the phylogeny under a wide range of phylogenetic models, including the multi-species coalescent. Users can thus apply `homologizer` to infer the phylogeny while treating the phasing as a nuisance parameter, to infer the phasing while integrating out the phylogeny, or for any combination of these goals. This method utilizes the full dataset (information about topology and subgenome identity from each locus is available to inform the phasing of the other loci), does not require any external data (such as a reference genome), and does not require that the progenitor diploids of the polyploid accessions be extant or sampled.

`homologizer` is implemented in the open-source Bayesian inference software `RevBayes` (Höhna et al., 2016)

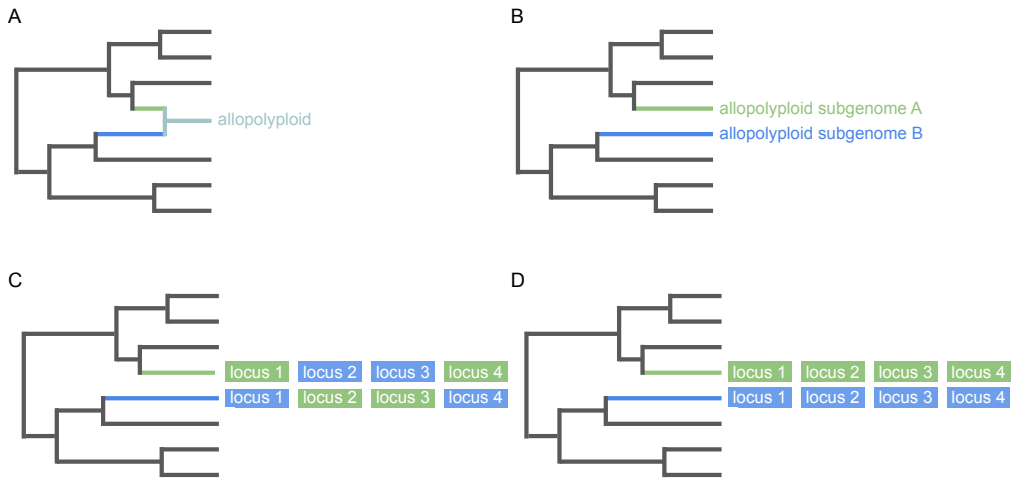

Figure 1: Phasing gene copies into polyploid subgenomes on a mul-tree. *A*: A phylogenetic network representing a single reticulation giving rise to an allopolyploid. *B*: The mul-tree representation of this phylogenetic network has two leaves (green and blue) representing the two subgenomes of the allopolyploid. *C*: Four loci were sequenced from the allopolyploid. Two copies (green and blue) of each locus were recovered. Loci 2 and 3 are incorrectly phased. *D*: After phasing, each locus is assigned to the correct subgenome.

and consists of dedicated MCMC proposals and monitors that switch tip assignments among gene copies and log the sampled phasing assignments, resulting in an approximation of the posterior distribution of the phasing of gene copies across loci. These `homologizer` features can be included in conjunction with any of the other modular phylogenetic inference components that `RevBayes` offers, enabling the phasing to be estimated under a large variety of models (*e.g.*, models that assume a shared gene tree for some or all loci, multi-species coalescent models that do or do not assume a constant population size, etc.). We anticipate that `homologizer` will be of greatest interest to phylogeneticists, but its applications extend beyond phylogenetics to other cases where it is important to determine from which subgenome a particular gene copy originates, for example, studies of expression-level dominance and subfunctionalization (Edger et al., 2018).

Description

`homologizer` makes use of multilocus sequence data, where a locus is a homologous genomic region present across the accessions sampled for the analysis (though the method allows for some proportion of missing sequences). The output of the method is the posterior distribution of phased homeologs, *i.e.*, the posterior distribution of the assignments of each gene copy, for each locus, into each of the available subgenomes. If joint inference of the phasing and phylogeny is performed, the posterior distribution of the multi-locus phylogeny is also inferred, along with all other parameters of the model.

The `homologizer` model

In the `homologizer` model we represent phylogenies containing polyploids as multilabeled trees ("mul-trees"), where each polyploid accession is present multiple times, once for each subgenome (each distinct evolutionary history; Huber et al., 2006). Gene copies are phased into subgenomes for each locus by swapping which gene copy is assigned to each of that accession's tips in the mul-tree (Figure 1). For each polyploid accession, there are $(n!)^g$ ways to phase the gene copies, where n is the number of gene copies per locus and g is the number of loci. The `homologizer` model thus adds a set of extra latent variables—gene-copy phase parameters—to the underlying phylogenetic model:

$$P(\xi, \tau, \theta \mid X) \sim P(X \mid \xi, \tau, \theta) P(\xi, \tau, \theta),$$

where ξ are the gene-copy phase parameters, τ is the tree, θ are all the other parameters of the model, and X is the set of sequence alignments.

The user defines the set of gene-copy phase parameters, which for each accession and each locus consists of $n-1$ parameters for n gene copies to be phased. Thus, the number of such parameters increases with both the number of polyploid subgenomes and with the number of loci. A given gene-copy phase parameter can take n values, corresponding to the n possible assignments of sequences to a given tip in the mul-tree; these values are

each assigned an equal prior probability and their posterior probability is estimated by MCMC (along with all other parameters of the model; Metropolis et al., 1953).

Setting up a `homologizer` analysis

Two types of information are required in order to run a `homologizer` analysis. First are the sequence data—the multiple-sequence alignments for each of the loci. Second is a determination, by the user, of how many tips each accession should get in the mul-tree. For example, if an accession has more than one copy at one or more loci, the user needs to decide whether those copies potentially represent multiple subgenomes or whether they instead represent allelic variation (in which case the user might split them into separate "individuals"—separate distinct tips that are not available for phasing). These determinations (the number of tips among which to phase gene copies) are fixed for a given analysis, although the tools of Bayesian model selection can be applied to determine how many tips are justified for each accession (see Model Selection section, below).

`homologizer` is implemented in `RevBayes` (Höhna et al., 2016), which is an modular program for implementing Bayesian phylogenetic analyses that uses an `R`-like syntax (R Core Team, 2013). Included in `RevBayes`' core functionality are tools for reading multiple sequence alignments and setting up all elements of a phylogenetic analysis (the substitution models, tree models, clock models, etc). For example, for the Cystopteridaceae analyses included in this paper, we set up an independent GTR substitution model for each locus, and ran a concatenated analysis (a single mul-tree topology shared by all loci). For the `homologizer` component of the model, a user initializes the MCMC analysis by first setting an initial phase assignment for each of the copies to be phased, using the `setHomeologPhase` command. For example, the Cystopteridaceae dataset has an accession, ×*Cystocarpium* #7974, that has four sequences for locus1, labelled "7974_copy1" through "7974_copy4", and we wish to phase those copies among four mul-tree tips, labelled "xCystocarpium_7974_A" through "xCystocarpium_7974_D". To set the initial phase assignment for these copies, one would use the following commands:

```
locus1.setHomeologPhase("7974_copy1", "
    xCystocarpium_7974_A")
locus1.setHomeologPhase("7974_copy2", "
    xCystocarpium_7974_B")
locus1.setHomeologPhase("7974_copy3", "
    xCystocarpium_7974_C")
locus1.setHomeologPhase("7974_copy4", "
    xCystocarpium_7974_D")
```

To successfully initialize the MCMC, each gene copy needs to have a unique initial phase assignment (each copy needs to be assigned to a mul-tree tip, and no two copies may be assigned to the same tip). This initial phase assignment can be randomly picked by the user and is only used to initialize the MCMC; it will not affect the final phasing estimate assuming the MCMC analysis converges.

The next command needed to set up a `homologizer` analysis is `mvHomeologPhase`, which creates an MCMC proposal that tries swapping gene copies among the mul-tree tips of a polyploid. For each locus with N gene copies to be phased, the analysis requires $\binom{N}{2}$ MCMC phase swapping proposals, so for ×*Cystocarpium*, with its four mul-tree tips, we need to apply the `mvHomeologPhase` command six times for each locus:

```
moves[++mvi] = mvHomeologPhase(locus1, "
    xCystocarpium_7974_A", "
    xCystocarpium_7974_B")
moves[++mvi] = mvHomeologPhase(locus1, "
    xCystocarpium_7974_A", "
    xCystocarpium_7974_C")
moves[++mvi] = mvHomeologPhase(locus1, "
    xCystocarpium_7974_B", "
    xCystocarpium_7974_C")
moves[++mvi] = mvHomeologPhase(locus1, "
    xCystocarpium_7974_A", "
    xCystocarpium_7974_D")
moves[++mvi] = mvHomeologPhase(locus1, "
    xCystocarpium_7974_B", "
    xCystocarpium_7974_D")
moves[++mvi] = mvHomeologPhase(locus1, "
    xCystocarpium_7974_C","
    xCystocarpium_7974_D")
```

The gene copy phase assignments sampled during the MCMC are logged to files using a monitor created with the `mnHomeologPhase` command.

There are complex interdependencies among the phasing estimation across accessions that can make it difficult for the MCMC to sample efficiently. For example, two closely related tips can get the phasing for a subset of their loci "stuck", because switching the phasing to the correct orientation requires traversing a low-probability valley in parameter space where one of the accessions has the correct phasing but the other does not. Users should run multiple independent MCMC analyses and check for convergence.

Summarizing `homologizer` inferences

The phasing estimates from a `homologizer` MCMC analysis can be summarized and plotted using the `R` (R Core Team, 2013) package `RevGadgets` (Tribble et al., 2020), which displays the joint maximum *a posteriori* (MAP) phasing assignment across the set of a polyploid's subgenomes for each locus. This joint MAP assignment is the highest probability assignment of each gene uniquely to a subgenome. Additionally, we show the marginal posterior probability for each locus and each subgenome. These marginal posterior probabilities are useful when quantifying the uncertainty in the joint MAP phasing assignment. For example, imagine a hexaploid with three

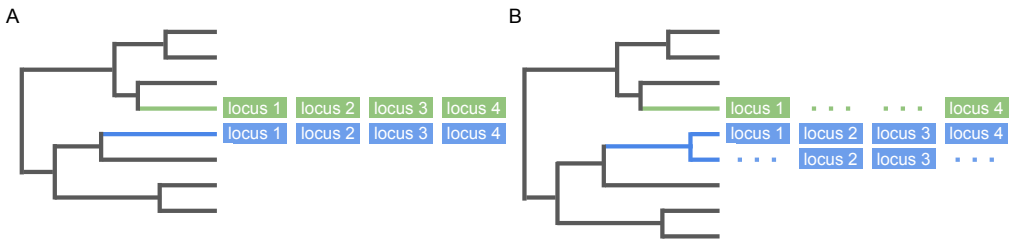

Figure 2: Model testing for allelic variants and/or homeologs. In this example, two copies of each locus were sequenced from a single allotetraploid. *A*: A phasing model which swaps the gene copies between two mul-tree tips (the green and the blue tips) assumes the two gene copies of each locus are homeologous. *B*: An alternative phasing model swaps gene copies between three mul-tree tips. This potentially allows for the gene copies of some loci to be inferred as allelic variation and the gene copies of other loci to be inferred as homeologs. In this example, the gene copies of locus 1 and locus 4 are homeologs and the gene copies of locus 2 and 3 are allelic variants. The fit of the data to these two models—mul-tree A and mul-tree B—can be compared through Bayes factors.

subgenomes (A, B, and C) and three gene copies of a particular locus. If subgenomes A and B are difficult to distinguish (*e.g.*,, they could be sister to each other), then `homologizer` may infer a low posterior probability for the joint MAP phasing assignment of this locus. However, examining the marginal probabilities of each phasing assignment would reveal that subgenome C was phased with high posterior probability and that the low joint posterior probability was due to the difficulties in phasing between subgenomes A and B. Furthermore, the mean of the marginal probabilities across loci of the phasing assignment for a subgenome summarizes the model's overall confidence in the phasing for that subgenome; this information is also incorporated into the `RevGadgets` visualization. Note that, in contrast to the joint MAP phasing, the marginal MAP phasing for each individual subgenome may result in assignment of the same gene copy to multiple subgenomes.

Comparing `homologizer` models

Users interested in distinguishing gene copies that evolved in separate polyploid subgenomes from those that arose from allelic variation within the same subgenome (or that are otherwise non-homeologous) can set up a series of `homologizer` models—which differ in the number of mul-tree tips available for phasing—and compare them using Bayes factors (Kass and Raftery, 1995). For example, consider the example in Figure 2. Here, the polyploid accession has two subgenomes, was sequenced for four loci, and for each of those loci, two copies were recovered. It could be that each locus is represented by copies from each of the subgenomes and thus the accession should be given two tips in the mul-tree (panel A in Figure 2). However, it's possible, instead, that the copies at one or more of the loci are allelic variants from one of the subgenomes. In this case, three mul-tips are needed—two tips for one subgenome, and one tip for the other subgenome (panel B in Figure 2). To compare these two models (one with two mul-tree tips for this accession, and one with three) using Bayes factors, the data must be kept constant, and only the number of phase assignment parameters changed. This condition is satisfied by running both models with a "third set" of blank sequences added for this accession, and assigned to a mul-tree tip using `setHomeologPhase`, but with that tip phased (using `mvHomeologPhase`) only in the "three-tip" analysis. In the two-tip analysis, conversely, the phase of this tip is not estimated and is fixed to blank sequences; after the analysis it can be pruned from the tree sample using the `RevBayes` `fnPruneTree` command.

Software availability

`RevBayes` (Höhna et al., 2016) is an open-source software package available at `http://revbayes.com`. `RevGadgets` (Tribble et al., 2020), the `R` package used for summarizing the output from `RevBayes` analyses, is also open-source and is available at `https://github.com/revbayes/revgadgets`. Further details, including an example analysis and all scripts and data used in this paper, are available at `https://github.com/wf8/homologizer`.

Simulation tests

To evaluate the performance of `homologizer`, we performed two separate simulation tests. The first tested the impact of different evolutionary factors on phasing accuracy and the second focused on whether applying Bayes factors to compare `homologizer` phasing models can suc-

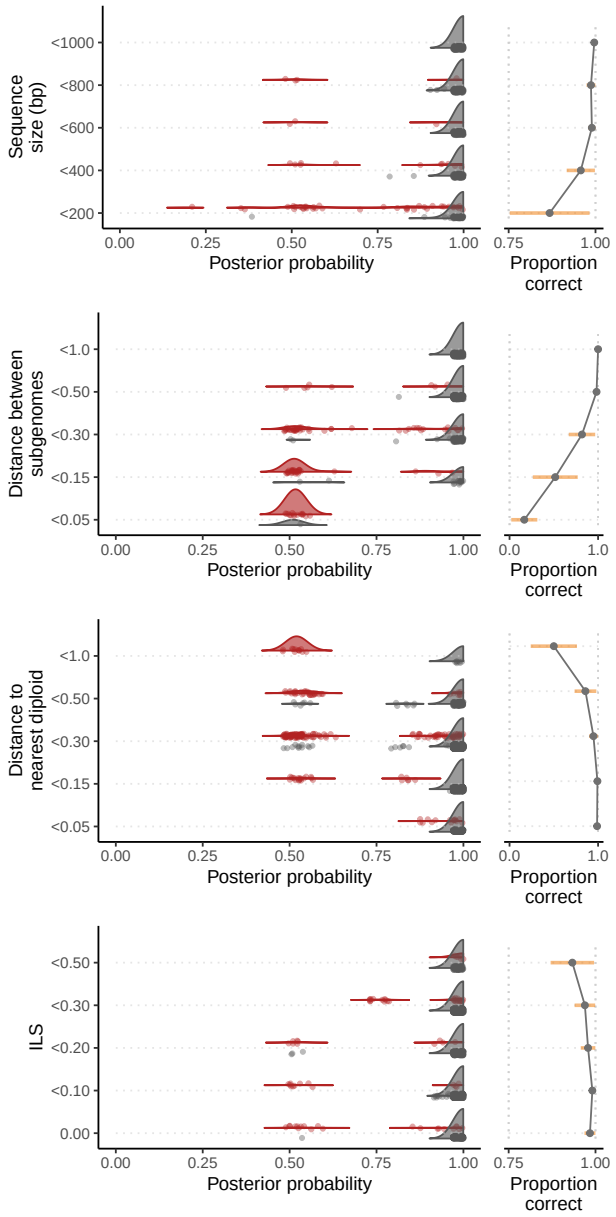

Figure 3: Performance of `homologizer` in phasing four loci under different simulated conditions. Each row shows how phasing performance was impacted by a factor expected to influence accuracy: sequence length; smallest distance between subgenomes; smallest distance between a polyploid subgenome and a sampled diploid; and degree of incomplete lineage sorting (ILS). See the main text for details on how each factor was quantified. The x-axis in the left panel of each row shows the mean marginal probability of the joint MAP phasing: each point represents an individual simulated polyploid. Points are colored red if the joint MAP phasing was estimated incorrectly and grey if the phasing was estimated correctly. The grey and red densities represent the distribution of the mean marginal posterior probabilities of correctly and incorrectly phased simulated polyploids, respectively. The right panel of each row summarizes how the proportion of times the model was correct changes with the focal factor. The orange bar is the variance.

cessfully distinguish between allelic versus homeologous variation.

Experiment 1: Testing phasing accuracy

To explore the behavior of `homologizer` under different evolutionary scenarios, we simulated datasets to test how the performance of `homologizer` was impacted by the following factors: (1) the phylogenetic informativeness of each locus; (2) the phylogenetic distance between subgenomes for each polyploid (the minimum distance in cases of more than two subgenomes); (3) the minimum phylogenetic distance from a polyploid subgenome to the nearest diploid, and; (4) the amount of incomplete lineage sorting (ILS) present. For each replicate we first simulated a multi-labeled species tree under a constant rate birth-death process with root age = 50.0, speciation rate = 0.2, extinction rate = 0.01, and the fraction of taxa sampled at the present = 0.01. Each multi-labeled species tree was conditioned on having twenty extant tips, five of which were randomly selected to represent allopolyploid subgenomes. Two of the allopolyploid tips were then randomly selected to represent the subgenomes of a tetraploid, and the other three allopolyploid tips represented subgenomes of a hexaploid. The remaining 15 tips were considered diploid lineages that did not need phasing. Each dataset thus included two allopolyploids that we could use to assess the performance of `homologizer`.

Once the multi-labeled species tree was simulated, we used the multi-species coalescent to simulate four gene trees, representing four loci. We varied the effective population size of these simulations as described below. Finally, over the simulated gene trees we simulated nucleotide sequences under an independent GTR substitution model (Tavaré, 1986) with exchangeability rates and stationary frequencies sampled from flat Dirichlet priors. The clock rate of the GTR process was constant over all branches of the tree and fixed to 0.01. We varied the length of the simulated sequences as described below.

All the simulated nucleotide sequences were then analyzed using `homologizer`. The phasing and the phylogeny were inferred simultaneously, linking the tree across loci (so these were "concatenated" analyses). Each locus was modelled by an independent GTR substitution model, with Dirichlet priors on the exchangeability parameters and stationary frequencies, a uniform prior on topology, and exponential priors (mean = 0.1) on branch lengths. The MCMC was run for 2000 generations where each generation consisted of 354 separate Metropolis-Hastings moves. The first 25% of samples were discarded prior to summarizing the posterior distribution. Effective sample sizes (ESS) of the posterior were consistently greater than 200; the mean ESS across simulation replicates was 355.6.

Based on this simulation design, we performed four tests, each varying one of the focal factors while keeping the others at fixed levels. Those fixed values were a sequence length of 800 base pairs, an effective pop-

ulation size of 0.0001 (which effectively meant there was zero ILS), a minimum distance between polyploid subgenomes of at least 0.25 (scaled by tree height such that a value of 1.0 indicates that two leaves of the tree share a most recent common ancestor at the root of the tree), and a minimum distance from a polyploid subgenome to the nearest diploid of no more than 0.25. For each experiment we tracked the proportion of simulated polyploids for which the phasing was correctly estimated (all four loci were phased correctly), and the mean marginal posterior probability of the joint MAP phasing for each polyploid.

To test the effect of phylogenetic informativeness on the performance of `homologizer`, we varied the length of the simulated nucleotide sequences. For 1000 simulation replicates we sampled the length of the sequence from a uniform distribution with a minimum of 1 and a maximum of 1000 base pairs. To test the effect of the phylogenetic distance between polyploid subgenomes we simulated 800 replicates, without placing any constraints on the maximum subgenome distance. To test the effect of the phylogenetic distance between each polyploid and the closest diploid tip we simulated 800 replicates, this time without constraints on the "diploid distance." Finally, to test the effect of ILS, we simulated five sets of 200 replicates, with effective population sizes of 0.001, 0.1, 0.5, 1.0, or 2.0, respectively. We then quantified the amount of ILS in each simulated dataset with the summary statistic R_g/R_s, where R_g is the mean Robinson-Foulds distance (Robinson and Foulds, 1981) between each gene tree and the species tree, and R_s is the mean Robinson-Foulds distance between each independently simulated species tree. This statistic provides an intuitive summary of the amount of gene discordance: when, *e.g.*, $R_g/R_s = 0.5$ the amount of discordance between the gene trees and the species tree is 50% of what one would observe between completely unlinked trees. Therefore, simulation replicates with $0.0 < R_g/R_s \leq 0.1$ represent low amounts of ILS, replicates with $0.1 < R_g/R_s \leq 0.3$ represent moderate amounts of ILS, and those with $0.3 < R_g/R_s \leq 0.5$ represent high levels of ILS.

Experiment 2: Testing Bayes factors to distinguish allelic variation from homeologs

Our second simulation test focused on testing whether Bayes factors could be used to compare `homologizer` phasing models and distinguish allelic variation from homeologs. For each replicate, we first simulated a multi-labeled species tree under a constant-rate birth-death process with root age = 50.0, speciation rate = 0.2, extinction rate = 0.01, and the fraction of taxa sampled at the present = 0.01. Each multi-labeled species tree was conditioned on have twenty extant tips, two of which were randomly selected to represent allopolyploid subgenomes of a tetraploid. The remaining 18 tips were considered

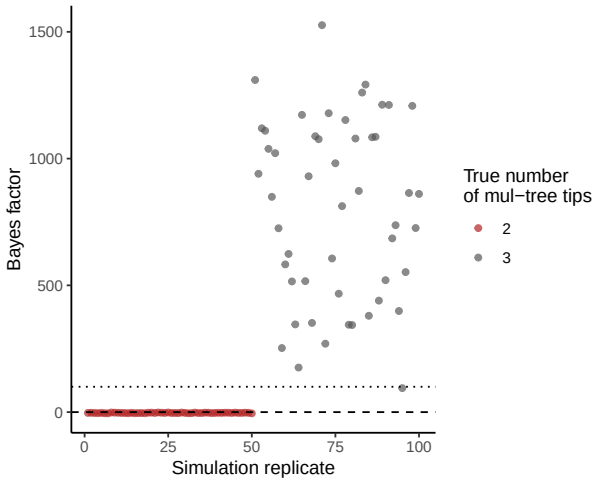

Figure 4: Effectiveness of model comparison tests to distinguish allelic variation from homeologs. Each dot represents a single simulated dataset as described in the main text. The y-axis shows the Bayes factor for the three mul-tree tip phasing model compared to the two mul-tree tip phasing model. All replicates had two gene copies simulated for each locus. Replicates 1–50 (red dots) were simulated with each of the two tips representing a different polyploid subgenome. For all replicates the Bayes factor was less than 0, indicating the three-tip model was not supported over the two-tip phasing model. Replicates 51–100 (grey dots) were simulated with three total tips representing two different subgenomes of a polyploid in some loci and a single subgenome with two allelic variants in other loci. For nearly all these replicates the Bayes factor was over 100 (dotted line), indicating "decisive" support (Kass and Raftery, 1995) for the three-tip phasing model over the two-tip model.

diploid lineages that did not need phasing.

Once the multi-labeled species tree was simulated, we used the multi-species coalescent to simulate four gene trees, representing four loci. In half of the replicates, all gene trees were simulated with one gene-tree tip for each species-tree tip. These replicates represent the case in which a two mul-tree tip phasing model is appropriate because there is no allelic variation (all gene copies for the tetraploid are true homeologs). For the other half of the replicates, gene trees were simulated with one gene-tree tip for each species-tree tip for two loci, and with two gene lineages from subgenome A and none from subgenome B for the remaining two loci. These replicates represent the case in which a three mul-tree tip phasing model is needed because the gene copies for two of the loci are allelic variants from the same subgenome but for the other two loci the gene copies are homeologs from different subgenomes.

The gene trees were simulated using an effective population size of 0.0001 (which, like in our first experiment, meant there was zero ILS), a minimum distance between polyploid subgenomes of at least 0.25 (scaled by tree height such that a value of 1.0 indicates two leaves of the tree share a most recent common ancestor at the root of the tree), and a minimum distance from a poly-

ploid subgenome to the nearest diploid of no more than 0.25. Over the simulated gene trees we simulated 800 nucleotide base pairs under an independent GTR substitution model (Tavaré, 1986) with exchangeability rates and stationary frequencies sampled from flat Dirichlet priors. The clock rate of the GTR process was constant over all branches of the tree and fixed to 0.01.

For each simulated dataset we ran stepping-stone analyses (Xie et al., 2011) to compute marginal likelihoods of the two mul-tree tip and three mul-tree tip `homologizer` phasing models. Our stepping-stone simulations used 50 stepping stones, sampling 1000 states from each step. We then compared the `homologizer` phasing models using Bayes factors (Kass and Raftery, 1995).

Simulation results and discussion

Experiment 1: Testing the phasing accuracy under simulated conditions

`homologizer` was able to correctly phase all four loci, with high posterior probability, over most of our simulation conditions, and in those cases where the joint MAP phasing was incorrect, the posterior probability was generally low (Figure 3). Even with minimal phylogenetic information available (sequence lengths less than 200 base pairs long), `homologizer` was able to phase the loci correctly over three-quarters of the time. The only cases where `homologizer` struggled to get the correct phasing was in situations where either the distance between subgenomes was very small, or the distance from the polyploid subgenomes to the nearest diploid was very large (Figure 3, middle two rows). In both those cases, the polyploid subgenomes are likely to be sister to each other and thus there is no topological information available to inform the phasing—the average marginal posterior probability in these cases converges on 50%, as expected (the phasing of each locus in the case of sister lineages being an effectively random choice; Figure 3).

Gene-tree incongruence (induced in our simulations by elevated degrees of ILS) had a subtly different impact on `homologizer` inference. While the method was generally robust to ILS (Figure 3, bottom row), high levels of incongruence resulted in an appreciable number of simulations where `homologizer` was "confidently wrong"—the joint MAP phasing was incorrect, but the average marginal phasing probability was high. This behavior is due to the model's assumptions being violated by the process that actually generated the data; the loci evolved along highly discordant gene trees and yet the `homologizer` model assumed a single shared topology for all gene trees. Stochastic variation in the "true" gene trees may result in one phasing having a slight advantage and thus high posterior probability, analogous to the high support inferred for arbitrary resolutions of polytomies by MCMC analyses that only visit fully resolved trees (the "star-tree paradox"; Lewis et al., 2005; Yang, 2007; Rothfels et al., 2012).

Table 1: Summary of Cystopteridaceae dataset. SEQ. #: the number of mul-tree tips represented by sequence data for that dataset; LEN.: the aligned length of that dataset in basepairs; %MISS.: the percentage of missing data (?s and -'s).

LOCUS	SEQ. #	LEN.	% MISS.
ApPEFP_C	44	1006	6.6%
gapCpSh	48	1068	7.7%
IBR3	45	862	4.2%
pgiC	48	1132	15.5%
total	53	4068	20.4%

Experiment 2: Testing Bayes factors to distinguish allelic variation from homeologs

Using Bayes factors to compare `homologizer` phasing models successfully distinguished allelic variation from homeologs (Figure 4). For the simulation replicates in which the three mul-tree tip phasing model was necessary to accommodate the presence of both homeologous gene copies and allelic variation, the Bayes factors for the three mul-tree tip phasing model compared to the two mul-tree tip phasing model were nearly always over 100, indicating "decisive" support (Figure 4, grey dots; Kass and Raftery, 1995). When all gene copies from the tetraploid were homeologs such that three mul-tree tips were not necessary, then the Bayes factors were always less than zero (Figure 4, red dots), indicating there was no support for the three mul-tree tip phasing model over the two mul-tree tip phasing model.

Cystopteridaceae analysis

Rothfels et al. (2017) analyzed a dataset of four single-copy nuclear loci (*ApPEFP_C*, *gapCpSh*, *IBR3*, and *pgiC*) for a sample of nine diploids and 19 tetraploids from the fern family Cystopteridaceae, using `alloPPnet` (Jones et al., 2013; Jones, 2017). We reanalyzed these data (Table 1) to compare the performance of `homologizer` with the explicit hybridization model of `alloPPnet`, and to see if the phasing information reported by `homologizer` informs our understanding of this dataset.

Because `alloPPnet` cannot accommodate ploidy levels higher than tetraploid, one accession (×*Cystocarpium roskamianum* Fraser-Jenk.; Fraser-Jenkins, 2008; Fraser-Jenkins et al., 2010), which is an intergeneric allotetraploid hybrid of two other allotetraploids and thus includes four distinct evolutionary histories (Rothfels et al., 2015), was treated by Rothfels et al. (2017) as two tetraploids (one for each of its parental genera). With `homologizer`, this ad hoc solution is not necessary, so we

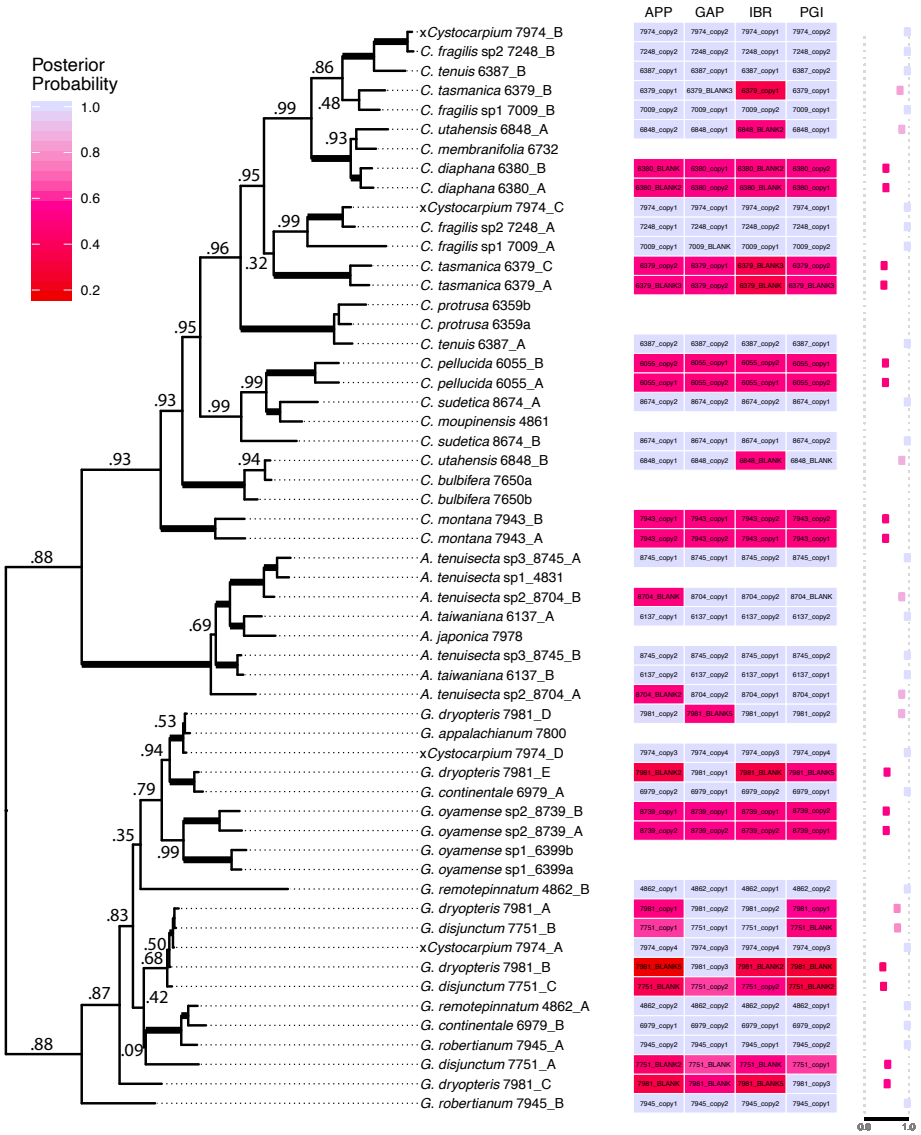

Figure 5: `homologizer` analysis of the Cystopteridaceae dataset. The phasing of gene copies into subgenomes is summarized on the maximum *a posteriori* (MAP) phylogeny. Thickened branches have posterior probabilities of 1.0; posterior probabilities < 1.0 are indicated. The columns of the heatmap each represent a locus, and the joint MAP phase assignment is shown as text within each box. Each box is colored by the marginal posterior probability of the phase assignment. Adjacent to the heatmap is a column that shows the mean marginal probability across loci of the phasing assignment per tip, which summarizes the model's overall confidence in the phasing of that tip. In the sample labels, "A." = *Acystopteris*, "C." = *Cystopteris*, "G." = *Gymnocarpium*, the four-digit numbers are Fern* Labs Database accession numbers (https://fernlab.biology.duke.edu/), capital A, B, etc, indicate subgenomes, sp1 and sp2 indicate undescribed cryptic species, and lowercase letters following the accession numbers indicate haploid "individuals" within the sampled diploids (*i.e.*, those diploids are heterozygous). Copy names with a "BLANK" suffix indicate missing sequences (*e.g.*, a subgenome that was present in some loci but not retrieved for others).

reanalyzed these data with ×*Cystocarpium roskamianum* allotted four tips in the mul-tree (to accommodate the homeologs of ×*Cystocarpium*), and most other tetraploids allotted two tips. The diploid samples were each given a single mul-tree tip, except for three (*Cystopteris bulbifera* 7650, *Cystopteris protrusa* 6359, and *Gymnocarpium oyamense* 6399) that had allelic variation and were each treated as two independent accessions. The final dataset comprises 4068 aligned sites (Table 1).

For three of the accessions—*Cystopteris tasmanica* 6379, *Gymnocarpium disjunctum* 7751, and *Gymnocarpium dryopteris* 7981—the number of mul-tree tips required was uncertain (for more discussion of this uncertainty, see the "Insights from `homologizer` model selection and data exploration" section, below). To determine the appropriate number of tips for these taxa, we tested seven differ-

Table 2: Cystopteridaceae phasing model tests. Bayes factors were used to compare phasing models for three polyploid accessions. Each model varied the number of mul-tree tips used to phase gene copies. The marginal log-likelihood was calculated four times independently for each phasing model; the mean and standard deviation is shown. The Bayes factor of the best supported model versus each other model is shown. The best supported model is indicated with -. Model comparisons with "strong" support according to Kass and Raftery (1995) are marked with **.

Accession	Number of mul-tree tips	Marginal log-likelihood (mean)	Marginal log-likelihood (SD)	Bayes factor
C. tasmanica 6379	2	-14519.12	2.27	**62.34****
	3	-14456.78	5.71	-
G. disjunctum 7751	2	-14157.80	1.81	1.90
	3	-14155.91	1.82	-
G. dryopteris 7981	3	-14216.27	2.11	**26.67****
	4	-14202.02	3.02	**12.42****
	5	-14189.61	2.80	-

ent phasing models (Table 2). Two of the models tested whether *C. tasmanica* should have two or three mul-tree tips, another two tested whether *G. disjunctum* should have two or three mul-tree tips, and, finally, three tested whether *G. dryopteris* should have three, four, or five mul-tree tips. We then calculated Bayes factors by stepping-stone analyses (Xie et al., 2011) to compare among the models for each of the three polyploids. We calculated the marginal log-likelihood of each model four times independently; each analysis comprised 50 stones, sampled 1000 time each.

We ensured that each model comparison tested a single polyploid accession in isolation by removing the other two uncertain accessions (i.e., when testing whether *C. tasmanica* should have two or three tips we removed all *G. disjunctum* and *G. dryopteris* tips). Additionally, we fixed the phasing of the other polyploid accessions to the MAP phase previously estimated in an analysis that excluded all three of the focal taxa (results not shown).

After the model comparisons were complete, we chose the number of mul-tree tips with the highest support for each polyploid and then jointly estimated the tree and the phase of all accessions together, linking the tree across loci. Each locus was modelled with an independent GTR substitution model, with Dirichlet priors on the exchangeability parameters and stationary frequencies, a uniform prior on topology, and exponential priors (mean = 0.1) on branch lengths. The MCMC was run seven times independently, each for 10000 generations where each generation consisted of 532 separate Metropolis-Hastings moves. Of these seven runs, three converged (see the "MCMC behaviour for `homologizer` models on empirical data" section, below); one of these runs was arbitrarily selected as our final analysis. The first 25% of samples were discarded prior to summarizing the posterior distribution. The ESS of the posterior was 252.

Cystopteridaceae results and discussion

Insights from `homologizer` model selection and data exploration

The ability to compare among different `homologizer` models with Bayes factors, and thus to use `homologizer` as a data-exploration tool, resulted in key insights about the data for three of the accessions in our dataset; these insights subsequently resulted in significantly altered results of the downstream inferences. The first case involves *Cystopteris tasmanica*, a tetraploid (Tindale and Roy, 2002; Brownlie, 1958; Brownsey and Perrie, 2018). From our *C. tasmanica* accession, we recovered two sequences for three of the four loci (we recovered a single copy from *IBR3*). This accession thus appears to be perfectly "well-behaved", with one copy of each of the two expected homeologs at most loci, which is how it was treated in the Rothfels et al. (2017) `alloPPnet` analyses. Preliminary `homologizer` analyses under this model (phasing two mul-tree tips for *C. tasmanica*), however, had an odd result: three of the loci phased with very high posterior support, but the phasing for the two *gapCpSh* copies was equivocal (each copy had ∼50% posterior probability of phasing to each of the mul-tree tips). To test the possibility that this strange result was due to the two *gapCpSh* copies being allelic rather than homoelogous (in which case we would expect each of the copies to fit equally to each mul-tree tip), we ran another `homologizer` model, this time with *C. tasmanica* allocated three mul-tree tips. This three-tip model strongly out-performed the two-tip one (Table 2), and as we had suspected, the two *gapCpSh* copies were resolved sister to each other, with a blank copy phased with high posterior probability to the other subgenome (Figure 5).

The other two insights were related to each other, and involve the allotetraploid *Gymnocarpium dryopteris* and one of its diploid progenitors, *G. disjunctum* (Pryer and Haufler, 1993; Rothfels et al., 2014, 2017). In our

dataset, *G. dryopteris* was represented by two sequences for two loci (as would generally be expected for an allotetraploid), but by three sequences for the other two loci. In the `alloPPnet` multispecies coalescent analyses of Rothfels et al. (2017), this accession was split into two "individuals" in order to accommodate its apparent heterozygosity. For our preliminary concatenation-based (instead of coalescent-based, as in `alloPPnet`) `homologizer` analyses, we ran a model that gave this accession three tips, in the thought that only one of the subgenomes had any allelic variation (*i.e.*, the pattern shown in Figure 2B), and a four-tip model, in case both subgenomes were heterozygous at at least one locus. The phasing for *G. dryopteris* in these preliminary analyses, however, had difficulty converging, and in addition, inspection of the *pgiC* gene tree showed that each of the three *G. dryopteris* copies were resolved in distinct places. Specifically, one copy grouped with each of *G. appalachianum* and *G. disjunctum*, as expected given these two diploids are the *G. dryopteris* progenitors (Pryer and Haufler, 1993; Rothfels et al., 2014, 2017); however, the third copy was in an isolated position. In response, we ran a five-tip model, and this model was strongly preferred over either of the three- or four-tip ones (Table 2). Similarly, *G. disjunctum* appeared to be heterozygous at two loci and in Rothfels et al. (2017) was split into two individuals; however this (two-tip) model converged poorly, and a three-tip model was weakly preferred by Bayes factor comparison (Table 2).

The final, integrated model (three tips for *C. tasmanica*, five tips for *G. dryopteris*, and three for *G. disjunctum*) resulted in a considerably revised understanding of our dataset. The *C. tasmanica* case was fairly straightforward, but also had a dramatic effect: by allowing the *gapCpSh* copies to be allelic rather than forcing them onto different subgenomes, the three-tip model inferred that *C. tasmanica* is an allopolyploid, with one subgenome from each of the major subclades of the core *fragilis* complex (*i.e.*, the *fragilis* complex minus *C. protrusa*; Rothfels et al., 2013). This conclusion is in strong contrast to that of Rothfels et al. (2017), which limited this accession to two tips and inferred that it was possibly an autopolyploid (the two tips were inferred to be sister to each other, albeit relatively deeply divergent). For *G. disjunctum*, both sequences of the two loci that had two copies were phased to tips consistent with their being "normal" alleles in a diploid genome—these two loci are heterozygous. However, the lone *pgiC* sequence did not phase to either of those tips; instead it was phased to its own tip, in an isolated portion of the phylogeny (Figure 5). Multiple lines of evidence suggest that this tip does not represent a subgenome history: 1) it is inconsistent with what is known of the ploidy of *G. disjunctum* (it's a diploid; Pryer and Haufler, 1993; Rothfels et al., 2014); 2) it is inconsistent with the phylogenetic history of this taxon (as inferred from multiple accessions of *G. disjunctum* itself as well as from the allopolyploids of which it is a genome donor (×*Cystocarpium* and *G. dryopteris*; Rothfels et al., 2014, 2017)), and; 3) this history is not represented by any of the other loci (Figure 5). Based on that evidence, and particularly in the context of the results for *G. dryopteris* (below), we suspect that this sequence represents a lineage-specific duplication of *pgiC* within the *G. disjunctum* lineage.

For *G. dryopteris*, while each locus had only one sequence from the *G. appalachianum* subgenome, the *gapCpSh* representative of this subgenome was sufficiently phylogenetically distinct (presumably due to coalescent variance) to be phased with high posterior probability to its own, closely related, tip. In the *G. disjunctum* subgenome, the results were more to be expected—all loci had a single copy, which phased to one of the tips, except for *gapCpSh*, which had two closely related alleles (Figure 5). However, the most important result was that one of the three *pgiC* copies was phased, with high posterior probability, to its own tip, phylogenetically distant from all other *G. dryopteris* tips. This pattern is very similar to that seen for *pgiC* in *G. disjunctum* (above), and we suspect these issues are related: *G. dryopteris* has inherited this putatively paralogous copy from *G. disjunctum*, and the *G. dryopteris* version in our dataset may additionally be a PCR recombinant. This conclusion is supported by the fact that *G. dryopteris* has two other *pgiC* copies that are cleanly phased to subgenomes corresponding to each of the putative progenitor diploids. By allowing this aberrant *pgiC* tip to be phased to its own branch, this three-tip `homologizer` model clarifies that the remaining *G. disjunctum* allele sequences are very closely related to each other, and provides strong support to the relationships within this area of the *Gymnocarpium* phylogeny (Figure 5.

MCMC behaviour for `homologizer` models on empirical data

The Cystopteridaceae analyses also demonstrated that inference under `homologizer` models, when applied to real datasets with a large number of polyploid accessions, involves complex interactions among the phasing estimates and between those estimates and those of the topology (and branch lengths). Particularly problematic is the case, described above (in the "Setting up a `homologizer` analysis" section), of two polyploids that have closely related subgenomes and can get their phasing for a subset of the loci "stuck" in the incorrect orientation. In the Cystopteridaceae dataset, for example, ×*Cystocarpium* and *Cystopteris fragilis* "sp 2" share two pairs of closely related subgenomes, each pair from a different major subclade of the *fragilis* complex (Figure 5). If a locus got out-of-phase for both accessions, it could get trapped in that orientation, presumably because a proposal to the phas-

ing for that locus for any one of the accessions would result in each pair of closely related tips containing very dissimilar sequences, which would have a low probability given the short phylogenetic distance between the two accessions. While we experimented with more complex joint phase moves, we found that simply increasing the frequency of the phase moves relative to the tree topology moves was more effective. However, while any given `homologizer` run was reasonably rapid (on the order of a "regular" analysis in `RevBayes`), they frequently failed to converge. Specifically, individual runs would get stuck in a suboptimal region of parameter space and appear to be sampling from the posterior distribution (effective samples sizes would be high, etc.), but multiple such runs would sample from different distributions. In our full Cystopteridaceae analysis, for example, we ran seven independent runs, only three of which converged to the posterior distribution. As a result, we recommend that users run preliminary analyses and inspect the results for areas where two closely related accessions are potentially interfering with each other. In addition, multiple independent runs are necessary to ascertain the efficiency of MCMC mixing for a given dataset.

A phased multi-locus Cystopteridaceae phylogeny

Our inferred Cystopteridaceae phylogeny under the best-fitting model (see Table 2) is well supported (Figure 5), and generally consistent with the inferences of Rothfels et al. (2017), although with the addition of all the phasing information.

The primary differences relate to the insights gained from our data exploration/model comparison (see above). Specifically, the strongly supported resolution of *Cystopteris tasmanica* as an allopolyploid with two fairly phylogenetically distant progenitors (Figure 5) is in strong contrast to the earlier inferences of this accession as a potential autopolyploid, with subgenomes that are sister to each other (Rothfels et al., 2017). Other secondary differences include the resolution of the *sudetica* clade sister to the *fragilis* complex instead of to the *bulbifera* clade, and a different placement of the morphologically anomalous *G. oyamense* within *Gymnocarpium*. It is difficult to determine what is driving these (secondary) topological differences, since, beyond the phasing component, the underlying analyses differ in their substitution model, clock model (or lack thereof), and in whether or not they model ILS (through the multi-species coalescent). Furthermore, while these relationships are all strongly supported in our `homologizer` analysis, `alloPPnet` doesn't report clade probabilities, so it is not clear how well supported the alternative relationships were in the earlier `alloPPnet`-based analysis.

The phasing is also generally well supported, despite the fact that considerable portions of the phylogeny don't have any diploid representation: by sharing phylogenetic information among polyploids we were able to phase them confidently even without closely related diploids being included. Those subgenome pairs that phased poorly (had low average marginal probabilities of phasing) were restricted to sister-pairs, where there is no topological phasing information available. Low marginal probabilities of phasing for individual loci (*e.g.*, where three loci are phased with high probability but the forth is not) nearly always reflect cases where that locus was not recovered for the given sample, such as IBR for *C. utahensis* (Figure 5).

Applications and future directions

`homologizer` for inference

The core application of `homologizer` is to infer multilocus multilabeled phylogenies and/or the phasing of gene copies to subgenomes. For this application, our simulation tests show that `homologizer` is robust to a wide range of factors that can potentially affect accuracy (Figure 3). When the `homologizer` model does not have enough information to confidently phase gene copies due to (1) phylogenetically uninformative sequences, (2) the polyploid subgenomes being too closely related, or (3) a lack of diploid lineages that help inform the phasing, then `homologizer` makes estimates with low posterior probability that correctly capture the uncertainty of the phasing.

Model violations, however, pose a greater challenge to `homologizer`. We see this sensitivity in our simulations with high levels of ILS (Figure 3, bottom panel). Here, since the `homologizer` model used in our simulation tests assumed a shared gene tree, the incongruence among gene trees was not modeled appropriately, and `homologizer` sometimes made incorrect estimates with high posterior probability ("confidently wrong" behavior). However, the modularity of `RevBayes` allows for the easy integration of `homologizer` with other model components. For example, if a user expects ILS to be important in their dataset, we recommend using `homologizer` within the full multi-species coalescent (MSC) framework, allowing for unlinked gene trees. This model, however, is computationally demanding (like all full MSC models), and can feasibly be applied to only relatively small datasets.

Other datasets might require different analysis approaches. For example, genome-scale target enrichment and transcriptome datasets often contain hundreds of loci. With these large datasets joint inference of the phylogeny and gene-copy phasing as presented here may not be computationally feasible. In those cases, it may be reasonable to adopt a sequential Bayesian approach: first infer a species-level mul-tree while phasing a subsample of

loci, and then condition the phasing of the remaining loci on that mul-tree. During the secondary phasing analyses one could optionally incorporate phylogenetic uncertainty in the mul-tree by integrating over the posterior distribution of mul-trees, as inferred in the first step. The `RevBayes` implementation of `homologizer` allows for a wide range of approaches, such as these, for scaling up phasing analyses that should be suitable for any sized dataset.

`homologizer` for data exploration

The other core functionality of `homologizer` is as a data exploration tool. With increasing numbers of taxa and loci, it can be difficult to detect non-orthologous sequences, such as hidden paralogs, contaminants, or allelic variation that is erroneously modelled as homeologous. `homologizer`, in conjunction with the `RevGadgets` visualization tools, allows for the convenient detection of cases where most loci phase easily but some do not, indicating a likely problem in the homology of the underlying sequences. This application can be extended as a hypothesis-testing tool via model comparison (see "Experiment 2: Testing Bayes factors to distinguish allelic variation from homeologs"), which provides an objective, powerful, and statistically rigorous method of understanding these complex patterns of homology. In the case of our empirical Cystopteridaceae dataset, this data exploration approach revealed cases of alleles masquerading as homeologs (*e.g.*, in *Cystopteris tasmanica*) and of hidden paralogy (in *Gymnocarpium disjunctum* and *dryopteris*). The identification and correction of these errors resulted in important changes to our downstream phylogenetic inference, demonstrating the impact that this form of data exploration can have.

Summary

`homologizer` provides a powerful and flexible method of phasing gene copies into subgenomes, allowing users to infer a multilocus phylogeny for lineages with polyploids (or F1 hybrids) while treating the phasing as a nuisance parameter, to infer the phasing while integrating out the phylogeny, or for any combination of these goals. Our simulations demonstrate that `homologizer` is able to confidently infer phasing even in cases of relatively weak phylogenetic signal (*e.g.*, short loci); in such cases where the signal is insufficient, `homologizer` returns an equivocal result (all potential phasings have lower posterior probability). In contrast, our simulations also demonstrate that `homologizer` can be sensitive to model violations. Specifically, if the true gene trees differ strongly from each other (*e.g.*, due to high levels of ILS) and yet the model assumes all loci share a single phylogeny, the method is forced to choose among a small set of more-or-less equally wrong phasing options, and may return the incorrect phasing with a high posterior probability (*i.e.*, the best of a bad lot). Our analyses of the empirical Cystopteridaceae dataset further demonstrate the power of the method; by simultaneously leveraging the phylogenetic information available in the full taxon sample (diploids and polyploids alike), across all loci, `homologizer` can confidently infer the phylogeny and phasing in areas of the tree where there is no diploid representation, and for markers where there is extensive missing data. As such, `homologizer` opens up the application of multilocus phylogenetics to groups containing polyploids or other individual organisms that may contain subgenomes with divergent evolutionary histories—such groups have been historically understudied, phylogenetically, which has limited our ability to explore evolutionary questions in general, and questions related to the macroevolution of polyploids specifically. Finally, `homologizer` may potentially be used beyond its core phasing functionality to identify non-homologous sequences more generally, with broad applications for phylogenetic and phylogenomic inference.

Acknowledgements

The authors would like to thank Bruce Baldwin for helpful discussions that inspired this work, and the Rothfels lab for comments that significantly improved the manuscript.

References

Alexandrou, M. A., Swartz, B. A., Matzke, N. J., and Oakley, T. H. (2013). Genome duplication and multiple evolutionary origins of complex migratory behavior in Salmonidae. *Molecular Phylogenetics and Evolution*, 69(3):514–523.

Beck, J. B., Windham, M. D., Yatskievych, G., and Pryer, K. M. (2010). A diploids-first approach to species delimitation and interpreting polyploid evolution in the fern genus *Astrolepis* (Pteridaceae). *Systematic Botany*, 35(2):223–234.

Bertrand, Y. J. K., Scheen, A.-C., Marcussen, T., Pfeil, B. E., de Sousa, F., and Oxelman, B. (2015). Assignment of homoeologues to parental genomes in allopolyploids for species tree inference, with an example from *Fumaria* (Papaveraceae). *Systematic Biology*, 64(3):448–471.

Bogart, J. P. (1980). *Evolutionary Implications of Polyploidy in Amphibians and Reptiles*, pages 341–378. Springer US, Boston, MA.

Bogart, J. P. and Licht, L. E. (1986). Reproduction and the origin of polyploids in hybrid salamanders of the genus *Ambystoma*. *Canadian Journal of Genetics and Cytology*, 28(4):605–617.

Brownlie, G. (1958). Chromosome numbers in New Zealand ferns. In *Transactions of the Royal Society of New Zealand*, volume 85, pages 213–216.

Brownsey, P. J. and Perrie, L. R. (2018). Fascicle 23, Cystopteridaceae. In *Flora of New Zealand: Ferns and Lycophytes*, pages 1–20. Manaaki Whenua Press, Lincoln, N.Z.

Chrtek, J., Mráz, P., Belyayev, A., Paštová, L., Mrázová, V., Caklová, P., Josefiová, J., Zagorski, D., Hartmann, M., Jandová, M., Pinc, J., and Fehrer, J. (2019). Evolutionary history and genetic diversity of apomictic allopolyploids in *Hieracium s.str.*: Morphological versus genomic features. *American Journal of Botany*, 107(1):66–90.

Dauphin, B., Grant, J., Farrar, D., and Rothfels, C. J. (2018). Rapid allopolyploid radiation of moonwort ferns (*Botrychium*; Ophioglossaceae) revealed by PacBio sequencing of homologous and homeologous nuclear regions. *Molecular Phylogenetics and Evolution*, 120.

Edger, P. P., McKain, M. R., Bird, K. A., and VanBuren, R. (2018). Subgenome assignment in allopolyploids: challenges and future directions. *Current opinion in plant biology*, 42:76–80.

Fortune, P. M., Pourtau, N., Viron, N., and Ainouche, M. L. (2008). Molecular phylogeny and reticulate origins of the polyploid *Bromus* species from section *Genea* (Poaceae). *American Journal of Botany*, 95(4):454–464.

Fraser-Jenkins, C. R. (2008). *Taxonomic revision of three hundred Indian subcontinental pteridophytes with a revised census-list: A new picture of fern-taxonomy and nomenclature in the Indian subcontinent*. Bishen Singh Mahendra Pal Singh.

Fraser-Jenkins, C. R., Pangtey, Y. P. S., and Kullar, S. P. (2010). *Asplenium laciniatum* D. Don (Aspleniaceae), a critical complex, and confused species in the Indian Subcontinent. *Indian Fern J*, 27:212–214.

Freyman, W. A. and Höhna, S. (2018). Cladogenetic and anagenetic models of chromosome number evolution: a Bayesian model averaging approach. *Systematic Biology*, 67(2):195–215.

Govindarajulu, R., Hughes, C. E., and Bailey, C. D. (2011). Phylogenetic and population genetic analyses of diploid *Leucaena* (Leguminosae; Mimosoideae) reveal cryptic species diversity and patterns of divergent allopatric speciation. *American Journal of Botany*, 98(12):2049–2063.

Griffin, P. C., Robin, C., and Hoffmann, A. A. (2011). A next-generation sequencing method for overcoming the multiple gene copy problem in polyploid phylogenetics, applied to *Poa* grasses. *BMC Biology*, 9(1):19.

Hedges, S. B., Bogart, J. P., and Maxson, L. R. (1992). Ancestry of unisexual salamanders. *Nature*, 356(6371):708–710.

Hénocq, L., Gallina, S., Schmitt, E., Castric, V., Vekemans, X., and Poux, C. (2020). A new tree-based methodological framework to infer the evolutionary history of mesopolyploid lineages: An application to the Brassiceae tribe (Brassicaceae). *bioRxiv*.

Höhna, S., Landis, M. J., Heath, T. A., Boussau, B., Lartillot, N., Moore, B. R., Huelsenbeck, J. P., and Ronquist, F. (2016). RevBayes: Bayesian phylogenetic inference using graphical models and an interactive model-specification language. *Systematic Biology*, 65(4):726–736.

Huber, K. T., Oxelman, B., Lott, M., and Moulton, V. (2006). Reconstructing the evolutionary history of polyploids from multilabeled trees. *Molecular Biology and Evolution*, 23(9):1784–1791.

Jones, G. (2017). Bayesian phylogenetic analysis for diploid and allotetraploid species networks. *bioRxiv*, 8(4):1015–1029.

Jones, G., Sagitov, S., and Oxelman, B. (2013). Statistical inference of allopolyploid species networks in the presence of incomplete lineage sorting. *Systematic Biology*, 62(3):467–478.

Kao, T.-T., Rothfels, C. J., Melgoza-Castillo, A., Pryer, K. M., and Windham, M. D. (2020). Infraspecific diversification of the star cloak fern (*Notholaena standleyi*) in the deserts of the United States and Mexico. *American Journal of Botany*, 107(4):658–675.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795.

Kates, H. R., Johnson, M. G., Gardner, E. M., Zerega, N. J. C., and Wickett, N. J. (2018). Allele phasing has minimal impact on phylogenetic reconstruction from targeted nuclear gene sequences in a case study of *Artocarpus*. *American Journal of Botany*, 105(3):404–416.

Lautenschlager, U., Wagner, F., and Oberprieler, C. (2020). AllCoPol: Inferring allele co-ancestry in polyploids. *BMC Bioinformatics*, pages 1–9.

Lee, J., Baldwin, B. G., and Gottlieb, L. D. (2002). Phylogeny of *Stephanomeria* and related genera (Compositae-Lactuceae) based on analysis of 18S-26S nuclear rDNA ITS and ETS sequences. *American Journal of Botany*, 89(1):160–168.

Lewis, P. O., Holder, M. T., and Holsinger, K. E. (2005). Polytomies and Bayesian phylogenetic inference. *Systematic biology*, 54(2):241–253.

Li, Z., Tiley, G. P., Galuska, S. R., Reardon, C. R., Kidder, T. I., Rundell, R. J., and Barker, M. S. (2018). Multiple large-scale gene and genome duplications during the evolution of hexapods. *Proceedings of the National Academy of Sciences*, 115(18):4713–4718.

Lowe, C. H. and Wright, J. W. (1966). Evolution of parthenogenetic species of *Cnemidophorus* (whiptail lizards) in western North America. *Journal of the Arizona Academy of Science*, 4(2):81–87.

Majidian, S., Kahaei, M. H., and de Ridder, D. (2020). Hap10: Reconstructing accurate and long polyploid haplotypes using linked reads. *bioRxiv*, 8(3):1520–1534.

Marchant, D. B., Soltis, D. E., and Soltis, P. S. (2016). Patterns of abiotic niche shifts in allopolyploids relative to their progenitors. *New Phytologist*.

Mayrose, I., Zhan, S., Rothfels, C. J., Arrigo, N., Barker, M., Rieseberg, L. H., and Otto, S. P. (2015). Methods for studying polyploid diversification and the dead end hypothesis: A reply to Soltis et al. (2014). *New Phytologist*, 206(1).

Mayrose, I., Zhan, S., Rothfels, C. J., Magnuson-Ford, K., Barker, M., Rieseberg, L. H., and Otto, S. (2011). Recently formed polyploid plants diversify at lower rates. *Science*, 333(6047).

McDade, L. (1990). Hybrids and phylogenetic systematics I. Patterns of character expression in hybrids and their implications for cladistic analysis. *Evolution*, pages 1685–1700.

McDade, L. A. (1992). Hybrids and phylogenetic systematics II. The impact of hybrids on cladistic analysis. *Evolution*, pages 1329–1346.

Melichárková, A., Španiel, S., Marhold, K., Hurdu, B.-I., Drescher, A., and Zozomová-Lihová, J. (2019). Diversification and independent polyploid origins in the disjunct species *Alyssum repens* from the Southeastern Alps and the Carpathians. *American Journal of Botany*, 106(11):1499–1518.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092.

Nauheimer, L., Weigner, N., Joyce, E., Crayn, D., Clarke, C., and Nargar, K. (2021). HybPhaser: A workflow for the detection and phasing of hybrids in target capture data sets. *Applications in Plant Sciences*, 9(7).

Oberprieler, C., Wagner, F., Tomasello, S., and 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(7):835–849.

Oxelman, B., Brysting, A. K., Jones, G. R., Marcussen, T., Oberprieler, C., and Pfeil, B. E. (2017). Phylogenetics of allopolyploids. *Annual Review of Ecology, Evolution, and Systematics*, 48(1):543–557.

Parks, M. B., Nakov, T., Ruck, E. C., Wickett, N. J., and Alverson, A. J. (2018). Phylogenomics reveals an extensive history of genome duplication in diatoms (Bacillariophyta). *American Journal of Botany*, 105(3):330–347.

Pryer, K. M. and Haufler, C. H. (1993). Isozymic and chromosomal evidence for the allotetraploid origin of *Gymnocarpium dryopteris* (Dryopteridaceae). *Systematic Botany*, 18(1):150–172.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ramsey, J. and Ramsey, T. S. (2014). Ecological studies of polyploidy in the 100 years following its discovery. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 369(1648):20130352.

Ramsey, J. and Schemske, D. W. (2002). Neopolyploidy in flowering plants. *Annual Review of Ecology and Systematics*, 33(1):589–639.

Robinson, D. F. and Foulds, L. R. (1981). Comparison of phylogenetic trees. *Mathematical Biosciences*, 53(1-2):131–147.

Rothfels, C. J. (2021). Polyploid phylogenetics. *New Phytologist*, 230(1):66–72.

Rothfels, C. J., Johnson, A., Windham, M., and Pryer, K. (2014). Low-copy nuclear data confirm rampant allopolyploidy in the Cystopteridaceae (Polypodiales). *Taxon*, 63(5).

Rothfels, C. J., Johnson, A. K., Hovenkamp, P. H., Swofford, D. L., Roskam, H. C., Fraser-Jenkins, C. R., Windham, M. D., and Pryer, K. M. (2015). Natural hybridization between parental lineages that diverged approximately 60 million years ago. *American Naturalist*, 185(3):433–442.

Rothfels, C. J., Larsson, A., Kuo, L.-Y., Korall, P., Chiou, W.-L., and Pryer, K. M. (2012). Overcoming deep roots, fast rates, and short internodes to resolve the ancient rapid radiation of eupolypod II ferns. *Systematic Biology*, 61(3):490.

Rothfels, C. J., Pryer, K., and Li, F.-W. (2017). Next-generation polyploid phylogenetics: rapid resolution of hybrid polyploid complexes using PacBio single-molecule sequencing. *New Phytologist*, 213(1).

Rothfels, C. J., Windham, M. D., and Pryer, K. M. (2013). A plastid phylogeny of the cosmopolitan fern family Cystopteridaceae (Polypodiopsida). *Systematic Botany*, 38(2):295–306.

Rousseau-Gueutin, M., Gaston, A., Aïnouche, A., Aïnouche, M. L., Olbricht, K., Staudt, G., Richard, L., and Denoyes-Rothan, B. (2009). Tracking the evolutionary history of polyploidy in *Fragaria* L. (strawberry): New insights from phylogenetic analyses of low-copy nuclear genes. *Molecular Phylogenetics and Evolution*, 51(3):515–530.

Såstad, S. (2005). Patterns and mechanisms of polyploid speciation in bryophytes. *Regnum Vegetabile*, 143:317–334.

Schrinner, S. D., Mari, R. S., Ebler, J., Rautiainen, M., Seillier, L., Reimer, J. J., Usadel, B., Marschall, T., and Klau, G. W. (2020). Haplotype threading: Accurate polyploid phasing from long reads. *bioRxiv*, 57(3-4):324–327.

Sessa, E. B., Zimmer, E. A., and Givnish, T. J. (2012a). Reticulate evolution on a global scale: A nuclear phylogeny for New World *Dryopteris* (Dryopteridaceae). *Molecular Phylogenetics and Evolution*, 64(3):563–581.

Sessa, E. B., Zimmer, E. A., and Givnish, T. J. (2012b). Unraveling reticulate evolution in North American *Dryopteris* (Dryopteridaceae). *BMC Evolutionary Biology*, 12(1):104.

Soltis, D. E., Albert, V. A., Leebens-Mack, J., Bell, C. D., Paterson, A. H., Zheng, C., Sankoff, D., DePamphilis, C. W., Wall, P. K., and Soltis, P. S. (2009). Polyploidy and angiosperm diversification. *American Journal of Botany*, 96(1):336–348.

Soltis, D. E., Segovia-Salcedo, M. C., Jordon-Thaden, I., Majure, L., Miles, N. M., Mavrodiev, E. V., Mei, W., Cortez, M. B., Soltis, P. S., and Gitzendanner, M. A. (2014). Are polyploids really evolutionary dead-ends (again)? A critical reappraisal of Mayrose et al.(2011). *New Phytologist*, 202(4):1105–1117.

Sousa, F., Bertrand, Y. J. K., and Pfeil, B. E. (2016). Patterns of phylogenetic incongruence in *Medicago* found among six loci. *Plant Systematics and Evolution*, pages 1–21.

Stebbins Jr, G. L. (1940). The significance of polyploidy in plant evolution. *American Naturalist*, pages 54–66.

Stebbins Jr, G. L. (1985). Polyploidy, hybridization, and the invasion of new habitats. *Annals of the Missouri Botanical Garden*, 72(4):824–832.

Tank, D. C., Eastman, J. M., Pennell, M. W., Soltis, P. S., Soltis, D. E., Hinchliff, C. E., Brown, J. W., Sessa, E. B., and Harmon, L. J. (2015). Nested radiations and the pulse of angiosperm diversification: Increased diversification rates often follow whole genome duplications. *New Phytologist*, 207(2):454–467.

Tavaré, S. (1986). Some probabilistic and statistical problems in the analysis of DNA sequences. *Lectures on mathematics in the life sciences*, 17(2):57–86.

Tiley, G. P., Crowl, A. A., Manos, P. S., Sessa, E. B., Solis-Lemus, C., Yoder, A. D., and Burleigh, J. G. (2021). Phasing alleles improves network inference with allopolyploids. *bioRxiv*.

Tindale, M. D. and Roy, S. K. (2002). A cytotaxonomic survey of the Pteridophyta of Australia. *Australian Systematic Botany*, 15:839–937.

Tribble, C. et al. (2020). RevGadgets: an R Package for visualizing Bayesian phylogenetic analyses from RevBayes. *MEE*.

Wagner Jr., W. H. (1970). Biosystematics and evolutionary noise. *Taxon*, 19:146–151.

Wood, T. E., Takebayashi, N., Barker, M. S., Mayrose, I., Greenspoon, P. B., and Rieseberg, L. H. (2009). The frequency of polyploid speciation in vascular plants. *Proceedings of the National Academy of Sciences, USA*, 106(33):13875–13879.

Xie, W., Lewis, P. O., Fan, Y., Kuo, L., and Chen, M.-H. (2011). Improving marginal likelihood estimation for Bayesian phylogenetic model selection. *Systematic biology*, 60(2):150–160.

Yang, Z. (2007). Fair-balance paradox, star-tree paradox, and Bayesian phylogenetics. *Molecular biology and evolution*, 24(8):1639–1655.

Zenil-Ferguson, R., Burleigh, J. G., Freyman, W. A., Igić, B., Mayrose, I., and Goldberg, E. E. (2019). Interaction among ploidy, breeding system and lineage diversification. *New Phytologist*, 224(3):1252–1265.

Zhan, S. H., Glick, L., Tsigenopoulos, C. S., Otto, S. P., and Mayrose, I. (2014). Comparative analysis reveals that polyploidy does not decelerate diversification in fish. *J Evolution Biol*, 27(2):391–403.