bioRxiv preprint doi: https://doi.org/10.1101/2020.10.22.351486; this version posted October 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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 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.

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

3. Through the use of simulation tests we demonstrate that `homologizer` is robust to a wide range of factors, such as the phylogenetic informativeness of loci and incomplete lineage sorting. 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, contaminants, or allelic variation that was erroneously modelled as homeologous.

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. 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 multi-species coalescent analysis), the underlying bifurcating model will be violated and the inferences unreliable (McDade, 1990, 1992; Oxelman et al., 2017; Rothfels, 2020). 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). 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 their 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

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.22.351486; this version posted October 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

(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 and don't attempt true multilocus analyses (*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're not just removing the information those accessions could provide about general evolutionary patterns. Rather, we're 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, 2020). 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 (Marchant et al., 2016; Ramsey and Schemske, 2002) or with dispersal and establishment ability (Stebbins Jr, 1985).

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 Hénocq et al. (2020), there are, to our knowledge, three currently available methods for phasing copies across loci. First, Bertrand et al. (2015) developed an approach to infer which homeolog from a low-copy nuclear locus shares an evolutionary history with the plastid (*i.e.*, is the maternal copy, assuming maternal plastid inheritance). Their method is based, first, on determining the largest sample of accessions for which there is a nuclear tree (considering each possible phasing of the sequences in that tree) that is not strongly supported as incongruent with the plastid tree (*i.e.*, where an SH test (Shimodaira and Hasegawa, 1999) fails to detect significant incongruence at an $P = .05$ level.) For accessions not included in this primary set, they inferred the optimal phasing as the one that, when added to the primary set of plastid and nuclear trees, had the closest match in likelihood scores when the two trees were constrained to have the same topology and when they were free to vary. These "secondary correspondence hypotheses" (the phasing of nuclear copies with the plastid copy for that accession) were then accepted if the incongruence between the two trees could be a result of stochastic substitution error or coalescent error (*i.e.*, if tests based on parametric bootstrapping were unable to reject both sources of error). This approach allowed Bertrand et al. (2015) to infer a two-locus phylogeny for 74 accessions of the genus *Fumaria*, including samples of ploidy levels up to 14x. However, it is combinatorically and computationally intensive, and would be difficult to extend beyond two markers, or potentially even to two low-copy nuclear markers (rather than one plastid and one nuclear).

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 deep coalescence in the context of the phylogeny of the diploid species. So if a tetraploid has four sequences at a locus—A1 through A4—the method considers three sets of subgenome "species" (A1+A2 and A3+A4; A1+A3 and A2+A4; etc) and for each runs a parsimony-based minimizing-deep-coalescence (MDC) analysis (Than and Nakhleh, 2009) of that set of subgenome "species" with the set of sampled diploids, to see which allocation of copies to subgenomes implies the least incomplete lineage sorting. This process is then repeated across loci—if the selected pairing for locus A was A1+A2 and A3+A4 and for locus B was B1+B3 and B2+B4, the MDC analyses would be run twice, with the species in the first run being A1+A2+B1+B3 and A3+A4+B2+B4, and in the second run being A1+A2+B2+B4 and A3+A4+B1+B3. This method is relatively fast, although the computational demands increase significantly with the number of polyploid accessions, and with the number of loci (Oberprieler et al.,

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.22.351486; this version posted October 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

2017). Perhaps more fundamentally, this method tackles each polyploid accession individually, so no information is available to be shared across polyploid samples, and suffers from some of the limitations of parsimony, such as the absence of a natural way to assess the statistical significance of the results.

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, does not provide an estimate of clade support, and is computationally demanding, so can only be used on relatively small datasets. Finally, in the specific context of this paper, `alloPPnet` solves the phasing-across-loci issue with a permutation move that switches the labels of the putative homeologous sequences for a given locus, and then either accepts or rejects that move as part of a Markov chain Monte Carlo (MCMC; Metropolis et al., 1953) analysis; unfortunately, however, it does not log samples of these phasing assignments during the MCMC, so `alloPPnet` runs cannot be used to provide phasing information for other analyses.

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) 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, 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 in a polyploid, 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 the other parameters of the model. The posterior distributions of phasing can be summarized in a number of ways useful for downstream analyses, as we describe below.

In the `homologizer` model we represent gene trees 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 accessions' tips in the mul-tree (Figure 1). The posterior probability of each possible assignment of gene copy to subgenome is estimated by MCMC. Higher ploidy levels can be accommodated by increasing the number of mul-tree tips (for that sample) among which gene copies may be swapped and multiple polyploids can be easily incorporated within a single analysis by defining different sets of mul-tree tips (subgenomes) within which gene copies may be swapped. For example, an analysis could include a hexaploid represented by three tips in the mul-tree and a tetraploid represented by two tips; the gene copies of each polyploid can only be swapped among that polyploid's respective tips.

Our approach allows for the topology and branch lengths of the gene trees to be either fixed *a priori* or estimated simultaneously with the phasing using standard Bayesian phylogenetic approaches as implemented in `RevBayes` (Höhna et al., 2016). Loci can be defined to share a single gene tree, share a set of gene trees,

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.22.351486; this version posted October 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

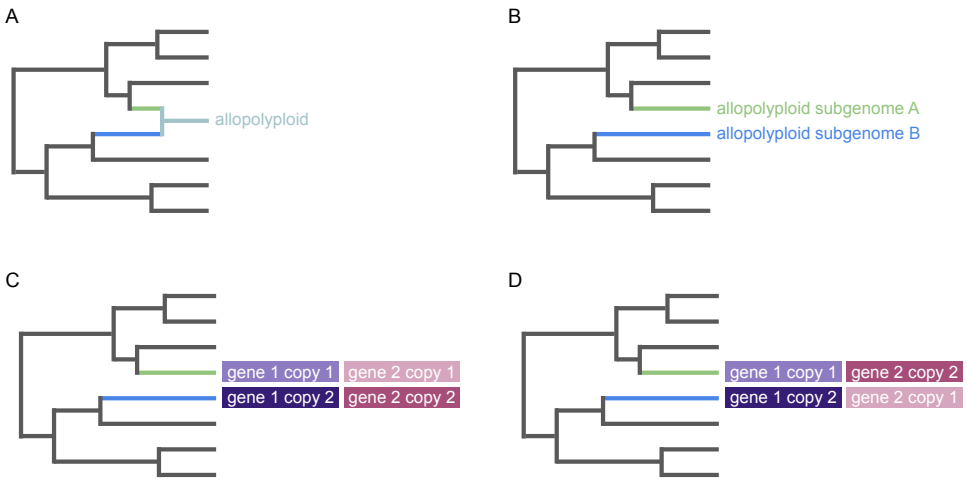

Figure 1: Phasing gene copies into polyploid subgenomes on a mul-tree. *A*: A phylogenetic network representing a single reticulation giving rise to an allopolyploid. *B*: The mul-tree representation of this phylogenetic network has two leaves representing the two subgenomes of the allopolyploid. *C*: Two loci (gene 1 and gene 2) were sequenced from the allopolyploid. Two copies of each locus were recovered. One possible phase of the gene copies into the allopolyploid subgenomes assigns gene 1 copy 1 and gene 2 copy 1 to subgenome A and gene 1 copy 2 and gene 2 copy 2 to subgenome B. *D*: Another possible phasing. For each polyploid, there are $(g!)^n$ ways to phase the gene copies, where g is the number of subgenomes and n is the number of loci.

or assumed to be unlinked, each with an independent gene tree. These gene trees can optionally be embedded within a species trees (that itself can be set *a priori* or co-estimated) under the multispecies coalescent (MSC; Yang and Rannala, 2010) to account for incomplete lineage sorting. It is important to note, however, that the use of the MSC with `homologizer` is an imperfect approximation of the full MSC because (1) the priors on the species tree do not explicitly model the hybridization process (see Zhang et al., 2018) and (2) the model is not aware of which lineages in the mul-tree are "homeologous" lineages that should share population size parameters (cf. Jones, 2017).

MCMC initialization and proposals

In `RevBayes` phylogenetic models are specified as graphical models (Höhna et al., 2014) under which inference is performed using MCMC. Aligned sequence data is modeled as a stochastic variable drawn from a phylogenetic continuous-time Markov chain (CTMC) process. In a typical phylogenetic analysis the value of this stochastic variable is fixed (or "clamped") to an observed sequence alignment, enabling inference of the tree topology, branch lengths, molecular substitution model parameters, etc. In a `homologizer` analysis an extra latent variable is estimated that indicates which observed sequence is assigned to each subgenome of the polyploid (different tips in the mul-tree). These assignments are sampled during the MCMC analysis to estimate the posterior distribution of gene copies phased into subgenomes.

To initialize a `homologizer` MCMC analysis, the user must use the `setHomeologPhase()` command to provide an initial assignment of each gene copy sequence to a subgenome tip in the phylogeny; this initial phasing assignment can be randomly picked by the user, and provides an opportunity to name the mul-tree tips such that they have different names from the gene-copy names in the alignment. For example, if there is a gene copy from a polyploid that is called `copy_1`, it could be assigned to a tip of the mul-tree representing the `A` subgenome of a polyploid with the command `setHomeologPhase("copy_1", "subgenome_A")`. Similar assignments are made so that each polyploid gene copy has a preliminary subgenome assignment. The observed sequence data alignment with its initial phase assignment is then clamped to a phylogenetic CTMC distribution.

MCMC operators that propose new phasing assignments for gene copies are then applied to the phylogenetic CTMC distribution. These proposals are created using the command `mvHomeologPhase`. For example, `mvHomeologPhase(locus_1, "subgenome_A", "subgenome_B")` creates a proposal that will swap the sequences assigned to the `subgenome_A` and `subgenome_B` tips of the mul-tree for the phylogenetic CTMC stochastic variable called `locus_1`. Each proposal enables a pair of tips to swap their assigned sequences for a given locus. This means that for samples with N subgenomes, $\binom{N}{2}$ proposals must be defined to enable all possible com-

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.22.351486; this version posted October 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

binations of gene copy to subgenome tip assignments to be explored by the MCMC. Multilocus analyses often include multiple phylogenetic CTMCs (*e.g.*, one for each locus), in which case these proposals are defined for each CTMC variable. Since this move swaps the sequences between a pair of tips with equal probability, it is a symmetric Metropolis move with a Hasting ratio of 1.

Samples drawn from the posterior distribution of phasing assignments are written to file during the MCMC using the command `mnHomeologPhase`, for example `mnHomeologPhase(filename="output/locus_1.log", ctmc_locus_1)`. One such monitor is defined for each locus being phased.

Summarizing posterior distributions

We summarize the posterior distribution of phasing assignments with 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. 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 and so is less useful.

The marginal posterior probability for each locus and each subgenome is, however, useful when quantifying the uncertainty in the joint MAP phasing assignment. For example, imagine a hexaploid with three 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.

Software availability

`homologizer` is implemented in the open-source Bayesian inference software `RevBayes` (`http://revbayes.com`; Höhna et al., 2016). New functionality will be added to the open-source R package `RevGadgets` so it can be used to summarize the output from `RevBayes` and generate summary plots similar to Figure 3. `RevGadgets` is available in the code repository: `https://github.com/revbayes/revgadgets`. Further details, including instructions for formatting input files and a full example analysis will be made available as a `RevBayes` tutorial at: `http://revbayes.com`.

Simulation tests

To explore the behavior of `homologizer` under different evolutionary scenarios, we used simulated datasets to test how the performance of `homologizer` was impacted by the following factors: (1) the phylogenetic informativeness of each locus; (2) the amount of incomplete lineage sorting (ILS) present; (3) the phylogenetic distance between each polyploid's subgenomes (the minimum distance in cases of more than two subgenomes), and; (4) the minimum phylogenetic distance from a polyploid subgenome to its nearest diploid. For each simulation 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 had twenty tips, five of which were randomly selected to represent allopolyploid subgenomes. Two of the allopolyploid tips were 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. 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.

To investigate each of the four factors listed above, we

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.22.351486; this version posted October 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

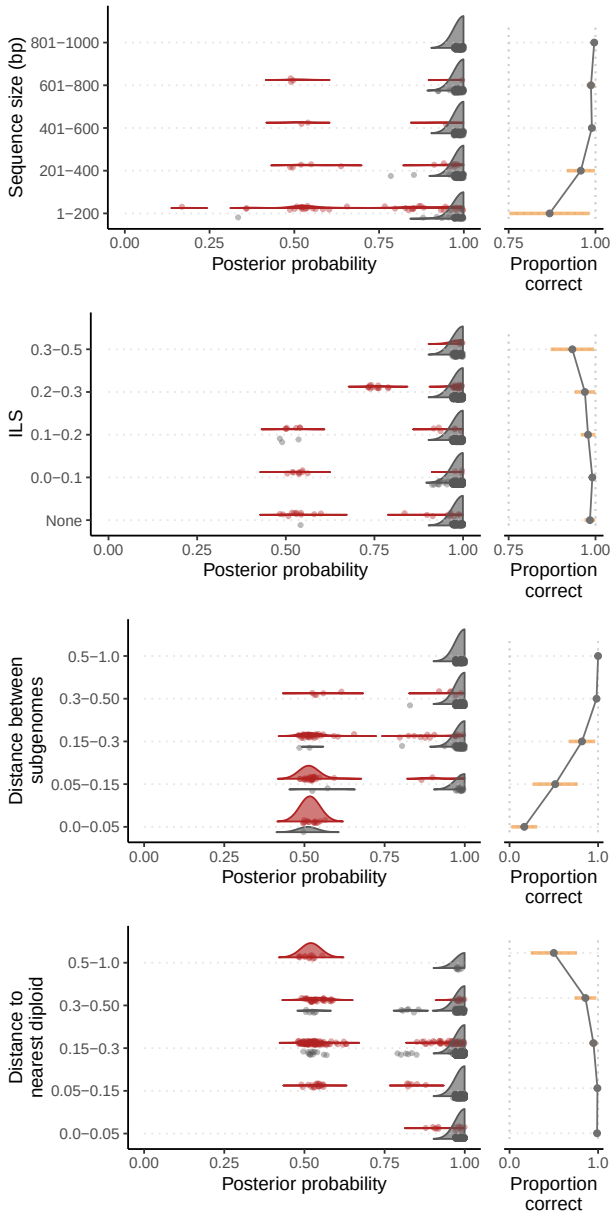

Figure 2: Performance of `homologizer` in phasing four loci under different simulated conditions. Each row shows how phasing performance was impacted by a factor expected to influence accuracy: sequence length; incomplete lineage sorting; smallest distance between subgenomes; and smallest distance between a polyploid subgenome and a sampled diploid. 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.

performed a series of simulations varying the focal factor while keeping the other factors at fixed levels. Those fixed values were a sequence length of 800 base pairs, an effective population 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 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. Finally, for each experiment we tracked the proportion of simulated polyploids for which the phasing was correctly estimated, 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 varying amounts 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. To test the effect of the phylogenetic distance between polyploid subgenomes we simulated 800 simulation replicates, without placing any constraints on the maximum subgenome distance. Finally, to test the effect of the phylogenetic distance between each polyploid and the closest diploid tip we simulated 800 simulation replicates, this time without constraints on this "diploid distance."

Simulation results and discussion

`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 2). 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

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.22.351486; this version posted October 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

subgenomes was very small, or the distance from the polyploid subgenomes to the nearest diploid was very large (Figure 2, bottom two rows). In both those cases, the polyploid subgenomes are likely to be sister to each other and thus there is little to no topological information 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 2).

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 2, second 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).

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 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 (Figure 3).

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 reanalyzed these data with ×*Cystocarpium roskamianum* and one accession of *Gymnocarpium dryopteris* allotted four slots in the mul-tree (to accommodate the homeologs of ×*Cystocarpium* and the allelic variation observed in the *Gymnocarpium*, where three sequences were recovered from each of two loci), and all other tetraploids allotted two slots. The diploid samples were each given a single slot, 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, and a fourth (*Gymnocarpium disjunctum* 7751) that had two copies at each of two loci, but it wasn't clear if these were alleles or not; to accommodate the possibility that this accession might actually be a polyploid with homeologous copies, it was allocated two swapping tips. The final dataset comprises 4068 aligned sites for 53 mul-tree tips (*i.e.*,, subgenomes; Table 1), with 27 of the 212 sequences being blank (composed entirely of "?"s added so that each locus had a sequence for each of the mul-tree tips, regardless of how many sequences were actually recovered from that sample for that locus).

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%

We inferred the phasing and the phylogeny simultaneously, linking the tree across loci. 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 branchlengths. This full `homologizer` model was run on one chain, for 10000 generations, with the first 50% of samples discarded prior to summarizing the posterior distribution.

Cystopteridaceae results and discussion

Our inferred Cystopteridaceae phylogeny is well supported (Figure 3), and generally consistent with the inferences of Rothfels et al. (2017). We do, however, find some differences, such as the *sudetica* clade sister to the *C. fragilis* complex instead of to the *bulbifera* clade, the *C. tasmanica* subgenomes somewhat phylogenetically separate from each other, rather than being sister lineages, and a different placement of the morphologically anomalous *G. oyamense* within *Gymnocarpium*. It is difficult to determine what is driving these topological differences, as the underlying analyses themselves 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

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.22.351486; this version posted October 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

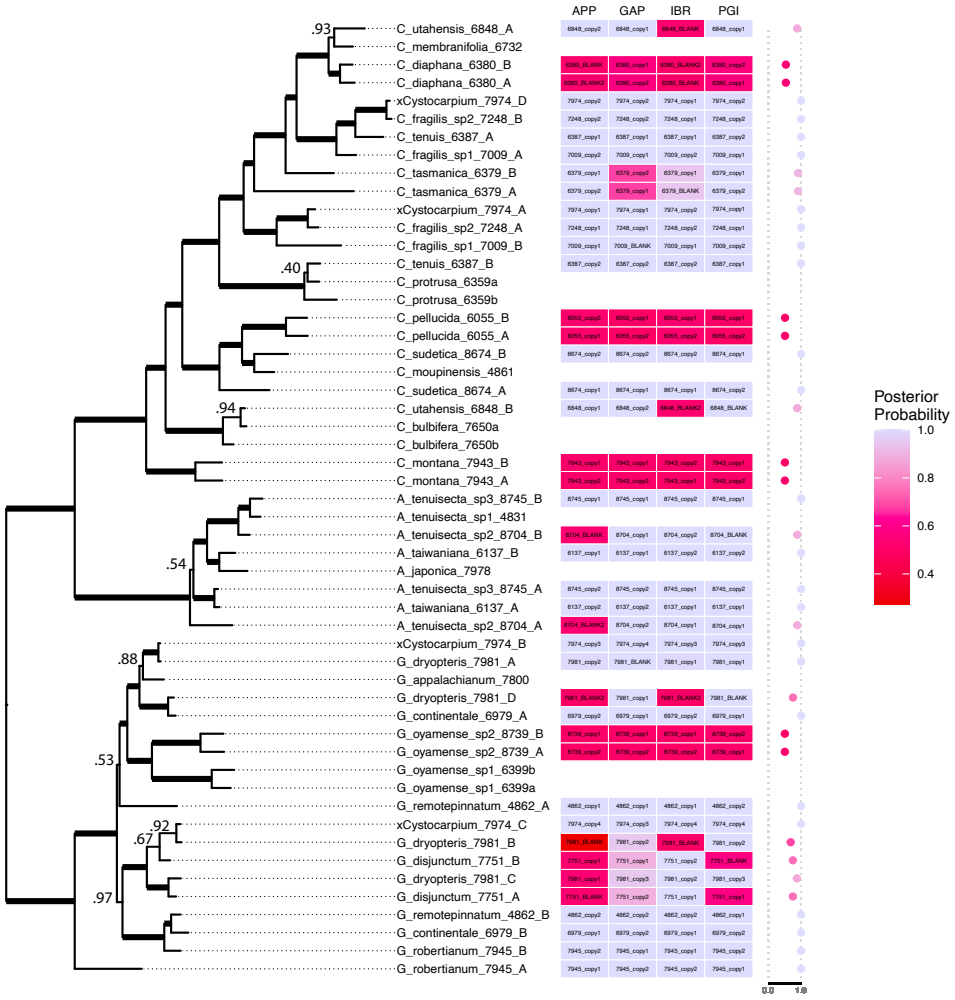

Figure 3: `homologizer` analysis of the Cystopteridaceae dataset. The phasing of gene copies into subgenomes is summarized on the maximum *a posteriori* (MAP) phylogeny. Thickened branches have posterior probabilities of 1.0; posterior probabilities < 1.0 are indicated. The columns of the heatmap each 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).

strongly supported in our analyses, `alloPPnet` doesn't report clade probabilities, so it is not clear how well supported the alternative relationships were in the earlier analysis.

The phasing is also generally well supported. 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 3). The two exceptions to this pattern are *C. tasmanica* for GAP and among the closely related *G. dryopteris* and *G. disjunctum* tips; in both cases `homologizer` does not confidently resolve the phasing, despite the presence of sequence data for at least one of the relevant copies. In the case of *C. tasmanica*, something appears unusual about the GAP sequences—while the other three loci phase well, this locus does not. This behaviour indicates another application of `homologizer`, which is as a data exploration tool: potentially the two *C. tasmanica* GAP copies are alleles from a single subgenome rather than homeologs representing two subgenomes. Similarly, there is uncertainty about the true number of evolutionary histories repre-

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.22.351486; this version posted October 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

sented by our *G. dryopteris* sequences. We gave this accession four tips in the mul-tree, even though no locus is represented by more than three sequences, given that it is not possible *a prior* to know whether or not the "third sequences" represent the same evolutionary history (are alleles of the same homeolog). One consequence of adding the fourth tip is that *G. dryopteris* has a lot of missing data, as do the closely related *G. disjunctum* tips—the uncertainty in phasing in this area of the tree may simply reflect a broader uncertainty in the phylogenetic relationships of these taxa (Figure 3).

Applications and future directions

The simulation tests performed here show that `homologizer` is robust to a wide range of factors that can potentially affect phasing accuracy. 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. The fourth factor impacting phasing accuracy that we examined was ILS. Unlike the other three factors, ILS can violate the model's assumptions. Since the `homologizer` model used in our simulation tests assumed a shared gene tree, under high levels of ILS the model sometimes made incorrect estimates with high posterior probability ("confidently wrong" behavior). We recommend that researchers assess the degree of gene-tree incongruence in their system independently before performing a `homologizer` analysis. If there is a low to moderate amount of incongruence then it is likely safe to proceed with the shared gene tree assumption presented here; if ILS, specifically, is thought to be important, 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.

Large genome-scale target enrichment and transcriptome datasets often contain hundreds of loci. When those datasets contain polyploids a crucial step for many downstream analyses is the phasing of hundreds of gene copies into their respective polyploid subgenomes. 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 for scaling up phasing analyses that should be suitable for any sized dataset.

Beyond its core functionalities of inferring multilocus mul-tree phylogenies while accounting for uncertainty in phasing, and inferring phasing while accounting for uncertainty in phylogeny, `homologizer` has other potentially useful applications. One application, as already discussed in the context of the Cystopteridaceae analyses, is as a data exploration tool. With increasing numbers of taxa and loci, it can be difficult to detect non-homologous 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 by statistically comparing the fit of `homologizer` models that differ in the number of mul-tree tips that represent the same accession. Extra "dummy tips" that potentially represent allelic variation or hidden paralogs may be added to the `homologizer` mul-tree, phasing gene copies into. Bayes factors can be used to compare `homologizer` models with and without these additional mul-tree tips to test for the presence of non-homeologous sequences.

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 fur-

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.22.351486; this version posted October 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

ther 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.

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., Heath, T. A., Boussau, B., Landis, M. J., Ronquist, F., and Huelsenbeck, J. P. (2014). Probabilistic graphical model representation in phylogenetics. *Systematic biology*, 63(5):753–771.

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.

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.

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.22.351486; this version posted October 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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.

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.

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. (2020). Polyploid phylogenetics. *New Phytologist*.

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).

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.

Shimodaira, H. and Hasegawa, M. (1999). Multiple comparisons of log-likelihoods with applications to phylogenetic inference. *Molecular biology and evolution*, 16(8):1114–1114.

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.

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.22.351486; this version posted October 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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.

Than, C. and Nakhleh, L. (2009). Species tree inference by minimizing deep coalescences. *PLoS Computational Biology*, 5(9):e1000501–12.

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.

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

Yang, Z. and Rannala, B. (2010). Bayesian species delimitation using multilocus sequence data. *Proceedings of the National Academy of Sciences*, 107(20):9264–9269.

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.

Zhang, C., Ogilvie, H. A., Drummond, A. J., and Stadler, T. (2018). Bayesian inference of species networks from multilocus sequence data. *Molecular Biology and Evolution*, 35(2):504–517.