Molecular Phylogenetics and Evolution 196 (2024) 108087

Contents lists available at ScienceDirect

Molecular Phylogenetics and Evolution

journal homepage: www.elsevier.com/locate/ympev

Shared single copy genes are generally reliable for inferring phylogenetic relationships among polyploid taxa

Jaells G. Naranjo [a], Charles B. Sither [b], Gavin C. Conant [a,c,d,*]

[a] Bioinformatics Research Center, North Carolina State University, Raleigh, NC, USA
[b] Department of Entomology and Plant Pathology, North Carolina State University, Raleigh, NC, USA
[c] Genetics and Genomics Academy, North Carolina State University, Raleigh, NC, USA
[d] Department of Biological Sciences, North Carolina State University, Raleigh, NC, USA

ARTICLE INFO

Keywords:
Polyploidy
Reciprocal gene loss
Phylogenetic inference
Synteny

ABSTRACT

Polyploidy, or whole-genome duplication, is expected to confound the inference of species trees with phylogenetic methods for two reasons. First, the presence of retained duplicated genes requires the reconciliation of the inferred gene trees to a proposed species tree. Second, even if the analyses are restricted to shared single copy genes, the occurrence of reciprocal gene loss, where the surviving genes in different species are paralogs from the polyploidy rather than orthologs, will mean that such genes will not have evolved under the corresponding species tree and may not produce gene trees that allow inference of that species tree. Here we analyze three different ancient polyploidy events, using synteny-based inferences of orthology and paralogy to infer gene trees from nearly 17,000 sets of homologous genes. We find that the simple use of single copy genes from polyploid organisms provides reasonably robust phylogenetic signals, despite the presence of reciprocal gene losses. Such gene trees are also most often in accord with the inferred species relationships inferred from maximum likelihood models of gene loss after polyploidy: a completely distinct phylogenetic signal present in these genomes. As seen in other studies, however, we find that methods for inferring phylogenetic confidence yield high support values even in cases where the underlying data suggest meaningful conflict in the phylogenetic signals.

1. Introduction

Polyploidy events, also known as whole genome duplications, have occurred across the eukaryotic tree of life (Van de Peer et al., 2017). Such events effectively duplicate every gene in the genome, but, due to the resulting genetic redundancy, many or most of these duplicates are lost soon after the polyploidy event. Ancient polyploidy events (known as paleopolyploidies) complicate the use of genetic data for the inference of relationships among the resulting descendant species for two reasons. Firstly, the presence of duplicated genes raises the problem of reducing the inferred gene tree with duplicated genes into a species tree (Salichos and Rokas, 2013). While there are several approaches to this reconciliation task (Chen et al., 2000; Smith et al., 2022), including ones that treat polyploidy (Thomas et al., 2017), none of them yet represent a general solution to the problem. For instance, while birth–death models have been employed to model paleopolyploidies on a phylogeny (Rabier et al., 2014), the models used treat the polyploidy as a point event and do not allow for species-specific duplicate gene losses among the genomes sharing that event (Chen and Zwaenepoel, 2023; Scannell et al., 2006). Second, even if one restricts the analysis to genes that are single copy across all the species considered (an approach which itself implicitly assumes the possession of a genome or transcriptome for each taxa), the occurrence of reciprocal gene loss (RGL) can give rise to single copy paralogous genes among the genomes, in which case the inferred gene tree will not necessarily be reflective of the species tree (Scannell et al., 2007).

Inferring species trees from genetic data is most conceptually straightforward when gene trees are inferred from orthologs (Koonin, 2005): genes in different species that last shared common ancestors at their respective speciation events. Even for orthologous genes, phenomena such as incomplete lineage sorting (Madison and Knowles, 2006), introgression, and methodological weaknesses (Felsenstein, 1978) can give rise to gene tree inferences that differ from the species tree. One solution to this problem is to use large datasets of genes, an approach known as phylogenomics (Philippe, 2005). Such analyses can be performed via the concatenation of individual gene alignments into a

* Corresponding author at: Bioinformatics Research Center, North Carolina State University, Raleigh, NC, USA.
E-mail address: gconant@ncsu.edu (G.C. Conant).

https://doi.org/10.1016/j.ympev.2024.108087
Received 12 December 2023; Received in revised form 22 March 2024; Accepted 24 April 2024
Available online 26 April 2024
1055-7903/© 2024 Elsevier Inc. All rights reserved.

single larger alignment (Rokas et al., 2003). Alternatively, gene trees can be inferred for each individual gene and those gene trees can be reconciled into a single species tree under the assumption that coalescent processes have given rise to the gene tree incongruences (Liu et al., 2009).

With either phylogenomic approach, there is an implicit assumption that the genes used are orthologs, a fact which is potentially problematic for polyploid taxa. As a result, a few approaches for identifying appropriate genes for analyses where polyploidy is a confound have been suggested. In principle, the use of single copy genes avoids the need to reconcile duplicated gene trees onto a species tree (Philippe, 2005). However, without a complete genome, one cannot know *a priori* which genes these are. One solution to this difficulty has been to use homologous genes that have been found to be single copy across a range of taxa. For instance, Duarte et al., published a list of 959 single copy genes from *A. thaliana*, *P. trichocarpa*, *V. vinifera* and *O. sativa* (Duarte et al., 2010). De Smet et al., (2013) inferred a list of homologous genes rapidly returned to single copy after multiple polyploidy events. Because such recurrent return to single copy after independent polyploidies implies that genes in this list are under natural selection against duplication, this list in theory provides a set of genes that are expected to be single copy in most angiosperms. With access to at least two genomes and many transcriptomes from the species under analysis, one can use a combination of synteny-based orthology prediction and transcriptomic methods to generate likely orthologous genes for analysis (Washburn et al., 2017). Finally, the gene-tree/species-tree reconciliation approach seeks to resolve the set of observed tip genes, each from a known species, onto a species tree through the placement of gene duplication events (Chen et al., 2000; Smith et al., 2022). In some variants of this approach, the set of taxa known to share a polyploidy can be specified, allowing the duplication placement to correspond to a shared polyploidy event (Thomas et al., 2017).

In prior work, the effects of phenomena like RGL on the accuracy of species tree inference has been explored by simulating it on known phylogenies (Xiong et al., 2022) or through the use of gene-tree/species-tree reconciliation approaches (Smith et al., 2022; Thomas et al., 2017). Such work has the implicit focus of whether or not polyploidy makes uncovering the underlying species tree more difficult. Here, we ask subtly different questions: what types of gene trees are produced through evolution after polyploidy and at what relative frequency are they produced? To address these questions, we employ information that was not included in earlier studies, namely the synteny relationships between the genes in question. Using those data gives us access to "ground truth" orthology information upon which to build our gene tree comparisons. The orthology inferences come from our analysis of the genomes in question with POInT (the Polyploid Orthology Inference Tool), which uses the shared gene order and patterns of duplicate losses to make probabilistic inferences of the orthology relationships of all genes in a clade of species sharing a paleopolyploidy (Conant and Wolfe, 2008). Fig. 1 gives an example of a syntenic region with orthology inferences from POInT. The tool couples to a phylogenetic model of paralog loss after polyploidy (Fig. 1B) similar to the models proposed by Lewis (2001) to a hidden Markov model. This combination allows us to estimate our confidence *c* in a particular orthology state (values at the top of the columns, or "pillars", in Fig. 1). Using these inferences, we assessed the degree to which duplication state and reciprocal gene loss are contributing to failures in phylogenetic inference. We found that gene tree inference from most single copy genes gives a similar and relatively robust phylogenetic signal of the species tree. More surprisingly, we found that even the inclusion of reciprocal gene losses into the set of loci used for inference does not contribute to phylogenetic error in a meaningful way. However, the inference problem is still not trivial: in none of the three datasets considered did all four of the inference methods used converge to the same species tree.

2. Methods

2.1. Data sources

We considered three ancient whole-genome duplications for which we have made orthologous gene inferences using POInT (v.1.61). POInT models the post-polyploidy loss of duplicated genes along a phylogeny,

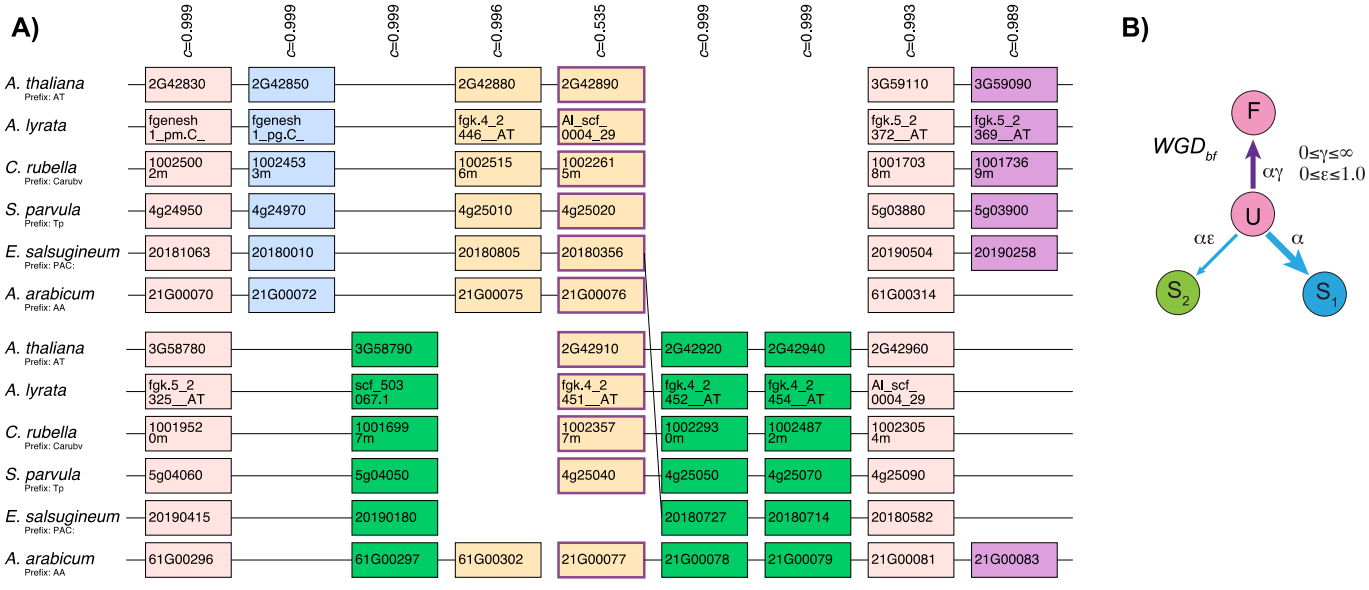


Fig. 1. (A) An example genomic region from six plant species sharing the At-α polyploidy event, visualized with POInT$_{browse}$ (Siddiqui and Conant, 2023). Confidence estimates (c) for the depicted orthology relationship relative to the other 2^6-1 = 63 possible relationships are shown at the top of each pillar. Single copy orthologs are shown in blue or green depending on which progenitor subgenome they descend from. Fully duplicated genes are shown in light pink. Genes with a mix of single copies and multiple copies are shown in tan and one RGL locus is shown in magenta. (B) A model of duplicate gene loss after polyploidy used by POInT to infer the orthology relationships in A. All loci start in duplicated state U. Over time, they may either become fixed at rate γ (state F) or lost, with subgenome 1 (S_1) being favored over subgenome two (S_2) when the ε parameter is less than one. Subgenome assignments, as well as γ and ε, are estimated from the pillar data by maximum likelihood. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

conditioning the corresponding orthology inferences at each ancestral locus on those loci in synteny with it. The events considered were the At-α event in *A. thaliana* and relatives (Emery et al., 2018); the ancient yeast WGD found in bakers' yeast and its relatives (Scannell et al., 2007) and the teleost genome duplication (TGD) shared by most bony fishes (Conant, 2020). The genomes used in these datasets are described in the respective manuscripts. Each ancestral gene duplicated at such a WGD creates a pair of *ohnologs* (Wolfe, 2000) that are inherited or lost in the descendant species; we call these loci *pillars* (Fig. 1). We consider different types of pillars for analyses here:

All pillars

Pillars where all taxa have only a single copy of the gene (single copy pillars),

Pillars where all taxa have a single copy of the gene and those genes are orthologs (orthologous pillars) and

Pillars where all taxa have a single copy of the gene, but those copies are reciprocally lost, meaning some of the genes are paralogs of each other (RGLs)

For the inference of the orthologous pillars (#3) and the pillars with RGLs (#4), we use POInT's orthology confidence score c ($0 \leq c \leq 1$; Fig. 1) to assess our confidence in that determination, selecting only cases of high confidence ($c \geq 0.7$ or $c \geq 0.9$, for Figs. 2 and 3,

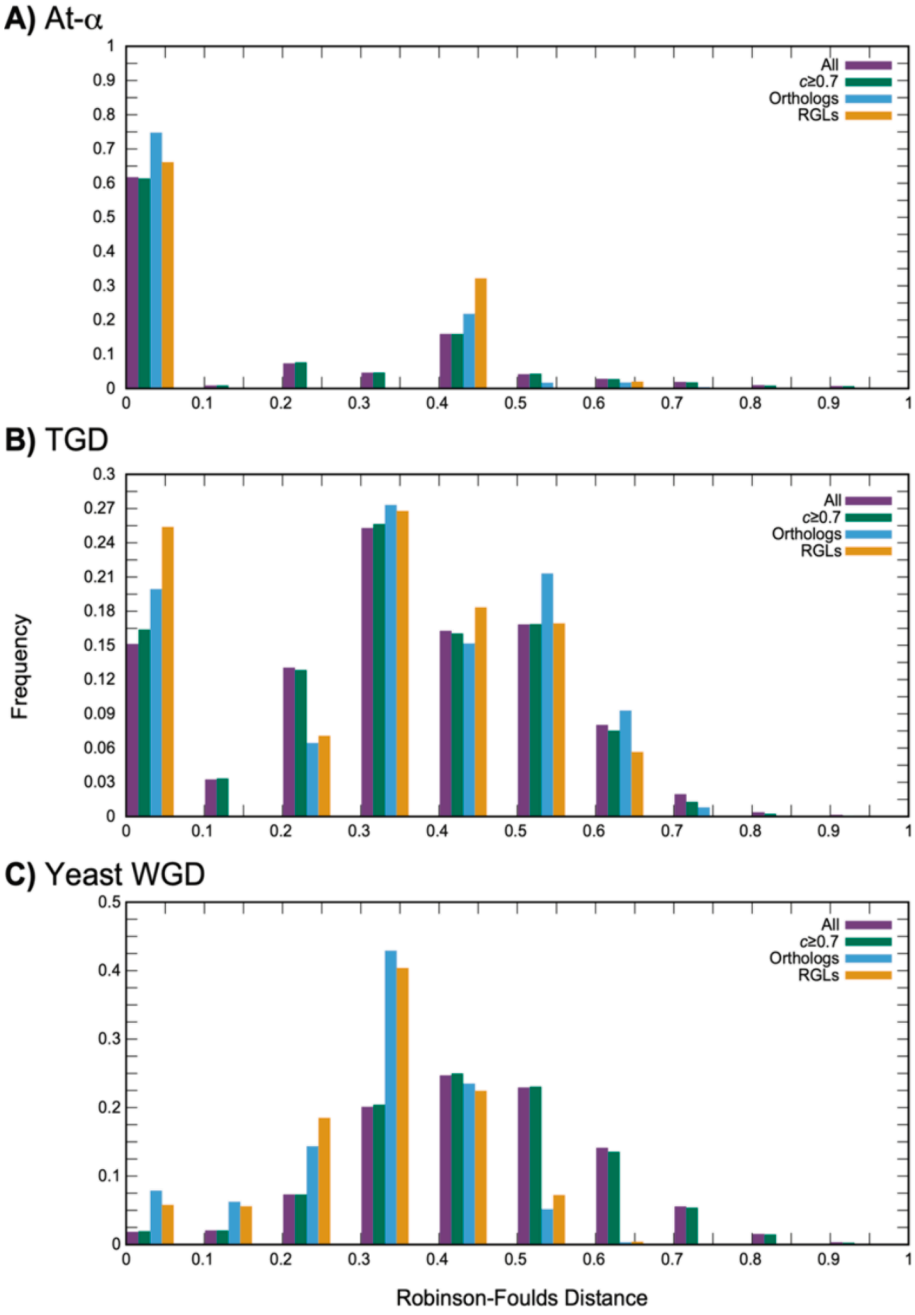
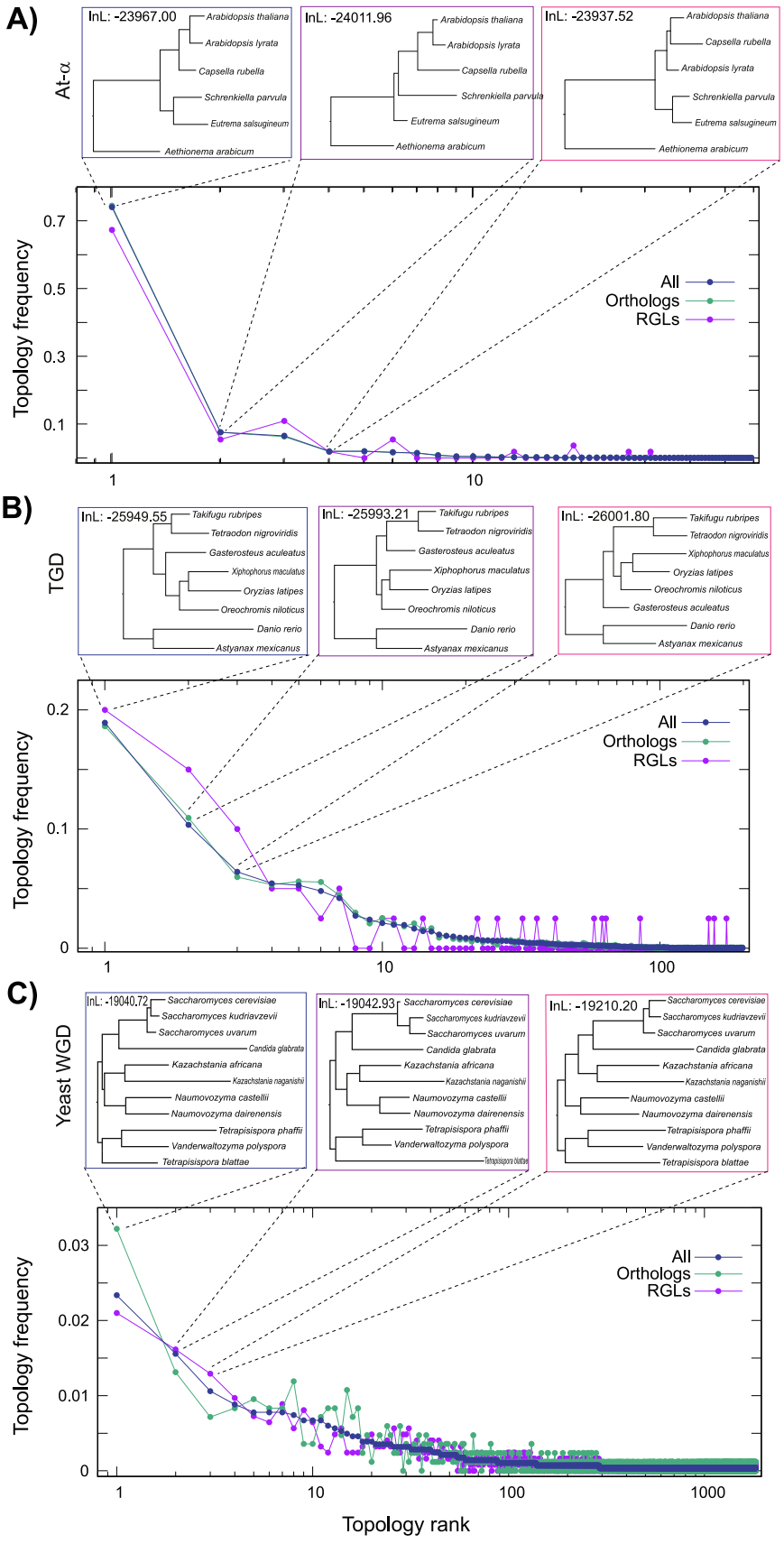


Fig. 2. Distribution of Robinson-Foulds distances between assumed species trees from POInT and the set of gene trees inferred with PAUP* for three different polyploidies. On the *x*-axis is the Robinson-Foulds distance between the POInT's expected gene tree and the inferred ML gene tree. On the *y*-axis is the proportion of gene trees in that distance interval. Four sets of gene trees are considered: all pillars in the data set (purple), all pillars with high confidence in the orthology inferences ($c \geq 0.7$, green) and all high confidence orthologs (blue) and RGLs (orange). **(A)** At-α. **(B)** TGD. **(C)** Yeast WGD. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)



(caption on next page)

Fig. 3. Distribution of single copy gene trees inferred from three paleopolyploidies. On *x* are the ranked topologies frequencies for all single copy pillars (note the log scale) and on *y* are their frequencies. Shown are the distributions for all single copy pillars (blue), single copy orthologs ($c \geq 0.9$, teal) and RGLs ($c \geq 0.9$, purple). **(A)** Distribution for At-α. Shown are the two most common gene tree topologies and their corresponding likelihoods under POInT and the *fourth* most common topology, which is the ML tree under POInT. **(B)** Distribution for the TGD. Shown are the likelihoods of the three most common gene tree topologies under POInT. Note that the most common gene tree topology (blue box) also gives a higher likelihood under POInT than does the previously assumed species tree from Near et al. (2013) (purple box). **(C)** Distribution for the yeast WGD, including the likelihood of the three most common topologies under POInT. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

respectively). In all cases, we obtained gene coding sequences and the orthology relationships from the $POInT_{browse}$ website (wgd.statgen.ncsu.edu: the genome release versions and sources are also listed here; Siddiqui and Conant, 2023); orthology confidences are computed as previously described (Hao and Conant, 2022).

2.2. Gene tree inference

For each pillar, we used the universal genetic code to translate the coding sequences into their corresponding amino acids sequences, which we then aligned using T-coffee (v13.45) under default parameters (Notredame et al., 2000). We then deduced the corresponding codon-preserving nucleotide alignments, which were used for gene tree inference using PAUP* (v4.0a build 168). We used the PAUP* "heuristic search" option under an HKY maximum likelihood model (Hasegawa et al., 1985) with empirical base frequencies and a transition/transversion ratio estimated from data. Rate variation was modeled with a four-category discrete gamma distribution with a shape parameter estimated by maximum likelihood (Yang, 1994). The At-α event contains 7243 pillars, of which 4225 are single copy genes. The TGD event contains 5589 pillars, of which 3322 are single copy genes. Finally, the yeast WGD event comprises 4065 pillars, of which 2966 are single copy genes.

2.3. Topological distance between gene trees and assumed species trees

For each pillar, we computed the topological distance between the gene tree inferred with PAUP* and the gene tree expected from the known species tree using the Robinson-Foulds metric (Robinson and Foulds, 1981). In all cases, we computed the expected gene tree using POInT's assumed species tree and orthology inferences, pruning out any lost ohnologs. This approach allowed us to analyze all gene trees, not only the single copy genes. The topological distances were computed with the TreeDist R package (Smith, 2020). For each event, we set up four different categories of pillar: (1) all pillars, (2) those with orthology confidence $c \geq 0.7$, (3) single copy orthologs with $c \geq 0.7$ and 4) RGLS with $c \geq 0.7$. We then calculated the distribution of distances for each category (Fig. 2).

2.4. Distribution of observed gene trees

Using only the single copy genes, we computed how often each different unrooted gene tree topology was observed, using the Robinson-Foulds distance to produce a histogram of all observed gene trees and their frequencies (Fig. 3). We computed this distribution for three groups of genes: (1) all single copy genes, (2) single copy orthologs with $c \geq 0.9$ and (3) RGLS with $c \geq 0.9$. For several of the common trees for each event, we also computed the log-likelihood of a rooted form of that topology from the gene order data using POInT. We note that for all three of these events, the earliest branching event among the polyploid taxa is uncontroversial: in other words, we are confident that *A. arabicum* is the earliest branching species for At-α, that *D. rerio* and *A. mexicanus* are basal relative to the other six taxa for the TGD and that *T. blattae, T. phaffii* and *V. polyspora* are basal relative to the other 8 yeast species. As a result, converting the unrooted gene trees into rooted species trees is trivial.

2.5. A reduced dataset from the yeast WGD

A previous phylogenomic analysis (Salichos and Rokas, 2013) using different taxa reported a phylogenetic position for *Candida galbrata* that conflicts with both the most common single copy gene tree and with the ML tree seen with POInT (Fig. 3C). Because we had previously seen similarly inconsistent placement of this taxon with a smaller sample of post-WGD yeasts, we reduced our 11 taxa yeast dataset to the five taxa we had previously analyzed (Conant and Wolfe, 2008) and inferred gene trees for the single copy genes, broken down into (1) all single copy genes (2) single copy orthologs ($c \geq 0.9$) and single copy RGLS ($c \geq 0.9$). We again ranked gene trees by their frequency for the three divisions.

2.6. Species tree inference and confidence estimation using Astral Pro2 and quartet sampling

We used Astral Pro2 (Mirarab et al., 2014; Zhang and Mirarab, 2022) and the set of all inferred single copy gene trees to infer a consensus species tree and corresponding support values for each of the three polyploidy events. We further applied 5000 replicates of quartet sampling (Pease et al., 2018) on the individual genes from the datasets to infer confidence measures on the species tree returned by Astral Pro2 (Fig. 4), using the quartetsampling (v. 1.3.1) package. We used a lnL cutoff of 1.0 for the QI (Quartet Informativeness) statistic. We also used quartetsampling with 100 replicates and the same likelihood cutoff to estimate confidence values from the concatenated single-copy alignments (see below).

2.7. Concatenation analyses

For each event, we concatenated the alignments from all single copy genes and analyzed these merged alignments using the partitioned GTR gamma model in RAxML (v 8.2.12). We assessed confidence in the inferred topology with 100 bootstrap replicates.

3. Results

3.1. Gene tree inference from synteny data and from standard approaches

In the absence of further information, the reconciliation of a number of duplicated genes onto a species tree is a challenging problem. Here, we have the advantage of using independent data on the orthology relationships between the duplicated genes, namely the synteny-based inferences from POInT. To assess both how the presence of duplicated genes and reciprocal gene loss (RGL) can adversely affect phylogenetic inference, we inferred gene trees from taxa sharing either the At-α event, the TGD, or the yeast WGD. POInT infers orthology between chromosomal segments duplicated by a polyploidy using a duplicate loss model along an assumed species phylogeny, conditioning the orthology inference at each pillar on all other pillars in the data (Conant and Wolfe, 2008). Hence, given an assumed species tree, we can take the POInT orthology inferences and prune out any gene losses to generate an expected gene tree. We compared these expected trees to gene trees inferred with standard approaches (see *Methods*). To that end, we obtained coding sequences of 16,897 sets of homologous genes (pillars in Fig. 1) created by these polyploidies. For each such pillar, we inferred an ML gene tree as described in the *Methods*.

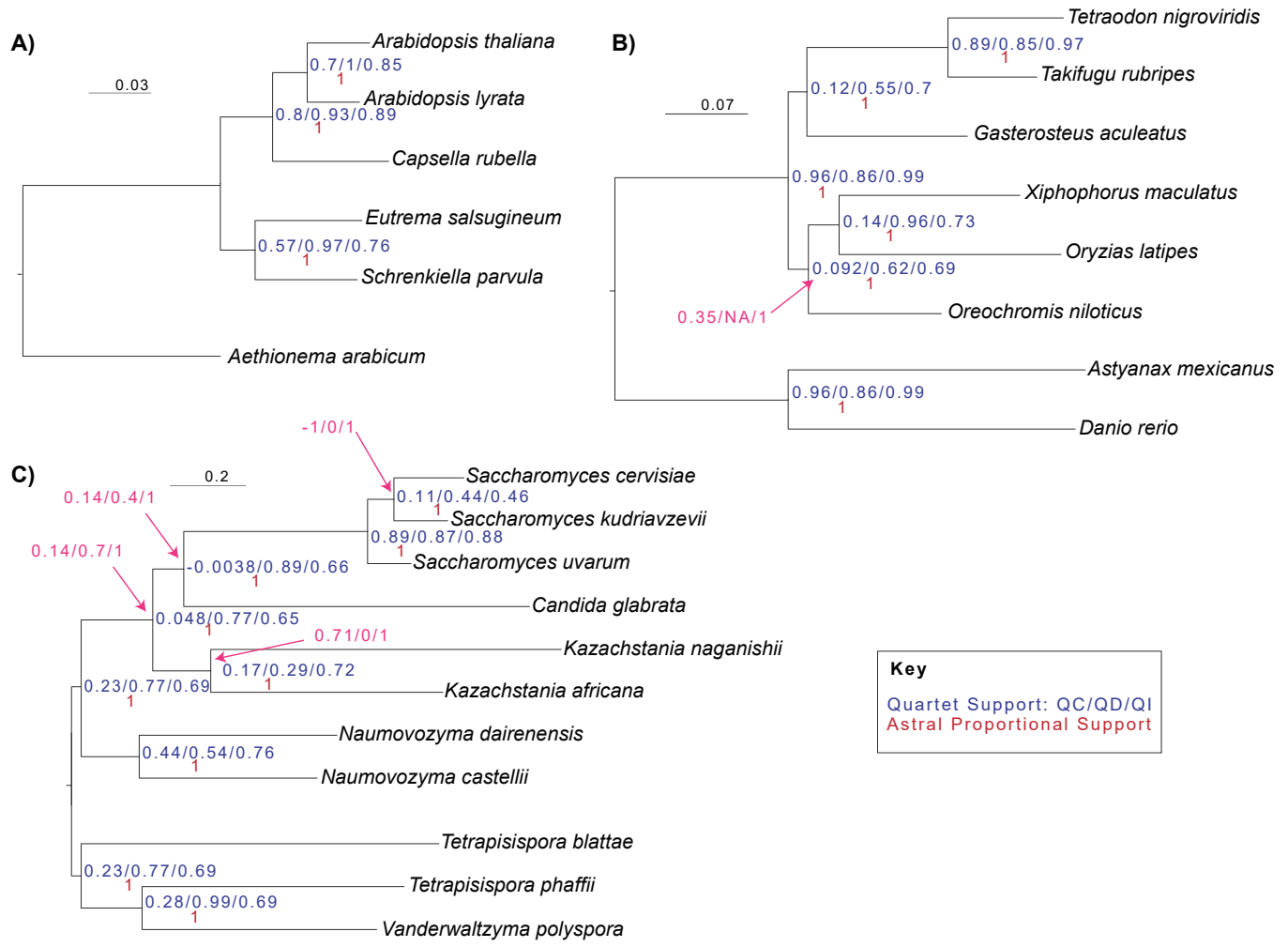


Fig. 4. Consensus trees for the three events inferred with all single copy pillars using the coalescent approach with Astral Pro 2 as well as the corresponding local posterior probability support values in red. Also shown are quartet support values from those gene trees with respect to the respective consensus trees from Astral (blue). These support values are reported as QC (Quartet Concordances, or proportion of resamplings yielding the topology indicated, if positive, or supporting the most common other topology, if negative), QD (Quartet Differential, how skewed the resampled quartets are between the alterative topologies, QD = 1 completely skewed to one topology, QD = 0, no skew) and QI (Quartet Informativeness: what proportion of the quartets show a topology with a likelihood meaningfully higher than the others: cutoff of 1 lnL unit, *Methods*). **(A)** At-α, made by coalescing 4225 single copy gene trees. **(B)** TGD, made by coalescing 3322 single copy gene trees. The single node without 100 % quartet sampling support when the concatenated alignment (*Methods*) is used is indicated in pink. **(C)** Yeast WGD, made by coalescing 2635 single copy gene trees. The four nodes without 100 % quartet sampling support when the concatenated alignment (*Methods*) is used are indicated in pink. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

3.2. Discordant species trees for the TGD

When examining the gene tree topologies seen from the TGD, we were surprised to find that the most common gene tree did not match the assumed species tree we had previously adapted from Near et al. (2013). When we analyzed this alternative topology with POInT, we were further surprised to find that it also had a higher likelihood of generating the synteny data for POInT. Hence, this alternative topology (shown in Fig. 2B) was used as the species tree for all further analyses. We have also incorporated pillar visualizations using this topology into POInT-browse (Siddiqui and Conant, 2023).

3.3. Phylogenetic uncertainty is larger for larger datasets

We compared the distribution of Robinson-Foulds distances between the POInT trees and the ML gene trees for four different classes of duplicated loci/pillars, namely: (1) all loci, (2) all single copy loci, (3) orthologous single copy loci and (4) single copy reciprocal gene losses (RGLs): the results of this analysis are shown in Fig. 2 (*Methods*)..

For At-α, the simplest phylogenetic problem with only six taxa, the modal distance between the expected and inferred gene trees was 0, implying that most gene tree topologies matched the assumed species tree. For the TGD and the yeast WGD, most gene trees did not match the species tree. However, these distances alone cannot distinguish between a highly favored alternative topology (most trees support a single alternative topology) and noise in the phylogenetic signal (many different alternative topologies). Somewhat surprisingly, the behavior of the RGL loci also differed between events. Recall that the expectation from RGL loci would be a mismatch between the species tree and the inferred gene tree, as was indeed seen for the At-α event. Curiously, however, for the TGD, the RGL loci are actually proportionally more likely to be close to the assumed species tree.

3.4. Distribution of single copy gene trees and comparisons to species tree inferences from POInT

To understand how the inferred single-copy gene trees differed amongst themselves, we computed the rank-order distribution of

topologies (i.e., a list of topologies ordered from most to least common, *Methods*). For the yeast WGD and the TGD, we took the top three most common gene tree topologies and computed their likelihoods under the gene loss data with POInT (Fig. 2B&C). For the At-α event, there were few enough taxa in the dataset to permit an exhaustive tree search with POInT.

For the yeast WGD and the TGD, the species tree with the highest likelihood from POInT is also the most common gene tree both when all single copy genes are considered and when considering only single copy orthologs. Strikingly, for At-α, the ML tree from POInT is only the *fourth* most common topology among the gene trees (Fig. 2A). However, this tree differs from the most common gene tree only in its placement of *A. thaliana* and *C. rubella* as sisters.

Given the expectation that RGL should confound species tree inference, it is somewhat surprising that, for all three events, the most frequent topology for RGL pillars is the same as the overall most frequent topology (which is also the most frequent topology for orthologous single copy pillars; Fig. 3). Among the lower-ranked topologies there are differences in frequency for the RGL pillars, with some topologies being more frequent with RGL pillars and some less. We note that because RGLs are quite rare for the At-α and TGD events, the trends for the orthologous pillars and for all single copy pillars are essentially identical because a very large proportion of those single copy genes are in fact orthologs rather than RGLs. This is not the case for the yeast WGD, where there are in fact more RGL pillars than orthologous pillars (1247 verses 838 for $c \geq 0.9$) Nonetheless, the RGLS for the yeast WGD do not present a single dominant topology that would tend to mislead a phylogenomic analysis (Fig. 3C).

The situation is less ideal when we consider a reduced five taxa dataset for the yeast WGD (Supplemental Fig. 1), which might be expected to show phylogenetic incongruence (*Methods*). In this case there are two topologies that are nearly equal in frequency; one that matches the assumed species tree from Fig. 3C, and one that places *C. glabrata* as sister to *V. polysporus*. This second topology is generally similar to the consensus topology of Salichos and Rokas (2013). POInT strongly prefers the assumed species tree to this alternative. In this case RGLs are contributing to the misleading phylogenetic signal, as more RGL gene trees prefer the alternative topology than the species tree, while the orthologous genes show the opposite pattern (Supplemental Fig. 1).

3.5. Comparisons of tree inferences between the full set of At-α single-copy orthologs and the Angiosperm 353 gene set

Johnson et al., (2019) have described of a set of 353 gene/probe pairs that are generally single-copy across the angiosperms. Of the corresponding 329 *Arabidopsis thaliana* genes, 160 are not present in our POInT dataset, 2 are a pair of preserved duplicates from At-α, 14 others are members of a pillar with at least one surviving duplicate, 14 have low orthology confidence values from POInT ($c < 0.9$) and the remaining 138 are high-confidence single-copy orthologs in the POInT dataset. None of the Angiosperm 353 are present in pillars we identify as cases of RGL. The distribution of gene trees from these 138 pillars that represent orthologs is effectively identical to that of all single-copy orthologs (Supplemental Fig. 2).

3.6. Concatenated analyses, coalescent inferences and node confidence estimates

To assess our confidence in the phylogenetic topologies estimated from these gene sequences, we used three approaches. First, we concatenated all of the single copy genes for each event and performed a partitioned maximum likelihood analysis with RAxML (*Methods*). We then used bootstrap replicates as a confidence measure (Supplemental Fig. 2). We also used Astral Pro to infer coalescent-based consensus trees for each event directly from its individual gene trees computed from the corresponding single copy genes (Fig. 4); we obtained local posterior probabilities as support values from this analysis. Using the Astral consensus trees, we also applied quartet sampling to the individual gene trees (Fig. 4; *Methods*). Both bootstrapping and Astral consensus approaches unsurprisingly yield very high node support, in keeping with the large numbers of gene trees used to produce the inferences. We note, however, that the TGD consensus tree seen with these two methods is *not* the same as the topology supported by POInT, which is also the most common gene tree. Likewise, the coalescent and concatenated methods disagree in the case of the yeast tree, with the concatenated approach yielding the *second* most common gene tree in Fig. 3C, unlike the coalescent approach, which, along with POInT, supports the most common gene tree (Supplemental Fig. 2). Notably, the reduced five-taxa yeast dataset shows very strong bootstrap support for the assumed species tree that POInT supports, even though two alternative gene trees have nearly equal frequencies (Supplemental Figs. 1 and 2).

More strikingly, the quartet support values taken from the gene trees are generally low (Fig. 4), suggesting that the quartet support within individual single-copy genes is surprisingly weak. When we perform quartet sampling on the complete concatenated alignments, the quartet concordance values are in general very high (see Fig. 4), with the exception of some nodes in the yeast phylogeny. Hence, the aggregate support for the topologies for both *A. thaliana* and its relatives and the teleost fishes is high. We attribute the instances where the quartet support values are low to reflect the long-branch problems inherent in these datasets, as illustrated with the 5-taxa yeast analysis in Supplemental Fig. 1.

4. Discussion

4.1. Use of single copy genes is a straight-forward approach to phylogenetic inference in polyploid taxa

Prior work suggests that polyploidy and its associated evolutionary patterns, such as RGL, are not expected to be a serious confound to phylogenetic inference (Smith et al., 2022; Thomas et al., 2017; Xiong et al., 2022). However, in the absence of a "gold standard" dataset of known orthologous genes, a full understanding of the behavior of phylogenomic datasets in the presence of polyploidy was lacking. Because prior studies relied on simulations and mathematical models in their treatment of RGL and post-polyploidy gene loss more generally (Xiong et al., 2022), there was still the possibility that real datasets might display some pathologies that were not yet understood. More generally, it would be desirable to understand the distribution of gene trees produced by polyploidy and subsequent gene loss in real genomes, since that information will shed light on why phenomena like RGL are generally not serving to confound inference procedures.

We had, in fact, hypothesized that difficulties in ortholog identification would prove to be a somewhat important confound phylogenetic inference in polyploid taxa, both due to difficulties in dealing with duplicated genes (Salichos and Rokas, 2013) and, more seriously, in dealing with reciprocal gene losses (Scannell et al., 2007). Instead, simply restricting the analysis to single copy genes and taking the most common gene tree provides a reasonable species tree inference in all datasets analyzed here. An implication of these findings the use of single copy genes, perhaps from the "universal single copy" gene lists (Duarte et al., 2010), may be a sufficient approach for inference with polyploid taxa, although synteny-based improvements to this approach could also be considered (Washburn et al., 2017). However, we emphasize that while the data analyzed here are in some sense the best available for the question, the phylogenetic problems we consider here are quite simple because the number of taxa is small.

The natural question arises as to why RGL seems to contribute so little to difficulties in phylogenetic inference. We think the answer lies in the fact that the loss events that generate RGLs are necessarily occurring on the same species tree that the genes are evolving along, giving the RGLs observed a particular character. The number of duplicate genes

lost per unit time decreases in time after a polyploidy for the simple reason that fewer genes remain to be lost the longer in the past the polyploidy is. As a result, the overwhelming proportion of RGL events involve reciprocal losses between the most diverged pair of clades in the rooted tree. For the At-α event, the result is that most RGLs give rise to a set of orthologous genes from *A. thaliana, A. lyrata, C. rubella, E. salsugineum* and *S. parvula* that are paralogs of a gene from *A. arabicum.* Similarly, for the TGD, RGLs tend to involve an orthologous pair from *D. rerio* and *A. mexicanus* that are paralogous to another set of orthologs from the other six species. What happens if one builds a gene tree on such data? Essentially, the inferred tree will tend to have the same topology as the species tree but will exaggerate the divergence between the two clades in question, because these two clades consist of paralogs that diverged at the ancient polyploidy rather than orthologs diverging at the more recent speciation. One might almost claim that this type of RGL gives the correct gene tree for the wrong reasons. Such events may also contribute to observed overestimates of divergence times from datasets that were not corrected for RGL (Siu-Ting et al., 2019). Even in the case of the yeast WGD, where RGL is very common, most of the events are of this form (separating *V. polyspora, T. phaffii,* and *T. blattae* from the remaining taxa), and so again the gene trees tend to support the species tree. Supposing that we included non-polyploid taxa in our analyses: would these types of RGL introduce significant biases in that case? We can only imagine one circumstance where they would: namely an allopolyploidy where some of the extant non-polyploid taxa are more closely related to one of the allopolyploid parents than the other. Exactly this circumstance is known for the yeast WGD (Marcet-Houben and Gabaldon, 2015). Whether it is a common situation, however, remains uncertain.

We note that the results presented speak to the question of how ancient polyploidies affect phylogenetic inference. More precisely, we are considering the case of a polyploidy followed by a least two speciation events, such that we would like to infer the relationships of the three or more species that descend from that polyploidy event. So-called *mesopolyploidies*, such as that in the Brassiceae (Hao et al., 2021), would fit into this framework despite their apparent continuing gene losses. A *neopolyploid* taxon that has, by definition, not produced more than a single new species (Ramsey and Schemske, 2002), presents a very different set of phylogenetic problems. For instance, in the case of an allopolyploid (Stebbins, 1947), the "duplicated" copies of a gene will likely (though not certainly; Gaeta et al., 2007) derive from different progenitor genomes with different evolutionary histories, meaning that the neopolyploid will not necessarily occupy a single location in a species tree, even were that tree known with complete confidence. A recent autopolyploid (Stebbins, 1947), on the other hand, might produce few difficulties in tree inference and might even be invisible to genome sequencing and assembly, due to the high identity of the homoeologous regions produced by that event.

One unusual feature of the data considered here is that we have the ability to make independent, maximum likelihood estimates of the species tree using gene loss information and POInT. In general, the species tree inferred with POInT agrees well with the most common gene tree topology. In the one case where agreement is not seen, namely the At-α event, we are virtually certain that this difference represents a failure of POInT to support the true species topology, namely that of *A. thaliana* and *A. lyrata* as sisters with *C. rubella* as an outgroup (Koenig and Weigel, 2015). POInT is not expected to perform well at distinguishing branching patterns long after a WGD because it uses gene losses as a signal and most of the duplicate losses occur early in the event's history. And indeed, POInT estimates that only about 36 gene losses occurred on the shared branch joining *A. thaliana* and *A. lyrata* (Fig. 2A; data not shown), suggesting low confidence in POInT's inferred topology.

However, while polyploidy on its own may not present insurmountable difficulties, the picture is not altogether rosy. For each of these three datasets, we have inferred the species tree with four methods: (1) POInT, (2) the most common gene tree, (3) consensus of gene trees via Astral Pro 2 and (4) concatenation. In no case do all four methods agree on the species tree. For the At-α dataset, only POInT yields a discordant topology, and this difference can generally be discounted. However, the other two datasets require more consideration. For the TGD, POInT agrees with the common gene tree, while the coalescent and concatenation inferences agree with each other but differ from that POInT tree (Figs. 3 and 4 and Supplemental Fig. 2). Notably, the tree inferred with coalescent and concatenation methods agrees with the published topology (Near et al., 2013). For the yeast WGD, the situation is perhaps even more perplexing, as POInT agrees with both the coalescent inference and with the most common gene tree, which is also the tree reported by Almeida et al (2014). However, the tree inferred with concatenation differs from that inferred with the other three methods (Figs. 3 and 4 and Supplemental Tree 2).

We are actually less concerned about the discordance itself than that the support values seen for both the TGD and yeast WGD under the coalescent and concatenation approaches suggest little phylogenetic uncertainty, when in fact there appears to be considerable conflict in the signals from the different gene trees. It is also striking that quartets that sample only a single taxon in each subtree and pillar do not show such high support, suggesting that some parts of these phylogenies are difficult to recover, perhaps due to long-branch effects.

From certain points of view, our results follow those of Salichos and Rokas (2013), showing that even phylogenetic inferences that show high support values from large datasets can hide underlying uncertainty. To explore this idea further, we analyzed a subset of our yeast species to see if we could replicate the highly supported but discordant yeast topology found by these authors. For our analysis of five post-WGD yeasts, we did not find the topology of Salichos and Rokas to be the most supported, but we did find that a relatively large minority of trees (and a majority of RGL trees) supported it. We suspect that the difficulty in placing *C. glabrata* in these datasets is probably a function of long branch attraction (Felsenstein, 1978): *C. glabrata* shows long branches with a variety of analyses, and an RGL event will also exaggerate the branch between *V. polyspora* and the other four taxa, amplifying the long-branch attraction difficulties. After adding more taxa to make a full analysis, we break up these long branches and recover a more plausible tree.

Broadly, our analyses reinforce both the power of phylogenetic methods, which appear to be robust even to changes as dramatic as a whole-genome duplication. At the same time, we should not imagine that deep sampling of sequences from relatively few taxa is a universal solution to difficult phylogenetic problems: the yeast example suggests that sampling more taxa may pay larger dividends. Beyond that, understanding the factors that contribute to phylogenetic difficulties will be important even in an era of complete genomes, given that even with full genomes there remains some doubt as to the relationships of the species sharing the TGD.

5. Data availability

Underlying data and software used for the analyses above are freely available at https://doi.org/10.5061/dryad.7d7wm3821. Genome accessions and versions are available at https://wgd.statgen.ncsu.edu/GenomeTable.html. All coding region sequences and POInT inferences are available from $\text{POInT}_{\text{browse}}$ (https://wgd.statgen.ncsu.edu).

CRediT authorship contribution statement

Jaells G. Naranjo: Writing – review & editing, Writing – original draft, Visualization, Software, Investigation. **Charles B. Sither:** Writing – review & editing, Supervision, Methodology. **Gavin C. Conant:** Writing – review & editing, Writing – original draft, Visualization, Supervision, Software, Project administration, Methodology, Investigation, Funding acquisition, Data curation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The underlying data are shared at: https://doi.org/10.5061/dryad.7d7wm3821

Acknowledgements

We would like to thank J. Thorne for helpful comments. JN and GCC were supported by National Science Foundation grant NSF-DEB-2241312.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ympev.2024.108087.

References

Almeida, P., Gonçalves, C., Teixeira, S., Libkind, D., Bontrager, M., Masneuf-Pomarède, I., Albertin, W., Durrens, P., Sherman, D.J., Marullo, P., 2014. A Gondwanan imprint on global diversity and domestication of wine and cider yeast Saccharomyces uvarum. Nat. Commun. 5, 4044.
Chen, K., Durand, D., Farach-Colton, M., 2000. NOTUNG: a program for dating gene duplicationsand optimizing gene family trees. J. Comput. Biol. 7, 429–447.
Chen, H., Zwaenepoel, A., 2023. Inference of ancient polyploidy from genomic data. Methods Mol. Biol. 2545.
Conant, G.C., 2020. The lasting after-effects of an ancient polyploidy on the genomes of teleosts. PLoS One 15, e0231356.
Conant, G.C., Wolfe, K.H., 2008. Probabilistic cross-species inference of orthologous genomic regions created by whole-genome duplication in yeast. Genetics 179, 1681–1692.
De Smet, R., Adams, K.L., Vandepoele, K., Van Montagu, M.C.E., Maere, S., Van de Peer, Y., 2013. Convergent gene loss following gene and genome duplications creates single-copy families in flowering plants. Biol. Sci. 110, 2898–2903.
Duarte, J.M., Wall, P.K., Edger, P.P., Landherr, L.L., Ma, H., Pires, J.C., Leebens-Mack, J., dePamphilis, C.W., 2010. Identification of shared single copy nuclear genes in Arabidopsis, Populus, Vitis and Oryza and their phylogenetic utility across various taxonomic levels. BMC Evol. Biol. 10, 61.
Emery, M.M., Willis, M.M.S., Hao, Y., Barry, K., Oakgrove, K., Peng, Y., Schmutz, J., Lyons, E., Pires, J.C., Edger, P.P., Conant, G.C., 2018. Preferential retention of genes from one parental genome after polyploidy illustrates the nature and scope of the genomic conflicts induced by hybridization. PLoS Genet. 14, e1007267em.
Felsenstein, J., 1978. Cases in which parsimony or compatibility methods will be positively misleading. Syst. Zool. 27, 401–410.
Gaeta, R.T., Pires, J.C., Iniguez-Luy, F., Leon, E., Osborn, T.C., 2007. Genomic changes in resynthesized Brassica napus and their effect on gene expression and phenotype. Plant Cell 19, 3403–3417.
Hao, Y., Conant, G., 2022. POInT: a tool for modeling ancient polyploidies using multiple polyploid genomes. Methods Mol. Biol. 2022, 81–91.
Hao, Y., Mabry, M.E., Edger, P., Freeling, M., Zheng, C., Jin, L., VanBuren, R., Colle, M., An, H., Abrahams, R.S., Washburn, J.D., Qi, X., Barry, K., Daum, C., Shu, S., Schmutz, J., Sankoff, D., Barker, M.S., Lyons, E., Pires, J.C., Conant, G.C., 2021. The contributions of the allopolyploid parents of the mesopolyploid Brassiceae are evolutionarily distinct but functionally compatible. Genome Res. 31, 799–810.
Hasegawa, M., Kishino, H., Yano, T., 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J. Mol. Evol. 22, 160–174.
Johnson, M.G., Pokorny, L., Dodsworth, S., Botigué, L.R., Cowan, R.S., Devault, A., Eiserhardt, W.L., Epitawalage, N., Forest, F., Kim, J.T., 2019. A universal probe set for targeted sequencing of 353 nuclear genes from any flowering plant designed using k-medoids clustering. Syst. Biol. 68, 594–606.
Koenig, D., Weigel, D., 2015. Beyond the thale: comparative genomics and genetics of Arabidopsis relatives. Nat. Rev. Genet. 16, 285–298.
Koonin, E.V., 2005. Orthologs, paralogs, and evolutionary genomics. Ann. Rev. Gene. 39, 309–338.
Lewis, P.O., 2001. A likelihood approach to estimating phylogeny from discrete morphological character data. Syst. Biol. 50, 913–925.
Liu, L., Yu, L., Kubatko, L., Pearl, D.K., 2009. Coalescent methods for estimating phylogenetic trees. Mol. Phylogene. Evol. 53, 320.
Madison, W.P., Knowles, L.L., 2006. Inferring phylogeny despite incomplete lineage sorting. Syst. Biol. 55, 21–30.
Marcet-Houben, M., Gabaldon, T., 2015. Beyond the whole-genome duplication: phylogenetic evidence for an ancient interspecies hybridization in the baker's yeast lineage. PLoS Biol. 13, e1002220.
Mirarab, S., Reaz, R., Bayzid, S., Zimmermann, T., Swenson, M.S., Warnow, T., 2014. ASTRAL: genome-scale coalescent-based species tree estimation. Bioinformatics 30, 541–548.
Near, T.J., Dornburg, A., Eytan, R.I., Keck, B.P., Smith, W.L., Kuhn, K.L., Moore, J.A., Price, S.A., Burbrink, F.T., Friedman, M., Wainwright, P.C., 2013. Phylogeny and tempo of diversification in the superradiation of spiny-rayed fishes. Proc. Natl. Acad. Sci. 110, 12738-12743.
Notredame, C., Higgins, D.G., Heringa, J., 2000. T-Coffee: a novel method for multiple sequence alignments. J. Mol. Biol. 302, 205–217.
Pease, J.B., Brown, J.W., Walker, J.F., Hinchliff, C.E., Smith, S.A., 2018. Quartet sampling distinguishes lack of support from conflicting support in the green plant tree of life. Am. J. Bot. 105, 385–403.
Philippe, H., 2005. Phylogenomics. Annu. Rev. Ecol. Evol. Syst. 36, 541–562.
Rabier, C.E., Ta, T., Ané, C., 2014. Detecting and locating whole genome duplications on a phylogeny: a probabilistic approach. Mol. Biol. Evol. 31, 750–762.
Ramsey, J., Schemske, D.W., 2002. Neopolyploidy in flowering plants. Annu. Rev. Ecol. Syst. 33, 589–639.
Robinson, D.F., Foulds, L.R., 1981. Comparison of Phylogenetic Trees. Math. Biosci. 53, 131–147.
Rokas, A., Williams, B.L., King, N., Caroll, S.B., 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature 425, 798–804.
Salichos, L., Rokas, A., 2013. Inferring ancient divergences requires genes with strong phylogenetic signals. Nature 497, 327.
Scannell, D.R., Byrne, K.P., Gordon, J.L., Wong, S., Wolfe, K.H., 2006. Multiple rounds of speciation associated with reciprocal gene loss in polyploid yeasts. Nature 440, 341–345.
Scannell, D.R., Frank, A.C., Conant, G.C., Byrne, K.P., Woolfit, M., Wolfe, K.H., 2007. Independent sorting-out of thousands of duplicated gene pairs in two yeast species descended from a whole-genome duplication. Proc. Natl. Acad. Sci. U.S.A. 104, 8397–8402.
Siddiqui, M., Conant, G.C., 2023. POInTbrowse: Orthology prediction and synteny exploration for paleopolyploid genomes. BMC Bioinf. 24, 174.
Siu-Ting, K., Torres-Sánchez, M., San Mauro, D., Wilcockson, D., Wilkinson, M., Pisani, D., O'Connell, M.J., Creevey, C.J., 2019. Inadvertent paralog inclusion drives artifactual topologies and timetree estimates in phylogenomics. Mol. Biol. Evol. 36, 1344–1356.
Smith, M.R., 2020. Information theoretic Generalized Robinson-Foulds metrics for comparing phylogenetic trees. Bioinformatics 36, 5007–5013.
Smith, M.L., Vanderpool, D., Hahn, M.W., 2022. Using all gene families vastly expands data available for phylogenomic inference. Mol. Biol. Evol. 39, msac112.
Stebbins Jr, G.L., 1947. Types of polyploids: their classification and significance. Adv. Gene. Elsevier 403–429.
Thomas, G.W., Ather, S.H., Hahn, M.W., 2017. Gene-tree reconciliation with MUL-trees to resolve polyploidy events. Syst. Biol. 66, 1007–1018.
Van de Peer, Y., Mizrachi, E., Marchal, K., 2017. The evolutionary significance of polyploidy. Nat. Rev. Genet. 18, 411–424.
Washburn, J.D., Schnable, J.C., Conant, G.C., Brutnell, T.P., Shao, Y., Zhang, Y., Ludwig, M., Davidse, G., Pires, J.C., 2017. Genome-guided phylo-transcriptomic methods and the nuclear phylogentic tree of the paniceae grasses. Sci. Rep. 7, 13528.
Wolfe, K.H., 2000. Robustness: It's not where you think it is. Nat. Genet. 25, 3–4.
Xiong, H., Wang, D., Shao, C., Yang, J., Ma, T., 2022. Species tree estimation and the impact of gene loss following whole-genome duplication. Syst. Biol. 71, 1348–1361.
Yang, Z., 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J. Mol. Evol. 39, 306–314.
Zhang, C., Mirarab, S., 2022. ASTRAL-Pro 2: ultrafast species tree reconstruction from multi-copy gene family trees. Bioinformatics 38, 4949–4950.