Gene-tree reconciliation with MUL-trees to resolve polyploidy events

Gregg W.C. Thomas[1], S. Hussain Ather, and Matthew W. Hahn

Department of Biology and School of Informatics and Computing, Indiana University, Bloomington, IN 47405.

[1]To whom correspondence may be addressed. Email: grthomas@indiana.edu

Keywords: Polyploidy, Reconciliation, MUL-trees

Abstract

Polyploidy can have a huge impact on the evolution of species, and is a common occurrence, especially in plants. Two types of polyploids – autopolyploids and allopolyploids – differ in the level of divergence between the genes brought together in the new polyploid lineage. Because allopolyploids are formed via hybridization, the homologous copies of genes within them are as divergent as the parental species that came together to form them. This means that common methods for estimating the timing of polyploidy events fail to correctly date allopolyploidy, and can lead to incorrect inferences about the number of gene duplications and losses. Here we have adapted the standard algorithm for gene-tree reconciliation to work with multi-labeled species trees (MUL-trees). MUL-trees are defined as having identical species labels, which makes them a natural representation of polyploidy events. Using this new reconciliation algorithm we can: accurately date allopolyploidy events on a tree, identify the parental lineages that hybridized to form allopolyploids, distinguish auto- from allopolyploidy, and correctly count the number of duplications and losses in a set of gene trees. We validate our method using gene trees simulated with and without polyploidy, and revisit the history of polyploidy in data from the clades including both baker's yeast and bread wheat. Our re-analysis of the yeast data confirms the allopolyploid origin of this group, but identifies slightly different parental lineages than a previous analysis. The method presented here should find wide use in the growing number of genomes from species with a history of polyploidy.

Introduction

Polyploidy as a result of whole genome duplication (WGD) can be a key evolutionary event. At least two ancient WGDs have been postulated at the origin of vertebrate animals (1, 2), with yet another occurring before the radiation of bony fishes (3). Polyploidy events are far more common in plants. It is estimated that approximately 25% of extant angiosperms have experienced a recent polyploidization, while 70% of angiosperm species show signs of a more ancient event (4). The prevalence of these events combined with the success of flowering plants suggests that they must confer some advantage, possibly by increasing speciation rates (5, 6; but see 7, 8), by decreasing extinction rates (9), or by providing species with a large amount of genetic material from which novel phenotypes can arise (10-12).

Because of the importance of polyploidy events in adaptation and speciation, multiple methods have been employed to detect and date them. The most commonly used method starts by identifying pairs of duplicates and measures their synonymous divergence (K_s). In a species that has not experienced polyploidy, the expectation is that most duplicates will be very recent – and have a low K_s – while very few duplicates will have high K_s (13). However, in a lineage in which polyploidy has occurred, peaks observed in this distribution can correspond to a burst of duplications from the WGD (14-16). When possible, these peaks are then placed in a phylogenetic context by comparing overlapping peaks of closely related species and mapping them onto a species tree with branch lengths scaled by K_s (e.g. 15, 17). Phylogenetic tree-based methods include reconciliation and species-overlap algorithms, both of which use gene-trees to map duplications onto a species tree in order to find a branch with a large number of events (16-19). Similarly, recent likelihood-based "count" methods use copy-number to identify a branch with more duplication than expected without polyploidy (20, 21).

None of these methods give a full picture of the evolutionary history of a polyploidy event, and may be positively misleading about multiple aspects of WGDs. This is because they do not differentiate between two types of polyploidy: auto- and allopolyploidy. Autopolyploidy occurs when an individual inherits sets of chromosomes from parents of the same species. Allopolyploidy occurs when an individual inherits sets of chromosomes from parents of different species. The distinction between these types has important implications for our inferences about polyploidy. While the aforementioned methods can help to tell us that a polyploidy event has occurred, none of the methods can differentiate auto- from allopolyploidy, and the dating schemes used by most of them implicitly assume that all WGDs are autopolyploid. In both K_s and gene-tree-based methods the "peak" of divergence or duplication events detected will not reliably identify the timing of allopolyploidy events, because these methods do not account for the reticulate nature of allopolyploid species. While they should work for autopolyploids (Figure 1a), for allopolyploids they incorrectly identify the most recent common ancestor of the species that hybridized to form the polyploid as the time that the WGD occurred (Figure 1b). When using standard reconciliation-based gene-tree methods, WGDs due to allopolyploidy can also lead researchers to incorrectly infer many additional duplications and losses when none have occurred (see below).

One major issue shared by these methods used to study polyploidy is that standard representations of species trees only show one of the multiple "sub-genomes" (sets of chromosomes) in any polyploid species (Figure 2a). Species networks are an alternative, useful representation for allopolyploids as they can highlight both the parental lineages involved and the timing of the hybridization event (Figure 2b; 22-24). However, networks still only represent a single sub-genome, and therefore can be less practically useful for analyses involving individual

genes in allopolyploid genomes. A more useful representation uses multi-labeled species trees (MUL-trees) for polyploidy events (Figure 2c; 25, 26). MUL-trees are trees in which the tip labels are not necessarily unique (27): this allows one to represent both sub-genomes in an allopolyploid as descendants of different parental lineages (Figure 2c), or as descendants of the same lineage for autopolyploids (e.g. Figure 1a).

Here, we have adapted the standard Least Common Ancestor (LCA) mapping algorithm for reconciliation (28, 29) for use with MUL-trees. This representation and algorithm allows us to correctly infer gene duplications and losses, and to identify the most likely placement of allopolyploid clades and their parental lineages. We demonstrate that this new method works on simulated data, and we revisit two different datasets that include allopolyploid species, providing a new perspective on the parental lineages leading to the clade that includes baker's yeast.

Methods

Algorithm

We have devised an LCA mapping algorithm that works with MUL-trees, a natural representation of polyploidy events. The standard LCA mapping algorithm is a method that places and counts duplication and loss events on a gene tree given an accepted species tree (28, 29). It can also be used for species tree inference by searching for the species tree that minimizes the total number of duplications and losses inferred given a set of gene trees (30). The main hurdle in applying LCA mapping to MUL-trees is that when reconciling to a MUL-tree, some nodes have more than one possible map. In particular, some tip nodes cannot be initialized with a single map (because tips are necessarily not uniquely labeled), which subsequently allows

internal nodes to also have more than one possible map. We side-step this problem by trying all possible combinations of initial tip maps and applying the parsimony assumption – that the correct map will have the lowest score.

In standard LCA mapping each node in the gene tree, n_g, is defined by the set of species at the tips below it in the tree. The same node is also associated with a node in the species tree, n_s, through a map, $M(n_g)$. $M(n_g) = n_s$, where n_s is the node in the species tree that is the least common ancestor of the species that define n_g. For example, in the gene tree depicted in Figure 3, node 1G is defined by the tips A1 and B1. These tips map to species A and B, respectively, and the first node in the standard species tree that includes species A and B is 1S. Therefore, $M(1G) = 1S$. This process is repeated for every internal node in the gene tree until all nodes are mapped. For each n_g there is a single possible node in the species tree to which it can map; however, nodes in the species tree can have multiple nodes map to them. Nodes in the gene tree are said to be duplication nodes when they map to the same species tree node for at least one of their descendants (for example, see maps for the top of Figure 3). Parsimony scores for a reconciliation are then calculated as the sum of the number of duplications and losses in that gene tree (see Supplementary Methods). This process hinges on the fact that the mapping function is initialized with the tips of the gene tree mapped to their corresponding species label in the species tree.

In a MUL-tree, repeated clades represent the sub-genomes of the polyploid species, and their placement in the species tree defines parental lineages of the polyploid event (Figure 2C). Given a gene tree, we proceed with the LCA mapping algorithm as described above, except that any tip that maps to a polyploid species now has two possible initial maps: to either of the sub-genomes represented in the MUL-tree (species "B" in Figure 3, bottom). We run LCA mapping

with a tip initialized to one sub-genome first, and then we run LCA mapping again with that same tip initialized to the other sub-genome, giving us two maps and two reconciliation scores for the single gene tree. We then apply the parsimony principle for these two possible maps: whichever initial mapping results in the lowest score is the correct map. If there is more than one gene in the gene tree from the polyploid clade we try all possible combinations of initial maps. The map to a MUL-tree results in fewer duplications and losses being counted because some mappings can now be accommodated by the extra identical taxa in the species tree (Figure 3).

This process is applicable for any number of genes in any number of polyploid species, but the algorithm becomes very slow for large polyploid clades because of the large number of combinations of initial maps to consider. Given that a gene tree has m genes represented from polyploid species, the run time for mapping this gene tree will be $O(2^m n)$, since the mapping algorithm itself is linear for a tree with n nodes (31) and we perform 2^m maps (Supplementary Figure 1b). However, we devised several methods to accelerate the process of choosing the correct map, using context in both the gene tree and MUL-tree (Supplementary Methods).

This method gives a reconciliation score for mapping a single gene tree to a single MUL-tree. We can also apply this to reconcile a set of gene trees to a single MUL-tree and sum all the scores to obtain a total reconciliation score for that MUL-tree. However, this assumes that the placement of the polyploidy event is already known, since a single MUL-tree represents a single polyploid scenario.

We have implemented a search strategy to find the most parsimonious placement of a polyploidy event, given a standard species tree. We define two nodes of interest in a standard species tree that we use to build a MUL-tree. Node H1 defines the location where the polyploid clade is represented in the standard species tree (as in Figure 2a). Node H2 defines the location

of the second, unrepresented parental lineage (shown in Figure 2c). When H1 is specified, the sub-tree that is rooted by it and the branch that it subtends are copied and placed on the branch that is subtended by H2. Our modified LCA mapping algorithm is run on the resulting MUL-tree and a set of gene trees, and a total reconciliation score is obtained by summing across scores for each gene tree. The algorithm can be limited to a specified pair of H1 and H2 nodes, or only a specified H1 node (searching for H2), or no nodes specified (searching for both H1 and H2). The MUL-tree defined by a particular H1 and H2 with the lowest total reconciliation score reveals the location and type of the most parsimonious polyploidy event. Placement of H1 and H2 on the same node in the species tree simply represents an autopolyploid event.

Results

Performance of algorithm

We have implemented our algorithm in the software package GRAMPA (Gene-tree Reconciliation Algorithm with MUL-trees for Polyploid Analysis; available at https://github.com/gwct/grampa). The main inputs of the program are a species tree (standard or MUL) and a set of gene trees. If a standard tree is input with H1 specified, GRAMPA will search for the optimal placement of the H2 node. If no H1 and H2 nodes are defined, GRAMPA will generate MUL-trees based on all possible H1 and H2 nodes. In either case GRAMPA will return a reconciliation score for each MUL-tree considered, including the total number of duplications and losses. If a MUL-tree is input (i.e. H1 and H2 specified), GRAMPA will return a total reconciliation score for the tree and individual duplication and loss scores for the gene trees.

We checked that our modified LCA mapping algorithm counted the correct number of duplications and losses by manually reconciling a small set (25) of gene trees onto 8 MUL-trees that represent varying cases of gain and loss (Supplementary Table S1). Our method always agrees with the expected counts for each type of event. GRAMPA's search method was then validated using larger sets of gene trees simulated using JPrIME (32). Every scenario under which we tested it returned the expected result (Supplementary Table S2). For example, when GRAMPA was given a set of gene trees simulated from a polyploidy event and a corresponding standard species tree with only one sub-genome represented, it always found the correct MUL-tree. We then assessed GRAMPA's performance when given a standard species tree and gene trees simulated from that species tree; in this scenario no polyploidy has occurred. However, since GRAMPA always returns a MUL-tree, we hypothesized that in this case it will simply return a MUL-tree with an autopolyploidy event occurring along the branch with the largest number of inferred duplicates. Indeed this turned out to be the case. This behavior will be a useful null hypothesis for cases of allopolyploidy. That is, if GRAMPA returns an allopolyploid MUL-tree then we can have more confidence that an allopolyploidy event has occurred (given accurate gene trees), as the MUL-tree returned when there is no WGD will be represented as an autopolyploid.

Analysis of baker's yeast

We revisited the interesting case of the WGD occurring in the ancestor of *Saccharomyces cerevisiae* (baker's yeast), a well-known example of polyploidy (33, 34). While early authors were circumspect about whether this WGD was an auto- or allopolyploid (33, 34), Marcet-Houben and Gabaldón (2015) recently postulated that this clade (labeled as the "BY" clade in

Figure 4) was the result of an allopolyploidy event. These authors detected a mismatch in the timing of duplications inferred by count methods and classic LCA reconciliation, as would be expected given allopolyploidy. This led to the conclusion that an ancient hybridization occurred to create an allopolyploid. This hybridization was inferred to have been between an ancestor of the ZT clade (Figure 4) and an extinct lineage sister to the KLE, ZT, and modern BY clades (node n3 in Figure 4). However, the phylogenetic methods employed by the authors to identify the parental lineages of the allopolyploid could not naturally deal with reticulation in a standard species tree, and thus may have been misled by problems similar to those outlined above.

We used 4,004 gene trees (after filtering, see Supplementary Methods) across 27 yeast species (Supplementary Figure S2) from ref. (19) to reinvestigate the polyploid history of baker's yeast using GRAMPA. We observed that the optimal MUL-tree inferred by GRAMPA has a reconciliation score of 153,368. We then compared this score to the scores of MUL-trees representing two alternative hypotheses. First, we wanted to confirm that this was an allopolyploidy event by comparing the optimal score reported above to the lowest scoring autopolyploid MUL-tree, which had a score of 159,937. Because the optimal MUL-tree has a score much lower than the autopolyploid MUL-tree, we confirm the result from ref. (19) that the modern baker's yeast clade is the result of an allopolyploidy event.

We then considered GRAMPA's optimal MUL-tree compared to the MUL-tree that corresponds to the allopolyploid scenario proposed by ref. (19). We find that the optimal MUL-tree scores much lower than the one proposed previously, which had a score of 166,898. This indicates that the hybridization event occurred between slightly different lineages than originally proposed. Our results suggest that the most probable parental lineages are the ancestor of a clade formed by *Z. rouxii* and *T. delbrueckii* (the so-called ZT clade) and an extinct lineage sister to

the ZT clade (Figure 4). GRAMPA reported a virtual tie between two MUL-trees (Supplementary Table S3), both supporting the extinct lineage sister to ZT as a parent, and differing as to whether the *Z. rouxii* or *T. delbrueckii* lineage was the other parent. Given the short branch subtending the ZT clade (19), and the generally low bootstrap support associated with most of the gene trees, we infer that the ancestor of this clade was one of the parental lineages. Our results support the claim of ref. (19) for an allopolyploid origin of the clade including baker's yeast, but conclude that the lineages that hybridized to form this clade are slightly different than previously proposed.

Analysis of wheat

We also applied GRAMPA to 9,147 gene trees from the clade including the hexaploid species, *Triticum aestivum*, commonly known as bread wheat. This species is the result of two hybridization events, each leading to a WGD (35). Analysis of this clade is an especially useful example to demonstrate the accuracy of GRAMPA because the relationships between sub-genomes are known, and genes can be assigned to their sub-genome of origin. With genes labeled according to sub-genome, standard reconciliation can be performed in an approach similar to ours but with a pre-labeled MUL-tree (36, 37). It also presents an interesting test because the current implementation of our algorithm is only designed to map one WGD per tree.

To show that GRAMPA is able to recover the correct MUL-tree for the clade including *T. aestivum* and nine other Poaceae species (Supplementary Figure S3), we started by analyzing genes from two sub-genomes at a time, and removed all labels associating genes with sub-genomes. When we allowed GRAMPA to search for the optimal MUL-tree we recovered the one

with correct sub-genome relationships every time (Supplementary Figure S4, Supplementary Table S4). Interestingly, when we searched for the optimal MUL-tree without removing any sub-genomes (i.e. analyzing all three at once), GRAMPA correctly recovered trees including the two WGDs as the top two MUL-trees, with an autopolyploid tree ranked third (Supplementary Figure S5, Supplementary Table S4). This behavior is especially useful because it means that GRAMPA could be used to search for multiple allopolyploidy events.

Discussion

We have developed a method to accurately identify whether an allpolyploidy event has occurred and to date it in a phylogenetic context. This allows us to identify the parental lineages of the polyploid species resulting from the hybridization. Our method also allows us to accurately infer the number of duplications and losses in a clade containing an allopolyploid. This is the first general method that we know of to perform these types of analyses, and it is applicable in a wide variety of contexts. Using our method to re-analyze bread wheat, the results of our algorithm always align with the accepted relationships between sub-genomes. Application to the WGD in the baker's yeast clade has confirmed an allopolyploid origin, but has identified different lineages involved in the hybridization event than the ones previously postulated (19).

K_s- or gene-tree-based methods for dating polyploidy events fail because they treat the two homologous copies of a gene arising from allopolyploidy as paralogs – that is, homologous genes related by a duplication event at their most recent common ancestor (38). However, these two genes are more akin to orthologs, as they are actually related by a speciation event at their most recent common ancestor (Figure 2c; 39). In fact, all of the genes in different sub-genomes

within an allopolyploid are related in this way, and the choice of which sub-genome to represent in a standard species tree as the "correct" set of relationships is arbitrary. Because K_s-based methods, reconciliation-based methods, and species overlap methods attempt to date WGDs as the time when a large number of paralogs are formed, they do not accurately capture the dynamics of allopolyploids (though they do work for autopolyploids). The term "homoeolog" was originally applied to relationships between chromosomes in allopolyploids (39), and seems most appropriate as a descriptor of the genealogical relationships between homologous genes within allopolyploids: not quite orthologs and not quite paralogs (39). Recognizing that homoeologs have relationships differing from those occurring between homologous genes within autopolyploids helps to prevent some of the incorrect inferences discussed in this paper.

It is important to repeat the point that our algorithm will always return a MUL-tree. This leads to several problems, most notably that we cannot directly compare the reconciliation score on a standard tree to that of one on a MUL-tree. Because a MUL-tree will naturally reduce the counts of duplications and losses in any reconciliation, our method cannot be used to determine whether a WGD has occurred. Nevertheless, we may be able to distinguish allopolyploidy from either no polyploidy or autopolyploidy if we assume that when no polyploidy has occurred, a MUL-tree that is autopolyploid like (i.e. H1 and H2 are the same node) will be returned. This approach treats autopolyploidy as the null hypothesis to be tested against, though there are cases of allopolyploidy that will appear to be autopolyploid. For instance, if the two parental lineages of an allopolyploid were sister to each other and both have subsequently gone extinct the resulting MUL-tree would look like an autopolyploid. It should also be stressed that none of the distance- or gene-tree-based methods discussed here can be used alone to determine whether a

WGD event occurred in the first place. Our algorithm can now be used alongside these other approaches in providing confidence that one has occurred.

Importantly, all reconciliation inferences are only as reliable as the underlying gene trees. Errors in gene tree reconstruction, or incongruence caused by biological phenomena such as incomplete lineage sorting (ILS), will cause problems for reconciliation algorithms (40). This problem is evident in our case studies, as gene trees from the relatively recent bread wheat allopolyploidy had individually much higher bootstrap support, and a single MUL-tree was clearly optimal (i.e. there was not a lot of disagreement among the relationships implied by gene trees). On the other hand, the WGD in the baker's yeast clade is approximately 100 million years old (19), and most gene trees used here had overall low bootstrap support, with conflicting signals as to the optimal MUL-tree. Though we can strongly reject the lineages proposed by ref. (19) as the ones that hybridized to form the allopolyploid, GRAMPA returned two MUL-trees with much lower scores as potential candidates for the second parental lineage of the clade. Our conclusion that an ancestor of the ZT clade was the second parent resolves conflicts between these two trees, but the low resolution in the gene trees may pose an absolute limit to the resolution of any approach. ILS will continue to be a problem for reconciliation algorithms, though solutions have been proposed to deal with this process (e.g. 41, 42). In the future it would be valuable to implement a similar solution for GRAMPA, or a solution based on ILS in species networks (24).

The algorithm and associated software presented here should allow researchers to re-examine many published cases of polyploidy, in order to determine whether these events were auto- or allopolyploidy. While many clades of plants often have multiple WGD events within them, our re-analysis of the wheat data gives us confidence that our method can be expanded to

identify multiple polyploidy events in the same tree. For cases with only a single WGD, our method provides accounting of duplication and loss, as well as the timing of these events.

Acknowledgements

We thank Clara Boothby for feedback on the manuscript, and Ben Moore for helpful information about the wheat data. This work was partially funded by National Science Foundation grant DBI-1564611 to MWH.

References

1. Ohno S (1970) Evolution by gene duplication. Springer-Verlag.
2. Dehal P, Boore JL (2005) Two rounds of whole genome duplication in the ancestral vertebrate. *PLoS Biol* 3(10):e314.
3. Van de Peer Y (2004) Computational approaches to unveiling ancient genome duplications. *Nat Rev Genet* 5(10):752-763.
4. Barker MS, Arrigo N, Baniaga AE, Li Z, Levin DA (2015) On the relative abundance of autopolyploids and allopolyploids. *New Phytol* 210:391-398.
5. Werth CR, Windham MD (1991) A model for divergent, allopatric speciation of polyploid Pteridophytes resulting from silencing of duplicate-gene expression. *The American Naturalist* 137(4):515-526.
6. Lynch M, Force A (2000) The probability of duplicate gene preservation by subfunctionalization. *Genetics* 154(1):459-473.
7. Mayrose I et al. (2011) Recently formed polyploid plants diversify at lower rates. *Science* 333:1257.
8. Muir CD, Hahn MW (2015) The limited contribution of reciprocal gene loss to increased speciation rates following whole genome duplication. *The American Naturalist* 185:70-86.
9. Crow KD, Wagner GP (2006) What is the role of genome duplication in the evolution of complexity and diversity? *Mol Biol Evol* 23(5):887-892.
10. Adams KL, Wendel JF (2005) Polyploidy and genome evolution in plants. *Curr Opin Plant Biol* 8(2):135-141.
11. Soltis PS, Soltis DE (2009) The role of hybridization in plant speciation. *Annu Rev Plant Biol* 60:561-588.
12. Edger PP et al. (2015) The butterfly plant arms-race escalated by gene and genome duplications. *Proc Natl Acad Sci USA* 112(27):8362-8366.
13. Lynch M, Conery JS (2000) The evolutionary fate and consequences of duplicate genes. *Science* 290(5495):1151-1155.
14. Blanc G, Wolfe KH (2004) Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. *The Plant Cell* 16(7):1667-1678.

15. Barker MS et al. (2008) Multiple paleopolyploidizations during the evolution of the Compositae reveal parallel patterns of duplicate gene retention after millions of years. *Mol Biol Evol* 25(11):2445-2455.
16. Li Z et al. (2015) Early genome duplications in conifers and other seed plants. *Sci Adv* 1(10):e1501084.
17. Jiao Y et al. (2011) Ancestral polyploidy in seed plants and angiosperms. *Nature* 473(7345):97-100.
18. Bowers JE, Chapman BA, Rong J, Paterson AH (2003) Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature* 422(6930):433-438.
19. Marcet-Houben M, Gabaldón T (2015) Beyond the whole genome duplication: Phylogenetic evidence for an ancient interspecies hybridization in the baker's yeast lineage. *PLoS Biol* 13(8):e1002220.
20. Rabier C-E, Ta T, Ané C (2014) Detecting and locationg whole genome duplications on a phylogeny: A probabilistic approach. *Mol Biol Evol* 31(3):750-762.
21. Tiley GP, Ané C, Burleigh JG (2016) Evaluating and characterizing ancient whole-genome duplications in plants with gene count data. *Genome Biol Evol* 8(4):1023-1037.
22. Linder CR, Rieseberg LH (2004) Reconstructing patterns of reticulate evolution in plants. *Am J Bot* 91(10:1700-1708.
23. Huber KT, Moultan V (2006) Phylogenetic networks from multi-labelled trees. *J Math Biol* 52(5):613-632.
24. Jones G, Sagitov S, Oxelman B (2013) Statistical inference of allopolyploid species networks in the presence of incomplete lineage sorting. *Syst Biol* 62(3):467-478.
25. Huber KT, Oxelman B, Lott M, Moultan V (2006) Reconstructing the evolutionary history of polyploids from multilabeled trees. *Mol Biol Evol* 23(9):1784-1791.
26. Lott M et al. (2009) Inferring polyploid phylogenies from multiply-labeled gene trees. *BMC Evol Biol* 9:216.
27. Huson DH, Rupp R, Scornavacca C (2010) Phylogenetic networks from trees. *Phylogenetic Networks,* Cambridge University Press.
28. Goodman M, Czelusniak J, William Moore G, Romero-Herrera AE, Matsuda G (1979). *Syst Biol* 28(2):132-163.
29. Page RDM (1994) Maps between trees and cladistic analysis of historical associations among genes, organisms, and areas. *Syst Biol* 43(1):58-77.
30. Guigó R, Muchnik I, Smith TF (1996) Reconstruction of ancient molecular phylogeny. *Mol Phylogenet Evol* 6(2):189-213.
31. Zmasek CM, Eddy SR (2001) A simple algorithm to infer gene duplication and speciation events on a gene tree. *Bioinformatics* 17(9):821-828.
32. Sjöstrand J, Arvestad L, Lagergren J, Sennblad B (2013) GenPhyloData: Realistic simulation of gene family evolution. *BMC Bioinformatics* 14:209.
33. Wolfe KH, Shields DC (1997) Molecular evidence for an ancient duplication of the entire yeast genome. *Nature* 287(6634):708-713.
34. Kellis M, Birren BW, Lander ES (2004) Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*. *Nature* 428(6983):617-624.
35. Petersen G, Seberg O, Yde M, Berthelsen K (2006) Phylogenetic relationship of *Triticum* and *Aegilops* and evidence for the origin of the A, B, and D genomes of common wheat (*Triticum aestivum*). *Mol Phylogenet Evol* 39(1):70-82.

36. Altenhoff AM et al. (2015) The OMA orthology database in 2015: Function predictions, better plant support, synteny view, and other improvement. *Nucleic Acids Res* 43(Database issue):D240-D249.
37. Bolser DM, Kerhornou A, Walts B, Kersey P. (2015) Triticeae resources in Ensembl Plants. *Plant Cell Physiol* 56(1):e3.
38. Fitch WM (1970) Distinguishing homologous from analogous proteins. *Syst Zool* 19(2):99-113.
39. Glover NM, Redestig H, Dessimoz C (2016) Homoeologs: What are they and how do we infer them? *Trends Plant Sci* S1360-1385(16):00059-5.
40. Hahn MW (2007) Bias in phylogenetic tree reconciliation methods: implications for vertebrate genome evolution. *Genome Biol* 8(7):R141.
41. Vernot B, Stolzer M, Goldman A, Durand D (2008) Reconciliation with non-binary species trees. *J Comput Biol* 15(8):981-1006.
42. Rasmussen MD, Kellis M (2012) Unified model of gene duplication, loss, and coalescence using a locus tree. *Genome Res* 22(4):755-765.
43. Durand D, Halldórsson BV, Vernot B (2006) A hybrid micro-macroevolutionary approach to gene tree reconstruction. *J Comput Biol* 13(2):320-335.
44. Kersey PJ et al. (2016) Ensembl Genomes 2016: more genomes, more complexity. *Nuc Acids Res* 44(D1):D574-D580.
45. Junier T, Zdobnov EM (2010) The Newick Utilities: high-throughput phylogenetic tree processing in the UNIX shell. *Bioinformatics* 26:1669-1670.

Figures

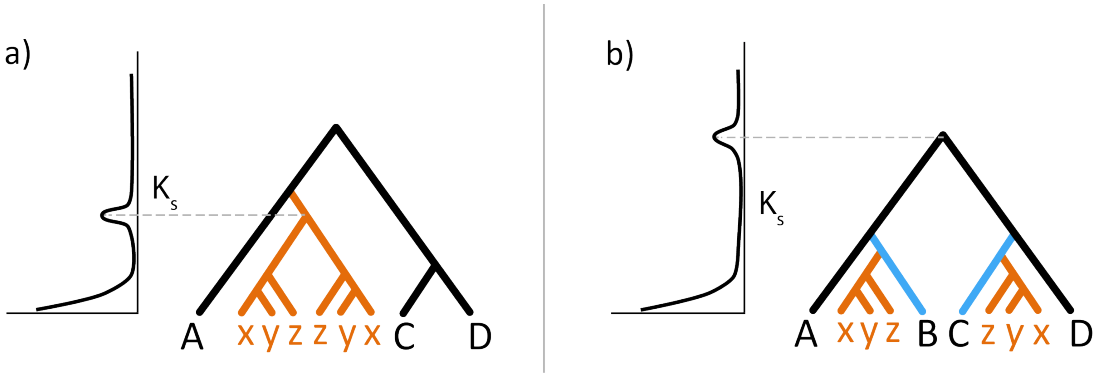

Figure 1. Methods to identify the timing of WGD. (***a***) K_s-based methods can correctly date cases of autopolyploidy. Here the peak in the distribution of K_s (shown as a density plot on the left) corresponds to the duplication node in the tree. (***b***) K_s-based methods incorrectly date cases of allopolyploidy. Here a hybridization event occurred between species B and C (in blue) resulting in the allopolyploid lineage that gave rise to the XYZ clade (orange). In cases like this the peak in the distribution of K_s corresponds to the most recent common ancestor of the two parental lineages, rather than the timing of the WGD. Similar results would be found using the reconciliation and species-overlap methods discussed in the text.

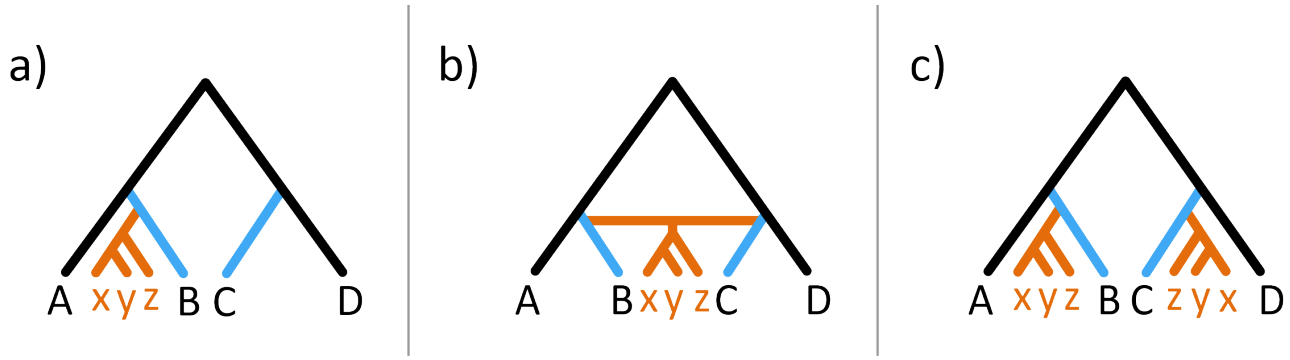

Figure 2. Representation of allopolyploid clades. Given that a hybridization event occurred between species B and C (in blue) resulting in the allopolyploid lineage that gave rise to the XYZ clade (orange), there are three ways of representing the species relationships. (***a***) A standard species tree does not show both sub-genomes present in the allopolyploid species, nor does it identify the parental lineages that hybridized. (***b***) Phylogenetic networks can correctly display the hybridization events and evolutionary history of allopolyploid species; however, they do not explicitly represent both sub-genomes. (***c***) Multi-labeled species trees (MUL-trees) are able to represent both sub-genomes of the polyploid species and to identify the parental lineages involved in the hybridization event. The timing of hybridization is implicit in this representation as the point at which both sub-genomes appear.

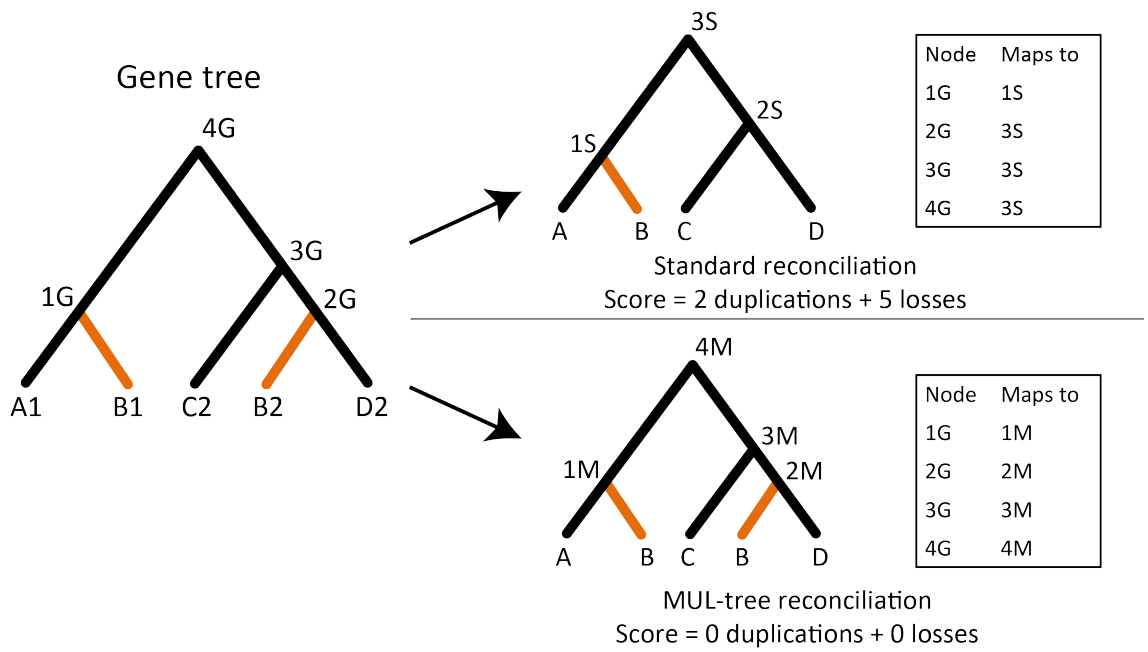

Figure 1. Reconciliation with a gene tree from an allopolyploid lineage. (***Left***) A representative gene tree is shown, with every gene labeled, including the two copies from allopolyploid species B. Internal nodes are also labeled to understand the mappings to the right. (***Top***) Standard reconciliation with only one sub-genome represented in the species tree, with maps between gene tree nodes and species tree nodes shown (species tree nodes are labeled 1S, 2S, and 3S). In this case extra duplications and losses are inferred, and the duplications are placed ancestral to the actual parental lineages (A and D) of the allopolyploid. (***Bottom***) Reconciliation with a MUL-tree, with maps between gene tree nodes and MUL-tree nodes shown (MUL-tree nodes are labeled 1M, 2M, 3M, and 4M). Using our algorithm no extra duplications or losses are inferred.

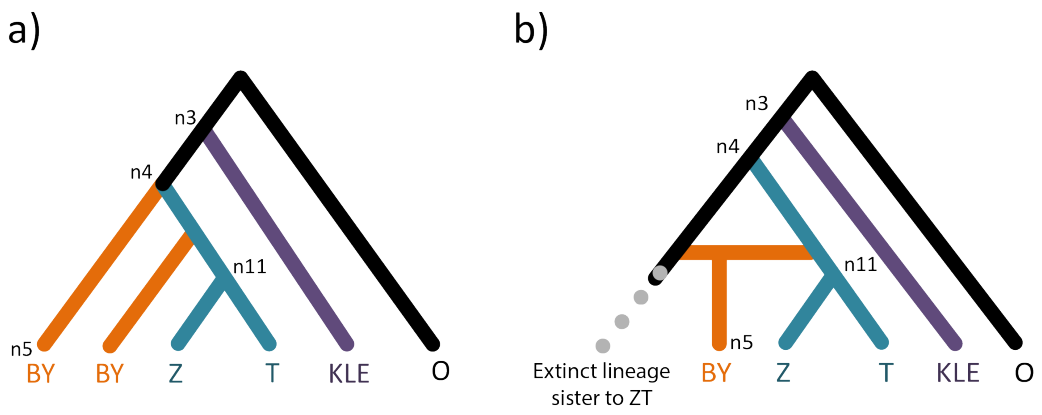

Figure 4. The inferred MUL-tree and species network for the baker's yeast data. BY: baker's yeast clade, Z: *Z. rouxii,* T: *T. delbruekii,* KLE: KLE clade, O: Outgroups. See Supplementary Figure S3 for full tree and all node labels. (***a***) The optimal MUL-tree inferred when searching for both H1 and H2. H1 was confirmed to be on the branch leading to the baker's yeast clade (as normally represented in species tree), while H2 was inferred to be ancestral to the ZT clade (including *Z. rouxii* and *T. delbruekii*). (***b***) The optimal MUL-tree as a network. This representation highlights the two parental lineages of the allopolyploid event: the common ancestor of *Z. rouxii* and *T. delbruekii*, and an extinct lineage sister to the ZT clade.