

# Lawrence Berkeley National Laboratory

## LBL Publications

### Title

Preferential retention of genes from one parental genome after polyploidy illustrates the nature and scope of the genomic conflicts induced by hybridization

### Permalink

<https://escholarship.org/uc/item/8mk7r5gw>

### Journal

PLOS Genetics, 14(3)

### ISSN

1553-7390

### Authors

Emery, Marianne  
Willis, M Madeline S  
Hao, Yue  
[et al.](#)

### Publication Date

2018

### DOI

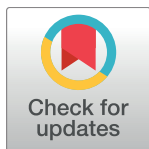
10.1371/journal.pgen.1007267

Peer reviewed

RESEARCH ARTICLE

# Preferential retention of genes from one parental genome after polyploidy illustrates the nature and scope of the genomic conflicts induced by hybridization

Marianne Emery<sup>1</sup>, M. Madeline S. Willis<sup>2</sup>, Yue Hao<sup>3</sup>, Kerrie Barry<sup>4</sup>, Khouanchy Oakgrove<sup>4</sup>, Yi Peng<sup>4</sup>, Jeremy Schmutz<sup>4,5</sup>, Eric Lyons<sup>6</sup>, J. Chris Pires<sup>1,7,8</sup>, Patrick P. Edger<sup>9,10</sup>, Gavin C. Conant<sup>3,11,12,13\*</sup>



**1** Division of Biological Sciences, University of Missouri-Columbia, Columbia, Missouri, United States of America, **2** Department of Biochemistry, University of Missouri-Columbia, Columbia, Missouri, United States of America, **3** Bioinformatics Research Center, North Carolina State University, Raleigh, North Carolina, United States of America, **4** Department of Energy Joint Genome Institute, Walnut Creek, California, United States of America, **5** HudsonAlpha Institute for Biotechnology, Huntsville, Alabama, United States of America, **6** School of Plant Sciences, University of Arizona, Tucson, Arizona, United States of America, **7** Informatics Institute, University of Missouri-Columbia, Columbia, Missouri, United States of America, **8** Bond Life Sciences Center, University of Missouri-Columbia, Columbia, Missouri, United States of America, **9** Department of Horticulture, Michigan State University, East Lansing, Michigan, United States of America, **10** Ecology, Evolutionary Biology and Behavior, Michigan State University, East Lansing, Michigan, United States of America, **11** Division of Animal Sciences, University of Missouri-Columbia, Columbia, Missouri, United States of America, **12** Program in Genetics, North Carolina State University, Raleigh, North Carolina, United States of America, **13** Department of Biological Sciences, North Carolina State University, Raleigh, North Carolina, United States of America

\* [gconant@ncsu.edu](mailto:gconant@ncsu.edu)

OPEN ACCESS

**Citation:** Emery M, Willis MMS, Hao Y, Barry K, Oakgrove K, Peng Y, et al. (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(3): e1007267. <https://doi.org/10.1371/journal.pgen.1007267>

**Editor:** Nathan M. Springer, University of Minnesota, UNITED STATES

**Received:** October 20, 2017

**Accepted:** February 21, 2018

**Published:** March 28, 2018

**Copyright:** This is an open access article, free of all copyright, and may be freely reproduced, distributed, transmitted, modified, built upon, or otherwise used by anyone for any lawful purpose. The work is made available under the [Creative Commons CC0](https://creativecommons.org/licenses/by/4.0/) public domain dedication.

**Data Availability Statement:** All data used was from published genomes except for the genome of *C. violacea*, which is freely available from the CoGe comparative genomics platform under accession number 23822. For access to the published genomes: 'S1 Data' lists the CoGe accession numbers for all of the plant genomes we use (<https://genomeevolution.org/coge/>). All of the yeast genomic data was retrieved from the YGOB project (<http://yglob.ucd.ie>).

## Abstract

Polyploidy is increasingly seen as a driver of both evolutionary innovation and ecological success. One source of polyploid organisms' successes may be their origins in the merging and mixing of genomes from two different species (e.g., allopolyploidy). Using POInT (the Polyploid Orthology Inference Tool), we model the resolution of three allopolyploidy events, one from the bakers' yeast (*Saccharomyces cerevisiae*), one from the thale cress (*Arabidopsis thaliana*) and one from grasses including *Sorghum bicolor*. Analyzing a total of 21 genomes, we assign to every gene a probability for having come from each parental subgenome (i.e., derived from the diploid progenitor species), yielding orthologous segments across all genomes. Our model detects statistically robust evidence for the existence of *biased fractionation* in all three lineages, whereby genes from one of the two subgenomes were more likely to be lost than those from the other subgenome. We further find that a driver of this pattern of biased losses is the co-retention of genes from the same parental genome that share functional interactions. The pattern of biased fractionation after the *Arabidopsis* and grass allopolyploid events was surprisingly constant in time, with the same parental genome favored throughout the lineages' history. In strong contrast, the yeast allopolyploid event shows evidence of biased fractionation only immediately after the event, with balanced gene losses more recently. The rapid loss of functionally associated genes from a single subgenome is difficult to reconcile with the action of genetic drift and suggests

**Funding:** This work was supported by the US National Science Foundation ([www.nsf.gov](http://www.nsf.gov)) under grant number NSF-IOS-1339156 (YH, EL, JCP and GCC) and by the US Department of Energy ([www.doe.gov](http://www.doe.gov)) under contract Number DE-AC02-05CH11231 (KB, KO, YP, JS). The work conducted by the U.S. Department of Energy Joint Genome Institute, a DOE Office of Science User Facility, is supported by the Office of Science of the U.S. Department of Energy under the contract number above. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Competing interests:** The authors have declared that no competing interests exist.

that selection may favor the removal of specific duplicates. Coupled to the evidence for continuing, functionally-associated biased fractionation after the *A. thaliana* At- $\alpha$  event, we suggest that, after allopolyploidy, there are functional conflicts between interacting genes encoded in different subgenomes that are ultimately resolved through preferential duplicate loss.

## Author summary

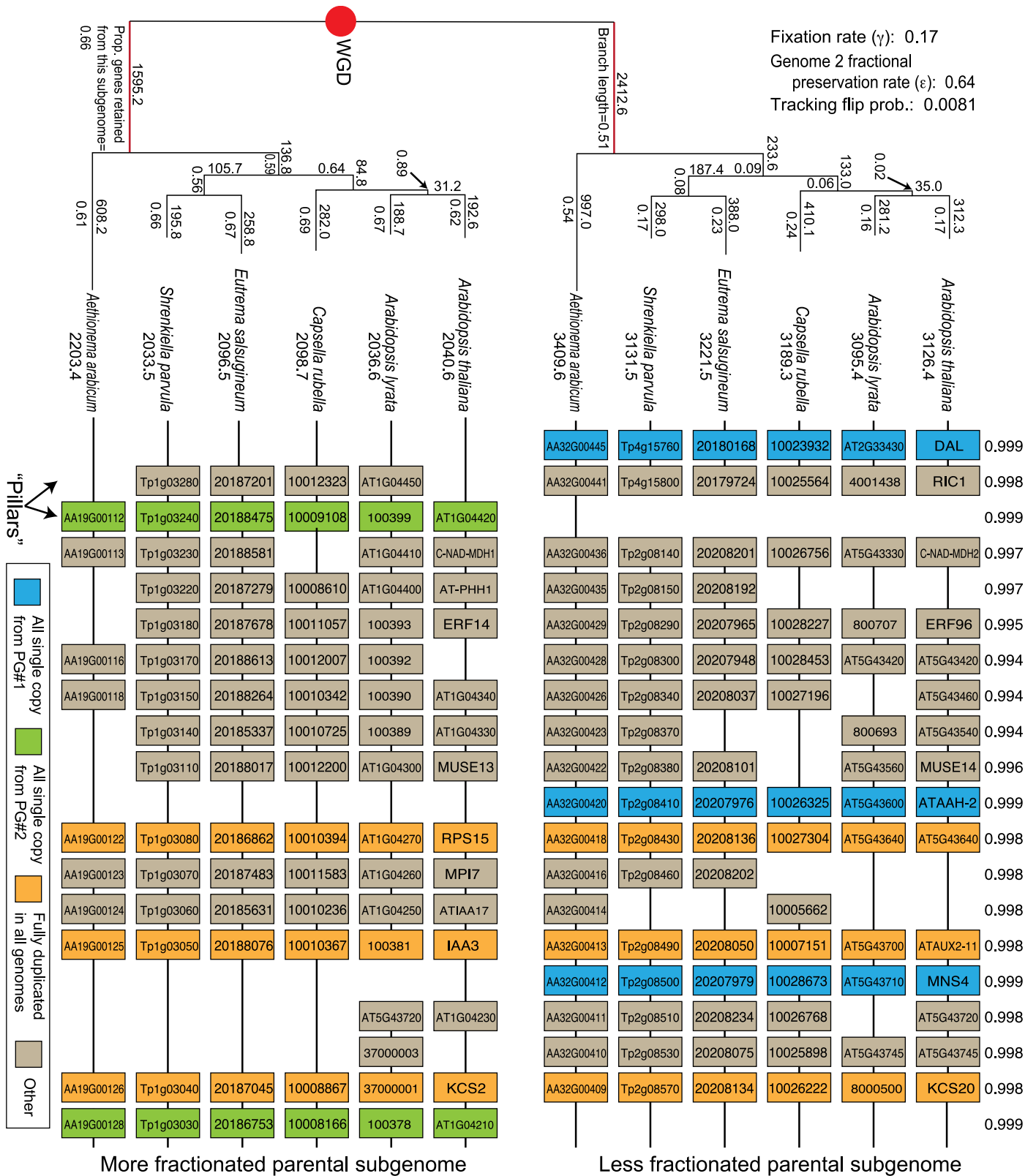
Genome duplications/polyploidies can transiently double an organism's gene content. However, this doubled condition is unstable and descendants of polyploid founders rapidly lose many of their duplicate genes. Here, we describe a phylogenomic pipeline that allows us to trace this history of gene loss across a set of modern genomes that all descend from the same three polyploidy events (two in flowering plants and one in yeasts). Polyploidy often occurs by the hybridization of related but not identical lineages, and previous studies in single polyploid genomes have identified a tendency for one lineage to lose fewer genes after polyploidy than the other (known as biased fractionation). However, single genome studies can be misled into inferring biased fractionation by incorrectly assuming that short regions of shared gene order are derived from the same parental genome. By phylogenetically modeling the resolution of these three polyploidy events across 21 genomes (10 plants and 11 yeasts), we confirm the existence of biased fractionation in plants and provide new evidence for it in yeasts (where it occurred only for a short interval post-polyploidy). We also show that genes from alternative parental genomes tend to encode products that do not physically interact, suggesting that selection to maintain function in co-adapted complexes helped to drive this bias in loss patterns.

## Introduction

Polyploidy events (also known as whole-genome duplications or WGDs) are widespread across the eukaryotic tree of life [1] and have long interested geneticists and evolutionary biologists for reasons varying from the nature of interspecific crosses to the organismal effects of changes in gene copy number to the origins of novel functions in evolution [2–5]. Recent work has associated genome duplications with evolutionary innovations [6–9] and with shifts in net diversification rates [10–13].

Understanding how polyploidy contributes to these biologically important processes requires coming to grips with three key patterns in the evolution of polyploid genomes. The first is the rapid loss of genetic redundancy after polyploidy. Most WGD-created duplicate genes, termed “ohnologs” [14], do not survive: their losses start very soon after WGD [15–17] and may be governed epigenetically in this period [18]. The net result of such losses can be dramatic: only 551 of an estimated 5000 duplicate gene pairs produced by the WGD in yeast survive in the *Saccharomyces cerevisiae* genome [19]. Nonetheless, the footprint of WGD is clear in the extant patterns of double-conserved synteny [DCS; 20, 21]: homologs of genes from a single genomic region in a non-polyploid relative will be split between two regions in the polyploid genomes (upper and lower blocks of Fig 1).

The second key trend is that, despite the rapidity of these duplicate losses, they are nonrandom, with certain functional classes of genes being overly frequent among surviving ohnologs and other being overly rare. In both yeasts and angiosperms, genes involved in DNA repair



**Fig 1. POInT's inferences regarding the loss of genes post-WGD.** The At- $\alpha$  duplication produced two sets of homoeologous regions, one from the parental subgenome with more surviving genes ("Less fractionated subgenome," upper track) and one with fewer ("More fractionated subgenome," lower track). Genes in these tracks may have surviving duplicates in at least some taxa (orange/tan), or they may be single-copy in all species (blue if derived from the less fractionated

subgenome and green if from the more fractionated one). Under each taxon name is the number of single-copy genes predicted to have been retained from that parental subgenome in that taxon. The branch length (numbers under the branches of the *upper* tree) gives the value of  $\alpha \times \text{time}$  in the model of Fig 2B: larger values correspond to a relatively higher chance that a position with a ohnolog pair present at the start of a branch will be single-copy by its end. Numbers above the branches give POInT's estimate of the number of genes returned to single copy deriving from the less fractionated (upper panel) and more fractionated (lower panel) subgenomes, respectively. Under the branches of the lower tree are the branch-specific ratio of genes retained from subgenome #2 relative to subgenome #1: these values can be compared to the overall estimate of this parameter, which is 0.64, shown in the upper left. POInT's estimates of the other global parameters for this model are also given here. Above each pillar of genes is POInT's estimate of the posterior probability of the set of subgenome assignments depicted, relative to the other  $2^n - 1$  possible assignments (where  $n$  is the number of genomes). The two root branches are shown in red: these correspond to branches where the biased fractionation parameter  $\epsilon$  was allowed to differ from the rest of the tree in our analyses of temporal patterns of biased fractionation (Methods). Similar trees depicting loss events for the grass and yeast WGDs are given as S1 Fig.

<https://doi.org/10.1371/journal.pgen.1007267.g001>

and those targeted to the organelles were rapidly returned to single copy after WGD [22, 23]. On the other hand, genes coding for transcription factors, ribosomal proteins and kinases were over-retained in duplicate after independent WGD events across a phylogenetically wide range of organisms from amoebae and plants to vertebrates and yeasts [24–28]. The force underlying these convergent patterns of loss/retention is most likely selection to maintain dosage balance among interacting gene products [29]. The *dosage balance hypothesis* explains a variety of observations about the evolution of both polyploid and non-polyploid genomes, including the pattern of post-WGD duplicate retentions [28, 30–34] and the tendency of these same gene families not to undergo single gene duplications, where balance would be perturbed [26, 35, 36]. Similarly, genes in central network positions or whose products are parts of protein complexes are likely to show dosage phenotypes [37] and are over-retained after WGD [22, 38].

The third and final trend in post-WGD evolution is that when genes are lost, they are apparently not always lost equally from the paired DCS regions. This pattern of *biased fractionation* has been observed across a range of WGD events, primarily in angiosperms [39–41] but also from other taxa [42]. The most plausible current hypothesis for why biased fractionation occurs is that the events in question were allopolyploidies [39, 43]. In the alternative case of autopolyploidy, the paired genomic regions created by polyploidy are identical, and we know of no mechanism by which these identical regions could be stably marked over evolutionary time so as to differ strongly in their duplicate retention patterns. However, the converse is *not* true: the absence of biased fractionation cannot be taken as evidence for autopolyploidy. If the genomes that merged were from closely related taxa, bias is not necessarily expected.

As for the genetic mechanism behind the bias in ohnolog losses, biases in gene expression between the two subgenomes in recent allopolyploids appear to be common [44, 45] and the chromosomal regions with lowered expression also appear more prone to ohnolog loss [41, 46], leading to the suggestion that biased fractionation might result from a tendency for the ohnolog with lower expression to be less likely to show a fitness defect when lost. One potential source of these initial differences in expression might then be the difference in transposon load between the subgenomes of an allopolyploid, with the transposon-rich genome facing greater silencing and hence higher rates of gene loss [41, 43].

A difficulty that arises in the analysis of biased fractionation (BF) is that there has been a degree of circularity in its detection. Because rearrangements occur after WGD events, the duplicated regions in a paleopolyploid genome, which are identified by shared gene order or *synteny*, will be separated from each other by breakpoints. Within each syntenic block, the identification of the homeologous region with more retained genes is straight forward. However, when comparing a single polyploid genome to a diploid outgroup, it is difficult to formally refute the possibility that the parent-of-origin of the highly retained subgenome in one block might be the same as that of the lowly retained subgenome in another [but see; 42]. This difficulty in fact motivates the phylogenetic approach to studying polyploidy that we use

below. There are also other potential factors that might be involved in driving BF that remain to be investigated. For instance, the convergent pattern of rapid losses in gene coding for the DNA repair enzymes [22, 23] suggests that there may be incompatibilities between the versions of these genes contributed by the two allopolyploid parents. If such incompatibilities were common, they could contribute to BF by favoring retention from a single subgenome once the symmetry of a particular genetic module has been broken by the first loss.

Using POInT, the Polyploid Orthology Inference Tool, we analyzed the resolution of three WGD events, one in yeasts [20], one in the grasses [the  $\rho$  event; 47, 48] and the most recent event (At- $\alpha$ ) in *Arabidopsis thaliana* and its relatives. Using POInT's synteny-based estimates of post-WGD gene losses, we show that BF was a genome-wide evolutionary pattern after the At- $\alpha$  and  $\rho$  WGD events and persisted over long periods. In contrast, in yeasts we find evidence for BF only in a very short time interval post-WGD. In *Arabidopsis*, we also find that there is preferential co-retention of genes from the same subgenome whose products interact, as opposed to interactions involving proteins from different parents. Collectively, these results suggest that biased fractionation is at least in part a relic of conflicts between the paralogous genes contributed by the two parents at the time of the allopolyploidy.

## Methods

### Identifying double-conserved synteny blocks in polyploid genomes

Our previous POInT analyses in yeast were based on human curated datasets [19, 49]. We do not have such inferences for either the At- $\alpha$  or the grass  $\rho$  event. Instead, using experience from previous projects [40, 50], we developed a new pipeline for inferring the paralogous genomic regions created by a WGD in the genomes sharing that event. We then merged these regions of DCS [20, 21] across all polyploid genomes and sought an ancestral gene order that minimized the number of synteny breaks. Fig 1 shows examples of such DCS blocks for At- $\alpha$ .

The goal of the pipeline is to find a common set of DCS blocks shared by the genomes of the six Brassicaceae species that possess At- $\alpha$ : *Arabidopsis thaliana* [51], *Arabidopsis lyrata* [52], *Capsella rubella* [53], *Shrenkiella parvula* [54], formerly known as *Thellungiella parvula* or erroneously as *Thellungiella halophila* [55], *Eutrema salsugineum* [56], and *Aethionema arabicum* [57] and for the four grasses with  $\rho$ : *Brachypodium distachyon* [58], *Oropetium thomaicum* [59], *Setaria italica* [60] and *Sorghum bicolor* [61]. To do so, we used outgroup genomes that lacked the WGD in question. For the At- $\alpha$  event, we used the draft genome of the outgroup plant *Cleome violacea*, which split from the six taxa studied prior to that event [11]: it likewise lacks the WGD found in other taxa in the Cleomaceae [9]. The *C. violacea* genome is available from the CoGe comparative genomics portal (<https://genomevolution.org/coge/>) under accession number 23822. For the grass  $\rho$  event, we used the genome of the pineapple *Ananas comosus* as an outgroup [62]. CoGe accession numbers for all plant genomes used are listed in S1 Data.

The product of a WGD is a set of duplicated genes in a genome that each originate from a single ancestral gene. Here, the *C. violacea* and pineapple genomes give us an estimate of these ancestral loci, and we seek to place either one (e.g., a duplicate loss has happened) or two genes (the ohnologs survive) from the duplicated genome in a “pillar” with each such ancestral gene (see Fig 1). Genome annotation files for these 12 plant genomes were obtained from CoGe [63]. With these data in hand, the inference of the shared DCS blocks that serve as POInT's input is a three step process: 1) a homology search of each polyploid genome against the diploid outgroup, 2) inference of species-specific DCS blocks and 3) inference of a common set of DCS blocks across all genomes along with an estimate of their ancestral order at the time of the polyploidy.



**Step 1: Homology search.** For At- $\alpha$ , we used a fast homology search program based on the SeqAn package [64, 65] to identify pairs of homologous genes, one from a genome with At- $\alpha$  and one from *C. violacea*. We defined a pair of genes as being homologous for the purposes of DCS inference if their protein sequences: 1) share two 7 amino acid residue exact matches, 2) have the shorter sequence having 80% of the length of the longer, and 3) show 70% amino acid identity overall. Because of the greater evolutionary distances involved in the grass  $\rho$  event, we used a slower but more sensitive BLAST-based search, employing our tool GenomeHistory to do so [66, 67]. In this case, we required a maximal BLAST E-value of  $10^{-8}$  to identify matches between the four duplicated grasses and pineapple: we then used the same 70% identity and 80% aligned length cutoffs as used with At- $\alpha$  to select homologs.

**Step 2: Genome-specific DCS inference.** Sequence homology alone is insufficient to identify the DCS blocks given the angiosperms' history of nested polyploidy [1]. Instead, for the second step of the pipeline, we used gene order information (synteny) to identify which of the potentially many homologs in each polyploid genome are the WGD-produced ohnologs. We frame this problem as follows. First, we define a set  $A$  of  $n$  DCS blocks that consists of ancestral pillars  $A_i$  such that  $A_i \in A | 1 \leq i \leq n$ . Each pillar is linked to a unique gene from *C. violacea* or pineapple and has elements  $A_i(p_1)$  and  $A_i(p_2)$ , which represent the potential homologous genes created by WGD. Each pillar  $A_i$  also has associated a set of genes  $\{h_1, \dots, h_n\}$  from the polyploid genome that are homologous to the pillar's ancestral gene. A maximum of two of these homologs can be assigned to  $A_i(p_1)$  and  $A_i(p_2)$ . We next define  $O(A_1, \dots, A_n)$  to be the order of the pillars in  $A$  for our analysis. Hence,  $A_{O(i)}$  represents the  $i^{th}$  pillar in this ordering. For a given  $A_{O(i)}(p_k) | 1 \leq k \leq 2$ , define  $A_{O(i+j)}(p_k)$  such that  $j = \min(x, i+1 \leq x \leq n)$  where  $A_{O(i+x)}(p_k) \neq \emptyset$ : in other words,  $i+j$  is the next pillar after  $i$  in  $O(A_1, \dots, A_n)$  with an assigned gene for parental genome  $k$ . We define the score  $s$  of such a combination of homolog assignments and pillar orders:

$$s = \sum_{i=1}^n \sum_{k=1}^2 \begin{cases} 1 & | A_{O(i)}(p_k) \text{ and } A_{O(i+j)}(p_k) \text{ are neighbors} \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

In other words, the score is the sum of the number of positions in  $O(A_1, \dots, A_n)$  where the genes in each pillar are the genomic neighbors of the genes in the next non-empty position. We cannot simply use the pillar order seen in the outgroup, because neither *C. violacea* nor pineapple is the true ancestor of the WGD events in question: both have evolved independently for many millions of years. Instead we must optimize  $O(A_1, \dots, A_n)$ . Note that, throughout this pipeline, neighbor is understood to exclude any genes that are not part of the current analysis set. For instance, a gene in *Arabidopsis thaliana* with no identified *C. violacea* homolog is ignored in the neighbor computation because it could never appear in an ancestral pillar. By the same logic, any position for which  $A_{O(i)}(p_k)$  and  $A_{O(i+j)}(p_k)$  are not neighbors is defined as a synteny break, and, if this situation is true for both  $k = 1$  and  $k = 2$ , we refer to position  $i$  as having a double synteny break.

To infer the combination of the homolog assignments  $A_i(p_k) | 1 \leq i \leq n, 1 \leq k \leq 2$  and the ordering  $O(A_1, \dots, A_n)$ , we used simulated annealing [68, 69]. This algorithm proposes random changes to either  $O(A_1, \dots, A_n)$  or to the  $A_i(p_k)$  assignments with the goal of maximizing  $s$ , which recomputed after each such change. We used the extant *C. violacea* and pineapple gene orders as our initial orders and made increasingly long runs until longer run times no longer produced meaningfully higher values of  $s$ .

*A. thaliana* and its relatives share a history of WGD [26]: prior to the WGD- $\alpha$  event modeled here there was another WGD, termed WGD- $\beta$ , which is shared with *C. violacea*. One might wonder if our simulated annealing algorithm has mistaken synteny blocks surviving from WGD- $\beta$  for the more recent products of WGD- $\alpha$ . We suspect that any such errors are

quite rare for two reasons. First, *C. violacea* also experienced WGD- $\beta$  and hence also possesses the corresponding synteny blocks, meaning that they are accounted for in the inputs to our simulated annealing routines. Second, we only considered homology relationships between genes in *C. violacea* and in *A. thaliana*, *A. lyrata*, *C. rubella*, *S. parvula* and *E. salsgineum* with nonsynonymous divergence ( $K_a$ ) less than 0.1 and between *C. violacea* and *A. arabicum* with  $K_a \leq 0.2$ . As a result, between 41% and 45% of the genes from *C. violacea* have only a single homolog identified in the other 6 genomes and hence cannot represent ambiguous surviving blocks from WGD- $\beta$  in *C. violacea*. Hence, it is difficult to see how ancestral WGD- $\beta$  blocks would have infiltrated our inferences in significant numbers.

**Step 3: Inferring a global ancestral ordering for POInT analyses.** Using the four/six individually optimized set of ancestral pillars (for  $\rho$  and At- $\alpha$ , respectively) with assigned genes (the  $A_i(p_k)$  values for each genome), we extracted, for each genome, only ancestral pillars for which each gene in the pillar had synteny support (i.e., each gene was a neighbor of at least one other gene in that pillar set). Using the outgroup gene from each ancestral pillar as an index, we then merged all of these inferences. Because we required that at least one gene from each genome be in each pillar, the effect of this merging was to limit our analyses to a set of  $m = 7243$  and  $= 3091$  ancestral pillars for At- $\alpha$  and  $\rho$ , respectively. However, those pillars have shared syntenic support across all genomes. The optimal ancestral order for each extant genome differs, so once the ancestral pillars were assembled, we inferred a globally-optimal ancestral order  $O(AG_1..AG_m)$ , again using simulated annealing. The optimality criterion here was to maximize the number of neighbor relationships, but in this case the  $A_i(p_k)$  assignments were held constant and only  $O(AG_1..AG_m)$  was changed.

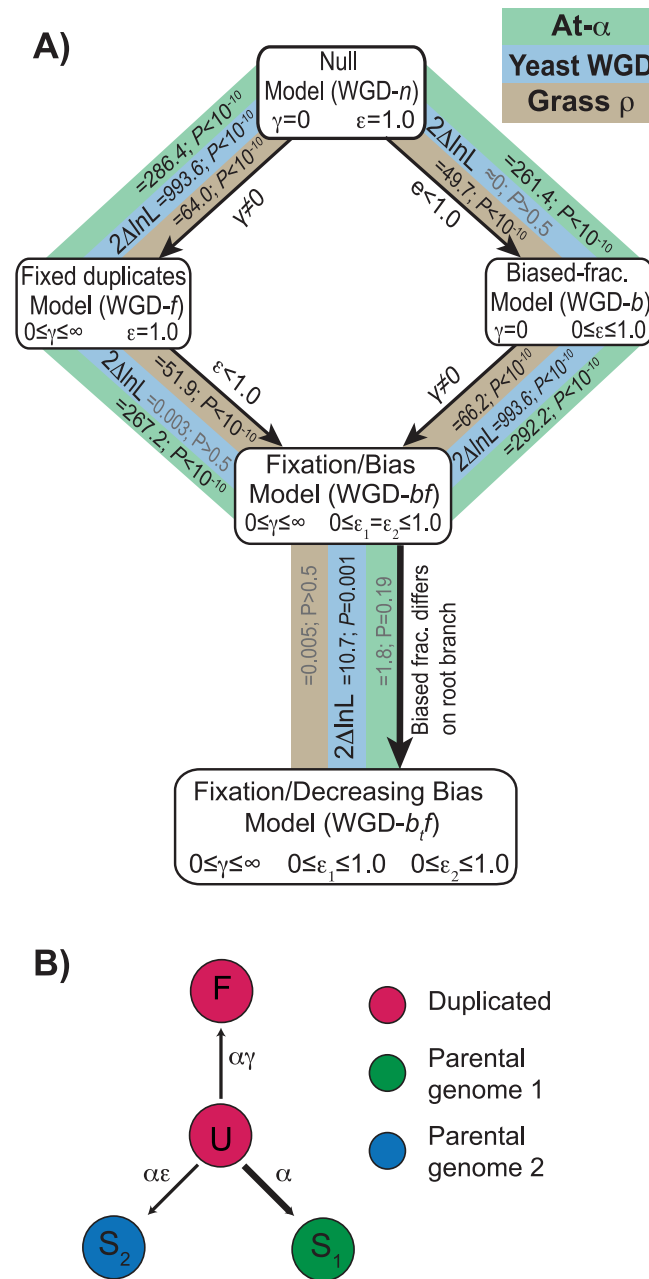
To assess the influence of the ancestral ordering on POInT's estimates, we fit the WGD-*bf* model (Fig 2B) to both the initial *C. violacea* order and to the 10 inferences of  $O(AG_1..AG_m)$  with the largest simulated annealing scores, using the order with the highest likelihood for further analyses (S1 Table). We similarly used the ancestral ordering of highest likelihood for our  $\rho$  analyses.

**Extracting a "high synteny" subset of ancestral pillars.** To assess if the fragmentation of synteny blocks was artificially leading us to invoke BF, we also extracted from our full At- $\alpha$  dataset a smaller set of ancestral loci with strong syntenic support, including only pillars with full syntenic support in at least one direction (e.g., two links per pillar per genome). The result was a dataset of  $m_h = 4556$  ancestral loci for which we also inferred an optimal ancestral ordering. No such analysis was performed for  $\rho$  due to the small total number of ancestral pillars found. Table 1 gives the parameter estimates from all four datasets for various ancestral orders.

## Modeling the evolution of WGD events with POInT

We have previously described POInT [22, 70], which fits a Markov model to duplicate loci created by WGD. The model has four states (Fig 2B), namely U (undifferentiated duplicated genes), F (fixed duplicate genes) and  $S_1$  and  $S_2$  (the single copy states): it is a generalization of a model proposed by Lewis [71]. Note that once the genes of each post-WGD genome have been assembled into ancestral pillars using the simulated annealing approach above, the *sequences* of the genes of the post-WGD genomes are never used again: all of POInT's inferences are based on shared DCS information. Since our prior work, we have completely re-written POInT to allow for user-defined evolutionary models, computing the resulting transition probabilities by exponentiating the user-supplied instantaneous rate matrix [72]. Using this new version of POInT, we fit five models to our four datasets (two from At- $\alpha$  and one each from the yeast and grass WGD events, Fig 2). We used likelihood ratio tests to assess whether more complex models better fit the loss data than did simpler models [73].





**Fig 2. Modeling WGD resolution with POInT.** We employed a number of models of the fates of the duplicates produced by WGD. **A)** Statistical relationships between the various models for the yeast WGD (blue), At- $\alpha$  (green) and  $\rho$  (brown) events. The simplest model (WGD-*n*) considers only a balanced process of gene loss. From this model, we can either allow duplicate genes to become fixed (for instance by neo- or sub-functionalization, WGD-*f*) or for one of the two parental subgenomes to lose more genes than the other (WGD-*b*). Using a likelihood ratio test (LRT), we find that, for all three WGD events, allowing duplicate fixation significantly improves the fit of the data to the models ( $P < 10^{-10}$ , LRT, [Methods](#)). However, for the yeast dataset, there is no significant evidence for biased fractionation ( $P > 0.5$ , LRT), while for the two plant WGDs, adding it significantly improves the fit ( $P < 10^{-10}$ ; LRT). From these two models, we can then allow the other process. Again, for yeast, there is significant evidence for fixation but not biased fractionation ( $P < 10^{-10}$  and  $P > 0.5$ , respectively, LRT) while for At- $\alpha$  and  $\rho$ , there is significant evidence for both ( $P < 10^{-10}$  in each case, LRT). We also tested a model where the biased fractionation parameter  $\epsilon$  (see panel **B**) was allowed to differ on the shared root branch of the tree (WGD-*bf*) compared to all of the other branches. For the two plant WGD events, there is no significant evidence that the level of biased fractionation differed early in history of the WGD relative to later in time ( $P \geq 0.19$ , [Results](#)). On the other hand, for the yeast WGD, biased fractionation was much more intense soon after the polyploidy event and weakened later ( $P = 0.001$ ; [Results](#)). **B)** Model states and parameters.

Our model has four states, two duplicated ones (U = undifferentiated duplicates and F = fixed duplicates) and two single copy states (S<sub>1</sub> and S<sub>2</sub>, corresponding to the two parental subgenomes). The base loss rate ( $\alpha$ ) is compounded with the estimated time to give the branch lengths of Fig 1. The relative fixation rate  $\gamma$  ( $0 \leq \gamma < \infty$ ) gives the rate of duplicate fixation relative to the loss rate  $\alpha$ . Likewise, the fractionation bias parameter  $\epsilon$  ( $0 \leq \epsilon \leq 1$ ) gives the excess of preservations from subgenome 1 relative to subgenome 2 (assumed to be the more fractionated subgenome).

<https://doi.org/10.1371/journal.pgen.1007267.g002>

POInT's focus on WGD has advantages over applying more general gene birth-death models to polyploid species [74, 75]. POInT models the process of duplicate loss and retention jointly across all genomes and along a phylogeny. Hence, the probability of a particular model state at a given ancestral locus is conditioned on all other loci and all other genomes. This conditioning is performed by analogy to the linkage analysis model of Lander and Green [76] using the hidden-Markov approach of Felsenstein and Churchill [77]. The states the Markov model considers are the set of  $2^n$  possible orthology relationships between the  $2n$  different loci (e.g., 2 duplicated loci in each of  $n$  genomes). The likelihood of site  $i+1$  having orthology state  $j$  given that site  $i$  has that orthology assignment is  $(1-\theta)$ , where  $\theta$  is a small constant estimated from data ( $0.0004 \leq \theta \leq 0.0081$  across these analyses). In cases where there is a double break in gene order in a particular genome,  $\theta = 0.5$ .

From this model structure, we can infer orthologous chromosomal regions produced by WGD between the genomes studied, along with confidence estimates in these inferences (Fig 1). The previous version of POInT did not distinguish between states S<sub>1</sub> and S<sub>2</sub>. The result was degeneracy in the inferences of orthologous regions. In other words, assigning the first member of each DCS pair to subgenome 1 and the second to subgenome 2 produced orthology assignment 111111 across the six genomes, which was identical in likelihood to assignment 222222. (The computation is completely analogous for the other two WGD events studied.) Effectively, this degeneracy corresponds to flipping the upper and lower panels of Fig 1, because each of the  $2^n$  possible orthology assignments has an equivalent assignment with all 1s converted to 2s and vice versa.

To model the process of BF, we relaxed this assumption by introducing parameter  $\epsilon$  (Fig 2B). This parameter makes losses to state S<sub>2</sub> potentially less common than to S<sub>1</sub>. If BF is present in the data, the maximum likelihood estimate of  $\epsilon$  will be less than 1.0, and the likelihood of orthology assignment 111111 will no longer be the same as 222222. We can then use the

**Table 1. POInT estimates for different datasets and ancestral orders.**

| Description   | Ancestral loci <sup>a</sup> | # breaks <sup>b</sup> | #double breaks <sup>c</sup> | WGD-bf lnL <sup>d</sup> | Fixation rate ( $\gamma$ ) <sup>e</sup> | Bias strength ( $\epsilon$ ) <sup>f</sup> |
|---|-----------------------------|-----------------------|-----------------------------|-------------------------|---|---|
| At- $\alpha$ , Full: <i>C. violacea</i> order         | 7243                        | 6614                  | 3021                        | -25357.46               | 0.160                                   | 0.538                                     |
| At- $\alpha$ , Full: Optimized order                  | 7243                        | 5468                  | 1129                        | -24497.04               | 0.169                                   | 0.645                                     |
| At- $\alpha$ , High-synteny: <i>C. violacea</i> order | 4556                        | 3544                  | 1039                        | -12837.67               | 0.205                                   | 0.718                                     |
| At- $\alpha$ , High-synteny: Optimized order          | 4556                        | 2266                  | 252                         | -12442.51               | 0.220                                   | 0.786                                     |
| Grass $\rho$ , Pineapple order                        | 3091                        | 4387                  | 2299                        | -8822.89                | 0.049                                   | 0.400                                     |
| Grass $\rho$ , Optimized order                        | 3091                        | 2457                  | 434                         | -8199.10                | 0.061                                   | 0.730                                     |
| Yeast WGD   | 4065                        | 4346                  | 796                         | -19374.10               | 0.137                                   | 0.955 <sup>g</sup>                        |

<sup>a</sup>: Number of ancestral loci studied.

<sup>b</sup>: Number of synteny breaks across the polyploid genomes.

<sup>c</sup>: Number of cases where both parental subgenomes showed a synteny break after an ancestral locus (see Methods).

<sup>d</sup>: ln-likelihood from fitting WGD-bf to this ancestral order.

<sup>e</sup>: Maximum likelihood estimate of the relative duplicate fixation rate for this ancestral order (see Fig 2).

<sup>f</sup>: Maximum likelihood estimate of the relative rate of retention from the more fractionated subgenome for this ancestral order (see Fig 2).

<sup>g</sup>:  $\epsilon$  not significantly different from 1.0; see Fig 2.

<https://doi.org/10.1371/journal.pgen.1007267.t001>

POInT model to estimate the posterior probability of the subgenome assignments (the numbers shown above every column in Fig 1) at every pillar. For convenience we refer to the resulting two regions as deriving from allopolyploid parents 1 and 2 [43], respectively, defining parent 1 as containing genes in state  $S_1$  (e.g., it is potentially less fractionated), similar to Thomas et al., [39].

In previous work in yeast [15, 22, 70], we found evidence for “convergent” gene losses that were phylogenetically independent and yet more often from the same subgenome than could be explained by chance. We modeled these events by adding two duplicated converging states to our model,  $C_1$  and  $C_2$ . Gene losses from  $C_1$  were always to  $S_1$  and similarly for  $C_2$ . We fit versions of this model both with ( $0 \leq \epsilon \leq 1.0$ ) and without ( $\epsilon = 1$ ) BF to our yeast, grass and *At- $\alpha$*  data: while these models improved the fit relative to the WGD-*bf* model used here, we present our results in terms of the WGD-*bf* model because both model classes give similar parameter estimates (S2 Table), and the more complex models do not add insight for the questions considered here.

### Dependence of POInT parameter estimates on the assumed phylogeny

Because we analyzed only four genomes sharing the grass  $\rho$  event, it was possible to use POInT to test all 15 possible rooted phylogenetic trees for these taxa to assess the dependence of our inferences on the inferred phylogeny. We present our results in terms of the optimal tree, but the global parameter estimates for the WGD-*bf* model were very similar for all topologies ( $0.061 \leq \gamma \leq 0.067$ ;  $0.719 \leq \epsilon \leq 0.739$ ;  $0.0061 \leq \theta \leq 0.068$ ; Fig 2).

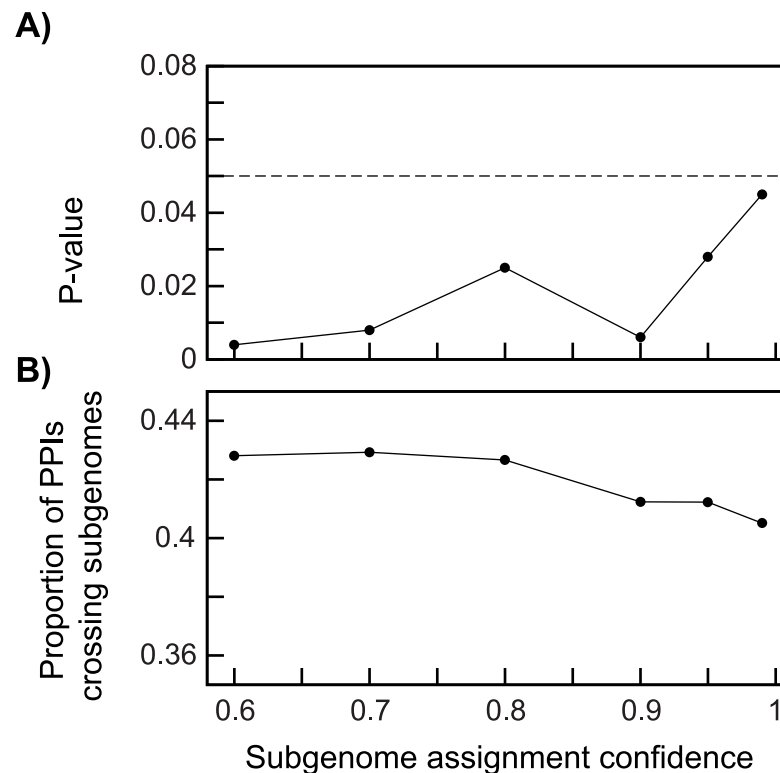
### Network analyses of biased losses

We asked whether genes surviving from one or the other of the subgenomes showed patterns of interconnection in the networks of *Arabidopsis thaliana*. We use the BioGrid database [78] to extract known protein-protein interactions [79]. We tested for paucity of interactions between the products of genes from different subgenomes with a randomization approach. We thus compared the number of interactions between gene products from alternative subgenomes in the actual data to this value computed after 1000 randomizations of the subgenome assignments. To assess the degree to which our conclusions were potentially affected by errors in the assignment of genes to subgenomes, we conducted our tests at a range of confidences in subgenome assignment (Fig 3).

### GO analyses of biased losses

We used the Gene List Analysis tool from the PANTHER classification system [80] to perform statistical overrepresentation tests to find over/under-represented Gene Ontology (GO) terms associated with biological processes, molecular functions, or cellular components. The input of our analysis consists of two sets of genes: the target list to analyze, and a reference list. The expected number of genes for a GO term in the target list was calculated based on the number of genes with that term in the reference list: binomial statistics for each GO term associated with genes in the target list were then computed from these expectations [81].

We first performed an overrepresentation test for 4,086 single copy genes from both subgenomes against the reference set of 4,152 surviving duplicated genes. The over/under represented GO terms in the analysis were filtered with a threshold  $P$ -value  $\leq 0.01$  after Bonferroni correction, and only terms with a fold-enrichment larger than 1.5 (overrepresented) or smaller than 0.67 (underrepresented) are reported. We next compared 2,552 single copy genes from subgenome 1 (dominant) relative to the terms for the 1,534 genes from subgenome 2 (more fractionated) with a similar approach. To compensate for the smaller number of terms found



**Fig 3. Protein interactions between single-copy genes from alternative subgenomes are rarer than expected.** We extracted single-copy genes for a range of values of POInT's overall confidence in pillar assignments to subgenomes (x-axis) and computed the *P*-value for the test of the null hypothesis of no fewer protein-protein interactions between products of genes from alternative subgenomes than expected (y-axis; panel A; see [Methods](#)). We also computed the frequency of such "crossing" interactions relative to interactions between products of the same subgenome (y-axis, panel B).

<https://doi.org/10.1371/journal.pgen.1007267.g003>

to be enriched in this second analysis, we used an FDR-corrected *P*-value of 0.05 as a threshold. Full lists of all significantly enriched terms for any comparison with associated GO identifiers are given as [S3–S5](#) Tables.

## Results

### Modeling WGD evolution with POInT

Using POInT, we analyzed the resolution of three phylogenetically widely-spaced polyploidy events: the WGD in the ancestor of *Saccharomyces cerevisiae* and relatives [20, 82], the  $\rho$  event found in the ancestor of the grasses [47, 48] and the At- $\alpha$  event shared by the model plant *Arabidopsis thaliana* and its relatives [26, 83]. Previous work has suggested that all of these WGDs were allopolyploid events [43, 82], meaning the duplicated regions in the extant polyploid genomes (hereafter subgenomes) derive from parental genomes from differing species. Whatever their origins, however, these subgenomes produced by polyploidy are now distinct due to their individual histories of gene loss. In order to assign the extant genes to one of the two subgenomes, we applied new duplicate resolution models that distinguished between a less fractionated genome (more surviving genes) and the more fractionated genome [fewer surviving genes; 39, 43].

As previously described [15, 22, 70], we used ohnologs from the Yeast Genome Order Browser project and an inferred ancestral genome order as POInT's inputs for the yeast

analyses [19, 49]. For the At- $\alpha$  and  $\rho$  events, no such data exist, so we developed a new pipeline that uses sequence homology and shared gene order (synteny) to assign genes from the polyploid genomes to a “pseudo-ancestral” gene from the extant outgroups *Cleome violacea* (for At- $\alpha$ ) and pineapple (for  $\rho$ ). First, we used simulated annealing to assign genes from each of the polyploid genomes to double-conserved synteny (DCS) blocks. These assignments were made forcing pairs of regions in the polyploid genomes to possess one or two homologous genes to one gene from a single region in outgroup genome: the simulated annealing algorithm then sought such assignments that maximized the shared gene order (see [Methods](#) for additional details). We then merged these single-genome inferences into a set of 7243 and 3091 (for At- $\alpha$  and  $\rho$ , respectively) ancestral gene pillars, each consisting of at least one gene from every genome that shared synteny with at least one other gene (see [Fig 1](#)). We then again used simulated annealing to optimize our estimate of ancestral genome order of these loci by maximizing the synteny among the pillars. [Fig 1](#) gives an example of the estimates made by POInT based on these inferred pillars: from the inferred pillar order, POInT is able to estimate the probability associated with assigning each genome segment from each species to either of the two subgenomes (numbers above the columns in that figure).

Using these data, we tested the hypothesis that biased fractionation (BF) was observed after the three WGD, explored its temporal characteristics and sought to associate it with functional properties of the genes in question.

**Biased fractionation was common after At- $\alpha$  and  $\rho$ .** By fitting nested models of evolution to these datasets, we tested for the presence of ohnolog fixation and biased fractionation after the three WGD events. Fixation (WGD-*f*, [Fig 2A](#)) is inferred when a WGD-produced duplicate pair has persisted across the tree longer than would be expected given the loss rates. There is evidence of such fixation events after all three WGDs ( $P < 10^{-10}$ , likelihood ratio test, [Fig 2](#) and [Methods](#)). We model biased fractionation (BF, WGD-*b*, [Fig 2A](#)) as a preference for losses of genes from subgenome 2 ( $0 \leq \epsilon \leq 1$ , [Fig 2B](#)) over subgenome 1. Note that the identity of subgenome 2 is inferred from the data and bespeaks no lack of generality in our model.

At- $\alpha$  and  $\rho$  show strong evidence of BF ( $P < 10^{-10}$ , likelihood ratio test, [Fig 2A](#) and [Methods](#)). However, similar to previous analyses of the yeast WGD [21], we find no statistical evidence for a *general* BF process after the yeast WGD ( $P > 0.5$ , likelihood ratio test, LRT, [Fig 2A](#)). Our estimate of the strength of BF after At- $\alpha$  is nearly identical to that found by Thomas and coauthors when considering only the *A. thaliana* genome [39], with the more fractionated subgenome showing approximately 2 single copy genes deriving from it for every 3 from the less fractionated subgenome. The bias estimated for the  $\rho$  event was slightly weaker: 3 genes from the more fractionated genome retained for every 4 from the other subgenome. We note that these estimates vary somewhat depending on the quality of the syntenic data used as the input for POInT: when we used the highly non-optimal *C. violacea* gene order (which has many more syntenic breaks), the estimated ratio of single copy genes from the more and less fractionated genomes was closer to 1:2 ([S1 Table](#)). However, it is unlikely that further order optimization would raise the estimates of the BF parameter  $\epsilon$  (e.g., imply less fractionation): all of the estimated ancestral orders gave similar estimates of  $\epsilon$ , with no trend of increasing  $\epsilon$  with smaller numbers of breaks ([S1 Table](#)). Likewise, we inferred a “highly syntenic” dataset of 4556 ancestral pillars for At- $\alpha$  that included only pillars with fully syntenic connections to at least one other pillar ([Methods](#)). While the estimate of  $\epsilon$  for this dataset is higher than that for the full dataset ([Table 1](#)), it is still significantly different from 1.0 ( $P < 10^{-10}$ ). Moreover, some of the increase in  $\epsilon$  here may be attributable to the greater number of surviving duplicates (larger  $\gamma$ , see [Table 1](#)).

**Biased fractionation occurred in a brief interval after the yeast WGD but has been a continuous process after At- $\alpha$  and  $\rho$ .** The process of duplicate loss immediately post-WGD

differs from that observed later [22, 23]. We hence fit a model where the strength of BF was allowed to differ on the shared root branch (Fig 1) relative to the remaining branches. For At- $\alpha$  and  $\rho$ , there is no significant evidence for such a difference ( $\epsilon_{\text{early}} = 0.67/0.74$ ,  $\epsilon_{\text{late}} = 0.63/0.73$ , for At- $\alpha$  and  $\rho$ , respectively;  $P \geq 0.19$ ). However, the strength of biased fractionation immediately after the yeast WGD was much higher than that seen later ( $\epsilon_{\text{early}} = 0.47$ ,  $\epsilon_{\text{late}} = 0.99$ ;  $P = 0.001$ ), showing that our initial conclusion of no BF in yeast was an artifact of low temporal resolution in the WGD-*bf* model. Approximately 277 single-copy genes from the less fractionated parent, and only 135 from the more fractionated one, were returned to single-copy along the shared root branch following the yeast WGD (S1 Fig). We note that it is difficult to directly compare the yeast and plant results because of the differing shape of the post-WGD phylogenies for the datasets. The yeast WGD was characterized by very rapid post-WGD speciation [15, 84]: thus only 412/4099 (10%) of the ohnolog pairs had lost a gene before the first speciation (S1 Fig). On the other hand, the taxa sharing At- $\alpha$  had undergone ohnolog losses at 4008/7243 (55%) of the ancestral positions before the speciation event that split *Aethionemae arabicum* from the other Brassicaceae (Fig 1), with a similar proportion of losses on the root branch after  $\rho$  (S1 Fig). The phylogenies reflect this difference, with POInT's estimate of the length of the root branch in the yeast analysis being 0.063 versus 0.55 for At- $\alpha$  and 0.63 for  $\rho$  (recall that branch lengths are proportional to the probability of an ohnolog loss along that branch). The tribe Aethionemae is sister to the remainder of all extant Brassicaceae species [85]. Hence, at least for At- $\alpha$ , there might have been short period of more intense biased fractionation that we cannot detect due to the lack of an extant early diverging lineage such as those we have studied in the yeasts.

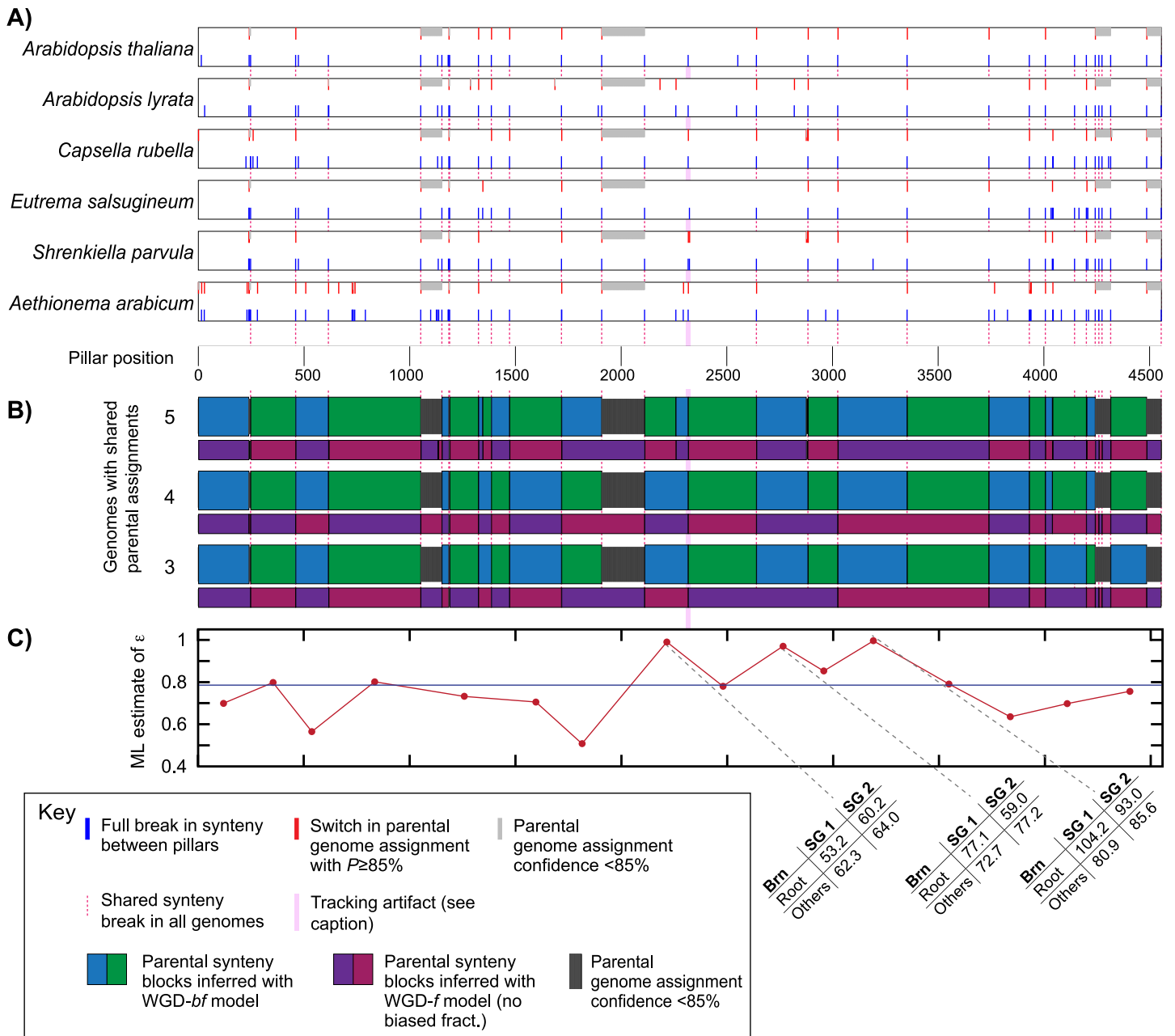
### Biased fraction is a genome-wide phenomenon

As mentioned, it is not guaranteed that two genomic regions each showing a higher retention rate than their homeologous partners necessarily originate from the same parental subgenome (the circularity problem in measuring BF). We used the high-synteny subset of the At- $\alpha$  data to assess the degree of this problem. From it, we produced a visual representation of the set of ancestral synteny blocks POInT was using for its inferences. In Fig 4B, we show how often 5, 4, or 3 genomes agree from pillar to pillar in their subgenome assignments. Notably, when only 3 of 6 genomes are required to agree at high probability, the model infers a relatively small number of ancestral syntenic blocks, consistent with a set of ancestral chromosomes prior to At- $\alpha$ . Moreover, these blocks are identifiable without the assumption of BF (e.g., they are also inferable from the WGD-*f* model, Fig 4B) and, at least for most of the larger blocks, give estimates of BF similar to the dataset as a whole (Fig 4C). Hence, it is clear that biased fractionation is not an artifact of synteny-block inference. Similar diagrams for the full At- $\alpha$  dataset, the  $\rho$  dataset and yeast are given in S2 Fig.

### Protein products of single copy genes from different subgenomes rarely physically interact

Using data from BioGrid [78, 79], we asked whether protein-protein interactions between the products of *A. thaliana* single-copy genes from alternate subgenomes were rarer than would be expected by chance. Across a large range of subgenome confidence estimates from POInT, there were fewer such “crossing” interactions than expected (Fig 3A), and the frequency of such interactions decreases as our confidence in the subgenome assignments increases (Fig 3B). Similar analyses were not performed for the  $\rho$  and yeast WGD events due to the lack of large-scale interaction data and the lack of substantial fractionation, respectively.





**Fig 4. Consistency across the ancestral genome of POINt's estimates of the subparental genome of origin.** **A)** In the six panels, we illustrate how often POINt's assignment of parental subgenome of origin for  $\alpha$ -changes between two successive pillars when considering the "high synteny" dataset. A red tick at position  $i$  corresponds to a situation where POINt assigned parents-of-origin to two chromosomal regions at position  $i-1$  with probability of  $\geq 85\%$  and either the *opposite* combination of parents at position  $i$  or with the same assignment but with confidence less than 85%. Gray ticks, in turn, correspond to those positions immediately after a red tick where the confidence in the parental assignments is less than 85%. The blue ticks in the lower half of each block indicate positions where there is a double synteny break after position  $i-1$  (see *Methods*). At these positions, the parental inferences at position  $i$  are independent of those at  $i-1$ . Locations where all 6 genomes have such breaks are shown with the pink dotted lines. **B)** Estimates of shared parental blocks across genomes. With very few exceptions, locations where POINt finds a change in subgenome assignments correspond to these six-fold synteny breaks from **A**. Each blue/green colored block corresponds to a situation where at least 5, 4, or 3 genomes (top, middle and bottom, respectively) agree between every neighbor as to the subgenome assignment at a confidence of 85% or more. Narrower black regions are regions where there is no position-to-position agreement in assignment for any number of genomes (e.g., these are regions where our confidence in subgenome assignments is low overall). Any shared loss of synteny can induce a new block: such synteny breaks might, for instance, reflect a shift to new ancestral chromosome. For reference, we also show the set of blocks inferred with the WGD- $f$  model as the smaller set of red/purple blocks. This model does not include BF, making it degenerate, so that subgenome 1 and 2 can be swapped. We therefore define one region of one genome as being subgenome #1 and make the block assignments correspondingly. Almost all of the phasing of blocks can be done without the assumption of BF, as is seen with the similarity between the blue/green and red/purple blocks. The implication of this fact is that the blocks are defined by the pattern of shared gene losses and that including BF in the model serves only to allow us to assign unlinked blocks to the same

subgenomes based on their BF patterns. C) For the 16 blocks with more than 100 pillars, we show the estimates of the strength of BF (maximum likelihood estimate of  $\epsilon$ ;  $y$ -axis) judged solely from that block (block mid-point on the  $x$ -axis). These values indicate strong BF in all but three cases: in most of the larger blocks the estimated strength of BF is nearly identical to that for the full dataset (blue line). For the three blocks with weak evidence for BF ( $\epsilon \approx 1.0$ ), we further interrogated the patterns of gene loss (tables at bottom). In two of three cases, the signal of BF is relatively strong along the shared root branch where most losses occurred, with conflicting patterns on other branches. We attribute these differences to sampling effects among the relatively small number of losses along each branch. For the final block, with coordinates from pillars 2113 to 2318, the inferred pattern of losses contradicts the subgenome assignment, with more inferred losses from subgenome 1. When we examined the pattern of synteny breaks in this region, we discovered an anomaly: all of the genomes except *Eutrema salsugineum* had a synteny break at the end of this block: *E. salsugineum* instead had a break six pillars later (the pink shaded region). Hence, this synteny pattern caused the block to be linked to the next, larger, block, giving rise to the incongruous gene loss inferences. Equivalent figures for the full *At- $\alpha$*  dataset, the yeasts and the grasses are given as [S2 Fig](#).

<https://doi.org/10.1371/journal.pgen.1007267.g004>

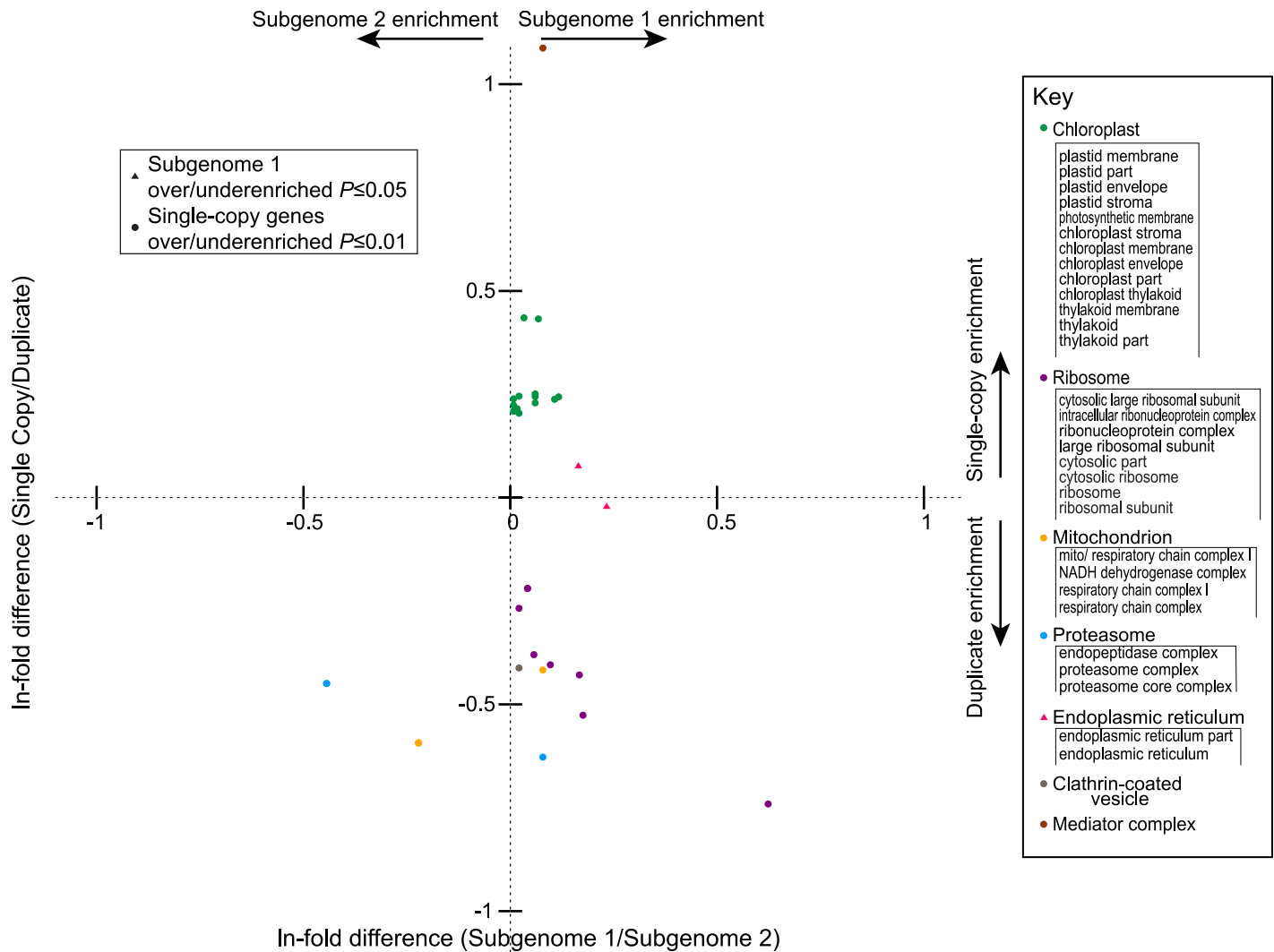
## BF has retained genes of distinct functions from each subgenome

As seen in previous analyses [23, 25, 26, 30], the surviving *At- $\alpha$*  ohnologs are enriched or depleted for a number of GO ontology categories (Fig 5 and S3 Fig). We had anticipated that those categories that were depleted for ohnolog pairs might represent a set of single-copy genes drawn preferentially from the dominant subgenome. However, such was not the case: even at a quite liberal FDR-corrected significant threshold ( $P \leq 0.05$ ), there are relatively few GO terms significantly differentially retained between the single copy genes of the two subgenomes. Moreover, these terms do not overlap with the ohnolog-depleted terms: instead the single copy genes operating in the endoplasmic reticulum more often derive from the less-fractionated subgenome (Fig 5). Similarly, genes involved in the cell cycle and circadian rhythm are preferentially drawn from the more fractionated subgenome and those for developmental genes in phloem or xylem from the less-fractionated subgenome (S3 Fig).

## Discussion

There is considerable and accumulating evidence for the actions of biased fractionation (BF) after WGD in angiosperms [39–42] and strong suggestions that allopolyploidy is more likely to produce such biases than autopolyploidy [43]. Nonetheless, there remains at least a theoretical danger that analyses of BF that consider only a single polyploid genome at a time [often by comparison to a diploid outgroup; 40, 41, 46, 86, 87] could mistake the random variation in preservation in small synteny blocks for biases in fractionation.

The results presented here refute this concern, and indicate that, at a minimum, BF acts consistently across regions at the chromosome scale. Our confidence in this conclusion is driven by the concordance of multiple lines of evidence as to the presence and strength of BF. At a methodological level, POInT integrates across multiple genomes, such that lineage-specific synteny breaks are passed through using data from genomes without such breaks (subject to limitations in genome assemblies and in the degree of shared history in the genomes). This approach dramatically increases synteny block size (see Fig 4). Moreover, POInT employs a very strict and transparent definition of synteny: only genomic neighbors are considered to be in synteny, meaning that POInT employs no parameters such as a window size that need to be tuned by the user and that could confound inferences. POInT also employs a robust modeling framework similar to those used in sequence evolution studies [88] and allows for explicit statistical tests for the presence of BF. Using this framework, we have shown very strong statistical support for BF after two independent WGD events: *At- $\alpha$*  and the grass  $\rho$  event, with a ratio of single copy genes from the less and more fractionated subgenomes somewhere between 3:2 and 5:4, in line with previous estimates [39]. This modeling approach has the further advantage of avoiding the circularity in block estimation: POInT infers parental genome assignments on the basis of shared gene losses, a point we have exploited previously [22, 89–91]. As a result, POInT effectively recovers the same shared parental genome assignments under a model without biased fractionation (red/purple blocks in Fig 4) as it does under the BF model.



**Fig 5. Statistically overrepresented GO terms from the cellular component hierarchy associated with *At-α* duplication status and parental subgenome of origin (see Methods).** On the y axis is the  $\ln(\text{fold-enrichment})$  of each GO terms among the single copy genes relative to the surviving duplicates from *At-α*. Dots represent cellular component terms that are significantly over (positive values) or underrepresented (negative values) among single copy genes relative to duplicates (Bonferroni corrected  $P$ -value  $\leq 0.01$  and a fold-enrichment of  $> \pm 1.5$ ). On the x axis is the  $\ln(\text{fold-enrichment})$  of GO terms of genes from subgenome 1 (the less fractionated genome) relative to those from subgenome 2 (the more fractionated one). GO terms that are overrepresented in genes from subgenome 1 with a  $P$ -value  $\leq 0.05$  after Bonferroni correction are shown as triangles. Points are colored based on the compartment in question, as indicated in the key at right. The patterns seen for the “Molecular Function” and “Biological Process” categories of terms are presented in S3 Fig.

<https://doi.org/10.1371/journal.pgen.1007267.g005>

Moreover, the simultaneous consideration of multiple genomes allows us to assess if the evidence for BF is consistent across those genomes: our loss estimates for each branch of the phylogeny all show BF of roughly similar magnitude, despite the fact that losses on the different tip branches of the phylogeny in Fig 1 are necessarily independent (an estimate of the BF ratio is given under each branch of the lower tree in that Figure). Finally, the absence of evidence for BF on most branches of the post-WGD yeast phylogeny [which was recently conclusively found to be an allopolyploidy; 82] illustrates that POInT is fully capable of rejecting the hypothesis of BF when evidence for it is weak (or temporally variable in this case).

One might argue instead that BF favored some chromosomes from one parental genome and some from another. However, this position is inconsistent with the results of our interaction

data and GO term analyses, since such interactions more often occur between products of genes from the same subgenome than between products of genes encoded on different subgenomes, and genes assigned to the same subgenome show consistency in low-level GO term associations. Likewise, there is a good accordance between the estimates of the strength of BF in three of the four largest synteny blocks of Fig 4 and the overall estimate: were BF a chromosome-by-chromosome phenomenon, it is difficult to understand why its strength would be so consistent across blocks.

While POInT represents a significant improvement over analyses of single polyploid genomes, there are always limitations to any modeling framework. From a practical point of view, our inferences are limited by the quality of the genomic data used as inputs: the more fragmented these genome assemblies, the less power POInT has to infer parental genomes of origin. The inference of DCS blocks by simulated annealing is a costly and computationally difficult problem, and while our scoring functions are reasonable, they may not be the optimal method for inferring ancestral genome orders [49]. As mentioned in the *Methods* section, there is also a potential for older polyploidies that are shared by the outgroup genome to mislead our scaffolding, although we do not believe this problem was significant here. Finally, POInT itself is imperfect in how it treats uncertainty in parental genome assignments: the error parameter  $\theta$  estimates the degree to which the input data fails to conform to POInT's underlying model. While our results above appear to be robust to these various sources of error, future studies of polyploid genomes with improved approaches could give more refined estimates of parental genomes of origin and fine-scale temporal patterns of post-polyploidy gene losses.

Having reaffirmed that BF is a robustly detectable phenomenon in the evolution of polyploid genomes, it is reasonable to try to better understand its origins. In this vein, several of our observations, which arise from POInT's unique capacity to probe polyploidy phylogenetically, serve to again suggest a link between BF and the hypothesized effects of allopolyploidy. The association of genes that physically interact with the same parental genome is one example of such an observation. Another is the conclusion that, after the At- $\alpha$  and  $\rho$  events, the strength of BF was uniform in time, but in yeast, BF was associated only with the very earliest stages of WGD resolution. We have previously found that a very particular group of genes, involved in DNA repair and mitochondrial function, were returned to single copy immediately after the yeast WGD [22]. Given the biases in those losses found here, it appears likely that BF in yeast was a result of selection for the removal of some ohnolog copies in order to prevent the mixing of genes for these two functions from the two diploid progenitor species. It is likely that the DNA repair enzymes and nuclear-encoded proteins targeted to the mitochondria have co-evolved separately in each parental genome (and that only one of the two parents contributed a mitochondrial genome to the hybridization). If true, these hypotheses would suggest that BF in yeast resulted from selection to maintain co-adapted genes after hybridization. Because these losses, in addition to being biased towards one subgenome and a limited set of functions, occurred very rapidly after the WGD event [15], it is difficult to believe they occurred through purely neutral processes: the proposal by De Smet et al., [23] that forces such as dominant negative interactions may have driven selection to favor certain losses seems increasingly plausible. These results also reinforce a point we have made several times before: one's understanding of the forces acting on a polyploid genome may depend on *when* in its history you look [22, 34, 92].

Our analyses are compatible with differences in gene expression driving BF [41, 93]. However, the BF process does not appear to be solely a product of expression: the presence of co-evolved modules in the two parental genomes also apparently plays a role. Not only do we see a strong bias in the retention of DNA repair enzymes and mitochondrially-targeted proteins

in yeast, but we also see a relative absence of protein-protein interactions between proteins encoded by different subgenomes in *A. thaliana*. This hypothesis would also explain our previous observation that both ribosomal proteins and histones underwent post-WGD gene conversions in yeasts [89, 90], as gene conversion represents a second mechanism for resolving parent-of-origin conflicts induced by polyploidy.

Returning to our point about the timing of post-WGD events, we propose that the process of BF and the selection that retains some ohnologs to preserve dosage balance are linked. In this view, some genetic modules [a vague but still useful concept; 94] do not tolerate being duplicated and are quickly returned to single-copy [23]. Others remain duplicated as predicted by the DBH [3, 30]. However, these duplications are not necessarily stable over long timescales [22, 34]: any incompatibilities between the subgenomes will favor one subgenome when duplicates are in the end lost. The origins of these conflicts most likely arise through co-evolution between genes in individual genomes [95]. From our GO analyses, it appears that the effects of this co-evolution decay quickly as one moves away from directly interacting genes: hence many biological processes have “mixed and matched” set of genes from the two subgenomes.

The three WGD events considered here cannot completely resolve these questions: the yeast WGD mostly lacks prolonged BF, while the early events after At- $\alpha$  and  $\rho$  are difficult to identify because of the long shared post-WGD branch. In the future, we will perform similar analyses with the recent *Brassica* hexaploidy to further refine our understanding of post-WGD functional evolution. So doing will not only improve our understanding of polyploidy but also of the nature of the functional links and the degree of co-evolution inherent in the interacting macromolecules that make up the cell.

## Supporting information

**S1 Fig. Gene loss patterns after the yeast (A) and grass  $\rho$  (B) WGD events.** Shown is the assumed 11 species topology for yeast and the maximum likelihood topology for  $\rho$ , with branch lengths estimated as in Fig 1. Above each branch is the estimated number of genes returned to single copy with the gene from parental subgenome 1 being retained, while below each branch is the corresponding number for subgenome 2. These data were inferred from model WGD-*bf* (e.g., a model with fixation and biased fractionation where the biased fractionation rate differs on the root branch, red:  $\epsilon_{\text{early}}$ , compared to the remainder of the tree, blue:  $\epsilon_{\text{late}}$ ) for yeast and WGD-*bf* for the grass  $\rho$  event (see Figs 1 & 2).  
(PDF)

**S2 Fig. Consistency of POInT's estimates of the parental genome of origin for each species individually (lower part) and for the combination of species (upper) for the full At- $\alpha$  dataset (panel A), the grass  $\rho$  event (panel B) and for the yeast WGD (panel C).** All details are otherwise as for Fig 4 in the main text. Blocks for the both the WGD-*bf* (blue/green) and the WGD-*f* (pink/purple) models are shown for the At- $\alpha$  and  $\rho$  events. Because the global WGD-*bf* model showed little evidence for BF in yeast, we illustrate the inferred blocks from the WGD-*bf* model for these taxa: note the lack of subgenome resolution due to the balanced gene losses seen on most of the branches of S1 Fig.  
(PDF)

**S3 Fig. Statistically overrepresented GO terms associated with At- $\alpha$  duplication status and parental genome of origin (see Methods).** A) Molecular function hierarchy. B) Biological processes hierarchy. On the y axis is the  $\ln(\text{fold-enrichment})$  of the term in question among the single copy genes relative to the duplicates. Blue dots represent terms that are significantly

over/underrepresented in single copy genes relative to duplicates with Bonferroni corrected P-value  $\leq 0.01$  and are  $>1.5$  fold over/underrepresented. On the  $x$ -axis is the  $\ln(\text{fold-enrichment})$  of GO terms of genes from subgenome 1 relative to those from subgenome 2. The GO terms that are over/underrepresented in genes from subgenome 1 (the less fractionated subgenome) with P-value  $\leq 0.05$  after Bonferroni correction are shown in triangles. GO terms that are significantly different both between the single copy and duplicate genes and between subgenome 1 and subgenome 2 are shown in dark purple.  
(PDF)

**S1 Table. Effect of inferred ancestral order on POInT estimates.**

(DOCX)

**S2 Table. Modeling convergent losses with POInT.**

(DOCX)

**S3 Table. Overrepresented cellular component GO terms.**

(DOCX)

**S4 Table. Overrepresented molecular function GO terms.**

(DOCX)

**S5 Table. Overrepresented biological process GO terms.**

(DOCX)

**S1 Data. CoGe accession numbers for all plant genomes analyzed (MS Excel).**

(XLSX)

**S2 Data. Underlying data for the plots in Figs 2, 3C and 5 (MS Excel).**

(XLSX)

**S3 Data. Gzipped tar file with files containing the conditional probability estimates for the timing of all gene losses for the At- $\alpha$ ,  $\rho$  and yeast WGD events as well as associated newick treefiles for these three events and a README file describing the data formats.**

(GZ)

## Acknowledgments

We would like to thank J. Washburn for assistance with the grass phylogeny, J. Birchler, J. Thorne, X. Ji and members of our fall 2016 Molecular and Network Evolution course for helpful discussions and M. Becchi, D. Li and H. Wu for computational assistance.

## Author Contributions

**Conceptualization:** Eric Lyons, J. Chris Pires, Patrick P. Edger, Gavin C. Conant.

**Data curation:** Marianne Emery, M. Madeline S. Willis, Kerrie Barry, Khouanchy Oakgrove, Yi Peng, Jeremy Schmutz, Eric Lyons.

**Funding acquisition:** Eric Lyons, J. Chris Pires, Gavin C. Conant.

**Investigation:** Marianne Emery, M. Madeline S. Willis, Yue Hao, Kerrie Barry, Khouanchy Oakgrove, Yi Peng, Jeremy Schmutz, Eric Lyons, J. Chris Pires, Patrick P. Edger, Gavin C. Conant.

**Methodology:** Marianne Emery, M. Madeline S. Willis, Yue Hao, Eric Lyons, Patrick P. Edger, Gavin C. Conant.



**Project administration:** J. Chris Pires.

**Resources:** Kerrie Barry, Khouanchy Oakgrove, Yi Peng, Jeremy Schmutz, Patrick P. Edger.

**Software:** Eric Lyons, Gavin C. Conant.

**Supervision:** J. Chris Pires, Gavin C. Conant.

**Validation:** Marianne Emery.

**Writing – original draft:** Yue Hao, J. Chris Pires, Gavin C. Conant.

**Writing – review & editing:** Marianne Emery, M. Madeline S. Willis, Yue Hao, Eric Lyons, J. Chris Pires, Patrick P. Edger, Gavin C. Conant.

## References

1. Van de Peer Y, Maere S, Meyer A. The evolutionary significance of ancient genome duplications. *Nat Rev Genet.* 2009; 10(10):725–32. <https://doi.org/10.1038/nrg2600> PMID: 19652647
2. Taylor JS, Raes J. Duplication and divergence: The evolution of new genes and old ideas. *Annual Review of Genetics.* 2004; 38:615–43. <https://doi.org/10.1146/annurev.genet.38.072902.092831> PMID: 15568988
3. Birchler JA, Veitia RA. Gene balance hypothesis: connecting issues of dosage sensitivity across biological disciplines. *Proc Natl Acad Sci U S A.* 2012; 109(37):14746–53. Epub 2012/08/22. <https://doi.org/10.1073/pnas.1207726109> PMID: 22908297.
4. Ohno S. *Evolution by gene duplication.* New York: Springer; 1970. 160pp. p.
5. Clausen R, Goodspeed T. Interspecific hybridization in *Nicotiana*. II. A tetraploid *glutinosa-tabacum* hybrid, an experimental verification of Winge's hypothesis. *Genetics.* 1925; 10(3):278. PMID: 17246274
6. Conant GC, Wolfe KH. Increased glycolytic flux as an outcome of whole-genome duplication in yeast. *Molecular Systems Biology.* 2007; 3:129. <https://doi.org/10.1038/msb4100170> PMID: 17667951
7. Merico A, Sulo P, Piškur J, Compagno C. Fermentative lifestyle in yeasts belonging to the *Saccharomyces* complex. *FEBS Journal.* 2007; 274:976–89. <https://doi.org/10.1111/j.1742-4658.2007.05645.x> PMID: 17239085
8. van Hoek MJ, Hogeweg P. Metabolic adaptation after whole genome duplication. *Mol Biol Evol.* 2009; 26(11):2441–53. Epub 2009/07/25. <https://doi.org/10.1093/molbev/msp160> PMID: 19625390.
9. Edger PP, Heide-Fischer HM, Bekaert M, Rota J, Glöckner G, Platts AE, et al. The butterfly plant arms-race escalated by gene and genome duplications. *Proceedings of the National Academy of Sciences.* 2015; 112(27):8362–6.
10. Mayrose I, Zhan SH, Rothfels CJ, Magnuson-Ford K, Barker MS, Rieseberg LH, et al. Recently formed polyploid plants diversify at lower rates. *Science.* 2011; 333(6047):1257. Epub 2011/08/20. <https://doi.org/10.1126/science.1207205> PMID: 21852456.
11. Schranz ME, Mohammadin S, Edger PP. Ancient whole genome duplications, novelty and diversification: the WGD Radiation Lag-Time Model. *Curr Opin Plant Biol.* 2012; 15(2):147–53. Epub 2012/04/07. <https://doi.org/10.1016/j.pbi.2012.03.011> PMID: 22480429.
12. Zhan SH, Drori M, Goldberg EE, Otto SP, Mayrose I. Phylogenetic evidence for cladogenetic polyploidization in land plants. *Am J Bot.* 2016; 103(7):1252–8. <https://doi.org/10.3732/ajb.1600108> PMID: 27466054
13. Kellogg EA. Has the connection between polyploidy and diversification actually been tested? *Curr Opin Plant Biol.* 2016; 30:25–32. <https://doi.org/10.1016/j.pbi.2016.01.002> PMID: 26855304
14. Wolfe KH. Robustness: It's not where you think it is. *Nat Genet.* 2000; 25:3–4. <https://doi.org/10.1038/75560> PMID: 10802639
15. Scannell DR, Frank AC, Conant GC, Byrne KP, Woolfit M, Wolfe KH. 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.* 2007; 104:8397–402. <https://doi.org/10.1073/pnas.0608218104> PMID: 17494770
16. Gaeta RT, Pires JC, Iniguez-Luy F, Leon E, Osborn TC. Genomic changes in resynthesized *Brassica napus* and their effect on gene expression and phenotype. *Plant Cell.* 2007; 19(11):3403–17. <https://doi.org/10.1105/tpc.107.054346> PMID: 18024568
17. Soltis DE, Soltis PS, PIRES JC, Kovarik A, Tate JA, Mavrodiev E. Recent and recurrent polyploidy in *Tragopogon* (Asteraceae): cytogenetic, genomic and genetic comparisons. *Biological Journal of the Linnean Society.* 2004; 82(4):485–501.

18. Edger PP, Smith RD, McKain MR, Cooley AM, Vallejo-Marin M, Yuan Y, et al. Subgenome dominance in an interspecific hybrid, synthetic allopolyploid, and a 140 year old naturally established neo-allopolyploid monkeyflower. *bioRxiv*. 2016:094797.
19. Byrne KP, Wolfe KH. The Yeast Gene Order Browser: Combining curated homology and syntenic context reveals gene fate in polyploid species. *Genome Res*. 2005; 15(10):1456–61. <https://doi.org/10.1101/gr.3672305> PMID: 16169922
20. Wolfe KH, Shields DC. Molecular evidence for an ancient duplication of the entire yeast genome. *Nature*. 1997; 387(#6634):708–13.
21. Kellis M, Birren BW, Lander ES. Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*. *Nature*. 2004; 428:617–24. <https://doi.org/10.1038/nature02424> PMID: 15004568
22. Conant GC. Comparative genomics as a time machine: How relative gene dosage and metabolic requirements shaped the time-dependent resolution of yeast polyploidy. *Mol Biol Evol*. 2014; 31(12):3184–93. <https://doi.org/10.1093/molbev/msu250> PMID: 25158798.
23. De Smet R, Adams KL, Vandepoele K, Van Montagu MC, Maere S, Van de Peer Y. Convergent gene loss following gene and genome duplications creates single-copy families in flowering plants. *Proc Natl Acad Sci U S A*. 2013; 110(8):2898–903. Epub 2013/02/06. <https://doi.org/10.1073/pnas.1300127110> PMID: 23382190.
24. Seoighe C, Wolfe KH. Extent of genomic rearrangement after genome duplication in yeast. *Proc Natl Acad Sci U S A*. 1998; 95(#8):4447–52.
25. Blanc G, Wolfe KH. Functional divergence of duplicated genes formed by polyploidy during Arabidopsis evolution. *Plant Cell*. 2004; 16(7):1679–91. <https://doi.org/10.1105/tpc.021410> PMID: 15208398.
26. Maere S, De Bodt S, Raes J, Casneuf T, Van Montagu M, Kuiper M, et al. Modeling gene and genome duplications in eukaryotes. *Proc Natl Acad Sci U S A*. 2005; 102(15):5454–9. <https://doi.org/10.1073/pnas.0501102102> PMID: 15800040.
27. Aury JM, Jaillon O, Duret L, Noel B, Jubin C, Porcel BM, et al. Global trends of whole-genome duplications revealed by the ciliate *Paramecium tetraurelia*. *Nature*. 2006; 444(7116):171–8. <https://doi.org/10.1038/nature05230> PMID: 17086204.
28. Makino T, McLysaght A. Ohnologs in the human genome are dosage balanced and frequently associated with disease. *Proc Natl Acad Sci U S A*. 2010; 107(20):9270–4. Epub 2010/05/05. <https://doi.org/10.1073/pnas.0914697107> PMID: 20439718.
29. Birchler JA, Riddle NC, Auger DL, Veitia RA. Dosage balance in gene regulation: biological implications. *Trends Genet*. 2005; 21(4):219–26. Epub 2005/03/31. <https://doi.org/10.1016/j.tig.2005.02.010> PMID: 15797617.
30. Freeling M. Bias in plant gene content following different sorts of duplication: tandem, whole-genome, segmental, or by transposition. *Annual Review of Plant Biology*. 2009; 60:433–53. Epub 2009/07/07. <https://doi.org/10.1146/annurev.arplant.043008.092122> PMID: 19575588.
31. Freeling M, Thomas BC. Gene-balanced duplications, like tetraploidy, provide predictable drive to increase morphological complexity. *Genome Res*. 2006; 16:805–14. <https://doi.org/10.1101/gr.3681406> PMID: 16818725
32. Birchler JA, Veitia RA. The gene balance hypothesis: from classical genetics to modern genomics. *Plant Cell*. 2007; 19(2):395–402. Epub 2007/02/13. <https://doi.org/10.1105/tpc.106.049338> PMID: 17293565.
33. Edger PP, Pires JC. Gene and genome duplications: the impact of dosage-sensitivity on the fate of nuclear genes. *Chromosome research: an international journal on the molecular, supramolecular and evolutionary aspects of chromosome biology*. 2009; 17(5):699–717. <https://doi.org/10.1007/s10577-009-9055-9> PMID: 19802709.
34. Conant GC, Birchler JA, Pires JC. Dosage, duplication, and diploidization: clarifying the interplay of multiple models for duplicate gene evolution over time. *Curr Opin Plant Biol*. 2014; 19:91–8. <https://doi.org/10.1016/j.pbi.2014.05.008> PMID: 24907529.
35. Dopman EB, Hartl DL. A portrait of copy-number polymorphism in *Drosophila melanogaster*. *Proc Natl Acad Sci U S A*. 2007; 104(50):19920–5. Epub 2007/12/07. <https://doi.org/10.1073/pnas.0709888104> PMID: 18056801.
36. Wapinski I, Pfeffer A, Friedman N, Regev A. Natural history and evolutionary principles of gene duplication in fungi. *Nature*. 2007; 449:54–61. <https://doi.org/10.1038/nature06107> PMID: 17805289
37. Papp B, Pal C, Hurst LD. Dosage sensitivity and the evolution of gene families in yeast. *Nature*. 2003; 424(6945):194–7. <https://doi.org/10.1038/nature01771> PMID: 12853957.
38. Bekaert M, Edger PP, Pires JC, Conant GC. Two-phase resolution of polyploidy in the *Arabidopsis* metabolic network gives rise to relative followed by absolute dosage constraints. *Plant Cell*. 2011; 23:1719–28. <https://doi.org/10.1105/tpc.110.081281> PMID: 21540436

39. Thomas BC, Pedersen B, Freeling M. Following tetraploidy in an Arabidopsis ancestor, genes were removed preferentially from one homeolog leaving clusters enriched in dose-sensitive genes. *Genome Res.* 2006; 16(7):934–46. Epub 2006/06/09. <https://doi.org/10.1101/gr.4708406> PMID: 16760422.
40. Tang H, Woodhouse MR, Cheng F, Schnable JC, Pedersen BS, Conant G, et al. Altered patterns of fractionation and exon deletions in *Brassica rapa* support a two-step model of paleohexaploidy. *Genetics.* 2012; 190(4):1563–74. Epub 2012/02/07. <https://doi.org/10.1534/genetics.111.137349> PMID: 22308264.
41. Schnable JC, Springer NM, Freeling M. Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss. *Proceedings of the National Academy of Sciences.* 2011; 108(10):4069–74.
42. Sankoff D, Zheng C, Zhu Q. The collapse of gene complement following whole genome duplication. *BMC genomics.* 2010; 11(1):313.
43. Garsmeur O, Schnable JC, Almeida A, Jourda C, D'Hont A, Freeling M. Two Evolutionarily Distinct Classes of Paleopolyploidy. *Mol Biol Evol.* 2013; 31(2):448–54. Epub 2013/12/04. <https://doi.org/10.1093/molbev/mst230> PMID: 24296661.
44. Buggs RJ, Chamala S, Wu W, Gao L, May GD, Schnable PS, et al. Characterization of duplicate gene evolution in the recent natural allopolyploid *Tragopogon miscellus* by next-generation sequencing and Sequenom iPLEX MassARRAY genotyping. *Molecular ecology.* 2010; 19(s1):132–46.
45. Wang J, Tian L, Lee H-S, Wei NE, Jiang H, Watson B, et al. Genomewide nonadditive gene regulation in Arabidopsis allotetraploids. *Genetics.* 2006; 172(1):507–17. <https://doi.org/10.1534/genetics.105.047894> PMID: 16172500
46. Cheng F, Wu J, Fang L, Sun S, Liu B, Lin K, et al. Biased gene fractionation and dominant gene expression among the subgenomes of *Brassica rapa*. *PLoS one.* 2012; 7(5):e36442. <https://doi.org/10.1371/journal.pone.0036442> PMID: 22567157
47. Paterson AH, Bowers JE, Chapman BA. Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics. *Proc Natl Acad Sci U S A.* 2004; 101(26):9903–8. <https://doi.org/10.1073/pnas.0307901101> PMID: 15161969.
48. Van de Peer Y, Mizrachi E, Marchal K. The evolutionary significance of polyploidy. *Nat Rev Genet.* 2017; 18(7):411–24. <https://doi.org/10.1038/nrg.2017.26> PMID: 28502977
49. Gordon JL, Byrne KP, Wolfe KH. Additions, losses and rearrangements on the evolutionary route from a reconstructed ancestor to the modern *Saccharomyces cerevisiae* genome. *PLoS Genetics.* 2009; 5(5):e1000485. <https://doi.org/10.1371/journal.pgen.1000485> PMID: 19436716
50. Joyce BL, Haug-Baltzell A, Davey S, Bomhoff M, Schnable JC, Lyons E. FractBias: a graphical tool for assessing fractionation bias following polyploidy. *Bioinformatics.* 2017; 33(4):552–4. <https://doi.org/10.1093/bioinformatics/btw666> PMID: 27794557.
51. The Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature.* 2000; 408(6814):796–815. Epub 2000/12/29. <https://doi.org/10.1038/35048692> PMID: 11130711.
52. Hu TT, Pattyn P, Bakker EG, Cao J, Cheng JF, Clark RM, et al. The Arabidopsis lyrata genome sequence and the basis of rapid genome size change. *Nat Genet.* 2011; 43(5):476–81. <https://doi.org/10.1038/ng.807> PMID: 21478890.
53. Slotte T, Hazzouri KM, Agren JA, Koenig D, Maumus F, Guo YL, et al. The *Capsella rubella* genome and the genomic consequences of rapid mating system evolution. *Nat Genet.* 2013; 45(7):831–5. <https://doi.org/10.1038/ng.2669> PMID: 23749190.
54. Dassanayake M, Oh DH, Haas JS, Hernandez A, Hong H, Ali S, et al. The genome of the extremophile crucifer *Thellungiella parvula*. *Nat Genet.* 2011; 43(9):913–8. <https://doi.org/10.1038/ng.889> PMID: 21822265.
55. Koch MA, German D. Taxonomy and systematics are key to biological information: Arabidopsis, *Eutrema* (*Thellungiella*), *Noccaea* and *Schrenkiella* (*Brassicaceae*) as examples. *Frontiers in plant science.* 2013; 4:267. <https://doi.org/10.3389/fpls.2013.00267> PMID: 23914192
56. Yang R, Jarvis DE, Chen H, Beilstein MA, Grimwood J, Jenkins J, et al. The Reference Genome of the Halophytic Plant *Eutrema salsugineum*. *Front Plant Sci.* 2013; 4:46. <https://doi.org/10.3389/fpls.2013.00046> PMID: 23518688.
57. Haudry A, Platts AE, Vello E, Hoen DR, Leclercq M, Williamson RJ, et al. An atlas of over 90,000 conserved noncoding sequences provides insight into crucifer regulatory regions. *Nat Genet.* 2013; 45(8):891–8. <https://doi.org/10.1038/ng.2684> PMID: 23817568
58. International Brachypodium Initiative. Genome sequencing and analysis of the model grass *Brachypodium distachyon*. *Nature.* 2010; 463(7282):763. <https://doi.org/10.1038/nature08747> PMID: 20148030

59. VanBuren R, Bryant D, Edger PP, Tang H, Burgess D, Challabathula D, et al. Single-molecule sequencing of the desiccation-tolerant grass *Oropetium thomaeum*. *Nature*. 2015; 527(7579):508. <https://doi.org/10.1038/nature15714> PMID: 26560029
60. Zhang G, Liu X, Quan Z, Cheng S, Xu X, Pan S, et al. Genome sequence of foxtail millet (*Setaria italica*) provides insights into grass evolution and biofuel potential. *Nature biotechnology*. 2012; 30(6):549. <https://doi.org/10.1038/nbt.2195> PMID: 22580950
61. Paterson AH, Bowers JE, Bruggmann R, Dubchak I, Grimwood J, Gundlach H, et al. The *Sorghum bicolor* genome and the diversification of grasses. *Nature*. 2009; 457(7229):551. <https://doi.org/10.1038/nature07723> PMID: 19189423
62. Ming R, VanBuren R, Wai CM, Tang H, Schatz MC, Bowers JE, et al. The pineapple genome and the evolution of CAM photosynthesis. *Nat Genet*. 2015; 47(12):1435. <https://doi.org/10.1038/ng.3435> PMID: 26523774
63. Lyons E, Freeling M. How to usefully compare homologous plant genes and chromosomes as DNA sequences. *The Plant Journal*. 2008; 53(4):661–73. <https://doi.org/10.1111/j.1365-3113X.2007.03326.x> PMID: 18269575
64. Taxis TM, Wolff S, Gregg SJ, Minton NO, Zhang C, Dai J, et al. The players may change but the game remains: Network analyses of ruminal microbiomes suggest taxonomic differences mask functional similarity. *Nucleic Acids Res*. 2015; 43(20):9600–12. <https://doi.org/10.1093/nar/gkv973> PMID: 26420832
65. Doring A, Weese D, Rausch T, Reinert K. SeqAn an efficient, generic C++ library for sequence analysis. *BMC Bioinformatics*. 2008; 9:11. Epub 2008/01/11. <https://doi.org/10.1186/1471-2105-9-11> PMID: 18184432.
66. Conant GC, Wagner A. GenomeHistory: A software tool and its application to fully sequenced genomes. *Nucleic Acids Res*. 2002; 30(15):3378–86. PMID: 12140322
67. Altschul SF, Madden TL, Schaffer AA, Zhang JH, Zhang Z, Miller W, et al. Gapped Blast and Psi-Blast: A new-generation of protein database search programs. *Nucleic Acids Res*. 1997; 25(#17):3389–402.
68. Kirkpatrick S, Gelatt CDJ, Vecchi MP. Optimization by simulated annealing. *Science*. 1983; 220(4598):671–80. <https://doi.org/10.1126/science.220.4598.671> PMID: 17813860
69. Conant GC, Wolfe KH. Functional partitioning of yeast co-expression networks after genome duplication. *PLoS Biol*. 2006; 4:e109. <https://doi.org/10.1371/journal.pbio.0040109> PMID: 16555924
70. Conant GC, Wolfe KH. Probabilistic cross-species inference of orthologous genomic regions created by whole-genome duplication in yeast. *Genetics*. 2008; 179:1681–92. <https://doi.org/10.1534/genetics.107.074450> PMID: 18562662
71. Lewis PO. A likelihood approach to estimating phylogeny from discrete morphological character data. *Systematic Biology*. 2001; 50:913–25. PMID: 12116640
72. Muse SV, Gaut BS. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Mol Biol Evol*. 1994; 11(5):715–24. <https://doi.org/10.1093/oxfordjournals.molbev.a040152> PMID: 7968485
73. Sokal RR, Rohlf FJ. *Biometry*: 3rd Edition. New York: W. H. Freeman and Company; 1995.
74. Rabier CE, Ta T, Ane C. Detecting and locating whole genome duplications on a phylogeny: a probabilistic approach. *Mol Biol Evol*. 2014; 31(3):750–62. <https://doi.org/10.1093/molbev/mst263> PMID: 24361993.
75. De Bie T, Cristianini N, Demuth JP, Hahn MW. CAFE: a computational tool for the study of gene family evolution. *Bioinformatics*. 2006; 22(10):1269–71. <https://doi.org/10.1093/bioinformatics/btl097> PMID: 16543274
76. Lander ES, Green P. Construction of multilocus genetic linkage maps in humans. *Proc Natl Acad Sci U S A*. 1987; 84:2363–7. PMID: 3470801
77. Felsenstein J, Churchill GA. A hidden markov model approach to variation among sites in rate of evolution. *Molecular Biology and Evolution*. 1996; 13(1):93–104. <https://doi.org/10.1093/oxfordjournals.molbev.a025575> PMID: 8583911
78. Stark C, Breitkreutz BJ, Chatr-Aryamontri A, Boucher L, Oughtred R, Livstone MS, et al. The BioGRID Interaction Database: 2011 update. *Nucleic Acids Res*. 2011; 39(Database issue):D698–704. Epub 2010/11/13. <https://doi.org/10.1093/nar/gkq1116> PMID: 21071413.
79. *Arabidopsis* Interactome Mapping Consortium. Evidence for network evolution in an *Arabidopsis* interactome map. *Science*. 2011; 333(6042):601–7. Epub 2011/07/30. <https://doi.org/10.1126/science.1203877> PMID: 21798944.
80. Mi H, Huang X, Muruganujan A, Tang H, Mills C, Kang D, et al. PANTHER version 11: expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool enhancements. *Nucleic Acids Res*. 2017; 45(D1):D183–D9. <https://doi.org/10.1093/nar/gkw1138> PMID: 27899595.

81. Mi H, Muruganujan A, Casagrande JT, Thomas PD. Large-scale gene function analysis with the PANTHER classification system. *Nat Protoc.* 2013; 8(8):1551–66. <https://doi.org/10.1038/nprot.2013.092> PMID: 23868073.
82. Marcet-Houben M, Gabaldon T. Beyond the Whole-Genome Duplication: Phylogenetic Evidence for an Ancient Interspecies Hybridization in the Baker's Yeast Lineage. *PLoS Biol.* 2015; 13(8):e1002220. <https://doi.org/10.1371/journal.pbio.1002220> PMID: 26252497.
83. Blanc G, Wolfe KH. Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. *Plant Cell.* 2004; 16:1679–91.
84. Scannell DR, Byrne KP, Gordon JL, Wong S, Wolfe KH. Multiple rounds of speciation associated with reciprocal gene loss in polyploid yeasts. *Nature.* 2006; 440:341–5. <https://doi.org/10.1038/nature04562> PMID: 16541074
85. Huang C-H, Sun R, Hu Y, Zeng L, Zhang N, Cai L, et al. Resolution of Brassicaceae phylogeny using nuclear genes uncovers nested radiations and supports convergent morphological evolution. *Mol Biol Evol.* 2016; 33(2):394–412. <https://doi.org/10.1093/molbev/msv226> PMID: 26516094
86. Schnable JC, Freeling M, Lyons E. Genome-wide analysis of syntenic gene deletion in the grasses. *Genome Biol Evol.* 2012; 4(3):265–77. Epub 2012/01/26. <https://doi.org/10.1093/gbe/evs009> PMID: 22275519.
87. Woodhouse MR, Schnable JC, Pedersen BS, Lyons E, Lisch D, Subramaniam S, et al. Following tetraploidy in maize, a short deletion mechanism removed genes preferentially from one of the two homeologs. *PLoS Biol.* 2010; 8(6):e1000409. <https://doi.org/10.1371/journal.pbio.1000409> PMID: 20613864
88. Liò P, Goldman N. Models of molecular evolution and phylogeny. *Genome Res.* 1998; 8:1233–44. PMID: 9872979
89. Scienski K, Fay JC, Conant GC. Patterns of Gene Conversion in Duplicated Yeast Histones Suggest Strong Selection on a Coadapted Macromolecular Complex. *Genome biology and evolution.* 2015; 7(12):3249–58. <https://doi.org/10.1093/gbe/evv216> PMID: 26560339
90. Evangelisti AM, Conant GC. Nonrandom survival of gene conversions among yeast ribosomal proteins duplicated through genome doubling. *Genome Biology and Evolution.* 2010; 2:826–34. <https://doi.org/10.1093/gbe/evq067> PMID: 20966100
91. Casola C, Conant GC, Hahn MW. Very low rate of gene conversion in the yeast genome. *Mol Biol Evol.* 2012; 29(12):3817–26. Epub 2012/07/31. <https://doi.org/10.1093/molbev/mss192> PMID: 22844073.
92. Mayfield-Jones D, Washburn JD, Arias T, Edger PP, Pires JC, Conant GC. Watching the grin fade: Tracing the effects of polyploidy on different evolutionary time scales. *Seminars in Cellular and Developmental Biology.* 2013; 24:320–31.
93. Chang PL, Dilkes BP, McMahon M, Comai L, Nuzhdin SV. Homoeolog-specific retention and use in allotetraploid *Arabidopsis suecica* depends on parent of origin and network partners. *Genome Biol.* 2010; 11(12):R125. <https://doi.org/10.1186/gb-2010-11-12-r125> PMID: 21182768.
94. Wisecaver JH, Borowsky AT, Tzin V, Jander G, Kliebenstein DJ, Rokas A. A global co-expression network approach for connecting genes to specialized metabolic pathways in plants. *Plant Cell.* 2017:tpc.00009.2017.
95. Codoner FM, Fares MA. Why should we care about molecular coevolution? *Evolutionary Bioinformatics.* 2008; 4:29–38. Epub 2009/02/11. PMID: 19204805.