Convergent evolution of polyploid genomes from across the eukaryotic tree of life

Yue Hao[1,*], Jonathon Fleming[2,*], Joanna Petterson[3], Eric Lyons[4], Patrick P. Edger[5,6], J. Chris Pires[7,8,9], Jeffrey L. Thorne[2,10-12], and Gavin C. Conant[2,10,12,†]

[1]Biodesign Center for Mechanisms of Evolution, Arizona State University, Tempe, AZ, U.S.A. [2]Bioinformatics Research Center and [3]Department of Biomedical Engineering, North Carolina State University, Raleigh, NC, U.S.A.; [4]School of Plant Sciences, University of Arizona, Tucson AZ, U.S.A.; [5]Department of Horticulture, Michigan State University, East Lansing, MI, U.S.A.; [6]Ecology, Evolutionary Biology and Behavior, Michigan State University, East Lansing, MI, U.S.A.;[7]Division of Biological Sciences, University of Missouri-Columbia, MO, U.S.A.; [8]Informatics Institute, University of Missouri-Columbia, MO, U.S.A.; [9]Bond Life Sciences Center, University of Missouri-Columbia, MO, U.S.A.; [10]Program in Genetics, [11]Department of Statistics and [12]Department of Biological Sciences, North Carolina State University, Raleigh, NC, U.S.A.

*These authors contributed equally to this work
†Correspondence: G. Conant, gconant@ncsu.edu

Running Head: Convergent patterns of evolution after polyploidy
Keywords: polyploidy, convergent evolution, reciprocal gene loss, evolutionary model

Abstract:

By modeling the homoeologous gene losses that occurred in fifty genomes deriving from ten distinct polyploidy events, we show that the evolutionary forces acting on polyploids are remarkably similar, regardless of whether they occur in flowering plants, ciliates, fishes or yeasts. The models suggest these events were nearly all allopolyploidies, with two distinct progenitors contributing to the modern species. We show that many of the events show a relative rate of duplicate gene loss prior to the first post-polyploidy speciation that is significantly higher than in later phases of their evolution. The relatively low selective constraint seen for the single-copy genes these losses produced lead us to suggest that most of the purely selectively neutral duplicate gene losses occur in the immediate post-polyploid period. We also find ongoing and extensive reciprocal gene losses (RGL; alternative losses of duplicated ancestral genes) between these genomes. With the exception of a handful of closely related taxa, all of these polyploid organisms are separated from each other by tens to thousands of reciprocal gene losses. As a result, it is very unlikely that viable diploid hybrid species could form between these taxa, since matings between such hybrids would tend to produce offspring lacking essential genes. It is therefore possible that the relatively high frequency of recurrent polyploidies in some lineages may be due to the ability of new polyploidies to bypass RGL barriers.

Introduction

That organisms with doubled genomes existed was evident early in the history of genetics (Kuwada 1911; Clausen and Goodspeed 1925), and a lively debate was entered as to the implications of this fact. Wagner (1970) declared polyploidy to be "evolutionary noise" the same year that Susumu Ohno (1970) was giving it pride of place among the forces generating evolutionary innovations. The advent of genome sequencing changed the ground of this debate, opening new horizons of time for studies of the prevalence and influence of polyploidy. We know now that great branches of the eukaryotic evolutionary tree, including the vertebrates, all flowering plants and many yeasts, descend from ancient polyploids (Van de Peer, et al. 2017), events that were difficult or impossible to detect with older data. For reasons that are not yet fully understood, many of these groups also show recurrent polyploidies, especially flowering plants (Soltis, et al. 2009) and teleost fishes (Braasch and Postlethwait 2012).

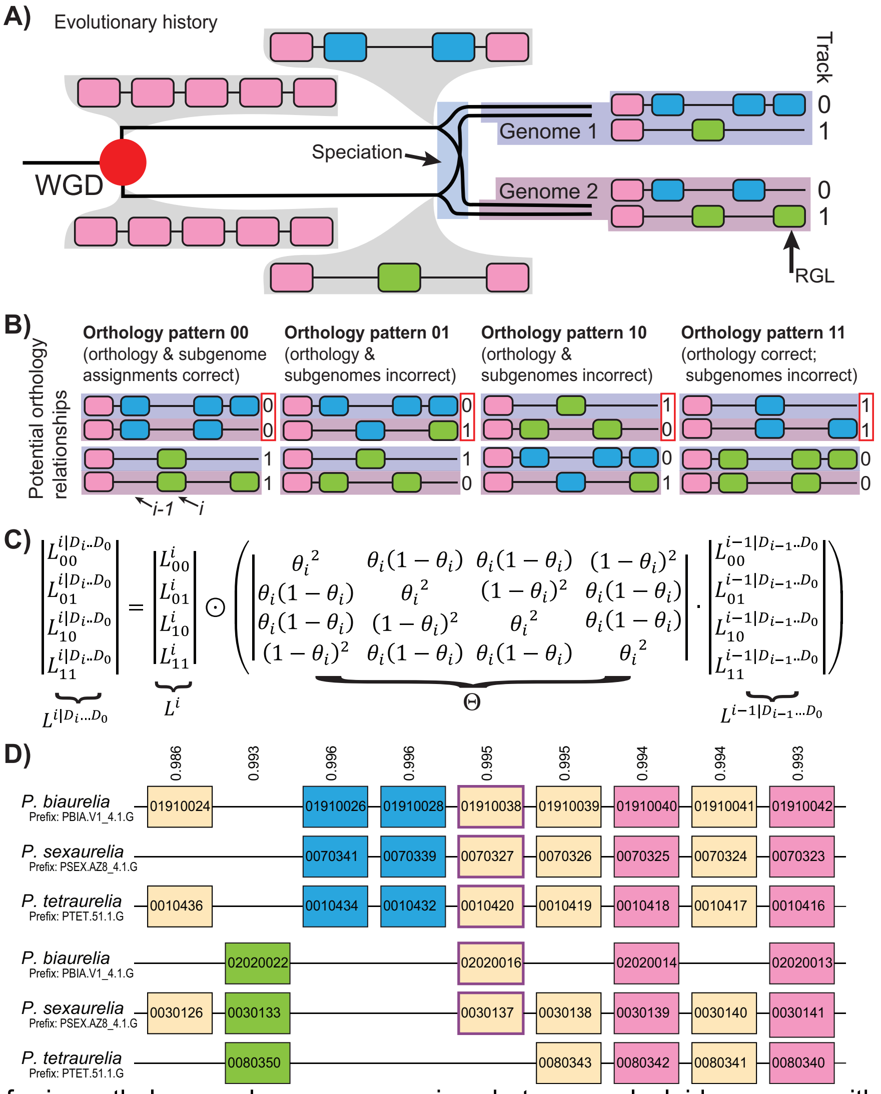
With this extensive new set of polyploidies as a resource, other old questions can also be revisited, such as the relative prevalence of auto- and allopolyploids (Stebbins Jr 1947). Allopolyploidy refers to hybridizations between distinct species that result in doubled (or more) genomes, while autopolyploids are derived from a single progenitor species (Kuwada 1911; Clausen and Goodspeed 1925; Stebbins Jr 1947). Analyses of several paleopolyploid genomes have shown that while gene losses are common after polyploidy, in many cases the losses are not experienced equally by the two parental subgenomes (Thomas, et al. 2006; Emery, et al. 2018), a pattern known as biased fractionation. These biases are plausible but not definitive indicators of allopolyploidy.

There has also been controversy as to whether and how polyploidy affects the rate of speciation. Werth and Windham (1991) proposed that reciprocal gene losses (RGLs), the alternative loss of one of the two duplicated genes from different populations, could create Bateson–Dobzhansky–Muller incompatibilities between populations, because matings between them would give rise to offspring with no copies of the genes. Were those genes essential, the offspring lacking them would be inviable (Werth and Windham 1991) Such incompatibilities have been observed both in the wild and the laboratory (Mizuta, et al. 2010; Maclean and Greig 2011). Muir and Hahn (2015) emphasize that RGL requires a period of reproductive isolation to form.

In the case of the yeast polyploidy, RGLs are commonly found between the descendant genomes, suggesting the potential for polyploidy to create new species by purely neutral means (Scannell, et al. 2006; Scannell, et al. 2007). However, direct analyses of the speciation and extinction rates of polyploid and nonpolyploid lineages has yielded inconclusive results, with some studies claiming reduced net diversification rates among polyploids and others disagreeing (Mayrose, et al. 2011; Soltis, Segovia-Salcedo, et al. 2014). More generally, the immediate and long-term adaptive value of polyploidy remains unclear: for instance, allopolyploids combine hybridizations with genome doubling and may derive immediate advantages from the hybridization effects rather than the doubling itself (Soltis, Visger, et al. 2014). Increased stress tolerance in polyploid organisms has also been invoked to argue for a radiation of polyploidy coincident with global catastrophes such as the KT mass extinction (Fawcett, et al. 2009).

Using our tool for modeling the evolution of polyploid genomes, POInT (the Polyploidy Orthology Inference Tool; Conant and Wolfe 2008), we explored the resolution of ten independent polyploidies. We adopt the term “homoeolog” below to refer to homologous genes produced by any type of polyploidy rather than “duplicate” or “ohnolog” because the events considered comprise several distinct types of polyploidy. The hallmark of polyploidy in a genome is a pattern of interleaved synteny, comprising not just the surviving homoeologs but also single-copy genes that are now found in interleaved positions on pairs (or more) of chromosomal segments homologous to the ancestral single-copy regions. In Figure 1A, we show an example of this evolutionary process, which yields conserved synteny blocks in the extant genomes. Those synteny blocks differ between genomes, meaning it is necessary to “phase” them into orthologous regions. As shown in Figure 1B, for a set of *n* tetraploid genomes, there are 2^n possible orthology relationships at each ancestral locus. We use the term “pillar” to denote all of the genes or lost homoeologs at such a locus. POInT computes the likelihood of the observed homoeolog presence/absence data at each pillar for each possible orthology relationship. Via a hidden Markov model (HMM) that combines the possible orthology relationships for each pillar with the syntenic organization among pillars (Figure 1C), POInT employs posterior decoding to infer orthology estimates for each pillar with associated posterior probabilities (top of Figure 1D) as well as estimates of the model parameters describing the process of homoeolog loss (Figure 2B and C).

Our analyses here encompass a total of 50 polyploid genomes and more than 460,000 individual genes (Figure 2A). We find that the patterns of gene loss after these different events show strikingly similar patterns, with strong evidence for biased fractionation and homoeolog fixation. Using synonymous substitutions as an evolutionary clock, we show that the rate of gene loss immediately after a polyploidy is generally higher than in later periods. RGL is also prevalent after all of these polyploidies, and we suggest it might introduce barriers to hybridization that could be overcome through subsequent allopolyploidy events.



$$\begin{vmatrix} L_{00}^{i|D_i..D_0} \\ L_{01}^{i|D_i..D_0} \\ L_{10}^{i|D_i..D_0} \\ L_{11}^{i|D_i..D_0} \end{vmatrix} = \begin{vmatrix} L_{00}^{i} \\ L_{01}^{i} \\ L_{10}^{i} \\ L_{11}^{i} \end{vmatrix} \odot \left(\begin{vmatrix} \theta_i^2 & \theta_i(1-\theta_i) & \theta_i(1-\theta_i) & (1-\theta_i)^2 \\ \theta_i(1-\theta_i) & \theta_i^2 & (1-\theta_i)^2 & \theta_i(1-\theta_i) \\ \theta_i(1-\theta_i) & (1-\theta_i)^2 & \theta_i^2 & \theta_i(1-\theta_i) \\ (1-\theta_i)^2 & \theta_i(1-\theta_i) & \theta_i(1-\theta_i) & \theta_i^2 \end{vmatrix} \cdot \begin{vmatrix} L_{00}^{i-1|D_{i-1}..D_0} \\ L_{01}^{i-1|D_{i-1}..D_0} \\ L_{10}^{i-1|D_{i-1}..D_0} \\ L_{11}^{i-1|D_{i-1}..D_0} \end{vmatrix} \right)$$

Figure 1: Inferring orthologous chromosome regions between polyploid genomes with POInT (the Polyploidy Orthology Inference Tool). **A)** Cartoon model of gene losses and a speciation event after a whole-genome duplication. Immediately after the WGD, all five genes are present in two homoeologous copies. Three homoeologous gene losses occur prior to the split of the two species, one in the less fractionated subgenome (Track "0;" yielding the green gene in the lower window) and two from the more fractionated subgenome (Track "1;" yielding the two blue genes in the upper window). After the speciation event, Genome 1 loses a homoeolog from the more

fractionated subgenome and Genome 2 loses one from the less fractionated subgenome, a case of reciprocal gene loss (RGL). **B)** There are 2^n=4 potential ways of phasing the two chromosomal regions from Genome 1 relative to Genome 2 (e.g., of assigning orthology between the two regions). We identify these 4 states with the subgenome assignment for the top track for each of the two genomes (00→11; red boxes at the right of each diagram). POInT uses a model of homoeolog loss to compute the likelihood of the observed gene presence/absence data at each locus (or "pillar") for each of these 2^n relationships. These relationships each constitute a hidden state of the HMM implemented by POInT whereas a likelihood of observed gene presence/absence data for a relationship represents an emission probability for the HMM. **C)** Recurrence equation for computing the likelihood of each orthology assignment at pillar *i* conditional on the data at pillars 0 through *i-1* (see **B**). For pillar *i*, we define a vector L^i to be the likelihood of the orthology states, with elements L_{00}^i, L_{01}^i, L_{10}^i and L_{11}^i being POInT's estimates of the likelihood of each such state based on the gene presence/absence data at that pillar. We then use a transition probability matrix Θ, with each entry representing the probability that pillar *i* has a particular orthology state conditional upon another orthology state at *i-1*. The probability that the orthology state is maintained between pillars *i-1* and *i* is $1-\theta_i$ for each genome (and $(1-\theta_i)^2$ in total); the chance that one genome changes orthology state is $\theta_i(1-\theta_i)$ and the chance that both change is θ_i^2. Here, $\theta_\iota=\theta$, a global constant estimated from the data by maximum likelihood, except when synteny is not maintained between pillars, in which case θ_i =0.5 (adjacent pillars do not inform on each other's orthology state; *Methods*). To compute a likelihood for the entire data set, POInT implements an HMM forward algorithm that expresses $L^{i|D_i \ldots D_0}$, the probabilities of orthology relationships for pillar *i* and the observed data at pillars *0* through *i* (denoted $D_i \ldots D_0$), in terms of the emission probabilities L^i, the transition probabilities Θ and the probabilities $L^{i-1|D_{i-1} \ldots D_0}$ that were already computed for pillar *i-1.* The vector of $L^{i|D_i \ldots D_0}$ is then the element-wise vector product (indicated with the "⨀") of $\Theta \cdot L^{i-1|D_{i-1} \ldots D_0}$ and L^i. This formula can be applied sequentially starting at pillar 0, with the base case $L^{0|D_0} = L^0$. For *m* pillars, the overall likelihood of the dataset is then the sum of the elements of $L^{m|D_m \ldots D_0}$. **D)** Given an inferred ancestral gene order prior to the polyploidy (*Methods*), POInT employs posterior decoding to infer the orthology relationships at each pillar. Here we illustrate a small region of such an ancestral order from the more recent *Paramecium* WGD (after phasing from the earlier duplication, see *Methods*), showing the set of orthology relationships inferred by posterior decoding. For reference, genes in adjacent pillars that are also neighbors in an extant genome are shown connected by lines. The number above each pillar is the posterior probability of the inferred orthology relationship. The upper set of three tracks correspond to the less-fractionated parental subgenome, the lower three to the more fractionated one, illustrating the possibility for local variation in biased fractionation. Gene retained from *only* the less-fractionated genome are colored blue, from *only* the more fractionated one green, and fully retained duplicates are shown in pink. All other patterns of duplicate retention are shown in beige for clarity. See also Figure 2B.

Results

Modeling evolution after ten independent polyploidies.

Using POInT, we assembled a set of ~70,000 homoeologous loci produced by ten different polyploidies. For each polyploidy, we inferred a set of pillars that it created and ordered them so as to maximize the retained synteny among the extant genes, approximating the ancestral order of the single-copy genes just prior to polyploidy (*Methods*). Six of the events are whole genome duplications (WGDs or tetraploidies): At-α in *Arabidopsis thaliana* and its relatives, a WGD found in legumes, the ρ event from grasses, the teleost-specific genome duplication (TGD), and WGDs from salmonids and yeasts. We further analyzed an asexual triploidy in nematodes, a hexaploidy (whole genome triplication; WGT) in cabbages and their relatives (*Brassica* WGT) and two octoploidies: the vertebrate 2R polyploidy and another in the paramecia (Figure 2A). Analyzing octoploidies in POInT is computationally expensive. As a result, we modeled the octoploidy among the paramecia as occurring via two sequential genome duplications and then extracted and analyzed only the more recent of these two events for the remainder of our work (*Methods*). This approach failed with the vertebrate 2R event, presumably because the two events are very ancient and closely spaced in time. A visual interface to these data is available from the POInT browser (http://wgd.statgen.ncsu.edu).

For the WGD events, we compared nested models of evolution (Figure 2B and Supplemental Table 1) that describe the process of homoeolog loss after polyploidy: these models differ as to whether they include biased fractionation, duplicate fixation and convergent homoeolog losses. For all seven tetraploidies, models that allow for the fixation of homoeologs after polyploidy fit the observed loss data better than models without such an effect ($\gamma \neq 0$; $P<10^{-10}$; likelihood ratio test or LRT; Figure 2). In addition, every event save that in yeast shows strong evidence for biased fractionation ($\varepsilon \neq 1$; $P<10^{-7}$; LRT; Figure 2), while all but the *Paramecium* event show a pattern of independent yet convergent losses to the same homoeolog in independent lineages ($\delta \neq 0$; $P<10^{-10}$; LRT; Figure 2). The nematode triploidy and the *Brassica* WGT also share similar patterns of biased fractionation (Figure 2 and Supplemental Table 2).

The fact that these events are of widely differing ages is evident from the different degrees of loss/resolution seen in the extant genomes. The branches of Figure 2A are color-coded by POInT's inferences of the proportion of single-copy genes (e.g., loci where all but one of the homoeologous genes have been lost) present at their beginning and ending. While the yeast WGD is inferred to be nearly "fully" resolved (nearly all homeologous loci reduced to single-copy), the tetraploidy in salmonid fishes and the nematode triploidy show proportionally few single-copy genes. The nematode triploidy differs from the remaining events in that these animals are asexual triploids and are likely under a different selective regime in their gene losses, (Schoonmaker, et al. 2020). The continued occurrence of meiotic chromosome pairings of homoeologous chromosomes created by the salmonid event may have reduced the rate of homoeolog loss in those genomes (Allendorf, et al. 2015).

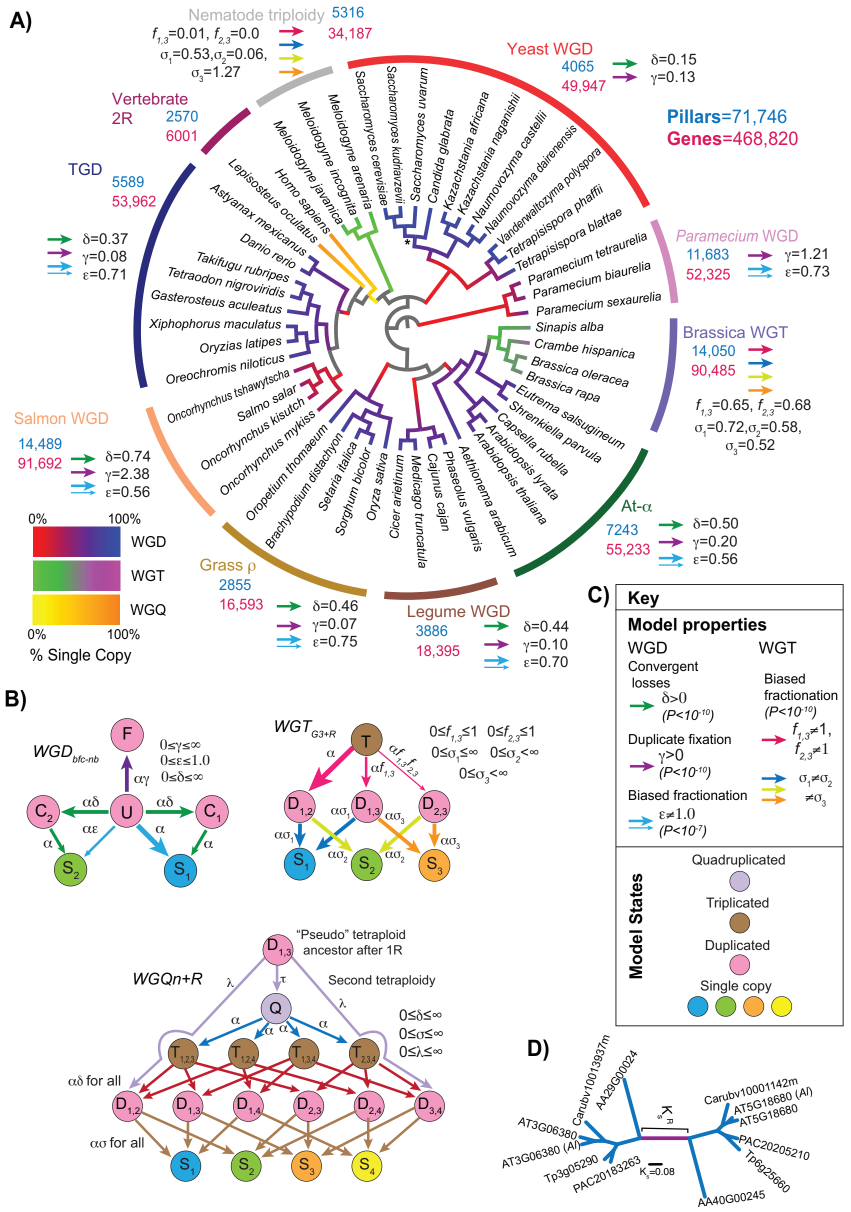


Figure 2: Modeling the resolution of ten polyploidy events with POInT. **A)** The assumed phylogenetic relationships of the ten polyploidies studied. Grey branches indicate where no polyploidy event was studied. The relationships of the taxa were inferred from the homoeolog loss data for the legume WGD, grass ρ, nematode triploidy, salmonid WGD and the Paramecium WGD. For the yeast WGD, At-α, the TGD and the Brassica WGT, the relationships were taken from published sources (the vertebrate 2R tree is trivial). Because the temporal divergences of various groups are not well established, the tree is illustrated in an ultrametric format with nonmeaningful branch lengths (Scaled topologies for each event are shown in Supplemental Figure 4). However, each polyploid branch is colored using POInT's estimates of the proportion of loci that were single-copy at its beginning and ending. Corresponding color keys for tetra, hexa and octoploidies are shown. The number of "pillars" (homoeologous loci) and the total number of gene models studied across each event are noted, as are the total number of loci and genes considered. The "*" on the yeast WGD branch indicates the branch where the proportion of genes returned to single-copy that are presently essential was tested (Supplemental Table 5). Next to each event, we show arrows and parameter estimates indicating post-polyploidy evolutionary processes such as biased

fractionation for which we found significant evidence in that event (see key in panel **C**). **B)** Nested models of post-polyploidy evolution for the three types of events (WGD: whole-genome duplication/tetraploidy, WGT: whole-genome triplication/hexaploidy and WGQ: whole-genome quadruplication/octoploidy). Using POInT, we fit nested models of gene loss after polyploidy with likelihood ratio tests (*Methods*). **WGD:** all pillars start in state **U** (Undifferentiated), from which they can transition to either the three other duplicated states, $\mathbf{C_1}$ (Converging state 1), $\mathbf{C_2}$ (Converging state 2) and **F** (Fixed) or to the two single-copy states $\mathbf{S_1}$ (Single-copy 1) and $\mathbf{S_2}$ (Single-copy 2). $\mathbf{C_1}$ and $\mathbf{S_1}$ are states where the gene from the less-fractionated parental subgenome will be or are preserved, and $\mathbf{C_2}$ and $\mathbf{S_2}$ the corresponding states for the more-fractionated parental subgenome. The null model has parameters $\gamma=\delta=0$ and $\varepsilon=1.0$. Duplicate fixation is inferred when $\gamma\neq0$, convergent losses when $\delta\neq0$ and biased fractionation when $\varepsilon<1.0$. **WGT:** in the base model all pillars start in state **T** (Triplicated) and transition first to duplicated states ($\mathbf{D_{x,y}}$) and hence to the single-copy states ($\mathbf{S_x}$). Genome 1 is assumed to be favored (fewer losses) and the identity of that genome inferred in the POInT computation. Losses from the triplicated state are then increasingly disfavored first to $\mathbf{D_{1,3}}$ (parameter $f_{1,3}$) and $\mathbf{D_{2,3}}$ (parameter $f_{2,3}$). There are also individual rates of loss from the duplicated to single-copy states (δ_x). In the null model, $f_{1,3}= f_{2,3}$=1.0 and $\delta_1= \delta_2= \delta_3$. We also fit a separate model that allow this set of parameters to take on separate values on the root branch and on the remaining branches (Supplemental Table 1). **WGQ:** Models of octoploid formation. The null model simply treats the four subgenomes as equivalent and as starting in the quadruplicated state (**Q**). This model has different loss rates from triplicated to duplicated loci ($\mathbf{T}_{x,y,z}$ to $\mathbf{D_{x,y}}$, parameter δ) and duplicated to single-copy loci ($\mathbf{D_{x,y}}$ to $\mathbf{S_x}$, parameter σ). A formation model for the octoploidy can then be added: all pillars start in state $\mathbf{D_{1,3}}$ and can symmetrically experience a gene loss from genome 1 or 3 (parameter λ) and transition to state $\mathbf{D_{1,2}}$ or $\mathbf{D_{3,4}}$ or become quadruplicated (null transition). The three models illustrated here are the most complex model fit to the various events, including the parameters associated and their numerical range. **C)** Description of the various modeled features from panels **A** and **B** (top) and the model states from **B** (bottom). **D)** An example mirrored gene tree for a completely retained set of homoeologs from At-α, illustrating the trees from which synonymous divergences were estimated. The branch lengths are given in number of synonymous substitution per synonymous site (e.g., K_s), with the shared internal (e.g., "root") branch shown in purple (K_s^R). For analysis purposes, the length of this branch was always divided by two to be comparable to the remaining branches (e.g., split at its midpoint).

Many events show rapid homoeolog loss immediately after polyploidy.

Loss of duplicate genes immediately after polyploidy can be rapid (Scannell, et al. 2006; Scannell, et al. 2007), and at least two non-exclusive hypotheses exist as to why. The first is that genetic drift should eliminate truly redundant gene copies quickly (Li 1980; Lynch and Conery 2000). The second is the potential for "selected" duplicate losses. These losses might occur if the increases in gene copy number after polyploidy induce disadvantageous dosage conflicts, such that natural selection acts to remove the homoeologous copies in question (Edger and Pires 2009; De Smet, et al. 2013).

To study the pattern of early losses, we examined the divergence that occurred immediately after the polyploidy and prior to any speciation events. In the context of a gene tree for a pair of homoeologous genes produced by a WGD, this period corresponds to the internal branch of the gene tree separating that pair of homoeologs. For a WGT, the situation is analogous except that there are three such branches separating the three homoeologous copies. For simplicity, we refer to these branch(es) as the "root" (purple in Figure 2D). For all branches in each polyploidy, we obtained a rough estimate of the time encompassed by that branch by using the mean number of synonymous substitutions per synonymous site ($\overline{K_s}$) across many homoeologous genes as a neutral clock (*Methods*). The rate of homoeolog loss for each branch is given by POInT's branch length estimate (αt), computed with an irreversible exponential loss model proportional to the number of homoeologous copies at the beginning of that branch (meaning that they are not biased by the fact that later branches have fewer total homoeologs available for loss, *Methods*). The ratio of $\alpha \cdot t/\overline{K_s}$ gives a sense of whether homoeolog losses per time are unusually high or low for a given branch relative to other branches in the same polyploidy. For the majority of the polyploidies, we found that the $\alpha \cdot t/\overline{K_s}$ ratio was higher for the root branch than any other branch, consistent with a more rapid loss of homoeologs along this branch (Figure 3). This result is the more striking because the inferred mean K_s value for the root branch ($\overline{K_s^R}$) should, in the case of an allopolyploidy, also include the pre-polyploidy progenitor divergences. Hence, the $\overline{K_s^R}$values for these events should be over-estimates, making the $\alpha \cdot t/\overline{K_s^R}$ ratio an underestimate of the relative homoeolog loss rate along the root branch.

If natural selection were actively favoring the loss of some homoeologous copies immediately after polyploidy, we might expect that the genes involved in those early losses would display a stronger selective constraint than do homoeologous copies lost later in the polyploidy's history. We hence compared the average selective constraint, measured as the ratio of nonsynonymous to synonymous substitutions, or K_a/K_s, of fully single-copy genes whose homoeologs were lost along the root branch to that of other fully single-copy genes where the preservation of homoeologous copies from alternative subgenomes means that the losses must have occurred after the first speciation event. For most events we observe little difference between these two groups, while for the Legume WGD the single-copy genes lost later are actually *more* constrained, the opposite of the prediction for selected losses (Supplemental Figure 1).

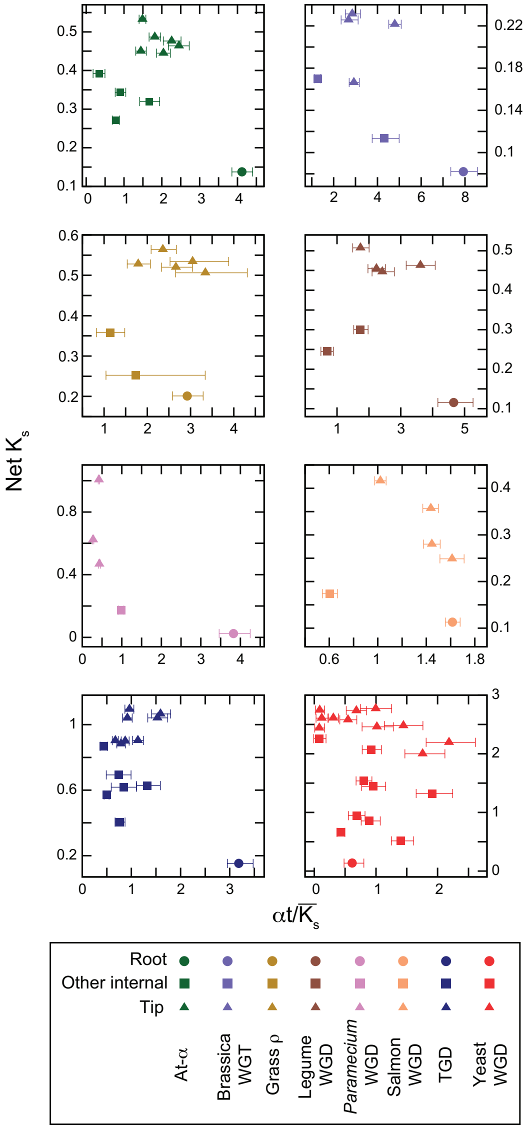


Figure 3: Rapid loss of homoeologs immediately after polyploidy. On the *x*-axis is the ratio of rate of homoeolog loss (the □t branch length estimate from POInT's models, see Figure 2) and the estimated mean synonymous divergence for that branch ($\overline{K_s}$; see *Methods*). Hence, larger values of this ratio indicate more homoeolog losses per unit K$_s$. For the At-α, Brassica WGT, Legume WGD, Paramecium WGD and the TGD, the $\alpha \cdot t/\overline{K_s}$ ratio for the root branch is significantly larger than seen on any other branch (c.f., the 95% confidence intervals shown, computed as described in the *Methods* section). For these panels, we used a model excluding duplicate fixation here because including fixation in the model occasionally results in very long estimates of tip branch lengths (*Methods*). However, our conclusions are similar under the fully WGD$_{\text{bfc-nb}}$ model (see Supplemental Figure 7). On the *y*-axis is the net synonymous divergence to the end of the branch in question: in other words, the sum of the synonymous divergence of that branch and all its ancestors back to the root branch. This net divergence value is a rough indicator of the time since the polyploidy for each branch. The root branch is indicated with a circle, other internal branches with squares and tip branches with triangles.

Extensive reciprocal gene loss between pairs of polyploid taxa.

Following Scannell and colleagues (2006; 2007), we searched for post-polyploidy reciprocal gene losses (RGL). We omitted the vertebrate 2R and nematode triploidy from this analysis due to the fragmented nature of the genomes used. With the exception of three closely related yeast species in the *Saccharomyces* genus, every pair of genomes in our remaining eight polyploidies were separated by at least 4 RGLs (this minimal number was seen in the platyfish, tilapia and medaka clade of the TGD; Figure 4C), with the number rising to over a thousand for a few of the yeast taxa pairs. These conclusions are also robust to the confidence cutoffs used to infer the RGLs (Supplemental Figure 2). Our results are in accord with previous work in yeasts and grasses (Scannell, et al. 2006; Scannell, et al. 2007; Schnable, et al. 2012), and there appears to be a relatively direct relationship between the synonymous divergence of a pair of taxa (a proxy for divergence time) and the number of RGLs separating them (Figure 4A and B). Such a relationship would be expected if both RGLs and synonymous substitutions were accumulating through neutral evolutionary processes (Figure 4A). However, the proportionality between synonymous substitutions and RGLs differs between polyploidy events, with the yeast WGD showing more RGLs per unit K_s than the other events. When we compared the genes involved in reciprocal losses in zebrafish, *A. thaliana* and bakers' yeast to other single-copy genes, there were no significant functional differences between these two sets, again as one would expect were RGL a neutral process (*Methods*).

The evolutionary importance of RGLs can be assessed by the biological role of the genes that experienced it. For instance, were only "non-essential" genes to experience RGL, then it might not present significant barriers to hybridization. We can use experimental data on gene essentiality from bakers' yeast, *A. thaliana* and zebrafish (*Methods*) to ask whether the proportion of RGLs that include an essential gene differs from the overall proportion of essential single-copy genes. For the At-α and TGD events, the proportion of RGLs where the surviving gene in *A. thaliana* or zebrafish is essential does not differ from the proportion of other single-copy genes that are essential (Supplemental Table 3). Curiously, the RGLs found when comparing bakers' yeast to some of its nearer relatives are actually *more* likely to be essential than other single-copy genes (Supplemental Table 3). This overrepresentation is likely attributable to the shared duplicate losses that occurred prior to the first speciation event being underrepresented in essential genes (Supplemental Table 4). As a result, RGLs, which must have occurred *after* the first speciation event (see the yeast clade of Figure 2A), would be enriched in essential genes simply because more essential genes survived in duplicate past that first speciation.

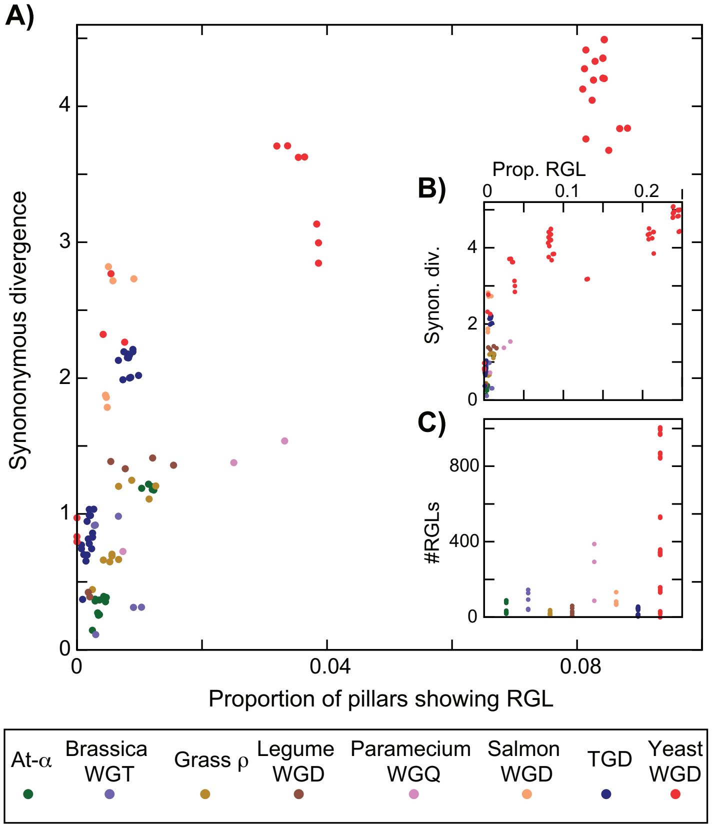


Figure 4: Reciprocal gene loss (RGL) after polyploidy. **A)** Reciprocal gene losses (RGLs) between pairs of polyploid taxa (*x*-axis, normalized by the total number of loci/pillars analyzed for that event) as a function of the inferred synonymous divergence of those taxa (*y*-axis). Panel **A** gives a cropped view that focuses on RGLs in the non-yeast taxa, while panel **B** shows how the RGL frequencies in yeast dramatically exceed those for the remaining events. For each pair of taxa from a given event, we identified all single-copy loci in the two genomes where POInT infers a 95% or greater confidence that those genes are paralogs created by the ancient polyploidy and not more recent orthologs produced by the post-polyploidy speciation events. There are roughly linear relationships between RGL frequency and synonymous divergence. Because the data points shown are phylogenetically dependent (different species pairs share considerable common evolutionary history), we have not attempted to fit regression lines to these data .Standard approaches to phylogenetically independent contrasts (Felsenstein 1985) do not apply here as the inferred RGLs are pairwise species traits and not independent measures on each taxon. It is however notable that the asexually reproducing yeasts appear to accumulate more RGLs per unit K_s than other taxa. **B)** As for **A** but including the full range of RGL prevalence in the taxa sharing the yeast WGD. **C)** Total numbers of RGLs inferred for each pair of taxa for each event.

The importance of RGL in driving speciation events among polyploid taxa has been questioned on theoretical grounds, as the appearance of RGLs is subject to the same requirement of reproductive isolation as are the appearances of other genetic incompatibilities among populations (Muir and Hahn 2015). This objection has more force for obligately sexual organisms than it does for organisms such as bakers' yeast, where it is estimated that there are 1000 mitotic cell divisions for every meiosis and that only about 1% of meioses are out-crosses (Tsai, et al. 2008). Indeed, Figure 4 suggests that RGL may occur more frequently in yeasts (and potentially in some plants, which may also reproduce asexually) than in the teleost fishes and particularly the salmonids.

Even if RGL does not drive speciation, it still represents a barrier to diploid hybrids: most of the taxa pairs for which essentiality data are available are separated from each other by at least one RGL for an essential gene, the exceptions being some of the closest relatives of *A. thaliana,* zebrafish and bakers' yeast studied (Supplemental Table 3). This observation is consistent with studies of the relative frequency of diploid and polyploid hybridizations in flowering plants. In these lineages, it is rare to find successful diploid hybrids involving distantly related parental species (where RGLs could be common). However, allopolyploid hybrids appear to form at similar rates across a much larger range of divergence times (Buggs, et al. 2009). A potential explanation for the frequency of recurrent polyploidy is therefore simply that a new allopolyploidy can allow paleopolyploids to again enjoy the benefits of hybridization (such as hybrid vigor and heterosis; Birchler, et al. 2006; Chen 2010) in the face of their isolation due to RGL.

Discussion

There are a surprising number of similarities seen in the manner of polyploidy resolution across these independent polyploidies. Biased fractionation and other patterns in the homoeolog losses are similar across many events: reciprocal gene losses are also present for most pairs of polyploid taxa. The rate of homoeolog loss immediately after polyploidy is very high for many, but not all, events (Figure 3).

Moreover, the differences in evolutionary patterns we do see are often in keeping with what we know about the history of the events themselves. For instance, the salmonid WGD is marked by continuing pairing of homoeologous chromosomes in meiosis (Allendorf, et al. 2015).

These pairings appear to limit the number of homoeolog losses and, for this event, loss rates at the phylogeny tips and root are similar (per unit K_s). The grass ρ and yeast events have loss rates that are roughly similar (again per unit K_s) across time, a fact for which we currently do not have an operating hypothesis.

For the events that do show rapid losses along the root branch, which of the two hypotheses mentioned, drift or selected losses, seems to best explain our data? The homoeologs lost along the root are not more selectively constrained than other purely single-copy genes known to have been lost later (Supplemental Figure 1). This fact probably speaks against any very large number of selected losses. The single-copy genes as a whole are also generally somewhat less selectively constrained than are genes with surviving homoeologs (Supplemental Figure 1). Moreover, there is a clear pattern in most events whereby most of the fully single-copy genes that exist are predicted to have been lost on the root branch (Supplemental Figure 3). The yeast, nematode, and Paramecium events may violate this pattern because the nematode event is an asexual triploidy while the other two involve lineages that have significant rates of asexual reproduction. In such cases, restoring proper meiotic pairing is less necessary than in taxa with primarily sexual reproduction. As a result, we expect that asexually reproducing lineages could more easily form viable new species immediately after polyploidy, meaning that the post-polyploid "lag" in speciation might be less evident (Schranz, et al. 2012). As a preliminary hypothesis, we therefore propose that, for most polyploidies in animals and plants, the majority of the purely neutral homoeolog losses occur prior to extensive species divergence in the polyploid clade. A natural extension to this proposal would be that the post-polyploidy lag represents this earlier period of neutral homoeolog loss, though the question of why speciation events might be rare during such a period is still to be answered. A further implication would be that later losses (including RGLs) would have occurred in homoeologous pairs that were initially preserved to maintain dosage balance. They are then only lost when later mutations, such as expression changes, release this dosage constraint and allow the loss of one of the copies (Birchler, et al. 2005; Conant, et al. 2014). The higher selective constraint of genes with surviving homoeologs is arguably also consistent with this hypothesis.

While the best-studied ancient polyploidy is in bakers' yeast, it is hence atypical in a number of respects. Biased fractionation is much less evident here (Emery, et al. 2018), losses are not heavily biased toward the earliest phases of the polyploidy (Figure 3) and RGL is much more prevalent. As mentioned above, one major source of these differences is likely the relative timing of the post-polyploidy speciations: the yeasts had almost no lag between their polyploidy and the first speciation (Supplemental Figure 4; Schranz, et al. 2012).

Other questions remain unanswered. The relative formation rates of allo- and autopolyploids are uncertain. While recent polyploids appear to be approximately equally divided between the two (Barker, et al. 2016), the potential selective advantages of being an allopolyploid, and hence a hybrid (Alix, et al. 2017; Blanc-Mathieu, et al. 2017), could result in a strong skew towards allopolyploids among the rare polyploidies that survive to became the ancient events of the kind studied here (Barker, et al. 2016). The results here are consistent with this hypothesis, but our sample of events is potentially biased by the available genome sequences. Across all of the events, we find that the ubiquity of homoeolog fixation and (except in paramecia) convergent homoeolog losses both speak to a common selective environment acting to maintain certain homoeologs after all of these events. The most obvious candidate for such a selective force is again the dosage balance hypothesis: it argues that highly interacting genes tend to remain in multiple copies post-polyploidy to preserve the stoichiometry of those interactions (Birchler, et al. 2005; Birchler and Veitia 2012; Tasdighian, et al. 2017). Whatever the role of RGL in speciation, it is clear that all of these polyploid organisms possess a degree of isolation due to it. The role of RGL in recurrent polyploidy is hence an important topic for future research. Biology has a history of viewing “rules” as being more honored in the breach, but the commonalities in post-polyploidy genome evolution across wide taxonomic distances are both interesting in their own right and for the insight they give on other aspects of biology (Pires and Conant 2016).

Methods

Synteny block inference.

Our three-step pipeline for inferring blocks of *n*-fold conserved synteny (NCS) produced by polyploidy (Conant 2020) first uses GenomeHistory (Conant and Wagner 2002) to find all pairs of homologous genes between each polyploid genome and a nonpolyploid outgroup (see Supplemental Table 5 for genome details and Supplemental Table 6 for parameters). The second step seeks to place these homologous genes into *N*:1 relationships between the polyploid genome and the outgroup (*N=2* for a WGD, *N*=3 for a hexaploidy and *N*=4 for an octoploidy). Using simulated annealing (Kirkpatrick, et al. 1983), this step proposes sets of ordered pillars, each of which contains a single gene from the nonpolyploid outgroup (*G*) and no more than *N* of the homologs of that gene from the polyploid genome. The annealing algorithm then seeks a combination of these assignments and a relative ordering of the *m* outgroup genes $G_1..G_m$ that maximizes the number of synteny relations. We define two genes to be in synteny if they are neighbors in the genome, ignoring any genes without homologs to the compared genome. In the third step, these NCS blocks for each polyploid genome are merged across all of the polyploid genomes. In this merging, only pillars where we have at least one homologous and syntenic gene from each polyploid genome are included. With the set of merged pillars, a further simulated annealing search is undertaken to infer a global pillar order that minimizes the number of synteny breaks. While not strictly an ancestral genome inference (Sankoff and Blanchette 1998), it is helpful to think of this optimal ordering as approximating the order of the genes just prior to the polyploidy. Our previous work has shown that this inference approach is highly specific, with no apparent cases of paralogous genes not created by the polyploidies in question being included in the pillars (Emery, et al. 2018; Conant 2020).

Modeling polyploidies with POInT

At each pillar, POInT calculates the probability of the observed gene presence-absence data conditional upon all possible orthology relationships and a tree. It carries this uncertainty in orthology through its likelihood computations using a hidden Markov model that resembles the Lander-Green approach for constructing linkage maps on a pedigree (Lander and Green 1987). The parameter θ_i corresponds to the probability that the inferred orthology relationships change between syntenic neighbors at pillars *i-1* and *i*. When a pair of pillars are separated by a synteny break, their orthology relationships are independent (i.e., θ_i=1/2). Otherwise, θ_i= θ, a global parameter estimated from the data by maximum likelihood.

This modeling framework allows for testing hypotheses about post-polyploidy gene losses. For tetraploidies, we analyzed three phenomena: duplicate fixation, biased fractionation and overly frequent parallel losses of the same homoeolog on independent branches of the phylogeny (Supplemental Table 1). For the triplication events, we focused on differences in homoeolog loss rates between the three subgenomes (Supplemental Table 2). We further allowed the root branch to have separate values of the model parameters to account for the two-step nature of hexaploidy formation (Figure 2; Tang, et al. 2012). For the Paramecium and vertebrate 2R octoploidies, we used a null model (WGQ$_n$; Figure 2) where losses occur equally from all four subgenomes, but where the loss rate from triplicated and duplicated loci can differ from that seen in quadruplicated loci. We also added an octoploid formation step to this model, with all pillars starting in state D$_{1,3}$ and then either experiencing a loss followed by the second tetraploidy (transitioning to D$_{1,2}$ or D$_{3,4}$) or becoming quadruplicated (Q).

Analyzing nested genome duplications with POInT.

The vertebrates and ciliates experienced two sequential genome duplications relative to the outgroup genome to which they were compared. They hence present a challenge because the POInT computation for such an octoploidy with *n* genomes scales as $O(24^{2n})$. As a result, it is only computationally feasible to analyze two such octoploid genomes. However, if the consecutive whole-genome doublings were sufficiently separated in time, POInT can separate them using the two-step model just described. This model assumes each locus starts as a *duplicated* one and then may either remain duplicated until the second polyploidy (and hence become quadruplicated) or experience a gene loss prior to that event, meaning that the second event only produces a *duplicate* gene pair (Figure 2). We thus sought to phase regions from both octoploidies into pairs of regions created by the most recent genome doubling. For the ciliate genomes, we were able to phase the quadruplicated loci into 11,683 pairs of duplicated loci with at least one gene from each genome and where our orthology assignment confidence for assigning extant genes to one of the two subgenomes from the *first* polyploidy event was ≥99%. For the vertebrate 2R events, a model that attempts to phase the 2R duplicates fit the data no better than did the null model (*P*=0.1, likelihood ratio test with 1 *d.f.*) and so no further phasing was attempted.

POInT and topological inference.

For the legume WGD, the grass ρ event, the Paramecium tetraploidy, the nematode triploidy and the salmonid WGD, we used POInT to infer the maximum likelihood phylogeny under the WGD$_{bfc\text{-}nb}$ or WGT$_{G3}$ models and an exhaustive tree search (Supplemental Figure 4). For the Brassica WGT, we assumed that *B. rapa* and *B. oleracea* were sister taxa and tested all three rooted topologies consistent with this constraint. The topology for the yeast WGD was taken from Kurtzman and Robnett (2003), for the TGD from Near et al., (2012) and for At-α from Huang et al., (2016). The vertebrate 2R topology is trivial.

For the salmonid WGD, the inferred topology differs significantly from others that have been published. We therefore fit the full POInT model under the topology published by Crespi and Fulton (2004). The orthology estimates and model parameters are largely unaffected by this topology change: the orthology relationships of only 106 (0.7%) pillars with posterior probability >80% differ when the topology is changed, and 91 of these changes simply swap the identities of the more and less fractionated genomes. The corresponding figures for 95% confidence are 9 and 7 pillars.

Orthology inferences and inference of synonymous distances.

Using high confidence orthologs estimated with POInT, we computed the mean synonymous divergence for every branch for each polyploidy. The nematode triploidy and vertebrate 2R events were omitted from this analysis due to their fragmented synteny blocks. For the tetraploidies, we considered "nearly fully duplicated" pillars: i.e., pillars with at most one missing gene copy from each of the two gene trees produced by the genome duplication (two total losses) for all events except the TGD and yeast WGDs, where we allowed two losses from each subtree (four total losses). For the Brassica hexaploidy, we analyzed only fully triplicated pillars. At each such pillar, we aligned amino acid sequences for the genes in question with T-coffee (Notredame, et al. 2000). We fit the Goldman and Yang codon model of evolution (Goldman and Yang 1994) to the corresponding codon-preserving alignments and mirrored gene trees and extracted the estimated synonymous divergence (K_s) for each branch from this codon model as described by these authors.

With the possible exception of the salmonids (Allendorf and Thorgaard 1984; Braasch and Postlethwait 2012), all of the events studied are believed to be allopolyploids. For a given pillar in set of allopolyploid taxa, the mean synonymous divergence observed along this root branch ($\overline{K_S^R}$; Figure 2D) should represent the sum of the pre-polyploidy divergence of the diploid progenitors as well as the divergence that occurred after the polyploidy but before the first speciation event among the polyploid taxa. However, recombination events could, through genetic drift, result in the replacement of alleles from one of the progenitors with those from the other (Wolfe 2001). These recombinations, or homoeologous exchanges (HE; Gaeta and Chris Pires 2010) are reasonably common in neopolyploid plants (Doyle, et al. 2008; Chalhoub, et al. 2014; Zhang, et al. 2020), but it is not clear whether they are frequent enough to effect the divergence seen along these root branches. We extracted the coding sequences for each pillar that had every homoeologous gene preserved. Post-polyploidy homoeolog displacement (Gaut and Doebley 1997; Wolfe 2001) will erase the divergence between the progenitor genomes, leaving only the post-displacement divergence to be observed. In such a case, we might expect to observe two modes in synonymous divergence, a larger value for homoeologs that did not experience displacement and a smaller one (lacking the progenitor divergence) for homoeologs that did. To test this hypothesis, we fit the set of estimated synonymous divergences (K_s) along the root branches to either one or two log-normal distributions using the R package mclust (Scrucca, et al. 2016) with the best-fit model (i.e., one or two distributions) chosen with the Bayesian information criterion (BIC; Schwarz 1978). Values of K_s less than $5x10^{-3}$ or greater than 2.0 were omitted from these analyses as representing either no synonymous divergence or saturated synonymous divergence, respectively. When two distributions were fit, a “weighting” *p* reflecting the mixing proportion of each component was also estimated. For a few root branches, a bimodal distribution is preferred. However, in most cases this bimodality is not consistent across different collections of pillars and, even when it is, the proportion of pillars belonging to one of the “modes” is generally very small (Supplemental Table 7). We hence see little suggestion of HE in these data.

Filtering for extreme instances of gene conversion.

Because gene conversion among homoeologs (as seen in yeasts; Evangelisti and Conant 2010; Scienski, et al. 2015) could confound our K_s estimates, we sought to filter out pillars that showed strong evidence of having experienced it. We created “gene conversion gene trees” for each pillar where each homoeologous gene was forced to be sister to its paralog(s). Any pillars where the likelihood of the sequence alignment under these gene conversion trees was higher than that seen in the mirrored species trees was omitted from our estimates of synonymous divergence (Supplemental Figure 5).

Comparing duplicate loss rates to estimated synonymous divergence.

Using the K_s inferences made above for each branch, we compared POInT's maximum likelihood estimate (MLE) of the rate of homoeolog loss (e.g., its estimated branch length, αt in Supplemental Figure 4) to each branch's mean synonymous divergence, $\overline{K_s}$, to see if the number of losses on any particular branch was unusually large or small. Estimating confidence intervals for these ratios of $\alpha \cdot t/\overline{K_s}$ is challenging. We treated the numerators and denominators of these ratios as being normally distributed and independent random variables. The maximum likelihood estimates (MLEs) of αt in the numerators should have asymptotically normal distributions with means that are equal to the true parameter values. The variances of these normal distributions were approximated by evaluating the inverse of the observed Fisher information (i.e., the Hessian of the negative log-likelihood; see Kendall and Stuart 1973). We estimated the observed Fisher information values via a single-dimension finite difference approximation that ignored covariances between the αt parameter and other parameters.

For each branch of the phylogeny, the K_s estimates that are in the denominator of the ratio $\alpha \cdot t/\overline{K_s}$ are obtained via a sample mean of the K_s estimates from the sequences of individual pillars (i.e., $\overline{K_s}$). Due to the Central Limit Theorem, this sample mean should be approximately normally distributed with mean equal to the true parameter value and with variance being approximately the sample variance among individual K_s estimates divided by the number of individual K_s estimates.

To infer confidence intervals for the ratio of $\alpha \cdot t/\overline{K_s}$ on each branch, we independently sampled from the aforementioned normal distributions that are used to approximate the uncertainty of αt and $\overline{K_s}$ estimates in the ratio. For each branch, we calculated the ratio of these sampled values for 1000 pairs of randomly sampled values. We then sorted the resulting ratios and set 95% confidence intervals by finding the ratio value that defined the lower and upper 2.5% of the sorted values.

Because the inclusion of fixation in our loss models can give rise to long tip branches (effectively the model suggests that all surviving duplicates in some genomes are now fixed), we present data using a model with convergent losses and biased fractionation but no fixation (WGD_{bc-nb}). However, our results are very similar with using the full WGD_{bfc-nb} model (Supplemental Figure 6).

Comparisons of selective constraint for different classes of polyploid loci

We examined the inferred average selective constraint (K_a/K_s, estimated as described above) for five classes of polyploid loci (e.g., pillars) across the seven WGD events: 1) Pillars that are single copy in all taxa and have a high probability of having returned to single-copy along the root branch, 2) Pillars that are completely single copy but where the genes did not return to single-copy on the root branch (e.g., where alternative copies of the duplicated genes are preserved in different genomes), 3) pillars with duplicates surviving in only a single species, 4) pillars where all but one species maintains the duplication and 5) pillars where all species maintain duplicate copies. Confidence intervals for these mean K_a/K_s estimates were estimated with the approach of described above.

Identifying reciprocal gene losses (RGLs) between polyploid taxa.

For a pair of single-copy genes from distinct genomes, the probability that these genes represent RGLs is simply the sum of the probabilities of the orthology relationships, estimated with POInT, that place them as paralogs rather than orthologs. We computed, for each pair of extant taxa in each polyploidy, the set of RGLs that we could identify with a confidence of ≥95% (Figure 4A). To avoid spurious inferences, we restricted our identification of RGL pairs to single-copy genes in each genome where either: a) both the gene and the "hole" corresponding to its lost homoeolog were in synteny with genes on either side or b) the single-copy gene in question was the only homolog of the outgroup gene used for the inference of the NCS blocks. In the first case, this filter corresponds to a clear absence of a corresponding homoeolog in the paralogous synteny block, in the second to the absence of a gene that could be the "missing" homoeolog. We then used TBLASTX (Altschul, et al. 1997) to search the non-coding regions of each genome for putative homoeologous copies of the inferred RGL gene that were missed in the genome annotations (e.g., the inference of RGL was spurious due to an annotation artifact). In Case "a" above, this search was restricted to the non-coding regions in the "hole" between the neighboring syntenic genes; in Case "b," we searched the entire genome for the potentially unannotated homoeolog. Only RGL genes with no such matching noncoding regions at an E-value cutoff of $\leq 10^{-10}$ were considered "true" RGLs. These secondary filters were not applied for the yeast WGD because those data were taken from the manually curated Yeast Genome Order Browser (YGOB, Byrne and Wolfe 2005).

Data on gene knockouts producing lethal phenotypes from zebrafish, *A. thaliana* and bakers' yeast were taken from ZFIN (Howe, et al. 2013; Conant 2020); a set of 510 "embryo-defective" genes identified by Meinke (2020); and Steinmetz et al., (2002), respectively. The proportion of RGLs in these "essential gene" lists was compared to the proportion of all other single-copy genes from the same organism in the list using Fisher's exact test (Sokal and Rohlf 1995). For these same three species, we used GeneOntology data (Gene Ontology Consortium 2015) and Panther Overrepresentation Tests (Release 20200728; Mi, et al. 2019) to ask if there were terms from the GO-Slim Biological Process, Cellular Compartment or Molecular Function ontologies that differed in their frequency between the RGL genes and other single-copy genes. After FDR correction (Benjamini and Hochberg 1995), no such terms were found for any of the three ontologies across any of the three genomes (FDR-corrected *P*-value > 0.05).

Data availability:

All underlying data are available from the POInT browser (wgd.statgen.ncsu.edu) and from figshare (DOI: https://doi.org/10.6084/m9.figshare.12750992.v4); the POInT package is available from GitHub (https://github.com/gconant0/POInT)

Acknowledgements:

We would like to thank K. Wolfe for helpful comments and K. Byrne for help with the YGOB datasets. YH, JCP and GCC were supported by National Science Foundation grant NSF-IOS-1339156. EL was supported by NSF-IOS-1339156 and NSF-IOS-1849708. JLT was supported by NSF-DEB-1754142 and by National Institutes of Health grant NIH-R01-GM118508.

Competing interests: The authors declare that they have no competing interests.

References:

Alix K, Gérard PR, Schwarzacher T, Heslop-Harrison J. 2017. Polyploidy and interspecific hybridization: partners for adaptation, speciation and evolution in plants. Annals of botany 120:183-194.

Allendorf FW, Bassham S, Cresko WA, Limborg MT, Seeb LW, Seeb JE. 2015. Effects of crossovers between homeologs on inheritance and population genomics in polyploid-derived salmonid fishes. Journal of Heredity 106:217-227.

Allendorf FW, Thorgaard GH. 1984. Tetraploidy and the evolution of salmonid fishes. In. Evolutionary genetics of fishes: Springer. p. 1-53.

Altschul SF, Madden TL, Schaffer AA, Zhang JH, Zhang Z, Miller W, Lipman DJ. 1997. Gapped Blast and Psi-Blast : A new-generation of protein database search programs. Nucleic acids research 25:3389-3402.

Barker MS, Arrigo N, Baniaga AE, Li Z, Levin DA. 2016. On the relative abundance of autopolyploids and allopolyploids. New Phytol 210:391-398.

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society, Series B (Methodological) 57:289-300.

Birchler JA, Riddle NC, Auger DL, Veitia RA. 2005. Dosage balance in gene regulation: biological implications. Trends Genet 21:219-226.

Birchler JA, Veitia RA. 2012. Gene balance hypothesis: connecting issues of dosage sensitivity across biological disciplines. Proc Natl Acad Sci U S A 109:14746-14753.

Birchler JA, Yao H, Chudalayandi S. 2006. Unraveling the genetic basis of hybrid vigor. Proceedings of the National Academy of Sciences 103:12957-12958.

Blanc-Mathieu R, Perfus-Barbeoch L, Aury J-M, Da Rocha M, Gouzy J, Sallet E, Martin-Jimenez C, Bailly-Bechet M, Castagnone-Sereno P, Flot J-F. 2017. Hybridization and polyploidy enable genomic plasticity without sex in the most devastating plant-parasitic nematodes. PLoS Genetics 13:e1006777.

Braasch I, Postlethwait JH. 2012. Polyploidy in fish and the teleost genome duplication. In. Polyploidy and genome evolution: Springer. p. 341-383.

Buggs RJ, Soltis PS, Soltis DE. 2009. Does hybridization between divergent progenitors drive whole-genome duplication? Molecular Ecology 18:3334-3339.

Byrne KP, Wolfe KH. 2005. The Yeast Gene Order Browser: Combining curated homology and syntenic context reveals gene fate in polyploid species. Genome research 15:1456-1461.

Chalhoub B, Denoeud F, Liu S, Parkin IA, Tang H, Wang X, Chiquet J, Belcram H, Tong C, Samans B. 2014. Early allopolyploid evolution in the post-Neolithic Brassica napus oilseed genome. Science 345:950-953.

Chen ZJ. 2010. Molecular mechanisms of polyploidy and hybrid vigor. Trends in plant science 15:57-71.

Clausen R, Goodspeed T. 1925. Interspecific hybridization in Nicotiana. II. A tetraploid glutinosa-tabacum hybrid, an experimental verification of Winge's hypothesis. Genetics 10:278.

Conant GC. 2020. The lasting after-effects of an ancient polyploidy on the genomes of teleosts. PloS ONE 15:e0231356.

Conant GC, Birchler JA, Pires JC. 2014. Dosage, duplication, and diploidization: clarifying the interplay of multiple models for duplicate gene evolution over time. Current Opinion in Plant Biology 19:91-98.

Conant GC, Wagner A. 2002. GenomeHistory: A software tool and its application to fully sequenced genomes. Nucleic acids research 30:3378-3386.

Conant GC, Wolfe KH. 2008. Probabilistic cross-species inference of orthologous genomic regions created by whole-genome duplication in yeast. Genetics 179:1681-1692.

Crespi BJ, Fulton MJ. 2004. Molecular systematics of Salmonidae: combined nuclear data yields a robust phylogeny. Molecular phylogenetics and evolution 31:658-679.

De Smet R, Adams KL, Vandepoele K, Van Montagu MC, Maere S, Van de Peer Y. 2013. Convergent gene loss following gene and genome duplications creates single-copy families in flowering plants. Proceedings of the National Academy of Sciences, U.S.A. 110:2898-2903.

Doyle JJ, Flagel LE, Paterson AH, Rapp RA, Soltis DE, Soltis PS, Wendel JF. 2008. Evolutionary genetics of genome merger and doubling in plants. Annual review of genetics 42:443-461.

Edger PP, Pires JC. 2009. Gene and genome duplications: the impact of dosage-sensitivity on the fate of nuclear genes. Chromosome Research 17:699-717.

Emery M, Willis MMS, Hao Y, Barry K, Oakgrove K, Peng Y, Schmutz J, Lyons E, Pires JC, Edger PP, 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 Genetics 14:e1007267em.

Evangelisti AM, Conant GC. 2010. Nonrandom survival of gene conversions among yeast ribosomal proteins duplicated through genome doubling. Genome biology and evolution 2:826-834.

Fawcett JA, Maere S, Van de Peer Y. 2009. Plants with double genomes might have had a better chance to survive the Cretaceous-Tertiary extinction event. Proc Natl Acad Sci U S A 106:5737-5742.

Felsenstein J. 1985. Phylogenies and the comparative method. American Naturalist:1-15.

Gaeta RT, Chris Pires J. 2010. Homoeologous recombination in allopolyploids: the polyploid ratchet. New Phytologist 186:18-28.

Gaut BS, Doebley JF. 1997. DNA sequence evidence for the segmental allotetraploid origin of maize. Proceedings of the National Academy of Sciences 94:6809-6814.

Gene Ontology Consortium. 2015. Gene ontology consortium: going forward. Nucleic acids research 43:D1049-D1056.

Goldman N, Yang Z. 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. Molecular biology and evolution 11:725-736.

Howe DG, Bradford YM, Conlin T, Eagle AE, Fashena D, Frazer K, Knight J, Mani P, Martin R, Moxon SA, et al. 2013. ZFIN, the Zebrafish Model Organism Database: increased support for mutants and transgenics. Nucleic acids research 41:D854-860.

Huang C-H, Sun R, Hu Y, Zeng L, Zhang N, Cai L, Zhang Q, Koch MA, Al-Shehbaz I, Edger PP. 2016. Resolution of Brassicaceae phylogeny using nuclear genes uncovers nested radiations and supports convergent morphological evolution. Molecular biology and evolution 33:394-412.

Kendall M, Stuart A. 1973. The advanced theory of statistics. London: Charles Griffen.

Kirkpatrick S, Gelatt CDJ, Vecchi MP. 1983. Optimization by simulated annealing. Science 220:671-680.

Kurtzman CP, Robnett CJ. 2003. Phylogenetic relationships among yeasts of the '*Saccharomyces* complex' determined from multigene sequence analyses. FEMS Yeast Research 3:417-432.

Kuwada Y. 1911. Maiosis in the Pollen Mother Cells of Zea Mays L.(With Plate V.). 植物学雑誌 25:163-181.

Lander ES, Green P. 1987. Construction of multilocus genetic linkage maps in humans. Proceedings of the National Academy of Sciences, U.S.A. 84:2363-2367.

Li W-H. 1980. Rate of gene silencing at duplicate loci: A theoretical study and interpretation of data from tetraploid fish. Genetics 95:237-258.

Lynch M, Conery JS. 2000. The evolutionary fate and consequences of duplicate genes. Science 290:1151-1155.

Maclean CJ, Greig D. 2011. Reciprocal gene loss following experimental whole-genome duplication causes reproductive isolation in yeast. Evolution: International Journal of Organic Evolution 65:932-945.

Mayrose I, Zhan SH, Rothfels CJ, Magnuson-Ford K, Barker MS, Rieseberg LH, Otto SP. 2011. Recently formed polyploid plants diversify at lower rates. Science 333:1257.

Meinke DW. 2020. Genome-wide identification of EMBRYO-DEFECTIVE (EMB) genes required for growth and development in Arabidopsis. New Phytologist 226:306-325.

Mi H, Muruganujan A, Ebert D, Huang X, Thomas PD. 2019. PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. Nucleic Acids Res 47:D419-D426.

Mizuta Y, Harushima Y, Kurata N. 2010. Rice pollen hybrid incompatibility caused by reciprocal gene loss of duplicated genes. Proceedings of the National Academy of Sciences 107:20417-20422.

Muir CD, Hahn MW. 2015. The limited contribution of reciprocal gene loss to increased speciation rates following whole-genome duplication. The American Naturalist 185:70-86.

Near TJ, Eytan RI, Dornburg A, Kuhn KL, Moore JA, Davis MP, Wainwright PC, Friedman M, Smith WL. 2012. Resolution of ray-finned fish phylogeny and timing of diversification. Proceedings of the National Academy of Sciences, U.S.A. 109:13698-13703.

Notredame C, Higgins DG, Heringa J. 2000. T-Coffee: A novel method for fast and accurate multiple sequence alignment. Journal of molecular biology 302:205-217.

Ohno S. 1970. Evolution by gene duplication. New York: Springer.

Pires JC, Conant GC. 2016. Robust Yet Fragile: Expression Noise, Protein Misfolding and Gene Dosage in the Evolution of Genomes. Annual review of genetics 50:113–131.

Sankoff D, Blanchette M. 1998. Multiple genome rearrangement and breakpoint phylogeny. Journal of Computational Biology 5:555-570.

Scannell DR, Byrne KP, Gordon JL, Wong S, Wolfe KH. 2006. Multiple rounds of speciation associated with reciprocal gene loss in polyploid yeasts. Nature 440:341-345.

Scannell DR, Frank AC, Conant GC, Byrne KP, Woolfit M, Wolfe KH. 2007. Independent sorting-out of thousands of duplicated gene pairs in two yeast species descended from a whole-genome duplication. Proceedings of the National Academy of Sciences, U.S.A. 104:8397-8402.

Schnable JC, Freeling M, Lyons E. 2012. Genome-wide analysis of syntenic gene deletion in the grasses. Genome Biol Evol 4:265-277.

Schoonmaker A, Hao Y, Bird D, Conant GC. 2020. A single, shared triploidy in three species of parasitic nematodes. G3: Genes, Genomes, Genetics 10:225-233.

Schranz ME, Mohammadin S, Edger PP. 2012. Ancient whole genome duplications, novelty and diversification: the WGD Radiation Lag-Time Model. Current Opinion in Plant Biology 15:147-153.

Schwarz G. 1978. Estimating the dimension of a model. Annals of statistics 6:461-464.

Scienski K, Fay JC, Conant GC. 2015. Patterns of Gene Conversion in Duplicated Yeast Histones Suggest Strong Selection on a Coadapted Macromolecular Complex. Genome biology and evolution 7:3249-3258.

Scrucca L, Fop M, Murphy TB, Raftery AE. 2016. mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. The R journal 8:289.

Sokal RR, Rohlf FJ. 1995. Biometry: 3rd Edition. New York: W. H. Freeman and Company.

Soltis DE, Albert VA, Leebens-Mack J, Bell CD, Paterson AH, Zheng C, Sankoff D, dePamphilis CW, Wall PK, Soltis PS. 2009. Polyploidy and angiosperm diversification. American Journal of Botany 96:336-348.

Soltis DE, Segovia-Salcedo MC, Jordon-Thaden I, Majure L, Miles NM, Mavrodiev EV, Mei W, Cortez MB, Soltis PS, Gitzendanner MA. 2014. Are polyploids really evolutionary dead-ends (again)? A critical reappraisal of Mayrose et al.(2011). New Phytologist 202:1105-1117.

Soltis DE, Visger CJ, Soltis PS. 2014. The polyploidy revolution then… and now: Stebbins revisited. American Journal of Botany 101:1057-1078.

Stebbins Jr GL. 1947. Types of polyploids: their classification and significance. In. Advances in genetics: Elsevier. p. 403-429.

Steinmetz LM, Scharfe C, Deutschbauer AM, Mokranjac D, Herman ZS, Jones T, Chu AM, Giaever G, Prokisch H, Oefner PJ, et al. 2002. Systematic screen for human disease genes in yeast. Nature genetics 31:400-404.

Tang H, Woodhouse MR, Cheng F, Schnable JC, Pedersen BS, Conant G, Wang X, Freeling M, Pires JC. 2012. Altered patterns of fractionation and exon deletions in Brassica rapa support a two-step model of paleohexaploidy. Genetics 190:1563-1574.

Tasdighian S, Van Bel M, Li Z, Van de Peer Y, Carretero-Paulet L, Maere S. 2017. Reciprocally retained genes in the angiosperm lineage show the hallmarks of dosage balance sensitivity. The Plant Cell 29:2766-2785.

Thomas BC, Pedersen B, Freeling M. 2006. Following tetraploidy in an Arabidopsis ancestor, genes were removed preferentially from one homeolog leaving clusters enriched in dose-sensitive genes. Genome research 16:934-946.

Tsai IJ, Bensasson D, Burt A, Koufopanou V. 2008. Population genomics of the wild yeast *Saccharomyces paradoxus*: Quantifying the life cycle. Proceedings of the National Academy of Sciences, U.S.A. 105:4957-4962.

Van de Peer Y, Mizrachi E, Marchal K. 2017. The evolutionary significance of polyploidy. Nature Reviews Genetics 18:411-424.

Wagner Jr W. 1970. Biosystematics and evolutionary noise. Taxon 19:146-151.

Werth CR, Windham MD. 1991. A model for divergent, allopatric speciation of polyploid pteridophytes resulting from silencing of duplicate-gene expression. The American Naturalist 137:515-526.

Wolfe KH. 2001. Yesterday's polyploids and the mystery of diploidization. Nat Rev Genet 2:333-341.

Zhang Z, Gou X, Xun H, Bian Y, Ma X, Li J, Li N, Gong L, Feldman M, Liu B. 2020. Homoeologous exchanges occur through intragenic recombination generating novel transcripts and proteins in wheat and other polyploids. Proceedings of the National Academy of Sciences.