Maintenance and Loss of Duplicated Genes by Dosage Subfunctionalization

Jean-Francois Gout*[,1] and Michael Lynch[1]
[1]Department of Biology, Indiana University, Bloomington
***Corresponding author:** E-mail: jgout@indiana.edu.
Associate editor: Jianzhi Zhang

Abstract

Whole-genome duplications (WGDs) have contributed to gene-repertoire enrichment in many eukaryotic lineages. However, most duplicated genes are eventually lost and it is still unclear why some duplicated genes are evolutionary successful whereas others quickly turn to pseudogenes. Here, we show that dosage constraints are major factors opposing post-WGD gene loss in several *Paramecium* species that share a common ancestral WGD. We propose a model where a majority of WGD-derived duplicates preserve their ancestral function and are retained to produce enough of the proteins performing this same ancestral function. Under this model, the expression level of individual duplicated genes can evolve neutrally as long as they maintain a roughly constant summed expression, and this allows random genetic drift toward uneven contributions of the two copies to total expression. Our analysis suggests that once a high level of imbalance is reached, which can require substantial lengths of time, the copy with the lowest expression level contributes a small enough fraction of the total expression that selection no longer opposes its loss. Extension of our analysis to yeast species sharing a common ancestral WGD yields similar results, suggesting that duplicated-gene retention for dosage constraints followed by divergence in expression level and eventual deterministic gene loss might be a universal feature of post-WGD evolution.

Key words: **duplication, expression level, subfunctionalization, *Paramecium*, whole-genome duplication, pseudogenization.**

Introduction

In all three domains of life, a substantial fraction of genes belong to multicopy gene families (Zhang 2003; Lynch 2007). Although gene duplication has long been seen as a major source of novelty (Ohno 1970), the fate of most duplicated genes is eventual loss (Lynch and Conery 2000), although it remains unclear why some duplicates are evolutionarily successful whereas others suffer elimination after degradation to pseudogenes. Several models have been proposed to explain the retention of duplicated genes. The neofunctionalization (Ohno 1970) and subfunctionalization (Force et al. 1999) (also called DDC for Duplication–Degeneration–Complementation) models require acquisition of new function or partitioning of ancestral functions (either at the level of the biochemical function of the encoded proteins or at the level of the expression pattern) for duplicated genes to be retained. On the other hand, absolute dosage and dosage-balance constraints have also been proposed to explain retention of duplicated genes without change of function (Papp et al. 2003; Gout et al. 2009, 2010; Qian et al. 2010; Birchler and Veitia 2012). These different models are not mutually exclusive (e.g., it has been suggested that dosage constraints allow genes to be retained for long enough before changes of function occur [Force et al. 1999; Rastogi and Liberles 2005]), but the relative contributions of each mechanism to duplicate gene retention remain unclear.

The observation that the fate of duplicated genes depends in part on whether they originate from small-scale duplications (SSD) or whole-genome duplication (WGD) (Davis and Petrov 2005; Maere et al. 2005; Hakes et al. 2007) may reveal the importance of dosage constraints on long-term duplicate-gene retention. Although the fixation of an SSD results in a modification of dosage, the opposite is true for WGDs, in which case the relative expression levels of all genes are initially preserved and subsequent gene losses cause dosage disruptions. Although far less frequent than SSDs, WGDs have been recurrent in the history of eukaryotes, with two rounds of WGDs thought to have arisen at the base of all vertebrates (Panopoulou and Poustka 2005; Hughes and Liberles 2008; Putnam et al. 2008; Decatur et al. 2013), and clear examples of ancient polyploidy in lineages leading to the yeast (Wolfe and Shields 1997), the African clawed frog (Morin et al. 2006), teleost fish (Postlethwait et al. 2000; Jaillon et al. 2004), flowering plants (Simillion et al. 2002; Doyle et al. 2008; Jiao et al. 2011), and the ciliated protozoan *Paramecium* (Aury et al. 2006; McGrath, Gout, Doak, et al. 2014; McGrath, Gout, Johri, et al. 2014).

One of the most valuable resources currently available for studying post-WGD evolution is provided by yeast, where whole-genome sequences are available for 12 species (including the highly studied *Saccharomyces cerevisiae*) sharing a common ancestral WGD as well as several outgroup species that have diverged before the WGD (Byrne and Wolfe 2005). More recently, the discovery of at least three successive rounds of WGDs in the lineage leading to *Paramecium tetraurelia* (Aury et al. 2006) has established *Paramecium* as

Article

another important model organism for studying post-WGD genomes evolution. *Paramecium tetraurelia* belongs to a group of at least 14 sibling species (*P. aurelia*) that are so similar in morphology that they were initially believed to be only one species (Sonneborn 1975). Given the ancient points of WGD in this lineage (Aury et al. 2006; McGrath, Gout, Doak, et al. 2014; McGrath, Gout, Johri, et al. 2014), this species complex provides a distinct contrast to the 2R hypothesis, which generally postulates that two rounds of genome duplication played a causal role in the morphological diversification of the vertebrates (Holland et al. 1994; Meyer and Van de Peer 2005; Freeling and Thomas 2006).

The sequencing and analysis of the *P. biaurelia* and *P. sexaurelia* genomes confirmed that the most recent WGD is basal to the *P. aurelia* group (McGrath, Gout, Johri, et al. 2014), and the *P. caudatum* genome allows us to analyze a species that diverged from the *P. aurelia* lineage before the two most recent WGDs (McGrath, Gout, Doak, et al. 2014). The average level of retention of WGD-derived paralogs (hereafter named ohnologs) is higher in *P. aurelia* (40–50% for the most recent WGD; McGrath, Gout, Johri, et al. 2014) than in yeast (8–14%; Wolfe and Shields 1997; Scannell et al. 2007). Therefore, *P. aurelia* offers a unique view of the evolutionary processes acting in earlier stages of the post-WGD gene loss process compared with yeast. However, this notion of "earlier stages" is relative, as the most recent *P. aurelia* WGD is estimated to be approximately 320 My old (McGrath, Gout, Johri, et al. 2014). Despite an extremely low mutation rate (Sung et al. 2012), the ohnologs in this lineage have had enough time to have incurred approximately 1.7 mutations per nucleotide site (dS) (McGrath, Gout, Johri, et al. 2014), which implies considerable intensity of selection for the conservation of the protein sequences of duplicate genes (a median level of ~0.05 nonsynonymous substitutions [d*N*] between ohnologs from the most recent WGD, with most ohnologs having d*N*/d*S* values $\ll 1$ [McGrath, Gout, Johri, et al. 2014]). In contrast to this overall pattern of strong purifying selection, the presence of numerous recent pseudogenes (Aury et al. 2006) and the observation that over a hundred gene losses occurred since the split of the two closely related *P. biaurelia* and *P. tetraurelia* species (McGrath, Gout, Johri, et al. 2014) indicate that duplicated-gene loss is still ongoing in the *P. aurelia* lineage. Therefore, it may seem paradoxical that purifying selection maintained duplicated genes during millions of years only to eventually allow their loss. Thanks to its exceptional characteristics, the *P. aurelia* species provide a unique opportunity for evaluating the mechanisms responsible for the differential on-going retention of gene duplicates over a considerable period of evolutionary time.

Here, we investigate the evolution of expression level between duplicated genes derived from the most recent *P. aurelia* WGD and explore the consequences of such changes for the fate of paralogous genes. We propose a refined version of the DDC model (Force et al. 1999; Qian et al. 2010) in which duplicated genes are initially retained by the partitioning of the dosage requirements between the two copies, with gradual stochastic changes in expression level between the two copies leading to an eventual imbalance in the selective pressures operating on the two copies, ultimately leading to the loss of the copy having the lowest contribution to total expression level.

Results

Expression Levels Are Conserved between Paralogs Originating from the Most Recent *Paramecium* WGD

Because both coding and regulatory regions and all *trans*-acting factors are duplicated together in the case of a WGD, it is likely that ohnologs (WGD-derived paralogs) initially share the exact same expression patterns. One possible exception is the case of allopolyploidization, which involves the conjoining of two different genomes, where ohnologs can show divergent expression patterns immediately after the WGD (Adams 2007). However, in the case of *P. aurelia*, it seems likely that the most recent WGD involved autopolyploidization (McGrath, Gout, Johri, et al. 2014), with pairs of ohnologs being born identical in all respects, an assumption that we will adhere to. Given previous analyses showing that the coding regions of ohnologous protein-coding genes in *Paramecium* evolve mostly under purifying selection (Aury et al. 2006; McGrath, Gout, Doak, et al. 2014; McGrath, Gout, Johri, et al. 2014), we reasoned that purifying selection likely constrains changes in expression levels as well, and that this should be reflected in restricted divergence in the transcript abundance between ohnologs. Recalling the high average level of synonymous substitution between ohnologs (McGrath, Gout, Johri, et al. 2014), regulatory regions and their resultant effects on expression levels should be highly divergent between ohnologs in the absence of purifying selection. Contrary to this null hypothesis, we found that expression levels are still significantly correlated between ohnologs within all three species (*P. biaurelia*: $r = 0.68$, *P. tetraurelia*: $r = 0.64$, and *P. sexaurelia*: $r = 0.68$, all $P < 0.0001$; fig. 1 and supplementary fig. S1, Supplementary Material online). This indicates that selection operates to prevent changes in expression level, confirming the importance of dosage in post-WGD paralog evolution.

Is Conservation of Expression Indicative of Conservation of Biochemical Function?

We reasoned that the strong conservation of expression level between ohnologs observed in *Paramecium* might indicate an overall conservation of their biochemical function. To test this hypothesis, we compared Gene Ontology (GO) annotations between ohnologs and found that 87–88% of ohnologs pairs in the three *P. aurelia* species analyzed share the same PANTHER (Mi et al. 2012) annotation. Interestingly, we observed that, in all three species, pairs of ohnologs with divergent PANTHER annotations have—on average—less conserved expression levels than those with identical PANTHER annotations ($P < 0.01$ in all three species, Mann–Whitney *U* test; supplementary fig. S2 and table S1,

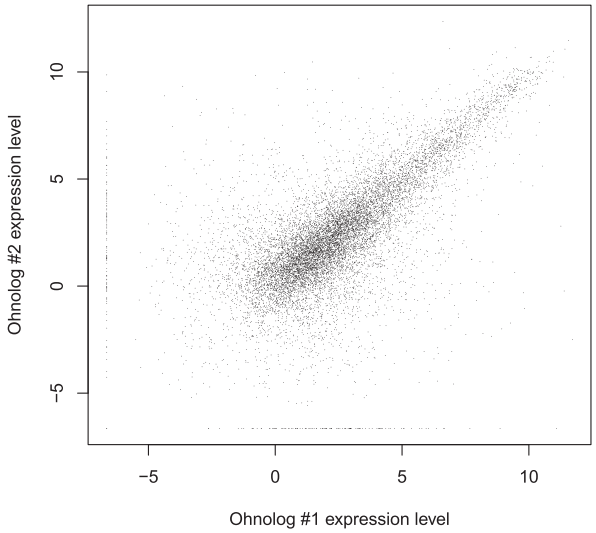

Fig. 1. Correlation of absolute expression level between ohnologs from the most recent WGD in *Paramecium biaurelia*. Expression level (log_2-transformed FPKM values, see Materials and Methods) in pairs of ohnologs in *P. biaurelia*. Pairs of ohnologs show a strong significant correlation of expression level ($r = 0.80$, $P < 0.0001$). Note that negative values result from the log-transformation of FPKM values lower than 1.

Supplementary Material online). Additional support for an association between conservation of function and conservation of expression level comes from the observation of a significant positive correlation between expression level divergence and nonsynonymous substitution level (d*N*) between ohnologs in *Paramecium* ($r = 0.30$, $r = 0.36$ and $r = 0.33$ for *P. tetraurelia*, *P. biaurelia* and *P. sexaurelia*, respectively, all $P < 0.001$). Similar correlations have been reported for ohnologs in *Arabidopsis* (Blanc and Wolfe 2004). However, in yeast, the correlation between expression divergence and coding sequence divergence is weak (Wagner 2000), possibly because ohnologs are more ancient. We investigated the correlation between biochemical function divergence and expression level divergence in two yeast species for which both GO annotation and expression data are available (*S. cerevisiae* and *Candida glabrata*). In both species, we found that ohnologs having at least one GO annotation difference show higher divergence of expression level than ohnologs sharing identical GO annotations (mean of absolute expression level difference: 1.39 vs. 1.73 log2-FPKM [fragment per kilobase per million reads], $P < 0.01$ for *S. cerevisiae* and 1.23 vs. 1.80, $P < 0.01$ for *C. glabrata*, see Materials and Methods). Although these observations do not allow us to draw conclusions on a gene-by-gene basis, they indicate that, on average, divergence of function is associated with divergence of expression in both yeast and *Paramecium*. Therefore, we interpret the strong conservation in expression level between *Paramecium* ohnologs as evidence against pervasive changes in the biochemical function in pairs of ohnologs.

Changes in Expression Levels Are Associated with Gene Loss

Although there is a clear, general trend in *Paramecium* for the maintenance of similar levels of transcript abundance between ohnologs, a number of gene pairs escape this pattern and show significantly different expression levels between the two ohnologs. We found 1,239, 1,177, and 926 pairs of duplicated genes in *P. biaurelia*, *P. tetraurelia*, and *P. sexaurelia*, respectively, with an over 10-fold difference in transcript abundance (measured as FPKM, see Materials and Methods) between the two ohnologs. We reasoned that changes in expression level could be tolerated as long as the total expression level (sum of the expression from both ohnologs) remains the same. This would allow the expression level of ohnologs to drift along a line of fixed total expression level (fig. 2, colored lines). Drift along such a line of equivalence can then lead to a situation in which one copy has a low enough expression level that total inactivation of this copy is effectively neutral (gray area in fig. 2).

To test this hypothesis, we investigated the relationship between the magnitude of expression differences within pairs of ohnologs in a given species and the probability of retention of the orthologous genes in the sister *P. aurelia* species. When both ohnologs have been retained and have similar expression levels in *P. biaurelia* (25% least divergent pairs, measured as absolute difference in log2-transformed FPKM value), the probability that one of the two orthologous ohnologs has been lost in *P. tetraurelia* is 7%. However, when the two *P. biaurelia* paralogs diverge in expression level (top 25% most divergent pairs, measured as absolute difference in log2-transformed FPKM value), the probability that one of the two orthologous ohnologs has been lost in *P. tetraurelia* rises to 16% ($P < 0.001$, Fisher's exact test). Moreover, the copy that is lost in *P. tetraurelia* is orthologous to the *P. biaurelia* copy with the lowest expression level in 81% of the cases, which is significantly more than the random expectation of 50% ($P < 0.001$, exact two-sided binomial test). We found similar results when investigating retention rates in *P. biaurelia* as a function of expression divergence in *P. tetraurelia* (table 1). When comparing *P. tetraurelia* and *P. biaurelia* to the more distantly related *P. sexaurelia* (see fig. 1 from McGrath, Gout, Johri, et al. [2014] for a phylogenetic tree displaying the relationship between *P. caudatum*, *P. biaurelia*, *P. tetraurelia*, and *P. sexaurelia*), we also found that expression divergence was associated with increased probability of gene loss in the sister species (table 1). However, in the comparisons involving *P. sexaurelia*, the fraction of cases for which the lost copy is orthologous to the one with the lowest expression was not always significantly different from 0.5 (table 1). This observation suggests that, although the fate of ohnologous pairs (whether both copies will be retained or not) was largely determined at the time of the WGD, the trajectory of the two copies within each pair (i.e., which copy will be retained and which one will be lost) was not yet determined at the time of the speciation between the *P. sexaurelia* and the *P. tetraurelia*/*P. biaurelia* lineages. It is also possible that difficulties in distinguishing orthologs from paralogs—because *P. sexaurelia* diverged from the lineage leading to *P. tetraurelia* and *P. biaurelia* very early after the WGD (McGrath, Gout, Johri, et al. 2014)—weaken the ability to ascertain the original bias. We obtained similar results when using the pairs of genes with the top 5% and bottom 5% most conserved expression

levels (supplementary table S2, Supplementary Material online). To avoid setting arbitrary limits when classifying pairs as divergent or similar in expression level, we also used the reverse approach: Comparing expression divergence of ohnologs whose orthologs have been lost or retained. We found that the average difference in expression-level difference between ohnologs is on average approximately 40% higher for pairs whose orthologous pair has lost one copy in the sister *P. aurelia* species, compared with those where the two copies have been retained (supplementary table S3, Supplementary Material online).

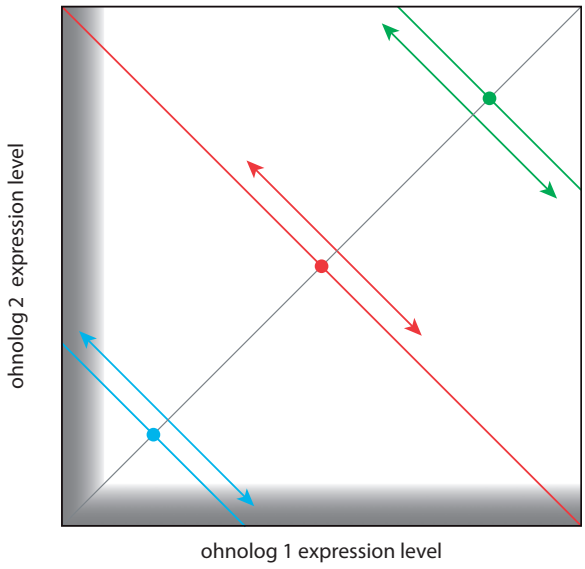

Fig. 2. A model of evolution by conservation of total expression. The central gray line corresponds to the hypothetical distribution of expression levels immediately following the WGD: Identical expression level between the two ohnologs for all pairs of genes. The colored lines correspond to trajectories of conserved total expression level (i.e., the sum of expression level from the two ohnologs remains constant). Pairs of genes can follow a random walk along these lines of conserved total expression until they reach a region where loss of the lowly expressed copy is effectively neutral (gray areas).

One prediction of our model is that, for ohnologs whose expression level drifted along lines of constant total expression level before loss of the lowly expressed copy, the remaining copy should have increased expression compared with its ancestral level immediately after the WGD. We used expression level of orthologous genes in *P. caudatum*, a species that diverged before the two most recent *Paramecium* WGDs as a rough proxy for ancestral (immediately after-WGD) relative expression level of genes in *P. aurelia* species. Briefly, we normalized the expression levels between the two species by either using the relative ranking of genes in each species or adjusting the distribution of log2-transformed FPKM values so that they have the same empirical distribution in both species (Materials and Methods). We found that the fraction of genes whose expression level has increased since the split with *P. caudatum* is higher among singletons than among pairs of conserved ohnologs ($P < 0.001$ in all three *P. aurelia* species) when using either a relative ranking (supplementary table S4, Supplementary Material online) or a quantile normalization (supplementary table S5, Supplementary Material online) method to perform the between-species comparisons (see Materials and Methods), supporting the idea that increased expression level of the copy that would eventually be retained compensated for decreased expression level of the copy that would eventually be lost.

A Similar Pattern in Yeast

We took advantage of publically available data for 12 yeast species that share an ancestral WGD with the model organism *S. cerevisiae*. Although it has been previously reported in yeast that post-WGD retained paralogs are biased toward highly expressed genes (Seoighe and Wolfe 1999), it is not clear whether the higher expression is a cause or a consequence of the retention. To answer this question, we used expression data from *Kluyveromyces lactis*, a species that diverged from *S. cerevisiae* before the WGD. In all 12 yeast

Table 1. Gene Loss Rates Depending on Expression Conservation of the Orthologous Pair in the Sister *P. aurelia* Species.

		Gene Loss Rate Depending on Expression Conservation of Orthologous Pair				
Expression Difference in	Gene Loss Rate in	Conserved Expression Level	Divergent Expression Level	P Value	Fraction of Gene Loss Corresponding to the Copy with the Lowest Expression Level	P value
Paramecium biaurelia	*Paramecium tetraurelia*	0.07	0.16	<0.001	0.81	<0.001
Paramecium biaurelia	*Paramecium sexaurelia*	0.30	0.44	<0.001	0.53	0.08
Paramecium tetraurelia	*Paramecium biaurelia*	0.08	0.18	<0.001	0.77	<0.001
Paramecium tetraurelia	*Paramecium sexaurelia*	0.29	0.46	<0.001	0.54	0.02
Paramecium sexaurelia	*Paramecium biaurelia*	0.15	0.25	<0.001	0.50	0.92
Paramecium sexaurelia	*Paramecium tetraurelia*	0.13	0.24	<0.001	0.54	0.12

Note.—Genes whose orthologs have conserved expression levels (top 25% more conserved pairs of ohnologs) in the sister *P. aurelia* species are more likely to have been retained in two copies following the recent WGD than those whose orthologs have divergent expression levels (top 25% most divergent pairs of ohnologs). Significant difference between the two fractions is tested by a Fisher's exact test (column 5). When one copy was lost following the WGD, the lost copy tends to be orthologous to the copy with the lowest expression level in the sister *P. aurelia* species (column 6). Significant deviation from random expectation of 50% of gene loss corresponding to the copy with the lowest expression level is tested by an exact binomial test (column 7).

species investigated, we found a significant positive correlation between the expression level of genes in *K. lactis* and the probability that their orthologs are retained in two copies following WGD (supplementary fig. S4, Supplementary Material online). This observation indicates that pre-WGD expression level influences the probability of post-WGD gene retention in yeast, similarly to what we have previously observed in *Paramecium* (McGrath, Gout, Doak, et al. 2014). Consistent with the more advanced level of ohnolog loss from the yeast WGD (Wolfe and Shields 1997; Aury et al. 2006; Scannell et al. 2007; McGrath, Gout, Johri, et al. 2014), we found that, although yeast ohnologs show significant conservation of expression level ($r = 0.38$, $P < 0.001$), the correlation is weaker than that of any *P. aurelia* species analyzed ($P < 0.001$ for all three comparisons, Fisher's z test). Despite this lower conservation of expression levels, the significant positive correlation between ohnologs indicates that, as observed in *Paramecium*, purifying selection has been operating to limit the divergence of expression level between ohnologs since the WGD.

We then asked whether changes in expression level between ohnologs are associated with increased probability of gene loss in yeast, as in *Paramecium*. We found that pairs of ohnologs in *S. cerevisiae* for which one orthologous copy was lost in at least one of the 11 other yeast species have an average difference in transcript abundance about 50% higher than pairs for which no copy was lost in the other yeast species (average difference in expression level: 1.63 vs. 1.11, $P = 0.01$, Man–Whitney *U* test). We found similar results when using expression data from two other yeast species (*Naumovozyma castellii* and *C. glabrata*, see Materials and Methods, table 2, and supplementary fig. S5, Supplementary Material online). These observations suggest that similar evolutionary forces tie the fate of duplicated genes to the evolution of their expression level in both yeast and *Paramecium*.

Table 2. Average Expression Level Difference between Ohnologs in Three Yeast Species for Pairs that Have Been Retained in All 12 Yeast Species or Have Lost One Copy in At Least One Species.

	Average Expression Level Difference between Ohnologs If the Orthologous Pair Was Lost in		
Species	None of the Other Yeasts	At Least One of the Other Yeasts	P value
Saccharomyces cerevisiae	1.11	1.63	0.01
Naumovozyma castellii	1.07	1.98	<0.001
Candida Glabrata	1.23	1.85	0.02

NOTE.—In all three species tested, pairs of genes for which one copy has been lost in at least one of the 12 yeast species studied show an average higher divergence of expression level than those for which both copies have been retained in all 12 yeast species. Statistical signification of the difference between the two means was tested by a Mann–Whitney *U* test.

Discussion

Dosage Constraints Limit Gene Expression Divergence Following WGD

In this study, we have observed significant correlations between the expression levels of duplicated genes originating from the most recent *Paramecium* WGD (ohnologs) and interpret this observation as evidence for pervasive purifying selection on the total expression level of pairs of ohnologs. Our interpretation is based on the assumptions that ohnologs initially (i.e., immediately after the WGD) shared identical expression levels and that the *Paramecium* WGD is old enough that expression levels between ohnologs would now be totally uncoupled in the absence of selective pressures. Because the average dS between ohnologs is greater than 1, it seems reasonable to assume that, in the absence of purifying selection, all regulatory signals would have been inactivated in at least one copy of each pair of genes and as a consequence, the expression levels between ohnologs would no longer be correlated. Although the correlations are clear and highly significant in all three *P. aurelia* species investigated, there are still a number of ohnologs with significantly different expression levels. Although it is possible that some of the observed differences are consequences of changes in function that have been promoted by selection, it seems more likely that they result from changes in expression level being effectively neutral (see below).

Preservation of Duplicated Genes in the Absence of Functional Changes

Our analysis of functional annotations suggests that almost 90% of ohnologs have retained the same biochemical function since the most recent *Paramecium* WGD. Although it is possible that bioinformatics predictions of biochemical functions miss subtle differences between ohnologs, this observation, coupled to the strong conservation of expression levels between ohnologs, suggests that most ohnologs are redundant in function, raising the question of the evolutionary mechanisms responsible for their retention.

A model for the maintenance of duplicated genes and their functional redundancy by reduced expression has been proposed recently by Qian et al. (2010). Under this model, the main selective constraint is on the total amount of the final product (i.e., protein) that is produced by the joint transcription of two duplicated genes. If gene duplication is followed by a reduction in absolute expression level of each copy, then losing one copy could result in an insufficient amount of the final product, which would be deleterious and therefore selected against. In the specific case of a WGD, it is apparent that this step of postduplication reduction in absolute expression level is unnecessary. Indeed, as all of the genes are duplicated together at the same time, it is likely that ohnologs initially are close to their optimal expression level and further reduction in final product expression through gene loss will be immediately detrimental (Gout et al. 2010). This is likely to be especially pertinent in the case of *Paramecium* where it is possible that, because of the nuclear dimorphism harbored by ciliates, the WGD did not result in an immediate change of mRNA concentrations in the post-WGD cells (McGrath, Gout, Doak, et al. 2014).

The Absolute-Dosage Subfunctionalization Model Allows for Extended Conservation of Duplicated Genes Followed by Gene Loss

With a majority of duplicated genes having been retained for hundreds of millions of years after the most recent *Paramecium* WGD, one can wonder why many of these genes are still being lost. What are the mechanisms capable of lifting the selective pressures that have been opposing gene loss for millions of years? Our analysis suggests that duplicated genes are initially retained because of dosage constraints (losing one copy would result in a deleterious decrease in the total amount of protein produced for a given pair of ohnologs). Our proposal that effectively neutral changes in expression level occur when the joint expression level of two ohnologs follows a random walk along a line of fixed total expression level implicitly assumes that individual genes can experience both increases and decreases in expression level. Support for this model derives from our observation that genes that have lost their ohnolog do indeed show a tendency for increased expression level since the WGD, ostensibly because of the need to compensate for the decreased expression level of the copy that is eventually lost.

We note that under this model the distance that must be traversed to enter the region where losing the lowly expressed copy becomes effectively neutral increases with the initial expression level of the pair of ohnologs (fig. 2). Although the exact shape of this region of effectively neutral gene loss is unknown, a reduction in the probability of gene loss with increasing initial expression level has been widely observed (Seoighe and Wolfe 1999; Aury et al. 2006; Gout et al. 2010; McGrath, Gout, Doak, et al. 2014; McGrath, Gout, Johri, et al. 2014). An alternative hypothesis to the walk along lines of fixed cumulative expression level is that one copy fixes by chance a succession of mutations reducing its expression in amounts small enough that each step behaves in an effectively neutral manner, slowly moving the pair away from the line of fixed total expression level and setting this copy on a trajectory for eventual gene loss. If one copy is more prone to mutations than the other, this process could result in driving the pair of ohnologs away from the line of constant total expression level. A recent study shows that most mutations in the promoter region of the yeast gene TDH3 change expression level by less than 10% (Metzger et al. 2015). If this is a general property of *cis*-regulatory mutations, then a large number of mutations are expected to affect expression levels by amounts small enough that they will be nearly neutral (invisible to selection). Under this scenario, small changes in expression level could accumulate over time, allowing the expression levels of ohnologs to slowly drift away and resulting in the expression level divergence observed today. In the absence of a precise knowledge of ancestral expression level, distinguishing between these two hypothesis remains difficult.

As previously noted (Gout et al. 2010), after numerous genes have been lost following WGD, some of the remaining pairs that maintain ancestral expression level are at risk of being overexpressed (relative to the other genes that have now lost one copy), so that a reduction in expression level of one or both copies could become advantageous. In support of this scenario, we observed—using both the ranking and the quantile normalization methods—in all three *P. aurelia* species more pairs where both copies have reduced expression level since the split from *P. caudatum* than pairs where both copies have increased expression level ($P < 0.001$ in all three *P. aurelia* species, Materials and Methods and supplementary tables S6 and S7, Supplementary Material online).

Although our model for duplicated gene retention corresponds to a specific case of the more general DDC model (Force et al. 1999), most of the subsequent empirical emphasis on this model has been on duplication-gene retention imposed by qualitative subfunctionalization, for example, losses of gene subfunctions or expression pattern across tissues in multicellular eukaryotes, rather than on the partial partitioning of the summed expression level of duplicated genes in the same context (quantitative subfunctionalization). Under the latter scenario, ancestral gene functions are simply partitioned between the two duplicated genes such that expression levels necessary to produce enough of the final product are retained; no change in biochemical functions or expression patterns (i.e., when/where the genes and their products are expressed) is necessary. Another important difference between our model and the traditional qualitative subfunctionalization model is that, although mutations that partition the function of the duplicated genes have to take place before any inactivating mutation in the classical subfunctionalization model, dosage constraints result in an immediate selective pressure against such inactivating mutations in our absolute-dosage subfunctionalization model. Because inactivating mutations might outnumber mutations that partition the function of a gene, this is an important point in understanding why so many genes are retained for so many years after a WGD.

One possible exception to this rule is when allopolyploidization creates duplicates that are born with different expression levels. Under this scenario, some mutations causing variation in expression levels may be present at the outset, which might facilitate potential paths toward subfunctionalization. We note that if the *Paramecium* or the yeast WGD had resulted from an allopolyploidization event (or if ohnologs were born with different expression levels for any other reason), our dosage-subfunctionalization model would still apply. However, because pairs of ohnologs would initially start away from the line of equal expression shown in figure 2, the degree to which one copy is predestined to eventual loss may be less stochastic than when genes are born with identical expression levels.

Finally, we note that by preserving duplicated genes for millions of years, dosage constraints may open a window for the evolution of new beneficial functions. Indeed, although beneficial mutations are believed to be rare, preserving duplicated genes for millions of years increases the likelihood that one copy will eventually experience one such rare beneficial mutation. If the gain in fitness from this beneficial mutation exceeds the loss in fitness caused by the reduced amount of the ancestral product produced (assuming that the new

beneficial function comes at the expense of the ancestral function), it could reach fixation with the unmodified copy then acquiring compensatory mutations to increase expression. Therefore, retention through absolute-dosage subfunctionalization could pave the way for long-term functional diversification of duplicated genes.

Materials and Methods

Paramecium Genome Data

Paramecium genomes sequence and annotations were downloaded from parameciumDB (Arnaiz et al. 2007; Arnaiz and Sperling 2011) at http://paramecium.cgm.cnrs-gif.fr/download/species/ (last accessed February 2015). Functional annotations, ohnologous and orthologous relationships were extracted from supplementary files provided in McGrath, Gout, Doak, et al. (2014) and McGrath, Gout, Johri, et al. (2014).

Paramecium Species Expression Level Measurements

Publically available RNAseq data for *Paramecium* were downloaded from the NCBI Sequence Read Archive (*P. caudatum*: SRP050987, *P. biaurelia*: SRP050163, *P. tetraurelia*: [SRP051035, ERX208798, ERX208797, ERX208794, ERX208793], and *P. sexaurelia*: SRP050164) and aligned to the corresponding genomes with tophat version 2.0.7 (Kim et al. 2013), using parameters: –min-intron-length (set at 15) –max-intron-length (set at 50), and –GTF (supplied with corresponding GTF file downloaded from parameciumDB). We then ran cufflinks version 2.1.1 (Trapnell et al. 2010) with –multi-read-correct and –frag-bias-correct options to obtain values of FPKM for each predicted protein-coding gene. In order as to allow log2-transformation of genes with FPKM values of zero we added a small value (0.01) to FPKM before log2-transformation. The values of expression level used in the analysis are the log2-transformed FPKM values.

Estimation of Ancestral Expression Level in *P. aurelia* Species

We compared the expression level of genes in extent *P. aurelia* species with that of their orthologs in *P. caudatum* to infer increase and decrease in expression level since the split between *P. caudatum* and *P. aurelia* species. We used two different methods to normalize and compare gene expression levels between *P. caudatum* and each *P. aurelia* species (as absolute values of FPKM depend on the depth of coverage from RNAseq data and the number of genes annotated in the genome and therefore cannot be readily used for interspecies comparisons). The first strategy consisted in transforming FPKM values into a relative rank. After removing genes with FPKM value of zero, genes are ordered by increasing FPKM value and the relative rank is computed as the absolute rank divided by the number of genes with FPKM greater than 0, therefore giving absolute ranks comprised between 0 and 1 for each species. The relative rank of genes in each *P. aurelia* species is then compared with that of its ortholog in *P. caudatum* to infer increased or decreased expression level. The second method consisted in normalizing the distributions of log2-transformed FPKM value between each *P. aurelia* species and *P. caudatum*, using the quantile normalization method provided in the limma package from Bioconductor (Smyth 2004). The rationale for this method is the assumption that the total amount of mRNA and the shape of the distribution of gene expression levels should be identical between the two species compared. This results in overlapping distributions of log2-transformed FPKM values between the *P. aurelia* species considered and *P. caudatum*. We then compare directly the normalized values of log2-transformed FPKM for a given *P. aurelia* gene with that of its ortholog in *P. caudatum* to infer increased or decreased expression level in the *P. aurelia* species.

Yeast Data and Expression Level

Orthologous and ohnologous relationships between genes from 12 yeast species sharing the same ancestral WGD (*Vanderwaltozyma polyspora, Tetrapisispora phaffii, Tetrapisispora blattae, N. dairenensis, N. castellii, Kazachstania naganishii, Kazachstania Africana, C. glabrata, S. bayanus, S. kudriavzevii, S. mikatae*, and *S. cerevisiae*) and one outgroup (*K. lactis*) were downloaded from the Yeast Gene Order Browser (Byrne and Wolfe 2005). Expression data were downloaded from NCBI Gene Expression Omnibus (*K. lactis*: GSE22198, *S. cerevisiae*: GSE36599 and GSE54300, *N. castellii*: GSE17870 and GSE22200, *C. glabrata*: GSE22194, GSE29855, GSE52382). Only samples corresponding to wild-type conditions were retained. Log-2 transformed values from different samples belonging to a given species were normalized using the quantile normalization method from the limma package (Smyth 2004) and we used the median value across all available samples as the expression level.

Supplementary Material

Supplementary figures S1–S5 and tables S1–S7 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgment

This work was supported by National Science Foundation grant EF-0328516-A006 to M.L.

References

Adams KL. 2007. Evolution of duplicate gene expression in polyploid and hybrid plants. *J Hered.* 98:136–141.

Arnaiz O, Cain S, Cohen J, Sperling L. 2007. ParameciumDB: a community resource that integrates the *Paramecium tetraurelia* genome sequence with genetic data. *Nucleic Acids Res.* 35: D439–D444.

Arnaiz O, Sperling L. 2011. ParameciumDB in 2011: new tools and new data for functional and comparative genomics of the model ciliate *Paramecium tetraurelia*. *Nucleic Acids Res.* 39:D632–D636.

Aury JM, Jaillon O, Duret L, Noel B, Jubin C, Porcel BM, Segurens B, Daubin V, Anthouard V, Aiach N, et al. 2006. Global trends of whole-genome duplications revealed by the ciliate *Paramecium tetraurelia*. *Nature* 444:171–178.

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.

Blanc G, Wolfe KH. 2004. Functional divergence of duplicated genes formed by polyploidy during *Arabidopsis* evolution. *Plant Cell* 16: 1679–1691.

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

Davis JC, Petrov DA. 2005. Do disparate mechanisms of duplication add similar genes to the genome? *Trends Genet.* 21:548–551.

Decatur WA, Hall JA, Smith JJ, Li W, Sower SA. 2013. Insight from the lamprey genome: glimpsing early vertebrate development via neuroendocrine-associated genes and shared synteny of gonadotropin-releasing hormone (GnRH). *Gen Comp Endocrinol.* 192:237–245.

Doyle JJ, Flagel LE, Paterson AH, Rapp RA, Soltis DE, Soltis PS, Wendel JF. 2008. Evolutionary genetics of genome merger and doubling in plants. *Annu Rev Genet.* 42:443–461.

Force A, Lynch M, Pickett FB, Amores A, Yan Y-I, Postlethwait J. 1999. The preservation of duplicate genes by complementary, degenerative mutations. *Genetics* 151:1531–1545.

Freeling M, Thomas BC. 2006. Gene-balanced duplications, like tetraploidy, provide predictable drive to increase morphological complexity. *Genome Res.* 16:805–814.

Gout J-F, Duret L, Kahn D. 2009. Differential retention of metabolic genes following whole-genome duplication. *Mol Biol Evol.* 26: 1067–1072.

Gout J-F, Kahn D, Duret L. 2010. The relationship among gene expression, the evolution of gene dosage, and the rate of protein evolution. *PLoS Genet.* 6:e1000944.

Hakes L, Pinney JW, Lovell SC, Oliver SG, Robertson DL. 2007. All duplicates are not equal: the difference between small-scale and genome duplication. *Genome Biol.* 8:R209.

Holland PW, Garcia-Fernandez J, Williams NA, Sidow A. 1994. Gene duplications and the origins of vertebrate development. *Development* (Supplement):125–133.

Hughes T, Liberles DA. 2008. Whole-genome duplications in the ancestral vertebrate are detectable in the distribution of gene family sizes of tetrapod species. *J Mol Evol.* 67:343–357.

Jaillon O, Aury JM, Brunet F, Petit JL, Stange-Thomann N, Mauceli E, Bouneau L, Fischer C, Ozouf-Costaz C, Bernot A, et al. 2004. Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype. *Nature* 431:946–957.

Jiao YN, Wickett NJ, Ayyampalayam S, Chanderbali AS, Landherr L, Ralph PE, Tomsho LP, Hu Y, Liang HY, Soltis PS, et al. 2011. Ancestral polyploidy in seed plants and angiosperms. *Nature* 473: 97–100.

Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. 2013. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* 14:R36.

Lynch M. 2007. The origins of genome architecture. Sunderland (MA): Sinauer Associates.

Lynch M, Conery JS. 2000. The evolutionary fate and consequences of duplicate genes. *Science (Washington)* 290:1151–1155.

Maere S, De Bodt S, Raes J, Casneuf T, Van Montagu M, Kuiper M, Van de Peer Y. 2005. Modeling gene and genome duplications in eukaryotes. *Proc Natl Acad Sci U S A.* 102:5454–5459.

McGrath CL, Gout JF, Doak TG, Yanagi A, Lynch M. 2014. Insights into three whole-genome duplications gleaned from the *Paramecium caudatum* genome sequence. *Genetics* 197: 1417–1428.

McGrath CL, Gout JF, Johri P, Doak TG, Lynch M. 2014. Differential retention and divergent resolution of duplicate genes following whole-genome duplication. *Genome Res.* 24:1665–1675.

Metzger BP, Yuan DC, Gruber JD, Duveau F, Wittkopp PJ. 2015. Selection on noise constrains variation in a eukaryotic promoter. *Nature* Advance Access published March 16, 2015, doi:10.1038/nature14244.

Meyer A, Van de Peer Y. 2005. From 2R to 3R: evidence for a fish-specific genome duplication (FSGD). *Bioessays* 27:937–945.

Mi H, Muruganujan A, Thomas PD. 2012. PANTHER in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. *Nucleic Acids Res.* 41:D377–D386.

Morin RD, Chang E, Petrescu A, Liao N, Griffith M, Chow W, Kirkpatrick R, Butterfield YS, Young AC, Stott J, et al. 2006. Sequencing and analysis of 10,967 full-length cDNA clones from *Xenopus laevis* and *Xenopus tropicalis* reveals post-tetraploidization transcriptome remodeling. *Genome Res.* 16:796–803.

Ohno S. 1970. Evolution by gene duplication. Berlin (Germany): Springer-Verlag.

Panopoulou G, Poustka AJ. 2005. Timing and mechanism of ancient vertebrate genome duplications—the adventure of a hypothesis. *Trends Genet.* 21:559–567.

Papp B, Pal C, Hurst LD. 2003. Dosage sensitivity and the evolution of gene families in yeast. *Nature* 424:194–197.

Postlethwait JH, Woods IG, Ngo-Hazelett P, Yan YL, Kelly PD, Chu F, Huang H, Hill-Force A, Talbot WS. 2000. Zebrafish comparative genomics and the origins of vertebrate chromosomes. *Genome Res.* 10: 1890–1902.

Putnam NH, Butts T, Ferrier DEK, Furlong RF, Hellsten U, Kawashima T, Robinson-Rechavi M, Shoguchi E, Terry A, Yu J-K, et al. 2008. The amphioxus genome and the evolution of the chordate karyotype. *Nature* 453:1064–1071.

Qian W, Liao BY, Chang AY, Zhang J. 2010. Maintenance of duplicate genes and their functional redundancy by reduced expression. *Trends Genet.* 26:425–430.

Rastogi S, Liberles DA. 2005. Subfunctionalization of duplicated genes as a transition state to neofunctionalization. *BMC Evol Biol.* 5:28.

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. *Proc Natl Acad Sci U S A.* 104:8397–8402.

Seoighe C, Wolfe KH. 1999. Yeast genome evolution in the post-genome era. *Curr Opin Microbiol.* 2:548–554.

Simillion C, Vandepoele K, Van Montagu C, Zabeau M, Van de Peer Y. 2002. The hidden duplication past of *Arabidopsis thaliana*. *Proc Natl Acad Sci U S A.* 99:13627–13632.

Smyth GK. 2004. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol.* 3:Article3.

Sonneborn TM. 1975. The *Paramecium aurelia* complex of fourteen sibling species. *Trans Am Microsc Soc.* 94:155–178.

Sung W, Tucker AE, Doak TG, Choi E, Thomas WK, Lynch M. 2012. Extraordinary genome stability in the ciliate *Paramecium tetraurelia*. *Proc Natl Acad Sci U S A.* 109:19339–19344.

Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L. 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol.* 28:511–515.

Wagner A. 2000. Decoupled evolution of coding region and mRNA expression patterns after gene duplication: implications for the neutralist-selectionist debate. *Proc Natl Acad Sci U S A.* 97:6579–6584.

Wolfe KH, Shields DC. 1997. Molecular evidence for an ancient duplication of the entire yeast genome. *Nature* 387:708–713.

Zhang JZ. 2003. Evolution by gene duplication: an update. *Trends Ecol Evol.* 18:292–298.