BMC Ecology and Evolution

RESEARCH

Open Access

Expectations of duplicate gene retention under the gene duplicability hypothesis

Amanda E. Wilson[1] and David A. Liberles[1*]

Abstract

Background Gene duplication is an important process in evolution. What causes some genes to be retained after duplication and others to be lost is a process not well understood. The most prevalent theory is the gene duplicability hypothesis, that something about the function and number of interacting partners (number of subunits of protein complex, etc.), determines whether copies have more opportunity to be retained for long evolutionary periods. Some genes are also more susceptible to dosage balance effects following WGD events, making them more likely to be retained for longer periods of time. One would expect these processes that affect the retention of duplicate copies to affect the conditional probability ratio after consecutive whole genome duplication events. The probability that a gene will be retained after a second whole genome duplication event (WGD2), given that it was retained after the first whole genome duplication event (WGD1) versus the probability a gene will be retained after WGD2, given it was lost after WGD1 defines the probability ratio that is calculated.

Results Since duplicate gene retention is a time heterogeneous process, the time between the events (t1) and the time since the most recent event (t2) are relevant factors in calculating the expectation for observation in any genome. Here, we use a survival analysis framework to predict the probability ratio for genomes with different values of t1 and t2 under the gene duplicability hypothesis, that some genes are more susceptible to selectable functional shifts, some more susceptible to dosage compensation, and others only drifting. We also predict the probability ratio with different values of t1 and t2 under the mutational opportunity hypothesis, that probability of retention for certain genes changes in subsequent events depending upon how they were previously retained. These models are nested such that the mutational opportunity model encompasses the gene duplicability model with shifting duplicability over time. Here we present a formalization of the gene duplicability and mutational opportunity hypotheses to characterize evolutionary dynamics and explanatory power in a recently developed statistical framework.

Conclusions This work presents expectations of the gene duplicability and mutational opportunity hypotheses over time under different sets of assumptions. This expectation will enable formal testing of processes leading to duplicate gene retention.

Keywords Gene duplication, Polyploidy, Probabilistic modeling, Molecular evolution, Comparative genomics, Mutational opportunity

Background

Gene duplication is an important process that gives rise to functional novelty in genomes through evolution [1–5]. Gene duplication can occur at a range of scales that are classified in two broad categories, whole genome duplication (WGD) and smaller scale duplication (SSD). While the nature of the duplicate gene fixation process

*Correspondence:
David A. Liberles
daliberles@temple.edu
[1] Department of Biology and Center for Computational Genetics and Genomics, Temple University, 1900 N. 12th Street, Philadelphia, PA 19122, USA

is different between the events, as SSD events begin with a frequency of 1/2N in diploid species, while the initial frequency of a WGD event in a population of competing and breeding individuals is more complex and may be affected by things like hybridization and genomic instability [6].

Another major difference between the events is the nature of selection from dosage balance in the two events [7–9]. After WGD events, all genes are duplicated together with the interacting partners of their protein products [5, 10, 11]. After SSD events, genes are typically duplicated without the interacting partners of their protein products, leading to unfavorable stoichiometric imbalances [7, 8, 12]. The role of dosage balance in generating duplicate gene retention through eventual subfunctionalization has been modeled elsewhere [9, 13]. Because individual genes are affected differently by the processes of gene dosage balance, neofunctionalization and subfunctionalization, some genes are observed to be more inherently likely to be retained through gene duplication than other genes, which is referred to as the gene duplicability hypothesis [14–16].

Duplicability is seen as an inherent property of the genes related to their response to expression changes as well as the modularity and nature of their functions. The number of functions in a gene affects the ability to be retained through the subfunctionalization mechanism, while the number of mutationally achievable non-existing functions affects the ability to be retained through neofunctionalization [17–23]. These functions affect the retention probabilities and time dependent loss-rates as well [18]. In plants, lowly duplicable functions include genome stability maintenance and organelle-specific function (which needs to interact with organelle-encoded genes that may not be duplicated), while highly duplicable functions include signaling, transport, and metabolism [22]. Essential housekeeping genes (with core metabolic and informational functions) in Angiosperms that tend to be highly conserved across eukaryotes are often not retained as duplicates [23]. Young duplicate genes in plant genes were enriched in gene categories involved in stress responses, reflecting non-essential genes that play roles in specific environments [20]. Genes expressed in the nervous system have higher rates of retention in vertebrates [17]. Overall, genes within each functional category tended to have similar patterns of retention across paramecium species, with ribosomal proteins, transcription factors and intracellular signaling proteins being highly duplicable [21].

The number of interactions, network complexity, and dosage constraints also affect gene retention and duplicability, with large protein complexes less likely to be duplicated [16, 24–28]. Genes involved in the same pathway and protein complex share loss patterns, and genes with high expression in general tend to be retained at higher rates, while genes with a lot of interactions are less likely to be retained [21, 25, 29]. This might suggest both a role for gene number and for organismal effective population size in driving differential genome-specific retention patterns [13].

Different genes are typically seen as duplicable after SSD events and after WGD events, being biased towards certain functions [30–34]. In *Arabadopsis*, whole genome duplication events favor transcription factors, signal transducers, and organismal development, while these genes were not favored by small-scale duplication events, while genes involved in secondary metabolism and stress response tended to be favored by both large and small-scale duplication events [30]. Intrinsically disordered proteins are more duplicable after a whole-genome duplication compared to a small-scale duplication [31]. In humans, essential genes were more duplicable after a WGD event than after a small-scale event, while in yeast, they were more duplicable after small-scale events than after WGD events [32, 35, 36]. Because of the differences in their effects on the stoichiometric balance between interacting partners, WGD events lead to a slower initial duplicate gene loss rate while SSD favors fast initial loss rate [11, 12, 17, 20, 37–41]. In this study, we will focus on WGD events. Different genomes are likely to differ in the composition of their genome that includes genes that are commonly duplicable after WGD events as well as those that are contextually duplicable for that species.

Genes that have a small number of functions, are expressed in a small number of tissues and are highly sensitive to dosage shifts, such as heteromultimers would be examples of genes that fall into our dosage balance category (Dos). Examples of genes that are differentially subject to dosage balance processes without changing function to enable retention would be enzymes of glycolysis that differ in their multimerization status across the tree of life [42]. Additionally, the *Paramecium* genome is particularly sensitive to dosage constraints compared to other genomes post-WGD [41, 43].

Promiscuous genes that are expressed in multiple tissues at multiple developmental stages may have more opportunity to develop an alternative function (Alt_func, subfunctionalization or neofunctionalization). Examples of known neofunctionalized and subfunctionalized genes include diverged homologs in Atlantic salmon [44, 45]; 13% of homolog gene pairs in maize showed evidence of neofunctionalization [46]; 25% of homolog gene pairs in *Cyprinus carpio* largely sub- and neofunctionalized [39], while the rest of the gene pairs retained were through dosage or chance. Specific examples of the many genes that are known to have subfunctionalized are a yeast

protein Orc1/Sir3 [47], transcription factor IIIA (gtf3a) and ovarian gtf3ab in teleost fish [48], and ruby2-ruby1 gene family in citrus [49]. Specific examples of genes known to have neofunctionalized are POLR3G and POLR3GL in mouse liver [50], and Retinoic Acid receptors in mammals, RARα, and RARγ [51]. Among angiosperm genes in the AP2/ERF gene family, which is involved in plant development and stress responses, those with broader expression patterns have higher rates of retention of duplicates than those with narrower expression profiles possibly suggesting they were retained through subfunctionalization [18]. Some genes have been found to neofunctionalize following subfunctionalization in a process coined subneofunctionalization, showing there isn't always a hard distinction between subfunctionalization and neofunctionalization as seen in yeast and humans, and where subfunctionalization serves as a transition state to neofunctionalization [52, 53].

Other genes are unlikely to be retained through acquiring alternative functionalization. These genes are typically only retained for short periods of time following a duplication event due to drift (Non). Examples of genes that are typically found as only 1:1 orthologs across species are single copy genes across angiosperms, many of which are more conserved genes with essential housekeeping functions including those involved in photosynthesis, core metabolic processes, and the cell cycle [54, 55].

Different eukaryotic genomes have different fractions of such genes depending upon the environment that they live in, their effective population size, and other molecular, population level, and life history characteristics. For example, the paramecium genome is more dosage sensitive than other genomes post-WGD. Of retained homolog pairs, maize has 13% neofunctionalized [46], and *Cyprinus carpio* 25% sub- and neofunctionalized [39]. Genome content can very greatly, where for example the trypanosome genome [56] is structured very differently than the mammalian genome [57]. Trypanosome genomes seem to use duplication more than transcription factor binding evolution to modulate functional activity when compared with other genomes [58].

One naïve expectation of the gene duplicability hypothesis (GD) is that when consecutive whole genome duplication events have been observed in the lineage of a species, the genes that were retained after the first duplication event would be more likely to be retained after the second whole genome duplication event. This expectation was not found to hold in two genomes where this analysis was performed, in Atlantic salmon and in the orchid *Phalaenopsis equestris* [44, 59]. In the analysis of Atlantic salmon, specific gene properties associated with dosage and the gene duplicability hypothesis, like co-retention of interacting partners, were associated with preferential retention through consecutive events [44, 59, 60]. From seeing this incomplete picture of the gene duplicability hypothesis, it was clear that a more detailed modeling framework was needed to characterize gene duplicability and its expectations.

In addition to the gene duplicability hypothesis, we propose a hypothesis called the Mutational Opportunity (MO) hypothesis. We propose that the subfunctionalization and neofunctionalization processes gives rise to fewer future opportunities for subsequent subfunctionalization and neofunctionalization. This is structured as a nested model which encompasses the gene duplicability hypothesis, as opposed to a separate hypothesis. The mutational opportunity hypothesis is based on the fact that after a gene copy neofunctionalizes there are fewer novel mutations that are accessible to that gene, and after subfunctionalization there are fewer functions to specialize between the gene copies [61]. It should be noted that neofunctionalization has the potential to recharge subfunctionalization as a counter-balancing effect.

Most models for duplicate gene retention use a time-independent loss rate as a Poisson process which is known to be an inaccurate representation of duplicate gene retention probabilities as different processes have given rise to different time-dependent expectations [37, 62]. While more mechanistic Markov models of increasing levels of sophistication have been built [13, 63], the survival analysis framework of Konrad et al. 2011 [37] presents time-dependent gene loss probabilities associated with process-specific hazard functions that can be used to evaluate time-dependent expectations of the gene duplicability hypothesis. The survival analysis curves reflect averages of genes with certain characteristics in a genome. Differences in the average and variance of loss behavior can be modeled by changing the underlying parameters of the Konrad model. Here, we present this analysis framework together with the resulting dynamics under different sets of assumptions. The dynamics used differ from the naïve expectation above and can be used for explicit hypothesis testing about processes affecting duplicate gene retention in genomes. In this scenario, the parameters that are being explored for their effects on retention properties would be optimized with retention data using (for example) maximum likelihood inference. They could alternatively be estimated from the types of genes present in particular genomes. Either way, an explicit understanding of the expectations of the popular gene duplicability hypothesis are necessary to ultimately evaluate this idea.

Additionally, to model the expectations of the mutational opportunity model, we add an additional parameter indicating how frequently neo or subfunctionalized

genes can no longer neo or subfunctionalize after the second whole genome duplication event. A mechanistic model for this process would have distributions for the numbers of subfunctionalizable functions and the potential number of unrealized mutationally accessible functions for a gene. Current understanding does not enable the construction of such models and the use of a category shifting parameter reflects a phenomenological approximation. The shifting parameter does not account for the increased hazard for genes with a reduced number of functions in the subfunctionalzation model but potentially averages this out with a shift of some genes to the nonfunctionalization model. Only a small number of sunfunctionalized genes with exactly 2 functions in the ancestor will be mechanistically described by this model. With this, models for gene duplicability and gene duplicability modulated by mutational opportunity are presented.

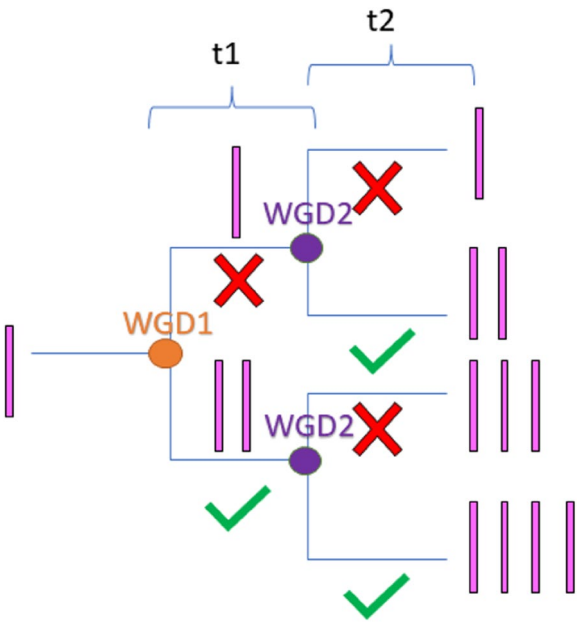

Fig. 1 Decision tree for duplicate gene copies (pink) to be retained (green check) or lost (red x) during time t1 after WGD1 (orange) and or t2 after WGD2 (purple)

Results and discussion

An experimental design has been created together with an associated probability ratio statistic to characterize patterns of duplicate gene retention in genomes with consecutive whole genome duplication events. The full dynamic behavior of gene retention under the gene duplicability hypothesis for consecutive whole genome duplication events (WGD1 and WGD2) with different time values for between the two duplication events (t1) and the time that has elapsed since the most recent duplication event (t2) will be examined in this work (Fig. 1). We used the survival analysis framework laid out in Konrad et al. 2011 [37] for the different behaviors of duplicate gene copy survival (Eq. 2) under different retention mechanisms(subfunctionalization, neofunctionalization, dosage balance, drift). In different genomes with different gene contents and multimerization patterns, we would expect these processes to play out with different proportions.

We therefore explored genome evolution with different percentages of the starting genomes following each pattern of survival under each mechanism of retention. While some of the percentages are extreme, the systematic exploration is meant to understand the behavior of the system across the full range of parameters. Although the Neofunctionalization and Subfunctionalization category of genes from the Konrad et al. 2011 [37] paper have slightly different behavior to their instantaneous rate of gene duplicate copy loss over time, they are similar enough in their behavior that they are hard to differentiate from one another [63]. We chose to combine them into one category called the Altered-function category (Alt_func). The similarities in their displayed behavior of the instantaneous rate of gene duplicate copy loss includes that they are both concave up and decrease over time, meaning the rate at which duplicate loss occurs slows and levels off. Because of this, the survival curves are not readily distinguishable from each other as described in the Konrad et al. 2011 [37] paper; therefore our combined (Alt_func) category retains these characteristics and has a survival curve that is concave up and decreasing, and has an asymptote at a value above zero, meaning the rate of duplicate loss slows as there are fewer duplicates left, but some portion of duplicates are retained more permanently (Fig. 2). Both of the two processes also result in terminal retention, having the same effect on the test statistic. The dosage balance (Dos) category of genes displays the behavior where the instantaneous rate of loss of gene copy duplicates increases as the duplication events age, and therefore the survival curve is concave down and decreasing, meaning gene duplicates get lost faster, until it hits an inflection point where the rate of gene loss slows and has an asymptote at zero (Fig. 2). The category where retention is purely by chance, and the genes can only nonfunctionalize (Non), has a time-homogeneous instantaneous rate of loss of gene duplicate copies. For this category, all of the gene copies will eventually reach the point where they nonfunctionalize with enough time following the duplication event, so its survival curve of gene duplicate copies has an asymptote at zero as in the Dos category. Here the survival curve is also concave up and decreasing, like the Alt_func category, and has a gene loss rate that

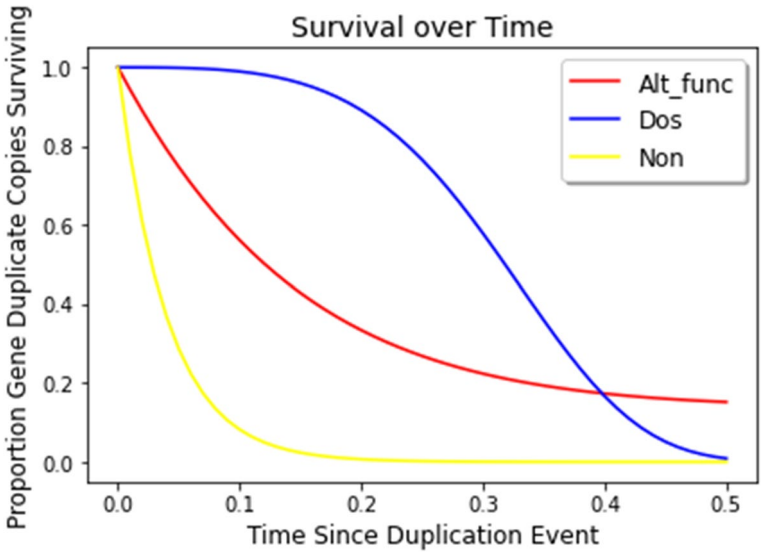

Fig. 2 Survival curves of duplicate gene copies from wgd1 during t1 for the Alt_func category of genes (red line), the Dos category of genes (blue line), and the Non category of genes (yellow line)

slows; however it decays much faster for reasons that are increasingly well understood [13] (Fig. 2). We calculated the probability ratio (p_{ratio}) for consecutive whole genome duplication events (Eq. 1). Incorporating these different percentages of different categories, using the α parameter that represents the proportion of the initial genome that fall into each category, as shown in Eq. 3. For small values of t1, the duplicate genes may not resolve into a terminal fate (Alt_functionalize or Nonfunctionalize) before another round of whole genome duplication. This would result in four entirely redundant copies. In this case, the model for retention during t2 is identical to the model as if they had resolved, but also includes a possibility of losing three of the four redundant copies. The probability of such an event can be explicitly calculated but has not been included here. The current model assumes all genes are simply duplicated. However, it is important to have a model that is applicable for small values of t1 because real fish and plant genomes can have small t1 values.

Figure 3 shows that the probability ratio depends on the t1 and t2 values. A p_{ratio} of 1 reflects that the genes are "unsorted", meaning the probability of a gene being retained after a second event is the same regardless of its probability of being retained after the first event. A high p_{ratio} reflects genes are "sorted" such that the probability of being retained after the second event is very likely if it was retained in the first event, and very unlikely to be retained if it was not retained in the first event. We chose not to include the p_{ratio} for a t1 and t2 value of zero because it is undefined in our model, and we can reasonably assume that the two events cannot happen simultaneously and happened with enough time that the events have fixed before the moment of acquiring the data. The p_{ratio} values are highest for very short t1 values and slowly get smaller as t1 values get larger. A very high p_{ratio} for very small t1 values reflects that almost all genes that are likely to be retained in t2 are probably retained immediately following the second whole genome duplication event. This effect is seen for genes in the Dos and Alt_func categories. The p_{ratio} starts at 1 for small t2 values because the gene retention pattern is unsorted, and the p_{ratio} gradually gets higher for older t2 values, the speed at which it gets bigger varies depending on the proportions of the starting genome in each category. While the magnitude of p_{ratio} peaks depend on the proportions of the starting genome in each category, there is consistently

(See figure on next page.)

Fig. 3 Gene duplicability – surface of the probability ratio over various times for t1 and t2 for given proportion of the starting genome in the Alt_func category, Dos category, and Non category (α_{Alt_func}, α_{Dos}, α_{Non} respectively). **a** $\alpha_{Alt\_func} = 0.75$, $\alpha_{Dos} = 0.0$, $\alpha_{Non} = 0.25$,**b** $\alpha_{Alt\_func} = 0.60$, $\alpha_{Dos} = 0.15$, $\alpha_{Non} = 0.25$. **c** $\alpha_{Alt\_func} = 0.45$, $\alpha_{Dos} = 0.3$, $\alpha_{Non} = 0.25$. **d** $\alpha_{Alt\_func} = 0.3$, $\alpha_{Dos} = 0.45$, $\alpha_{Non} = 0.25$. **e** $\alpha_{Alt\_func} = 0.15$, $\alpha_{Dos} = 0.6$, $\alpha_{Non} = 0.25$. **f** $\alpha_{Alt\_func} = 0.0$, $\alpha_{Dos} = 0.75$, $\alpha_{Non} = 0.25$. **g** $\alpha_{Alt\_func} = 0.5$, $\alpha_{Dos} = 0.0$, $\alpha_{Non} = 0.5$. **h** $\alpha_{Alt\_func} = 0.4$, $\alpha_{Dos} = 0.1$, $\alpha_{Non} = 0.5$. **i** $\alpha_{Alt\_func} = 0.3$, $\alpha_{Dos} = 0.2$, $\alpha_{Non} = 0.5$. **j** $\alpha_{Alt\_func} = 0.2$, $\alpha_{Dos} = 0.3$, $\alpha_{Non} = 0.5$. **k** $\alpha_{Alt\_func} = 0.1$, $\alpha_{Dos} = 0.4$, $\alpha_{Non} = 0.5$. **l** $\alpha_{Alt\_func} = 0.0$, $\alpha_{Dos} = 0.5$, $\alpha_{Non} = 0.5$. **m** $\alpha_{Alt\_func} = 0.25$, $\alpha_{Dos} = 0.0$, $\alpha_{Non} = 0.75$. **n** $\alpha_{Alt\_func} = 0.1$, $\alpha_{Dos} = 0.15$, $\alpha_{Non} = 0.75$. **o** $\alpha_{Alt\_func} = 0.0$, $\alpha_{Dos} = 0.25$, $\alpha_{Non} = 0.75$

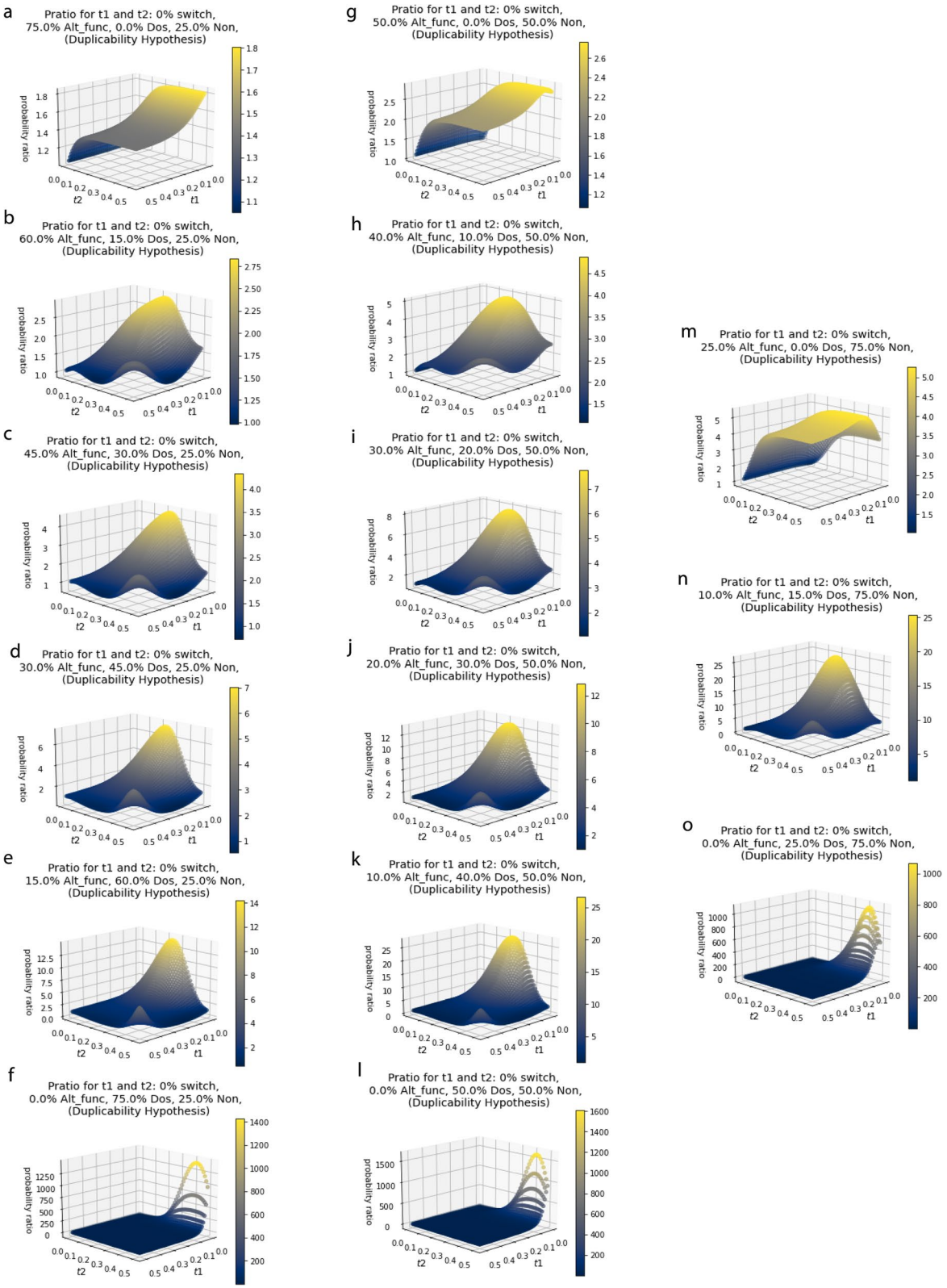

Fig. 3 (See legend on previous page.)

the highest peak at short t1 values and medium to long t2 values. There is a secondary smaller peak at very long t1 and t2 values.

Figure 3 shows the probability ratio over a range of t1 and t2 values depend on the starting proportion of the genome in the Alt_func category, Dos category, and Non category (α_{Alt_func}, α_{Dos}, α_{Non} respectively). Three percentages from the Non category were used, 25% (Fig. 3a-f), 50% (Fig. 3g-l) and 75% (Fig. 3m-o). The remaining percent was then split between the Alt_func and Dos categories, with decreasing Alt_func and increasing Dos by 10–15%, including 0% of each. Figure 3 shows that with increasing Dos and decreasing Alt_func, the peak probability ratio reached is higher, particularly in those with a short to moderate t1 and a moderate to long t2, and those with long t1 and t2 values, with a greater distribution of probability ratios that can be achieved. These very high peaks are driven by high dosage because it leads to the retention of more of the genes for longer in both t1 and t2, creating a more sorted effect. For those with none in the Dos category, the peak has the appearance of a raised plateau, and never peaks as high as with genomes containing genes in the Dos category. Of the graphs that have no Dos genes (Fig. 3a, g and m), as the percentage of initial genes in the Non increases and Alt_func decreases, the higher the probability ratio plateau peaks, particularly for younger t1 events.

Importantly to note, we surprisingly see some p_{ratio} values below one, even for those under the gene duplicability hypothesis. This happens for those with relatively large percentage in the Dos category for shorter t1 values but long t2 values. This is because at shorter time values, those in the Dos category are significantly more likely to be retained than at long time values, so it leads to "sorting" in the opposite direction, where those that were retained at short time values in t1 are more likely to be lost in the long t2 value. A similar, but opposite pattern is observed short t2 values and long t1 values. These effects occur because Dos genes are retained over shorter time periods but are ultimately lost after long periods while Alt-func genes that are lost are lost over shorter time periods but ultimately retained over long periods with a reduced loss rate.

Figure 4 models the independence hypothesis and shows the probability ratio over a range of t1 and t2 values if 100% of the genes in the genome had the same probability of being retained. This example has all the genes being in the Non category. The figure confirms expectation, that the probability ratio is equal to one regardless of how old either duplication event is. These findings show that the gene duplicability hypothesis expectations are distinct from the expectations under the independence hypothesis over the range of t1 and t2 values; however, the gene duplicability hypothesis

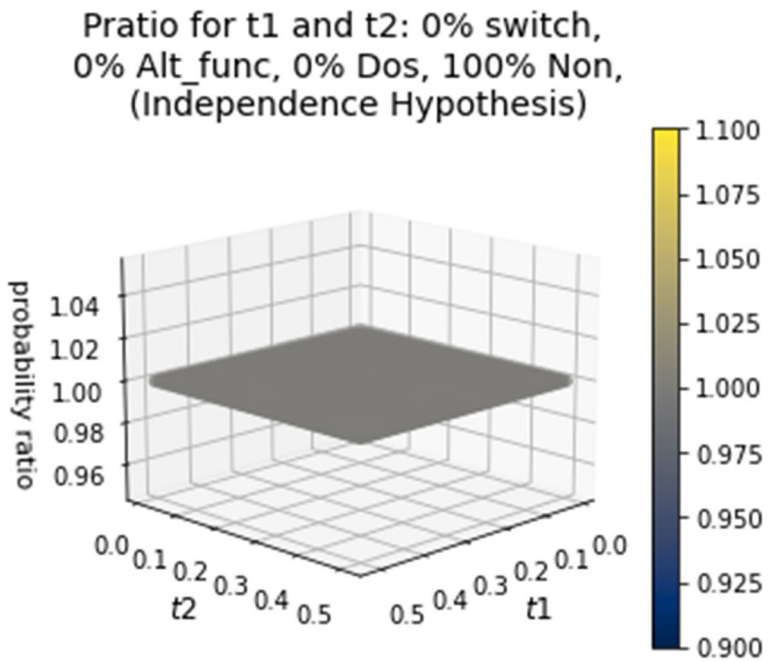

Fig. 4 Independence – probability ratio over various times for t1 and t2 if 100% of the genes in the genome had the same probability of being retained. This example has all the genes being in the Non category. This is a model of the independence hypothesis, the null hypothesis to the gene duplicability hypothesis. It is presented as a confirmation of expectations

and mutational opportunity hypothesis have t1 and t2 values that can lead to a p_{ratio} of 1.

In addition to the treatment of duplicates that have not fully resolved in the Alt_func category after a short t1 period, the explicit gene duplicability model assumes that the retention probabilities do not change in a second event following retention in the first event. Subfunctionalization reduces the probability of further subfunctionalization because there are fewer functions to subfunctionalize in the second round. The same is true for neofunctionalization reducing the probability of future neofunctionalization. However, it is also true that neofunctionalization can enable future retention by subfunctionalization. We describe this model, with these added layers of complexity, as a distinct hypothesis we call, mutational opportunity. Mutational opportunity was modeled by the introduction of a parameter that shifts the category for some of the retained pairs in the Alt_func category for the second retention period (Eq. 4). Like in the gene duplicability model, it becomes more complicated if accounting for duplicates that have not fully resolved after t1. We see a similar pattern for those under the mutational opportunity hypothesis as we do for those in the gene duplicability hypothesis, but the 3D surface is suppressed with lower p_{ratio} values, especially at long values of t1 and t2 (Fig. 5). There is a bigger suppression effect for those with more in the Alt_func category, and for those that have more of the Alt_func category switching to the nonfunctionalization category in the second WGD event (Fig. 6). For those without genes in the Alt_func category we do not see any effect, which is to be expected (Figs. 5 and 6). We only see a topological change when Alt_func is very large and the percentage that switch to nonfunctionalization in t2 (β_{switch_mo}) is also exceptionally large, especially when there are no genes in the Dos category (Fig. 6). For example, the genome with 75% in the Alt_func category and 25% in the Non category, and a β_{switch_mo} of 75%, the p_{ratio} actually peaks close to one for short t2 and the p_{ratio} gets smaller as t2 gets larger, pretty much regardless of t1 values, however these values are likely outside the realm of realism (Fig. 6).

Conclusions

We characterized the behavior of gene duplicate copy survival for genomes that experienced consecutive whole genome duplication events under the gene duplicability hypothesis mutational opportunity hypotheses. We modeled the gene duplicability hypothesis as the starting genome having some proportion of genes that are dosage sensitive and some having opportunity for neofunctionalization or subfunctionalization. We modeled the mutational opportunity hypothesis as also incorporating a proportion of genes that neofunctionalized or subfunctionalized as not being able to subsequently neo or subfunctionalize again. The model predicted expected probability ratios for survival of gene copies over different t1 and t2 values and predicts that the probability ratio (p_{ratio}) expected will depend on the proportion of the starting genome that are sensitive to dosage or having opportunity for neo- or subfunctionalization. In addition, the expected p_{ratio} depends on the t1 and t2 values. Our finding shows that the gene duplicability hypothesis predicts distinct expectations for p_{ratios} from the independence hypothesis. The independence hypothesis predicts a p_{ratio} of one regardless of t1 and t2 values or the proportion of the genome sensitive to dosage pressure and with opportunity to neo- or subfunctionalize, while the gene duplicability hypothesis predicts the p_{ratio} to be different for different t1 and t2 values and depending on the proportion of starting genes in each category. It is unclear if the gene duplicability hypothesis is distinguishable from the mutational opportunity hypothesis, except under extreme circumstances, but does result in lower p_{ratio} values than what would be expected under the gene duplicability hypothesis.

The p_{ratio} of less than one is possible under the gene duplicability models, although most are greater than 1 for those that include a terminal retention process. Some time-point combinations give rise to an expectation of or close to 1. This can happen for long t2 values and will depend upon if the hazard rate of the Alt_func model decays to zero or a value above zero. Interestingly, different sets of time points give rise to expectations of a ratio greater than 1 when there is no terminal retention process available to any of the genes. Overall,

(See figure on next page.)

Fig. 5 Mutational opportunity – surface of the probability ratio over various times for t1 and t2 for given proportion of the starting genome in the Alt_func category, Dos category, and Non category (α_{Alt_func}, α_{Dos}, α_{Non} respectively) and β_{switch_mo} equal to 0.25, which is the proportion of the Alt_func category that switches to the Non category during t2 because they cannot neofunctionalize or subfunctionalize again. **a** $\alpha_{Alt\_func}=0.75$, $\alpha_{Dos}=0.0$, $\alpha_{Non}=0.25$,**b** $\alpha_{Alt\_func}=0.60$, $\alpha_{Dos}=0.15$, $\alpha_{Non}=0.25$. **c** $\alpha_{Alt\_func}=0.45$, $\alpha_{Dos}=0.3$, $\alpha_{Non}=0.25$. **d** $\alpha_{Alt\_func}=0.3$, $\alpha_{Dos}=0.45$, $\alpha_{Non}=0.25$. **e** $\alpha_{Alt\_func}=0.15$, $\alpha_{Dos}=0.6$, $\alpha_{Non}=0.25$. **f** $\alpha_{Alt\_func}=0.0$, $\alpha_{Dos}=0.75$, $\alpha_{Non}=0.25$. **g** $\alpha_{Alt\_func}=0.5$, $\alpha_{Dos}=0.0$, $\alpha_{Non}=0.5$. **h** $\alpha_{Alt\_func}=0.4$, $\alpha_{Dos}=0.1$, $\alpha_{Non}=0.5$. **i** $\alpha_{Alt\_func}=0.3$, $\alpha_{Dos}=0.2$, $\alpha_{Non}=0.5$. **j** $\alpha_{Alt\_func}=0.2$, $\alpha_{Dos}=0.3$, $\alpha_{Non}=0.5$. **k** $\alpha_{Alt\_func}=0.1$, $\alpha_{Dos}=0.4$, $\alpha_{Non}=0.5$. **l** $\alpha_{Alt\_func}=0.0$, $\alpha_{Dos}=0.5$, $\alpha_{Non}=0.5$. **m** $\alpha_{Alt\_func}=0.25$, $\alpha_{Dos}=0.0$, $\alpha_{Non}=0.75$. **n** $\alpha_{Alt\_func}=0.1$, $\alpha_{Dos}=0.15$, $\alpha_{Non}=0.75$. **o** $\alpha_{Alt\_func}=0.0$, $\alpha_{Dos}=0.25$, $\alpha_{Non}=0.75$

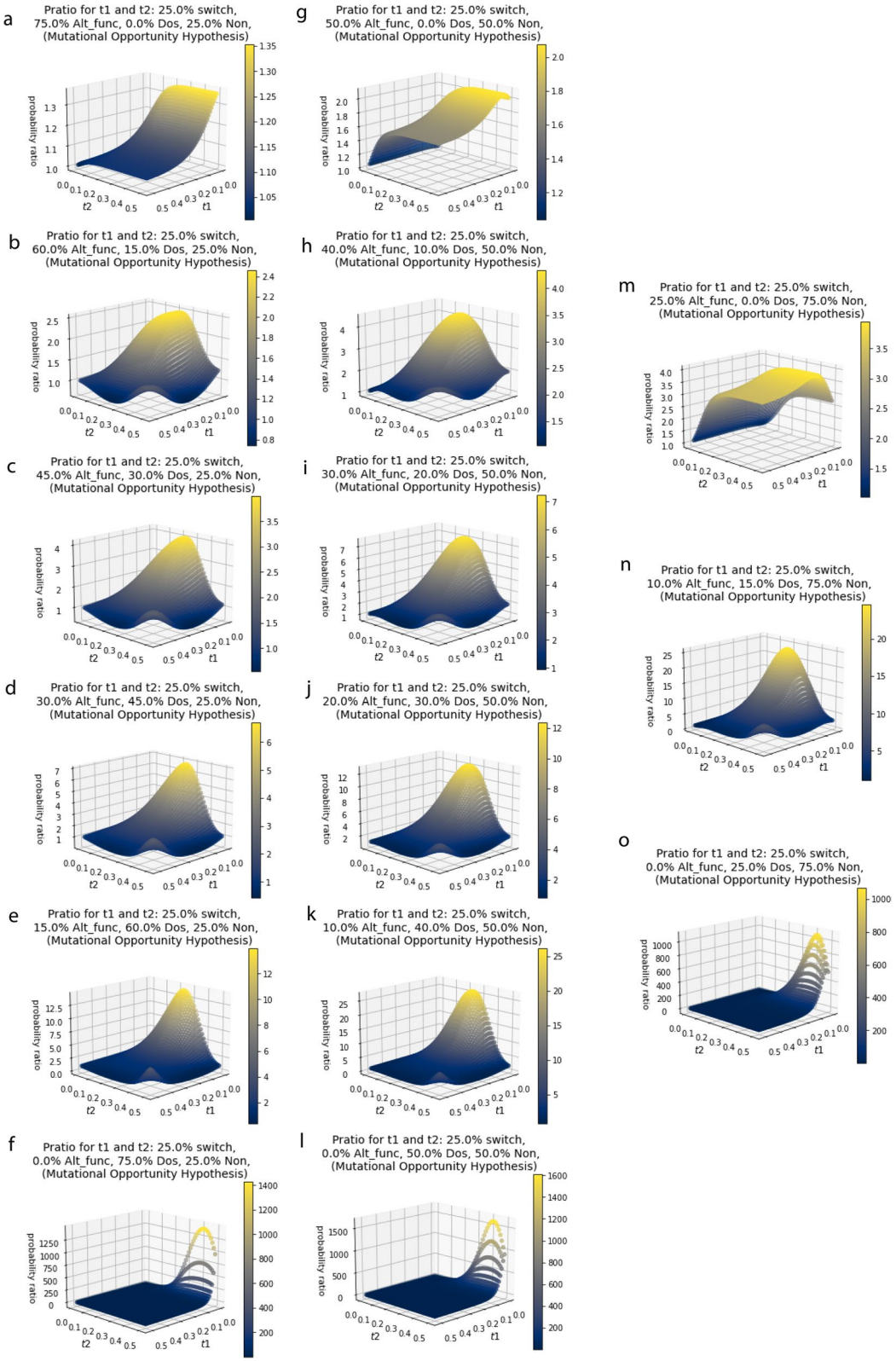

Fig. 5 (See legend on previous page.)

the set of genes subject to dosage balance processes (for example those that obligately form heteromultimers) leads to a very large expected p_{ratio} with short t1 values, which decay with increasing t2 as described. The pattern described is the same under the mutational opportunity model, although there are smaller expected p_{ratios} across the board.

The surface is flatter and closer to one for genomes with a smaller proportion of starting genes under dosage balance forces. Genomes with a lower proportion of genes with opportunity for neofunctionalization or subfunctionalization and higher proportion of genes under dosage balance effects and can only be retained by chance have a steeper surface with higher peaks, particularly for short to moderate values of t1 and long t2, however long t1 and t2 values also have a small peak. However, the gene duplicability hypothesis does give rise to a p_{ratio} of one and less than one for certain t1 and t2 values and different proportions of starting genes in each category.

In applying models to genomic data, it is assumed that the clades being examined will have similar starting fractions of genes in each category, which becomes a set of parameters to estimate using likelihood-based methods. This is a reasonable assumption for comparing species that are related and have broadly similar lifestyles, such as monocot species together or teleost fish together.

Methods

To explore patterns of duplicate gene retention, we examined consecutive whole genome duplication events and the probability of both gene duplicate copies being retained after a second duplication event conditional on whether they were both retained after the first duplication event, using a summary statistic called the probability ratio (p_{ratio}) given by Eq. 1. The first duplication event we refer to as WGD1 and the second WGD. The time between the two events is t1 and the time since the most recent event is t2.

$$pratio = \frac{probability\ of\ survival\ in\ t2 \mid survived\ in\ t1}{probability\ of\ survival\ in\ t2 \mid lost\ in\ t1} = \frac{2 * S(t_1) * S(t_2)}{(1 - S(t_1)) * S(t_2)} \tag{1}$$

More work is needed to identify more p_{ratio} data points from additional genomes with different t1 and t2 values to determine which hypothesis best explains observed data in genomes to illuminate evolutionary mechanisms.

$$s(t) = e^{(-dt - f\sum_{n=0}^{\infty} \frac{(-b)^n * t^{c*n+1}}{c*n(n!)+n!})} \tag{2}$$

$$\begin{aligned} pratio_{gene\ duplicability} = \\ \frac{(P(survival\ in\ t2 \mid survived\ in\ t1)_{Alt_{func}} * \%_{Alt_{func}}) + (P(survival\ in\ t2 \mid survived\ in\ t1)_{Dos} * \%_{Dos}) + (P(survival\ in\ t2 \mid survived\ in\ t1)_{Non} * \%_{Non})}{(P(survival\ in\ t2 \mid lost\ in\ t1)_{Alt_{func}} * \%_{Alt_{func}}) + (P(survival\ in\ t2 \mid lost\ in\ t1)_{Dos} * \%_{Dos}) + (P(survival\ in\ t2 \mid lost\ in\ t1)_{Non} * \%_{Non})} \\ = \frac{2*\alpha_{Alt\_func}*S_{Alt\_func}(t_1)*S_{Alt\_func}(t_2)+2*\alpha_{Dos}*S_{Dos}(t_1)*S_{Dos}(t_2)+2*\alpha_{Non}*S_{Non}(t_1)*S_{Non}(t_2)}{(1-S_{Alt\_func}(t_1))*\alpha_{Alt\_func}*S_{Alt\_func}(t_2)+(1-S_{Dos}(t_1))*\alpha_{Dos}*S_{Dos}(t_2)+(1-S_{Non}(t_1))*\alpha_{Non}*S_{Non}(t_2)} * \frac{(1-S_{Alt\_func}(t_1))*\alpha_{Alt\_func}+(1-S_{Dos}(t_1))*\alpha_{Dos}+(1-S_{Non}(t_1))*\alpha_{Non}}{2*\alpha_{Alt\_func}*S_{Alt\_func}(t_1)+2*\alpha_{Dos}*S_{Dos}(t_1)+2*\alpha_{Non}*S_{Non}(t_1)} \end{aligned} \tag{3}$$

(See figure on next page.)

Fig. 6 Comparison of the expected p_{ratio} values under the **a** gene duplicability hypothesis and **b**-**d** mutational opportunity hypothesis with different values of β_{switch_mo} (0.25, 0.50, 0.75 respectively), which is the proportion of the Alt_func category that switches to the Non category during t2 because they cannot neofunctionalize or subfunctionalize again. These were shown for High Alt_func, No Dos, and Low Non (75% Alt_func, 0% Dos, 25% Non), Low Alt_func, No Dos, and High Non (25% Alt_func, 0% Dos, 75% Non), Medium Alt_func, Dos and Non (30% Alt_func, 20% Dos, 50% Non), and No Alt_func, Half Dos and Half Non (0% Alt_func, 50% Dos, 50% Non). This shows that genomes with genes in the Alt_func category have smaller p_{ratio}s for those that are more likely to lose their ability to be retained again in t2. Genomes without genes in the Alt_func do not change between the gene duplicability hypothesis and the mutational opportunity hypothesis. Genomes with extremely large Alt_func categories and are extremely unlikely to be retained in t2 under this mechanism have p_{ratio}s exclusively less than one, and with a surface topology that changes the most from the gene duplicability hypothesis

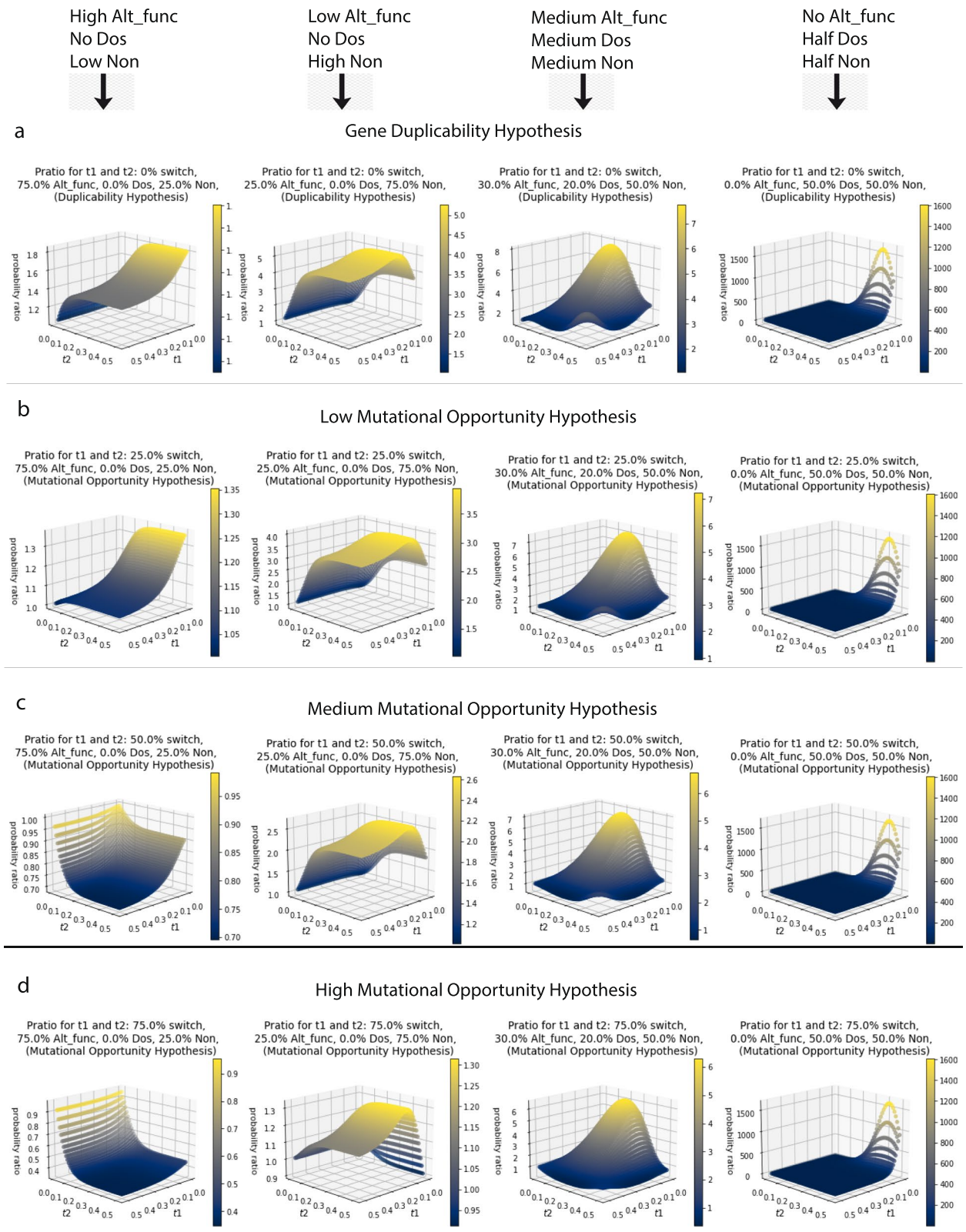

Fig. 6 (See legend on previous page.)

$$\frac{\begin{array}{c}(P(survival\ in\ t2\ |survived\ in\ t1)_{Alt_{func}} * \%_{Alt_{func}} * (1-\%_{switch_{mo}})) + (\ P(survival\ in\ t2|survived\ in\ t1)_{Alt_{func}\to Non} * \%_{Alt_{func}} * \%_{switch_{mo}}) \\ + (P(survival\ in\ t2\ |survived\ in\ t1)_{Dos} * \%_{Dos}) \ + \ (P(survival\ in\ t2\ |survived\ in\ t1)_{Non} * \%_{Non})\end{array}}{(P(survival\ in\ t2\ |lost\ in\ t1)_{Alt_{func}} * \%_{Alt_{func}}) \ + \ (P(survival\ in\ t2\ |lost\ in\ t1)_{Dos} * \%_{Dos}) \ + \ (P(survival\ in\ t2\ |lost\ in\ t1)_{Non} * \%_{Non})}$$

$$= \frac{\begin{array}{c}\left(2*\alpha_{Alt_{func}}*S_{Alt_{func}}(t_1)*S_{Alt_{func}}(t_2)*(1-\beta_{switch\ \ mo})\right)+\left(2*\alpha_{Alt_{func}}*S_{Alt_{func}}(t_1)*S_{non}(t_2)*\beta_{switch_{mo}}\right)+ 2*\alpha_{Dos}*S_{Dos}(t_1)*S_{Dos}(t_2)+ 2*\alpha_{Non}*S_{Non}(t_1) \\ *S_{Non}(t_2)\end{array}}{\left(1-S_{Alt\_func}(t_1)\right)*\alpha_{Alt\_func}*S_{Alt\_func}(t_2)+ \left(1-S_{Dos}(t_1)\right)*\alpha_{Dos}*S_{Dos}(t_2)+ \left(1-S_{Non}(t_1)\right)*\alpha_{Non}*S_{Non}(t_2)} \quad (4)$$

$$* \frac{\left(1-S_{Alt_{func}}(t_1)\right)*\alpha_{Alt_{func}}+ \left(1-S_{Dos}(t_1)\right)*\alpha_{Dos}+ \ \left(1-S_{Non}(t_1)\right)*\alpha_{Non}}{2*\alpha_{Alt_{func}}*S_{Alt_{func}}(t_1)+ 2*\alpha_{Dos}*S_{Dos}(t_1)+ 2*\alpha_{Non}*S_{Non}(t_1)}$$

The process of gene duplicate copy retention is time heterogeneous and differs depending on the mutation and selective forces on each specific gene. To model this, we used the survival analysis framework designed by Konrad et al. 2011 [37], that describes process-specific hazard functions of duplicate gene copies (Eq. 2). To model the process-specific time-heterogeneity of duplicate copy retention, they gave different parameter values to use for categories of genes with different mutation and selective forces that affect the processes of retention available to them. The parameters used were b, c, d, and f. Parameters f + d represent the rate at which fully redundant genes get lost from the genome, and this diminishes to d, which is the rate at which non-duplicated genes are lost. Parameters b and c describe the shape of the curve, i.e. the dynamics/behavior of the process when moving from the instantaneous rate to the asymptotic rate. These parameters are not explicit values that can be experimentally determined, but they are summaries of the underlying biological processes, and can be determined through model fitting to existing data. The four categories of genes described in the Konrad et al. 2011 [37] model were (1) those that are sensitive to stochiometric balance effects that would lead to selective pressure for both copies to be retained by dosage balance/compensation or one of the copies lost through nonfunctionalization (Dos), (2) those that cannot be retained through any given process, and therefore both can only be retained by chance that one has yet to nonfunctionalize (Non), (3) those that have the potential for both copies to be retained through subfunctionalization or one of the copies lost through nonfunctionalization, and (4) those that have the potential for both copies to be retained through neofunctionalization of one of the copies or one of the copies lost through nonfunctionalization. The survival curves are not easily differentiable between gene copy pairs retained through the process of subfunctionalization and the neofunctionalization process, so we combined these retention mechanisms in our model (Alt_func) [38].

We incorporated the survival analysis framework into our probability ratio statistic (Eq. 1). Then, using this framework, we could test the gene duplicability hypothesis (Eq. 3), where we assume that some percentage of the genome has genes that fit into one of the three categories. Both gene copies can be retained through some mechanism of retention, Alt_func and Dos, or they can both be retained only by chance that one of them has yet to have nonfunctionalized, Non. This models the gene duplicability hypothesis because it assumes that genes have some inherent predisposition to be retained through specific mechanisms, either through their function (for example: GO terms) or number of interacting partners. The α parameter represents the proportion of the initial genome that fall into each category. Each surviving proportion of duplicate gene copies in t1 (s(t1)) is in Eqs. 1 and 3 are multiplied by 2 because there will be two of those copies that can be duplicated in the second

Table 1 Values of the α parameter which represents the proportion of the initial genome for each category, Alt_func, Dos and Non, where $\alpha_{Alt\_func} + \alpha_{Dos} + \alpha_{Non} = 1$.

	A_{Alt_func}	α_{Dos}	α_{Non}
Figures 3a and 5a	0.75	0.0	0.25
Figures 3b and 5b	0.60	0.15	0.25
Figures 3c and 5c	0.45	0.3	0.25
Figures 3d and 5d	0.3	0.45	0.25
Figures 3e and 5e	0.15	0.6	0.25
Figures 3f and 5f	0.0	0.75	0.25
Figures 3g and 5g	0.5	0.0	0.5
Figures 3h and 5h	0.4	0.10	0.5
Figures 3i and 5i	0.3	0.2	0.5
Figures 3j and 5j	0.2	0.3	0.5
Figures 3k and 5k	0.1	0.4	0.5
Figures 3l and 5l	0.0	0.5	0.5
Figures 3m and 5m	0.25	0.0	0.75
Figures 3n and 5n	0.1	0.15	0.75
Figures 3o and 5o	0.0	0.25	0.75

wgd event. The probability of those lost in t1 are 1– the surviving proportion of duplicate gene copies (1-s(t1)).

Using a nested framework, we added an additional parameter for the mutational opportunity hypothesis that represents the portion of genes that have subfunctionalized and/or neofunctionalized lose the ability to subfunctionalize or neofunctionalize again after the second duplication event (Eq. 4). The percentage of retained genes that "switch" from this alternative-functionalization category in the second duplication event is represented by the beta value (β_{switch_mo}). Therefore, the probability of surviving duplicate gene copies in the alt_func category is split into two parts, the first part being the percentage of genes that can be retained again, and the second part being the percentage of genes that cannot be retained again, therefore they switch to the non-functionalization (non) category for t2.

We designed a python script, which can be found at https://github.com/aewilson96/Gene_Duplicability_Models, to calculate the p_{ratio} for the gene duplicability hypothesis and mutational opportunity hypothesis (Eqs. 3 and 4), using this survival analysis framework. The parameters used to generate the figures provided were chosen from the Konrad et al. 2011 [37] paper, and adjusted for visual clarity. These included n_max = 100, b_Alt_func = 10.0, c_Alt_func = 2.37, d_Alt_func = 0.00054, f_Alt_func = 5.84, b_Dos = -17.0, c_Dos = 0.2573, d_Dos = -0.000028, f_Dos = 0.000028, b_Non = 0, c_Non = 1, d_Non = 20, and f_Non = 5. The script calculates the p ratio for genomes with different starting proportions in each category (Alt_func, Dos, or Non) for 50 time points from t = 0.0 to t = 0.5 for both t1 and t2, so it includes every combination of these time points to the hundredths. In Fig. 4, we showed the independence hypothesis, so we assumed 100% of the genome was acting under the same model. For Fig. 3a-o, α_{Alt_func}, α_{Dos}, α_{Non} used were as shown in Table 1. For over fifty time points, 3D surface plots were made for each combination of α values. These combinations help provide a visual range of what the surface figures look like for the gene duplicability hypothesis and how it can be differentiated from the independence hypothesis. The same α_{Alt_func}, α_{Dos}, α_{Non} (as shown in Table 1) and fifty time points were used to generate 3D surface plots for the mutational opportunity hypothesis with a β_{switch_mo} value of 25% (Fig. 5a-o). Figure 6 shows a few select comparisons between the gene duplicability and mutational opportunity expectations with β_{switch_mo} value of 25%, 50% and 75%.

Abbreviations

Alt_Func	Is the altered function model for duplicate gene retention consistent with neofunctionalization or subfunctionalization
Dos	Is the dosage balance model for duplicate gene retention
Non	Is the nonfunctionalization model for duplicate genes
P_{ratio}	Is the conditional probability ratio
SSD	Refers to smaller scale duplication
T1	Is the time between consecutive WGD events
T2	Is the time since the most recent WGD event
WGD	Refers to whole genome duplication
Mut Op or MO	Refers to the Mutational Opportunity Hypothesis
Gene Dos or GD	Refers to the Gene Duplicability Hypothesis

Acknowledgements
AEW is grateful to Mathias Fuelling for coaching through the writing process in generating the initial draft. AEW is also grateful to Megan Lynn Wilson for her mentorship in programming.

Authors' contributions
AEW and DAL conceived the study and wrote the manuscript. AEW wrote all computer code and performed the analysis under the supervision of DAL. All authors read and approved the final manuscript.

Funding
AEW was supported by funding from Temple University.

Availability of data and materials
All code to enable the work presented here is available at https://github.com/aewilson96/Gene_Duplicability_Models. No other data was used for this study.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 14 December 2022 Accepted: 2 November 2023
Published online: 14 December 2023

References

1. Ohno S. Evolution by gene duplication. Berlin Heidelberg: Springer-Verlag; 1970.
2. Lynch M, Conery JS. The evolutionary fate and consequences of duplicate genes. Science. 2000;290:1151–5.
3. Freeling M, Thomas BC. Gene-balanced duplications, like tetraploidy, provide predictable drive to increase morphological complexity. Genome Res. 2006;16:805–14.
4. Jin G, Ma P-F, Wu X, Gu L, Long M, Zhang C, et al. New genes interacted with recent whole-genome duplicates in the fast stem growth of Bamboos. Mol Biol Evol. 2021;38:5752–68.
5. Roth C, Rastogi S, Arvestad L, Dittmar K, Light S, Ekman D, et al. Evolution after gene duplication: models, mechanisms, sequences, systems, and organisms. J Exp Zool B Mol Dev Evol. 2007;308:58–73.
6. Marsit S, Hénault M, Charron G, Fijarczyk A, Landry CR. The neutral rate of whole-genome duplication varies among yeast species and their hybrids. Nat Commun. 2021;12:3126.
7. Veitia RA. Exploring the etiology of haploinsufficiency. BioEssays. 2002;24:175–84.
8. Birchler JA, Veitia RA. The gene balance hypothesis: from classical genetics to modern genomics. Plant Cell. 2007;19:395–402.
9. Teufel AI, Liu L, Liberles DA. Models for gene duplication when dosage balance works as a transition state to subsequent neo-or sub-functionalization. BMC Evol Biol. 2016;16:45.
10. Veitia RA. Gene dosage balance in cellular pathways: implications for dominance and gene duplicability. Genetics. 2004;168:569–74.

11. Liang H, Plazonic KR, Chen J, Li W-H, Fernández A. Protein under-wrapping causes dosage sensitivity and decreases gene duplicability. PLoS Genet. 2008;4:e11.
12. Papp B, Pál C, Hurst LD. Dosage sensitivity and the evolution of gene families in yeast. Nature. 2003;424:194–7.
13. Wilson AE, Liberles DA. Dosage balance acts as a time-dependent selective barrier to subfunctionalization. BMC Ecol Evo. 2023;23:14. https://doi.org/10.1186/s12862-023-02116-y.
14. Davis JC, Petrov DA. Preferential duplication of conserved proteins in eukaryotic genomes. PLoS Biol. 2004;2:E55.
15. Lynch M, Force A. The probability of duplicate gene preservation by subfunctionalization. Genetics. 2000;154:459–73.
16. Yang J, Lusk R, Li W-H. Organismal complexity, protein complexity, and gene duplicability. Proc Natl Acad Sci U S A. 2003;100:15661–5.
17. Roux J, Liu J, Robinson-Rechavi M. Selective constraints on coding sequences of nervous system genes are a major determinant of duplicate gene retention in vertebrates. Mol Biol Evol. 2017;34:2773–91.
18. Wang L, Ma H, Lin J. Angiosperm-wide and family-level analyses of AP2/ERF genes reveal differential retention and sequence divergence after whole-genome duplication. Front Plant Sci. 2019;10:196.
19. Woods S, Coghlan A, Rivers D, Warnecke T, Jeffries SJ, Kwon T, et al. Duplication and retention biases of essential and non-essential genes revealed by systematic knockdown analyses. PLoS Genet. 2013;9:e1003330.
20. Geiser C, Mandáková T, Arrigo N, Lysak MA, Parisod C. Repeated whole-genome duplication, karyotype reshuffling, and biased retention of stress-responding genes in Buckler mustard. Plant Cell. 2016;28:17–27.
21. McGrath CL, Gout J-F, Johri P, Doak TG, Lynch M. Differential retention and divergent resolution of duplicate genes following whole-genome duplication. Genome Res. 2014;24:1665–75.
22. Li Z, Defoort J, Tasdighian S, Maere S, Van de Peer Y, De Smet R. Gene duplicability of core genes is highly consistent across all angiosperms. Plant Cell. 2016;28:326–44.
23. De Smet R, Adams KL, Vandepoele K, Van Montagu MCE, Van de Maere S. Convergent gene loss following gene and genome duplications creates single-copy families in flowering plants. Proc Natl Acad Sci U S A. 2013;110:2898–903.
24. Ascencio D, Diss G, Gagnon-Arsenault I, Dubé AK, DeLuna A, Landry CR. Expression attenuation as a mechanism of robustness against gene duplication. Proc Natl Acad Sci U S A. 2021;118:e2014345118.
25. Aury J-M, Jaillon O, Duret L, Noel B, Jubin C, Porcel BM, et al. Global trends of whole-genome duplications revealed by the ciliate Paramecium tetraurelia. Nature. 2006;444:171–8.
26. Qian W, Zhang J. Gene dosage and gene duplicability. Genetics. 2008;179:2319–24.
27. Kuzmin E, Taylor JS, Boone C. Retention of duplicated genes in evolution. Trends Genet. 2022;38:59–72.
28. Zhang Z, Luo ZW, Kishino H, Kearsey MJ. Divergence pattern of duplicate genes in protein-protein interactions follows the power law. Mol Biol Evol. 2005;22:501–5.
29. Kuzmin E, VanderSluis B, Nguyen Ba AN, Wang W, Koch EN, Usaj M et al. Exploring whole-genome duplicate gene retention with complex genetic interaction analysis. Science. 2020;368:eaaz5667
30. Maere S, De Bodt S, Raes J, Casneuf T, Van Montagu M, Kuiper M, et al. Modeling gene and genome duplications in eukaryotes. Proc Natl Acad Sci U S A. 2005;102:5454–9.
31. Banerjee S, Feyertag F, Alvarez-Ponce D. Intrinsic protein disorder reduces small-scale gene duplicability. DNA Res. 2017;24:435–44.
32. Acharya D, Ghosh TC. Global analysis of human duplicated genes reveals the relative importance of whole-genome duplicates originated in the early vertebrate evolution. BMC Genomics. 2016;17:71.
33. Van de Maere S. Retention after small- and large-scale duplications. In: Dittmar K, Liberles D, editors. Evolution after gene duplication. 2011. p. 31–56.
34. Hughes T, Liberles DA. Whole-genome duplications in the ancestral vertebrate are detectable in the distribution of gene family sizes of tetrapod species. J Mol Evol. 2008;67:343–57.
35. Guan Y, Dunham MJ, Troyanskaya OG. Functional analysis of gene duplications in Saccharomyces cerevisiae. Genetics. 2007;175:933–43.
36. Hakes L, Pinney JW, Lovell SC, Oliver SG, Robertson DL. All duplicates are not equal: the difference between small-scale and genome duplication. Genome Biol. 2007;8:R209.
37. Konrad A, Teufel AI, Grahnen JA, Liberles DA. Toward a general model for the evolutionary dynamics of gene duplicates. Genome Biol Evol. 2011;3:1197–209.
38. Teufel AI, Zhao J, O'Reilly M, Liu L, Liberles DA. On mechanistic modeling of gene content evolution: birth-death models and mechanisms of gene birth and gene retention. Computation. 2014;2:112–30.
39. Li JT, Hou GY, Kong XF, Li CY, Zeng JM, Li HD, et al. The fate of recent duplicated genes following a fourth-round whole genome duplication in a tetraploid fish, common carp (Cyprinus carpio). Sci Rep. 2015;5:8199.
40. Gillard GB, Grønvold L, Røsæg LL, Holen MM, Monsen Ø, Koop BF, et al. Comparative regulomics supports pervasive selection on gene dosage following whole genome duplication. Genome Biol. 2021;22:103.
41. Hughes T, Ekman D, Ardawatia H, Elofsson A, Liberles DA. Evaluating dosage compensation as a cause of duplicate gene retention in Paramecium tetraurelia. Genome Biol. 2007;8:213.
42. Lynch M. Evolutionary diversification of the multimeric states of proteins. Proc Natl Acad Sci U S A. 2013;110:E2821–8.
43. Gout J-F, Hao Y, Johri P, Arnaiz O, Doak TG, Bhullar S, et al. Dynamics of gene loss following ancient whole-genome duplication in the cryptic Paramecium complex. Mol Biol Evol. 2023. https://doi.org/10.1093/molbev/msad107.
44. Lien S, Koop BF, Sandve SR, Miller JR, Kent MP, Nome T, et al. The Atlantic salmon genome provides insights into rediploidization. Nature. 2016;533:200–5.
45. Warren IA, Ciborowski KL, Casadei E, Hazlerigg DG, Martin S, Jordan WC, et al. Extensive local gene duplication and functional divergence among paralogs in Atlantic salmon. Genome Biol Evol. 2014;6:1790–805.
46. Hughes TE, Langdale JA, Kelly S. The impact of widespread regulatory neofunctionalization on homeolog gene evolution following whole-genome duplication in maize. Genome Res. 2014;24:1348–55.
47. Hickman MA, Rusche LN. Transcriptional silencing functions of the yeast protein Orc1/Sir3 subfunctionalized after gene duplication. Proc Natl Acad Sci U S A. 2010;107:19384–9.
48. Rojo-Bartolomé I, Santana de Souza JE, Diaz de Cerio O, Cancio I. Duplication and subfunctionalisation of the general transcription factor IIIA (gtf3a) gene in teleost genomes, with ovarian specific transcription of gtf3ab. PLoS ONE. 2020;15:e0227690.
49. Huang D, Wang X, Tang Z, Yuan Y, Xu Y, He J, et al. Subfunctionalization of the Ruby2–Ruby1 gene cluster during the domestication of citrus. Nat Plants. 2018;4:930–41.
50. Renaud M, Praz V, Vieu E, Florens L, Washburn MP, l'Hôte P, et al. Gene duplication and neofunctionalization: POLR3G and POLR3GL. Genome Res. 2014;24:37–51.
51. Escriva H, Bertrand S, Germain P, Robinson-Rechavi M, Umbhauer M, Cartry J, et al. Neofunctionalization in vertebrates: the example of retinoic acid receptors. PLoS Genet. 2006;2:e102.
52. He X, Zhang J. Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution. Genetics. 2005;169:1157–64.
53. Rastogi S, Liberles DA. Subfunctionalization of duplicated genes as a transition state to neofunctionalization. BMC Evol Biol. 2005;5:28.
54. Li Z, De La Torre AR, Sterck L, Cánovas FM, Avila C, Merino I, et al. Single-copy genes as molecular markers for phylogenomic studies in seed plants. Genome Biol Evol. 2017;9:1130–47.
55. Han F, Peng Y, Xu L, Xiao P. Identification, characterization, and utilization of single copy genes in 29 angiosperm genomes. BMC Genomics. 2014;15:504.
56. Daniels J-P, Keith G, Bill W. Cell Biology of the trypanosome genome. Microbiol Mol Biol Rev. 2010;74:552–69.
57. Christmas MJ, Kaplow IM, Genereux DP, Dong MX, Hughes GM, Li X, et al. Evolutionary constraint and innovation across hundreds of placental mammals. Science. 2023;380:eabn3943.
58. Reis-Cunha JL, Valdivia HO, Bartholomeu DC. Gene and chromosomal copy number variations as an adaptive mechanism towards a parasitic lifestyle in trypanosomatids. Curr Genomics. 2018;19:87–97.
59. Henry CN, Piper K, Wilson AE, Miraszek JL, Probst CS, Rong Y, et al. WGDTree: a phylogenetic software tool to examine conditional probabilities of retention following whole genome duplication events. BMC Bioinformatics. 2022;23:505.
60. Hermansen RA, Hvidsten TR, Sandve SR, Liberles DA. Extracting functional trends from whole genome duplication events using comparative genomics. Biol Proced Online. 2016;18:11.

61. Assis R, Conant G, Holland B, Liberles DA, O'Reilly MM, Wilson AE. Models for the retention of duplicate genes and their biological underpinnings. F1000Res. 2023;12:1400.
62. Arvestad L, Lagergren J, Sennblad B. The gene evolution model and computing its associated probabilities. J ACM. 2009;56:1–44.
63. Stark TL, Liberles DA, Holland BR, O'Reilly MM. Analysis of a mechanistic Markov model for gene duplicates evolving under subfunctionalization. BMC Evol Biol. 2017;17:38.

Publisher's Note