Classifying the Post-duplication Fate of Paralogous Genes

Reza Kalhor, Guillaume Beslon, Manuel Lafond, Celine Scornavacca

Classifying the Post-Duplication Fate of Paralogous Genes

Reza Kalhor[1*], Guillaume Beslon[2], Manuel Lafond[1], Celine Scornavacca[3]

[1]Department of Computer Science, Université de Sherbrooke,
Sherbrooke, Canada

[2]Université de Lyon, INSA-Lyon, INRIA, CNRS,
LIRIS UMR5205, Lyon, France

[3]Institut des Sciences de l'Evolution de Montpellier
(Université de Montpellier, CNRS, IRD, EPHE), Montpellier, France

*E-mail: Reza.Kalhor@USherbrooke.ca

October 9, 2023

Abstract: Gene duplication is one of the main drivers of evolution. It is well-known that copies arising from duplication can undergo multiple evolutionary fates, but little is known on their relative frequency, and on how environmental conditions affect it. In this paper we provide a general framework to characterize the fate of duplicated genes and formally differentiate the different fates. To test our framework, we simulate the evolution of populations using aevol, an *in silico* experimental evolution platform. When classifying the resulting duplications, we observe several patterns that, in addi-

tion to confirming previous studies, exhibit new tendencies that may open up new avenues to better understand the role of duplications.

Keywords: Gene duplication, Duplication fates, Classification, Paralogy and Simulation

1 Introduction

Gene duplication is largely responsible for boosting the innovation and function variation of genomes (Carvalho et al., 2010; Kuzmin et al., 2021; Vosseberg et al., 2021), and plays a central role in the evolution of gene families (Demuth and Hahn, 2009). Copies of genes arising from duplication can undergo multiple evolutionary fates (Ohno, 2013). For instance, the copies may perform the same role, share functions, or one of them could accumulate mutations while the other maintains the original function (Ohno, 1999). The more commonly-studied fates, described in detail in the following section, are pseudogenization (one gene is lost), (double)-neofunctionalization (both/one gene diverges in function), conservation (both genes preserve functions), subfunctionalization (genes split the functions) and specialization (genes split functions and acquire novel ones).

Still, little is known on whether some of these fates are more frequent than others, and on how environmental conditions affect their relative frequency. Inferring the fate of paralogous genes is a difficult task for two main reasons. First, the functions of their lowest common ancestor is usually unknown, making it difficult to predict how the roles of each gene evolved. Second, even if the ancestral functions were known, their evolution may not fit perfectly into one of the established classes. Several works have focused on understanding the role of duplications (see e.g. (Ascencio et al., 2021)), but to our knowledge, no rigorous framework has been developed to classify these roles. Here, we aim at providing a general framework to formally characterize the possible fates of duplicated

genes to be able to discriminate them using phylogenetic data. Our approach is based on comparison of the biological functions of the original gene and the duplicated ones, and provides a continuum between the different fates.

Most research works on the topic are theoretical and propose statistical fate models to make predictions. For example, Lynch et al. (Lynch and Force, 2000; Lynch et al., 2001) model genes as discrete sets of functions and propose a population-based model of subfunctionalization that considers mutation rates at regulatory regions. They notably show that the probability of subfunctionalization tends to 0 as population sizes increase. Using similar ideas, Walsh (Walsh, 2003) compares pseudogenization against other fates, showing that predictions depend on mutation rates. In (Stark et al., 2017), the authors also compare subfunctionalization and pseudogenization using a mechanistic model based on Markov chains, which allows for data fitting and improved characterizations of hazard rates of pseudogenization. Markov chains were also used in (Diao et al., 2020) to predict the evolution of gene families undergoing duplications, loss, and partial gain/loss of function. Also, the theoretical impacts of neofunctionalization on orthology prediction were discussed in (Lafond et al., 2018). Classification tools based on gene-species reconciliation have also been proposed, e.g. for xenologs (Darby et al., 2017), which are pairs of genes whose divergence includes a horizontal gene transfer.

In more practical settings, perhaps the closest work to ours is that of Assis and Bachtrog (Assis and Bachtrog, 2013). Based on the ideas of (Otto and Yong, 2002), they used Euclidean distances between gene expression profiles to distinguish between neofunctionalization, subfunctionalization, conservation and specialization. Using drosophilia data, they show that neofunctionalization is the dominant fate, followed by conservation and specialization, and they find very few cases of subfunctionalization. In (He and Zhang, 2005), the authors use

d_N/d_S ratios and expression data to distinguish subfunctionalization and neofunctionalization. They notably conclude that such dichotomic fate models are insufficient to explain the variety of functional patterns of duplicate genes. This motivates the need to develop classification methods that account for hybrid fates. Several works have also focused on pseudogenization, based on sequence comparisons and homology detection, showing that it is very likely in certain species (Jaillon et al., 2004; Brunet et al., 2006). For instance in Zebrafish, it is estimated that up to 20% of duplicated genes are retained and the rest are non-functional (Woods et al., 2005). Neofunctionalization has also been studied in practice. It can occur through changes in the biological processes of a copy, but also in the expression at the transcriptional level. The latter was argued to play an important role in evolution (Gu et al., 2004; Huminiecki and Wolfe, 2004; Gu et al., 2005). Functional changes can occur at the enzymatic level (Conant and Wolfe, 2008) and, more recently, were shown to also occur at the post-translational level (Nguyen Ba et al., 2014). This was achieved by comparing one fate against another for three species in which short regulatory motifs were identified and statistically correlated with observed post-translational changes.

Our framework aims at generalizing the approaches developed in these experimental studies. To test our framework, we use an *in silico* experimental evolution platform that enable to simulate the evolution of a population of individuals under the combined effect of selection and variation (Hindré et al., 2012; Batut et al., 2013). Specifically, we used the aevol platform (Knibbe, 2006), a computing platform where populations of digital organisms can evolve under various conditions, enabling to experimentally study the effect of the different evolutionary forces on genomes, gene repertoire and phenotypes. Aevol has already been used to study the direct and indirect effect of segmental duplications/deletions, showing that their mutational effect is likely to regulate the amount of non-

coding sequences due to robustness constraints (Knibbe et al., 2007a; Rutten et al., 2019). The platform has also been used to show that genetic association can help maintaining cooperative behaviour in bacterial populations (Frénoy et al., 2013). More recently, aevol has been used to study the "complexity ratchet", showing that epistatic conflicts between genes duplication-divergence (i.e. neofunctionalization or double-neofunctionalization fates) and local events (i.e. allelic variation of a single gene) opens the route to biological complexity even in situations where simple phenotypes would easily thrive (Liard et al., 2020). However, although it as been shown that gene duplications is a rather frequent event in aevol, (almost half of the gene families being created by a segmental event (Knibbe, 2014)), the precise fate of gene duplicates has never been specifically studied in the model.

In this paper, we fill this gap by simulating the evolution of populations of individuals via aevol and classifying the resulting duplications using our framework. Our tests on aevol confirm the experimental studies on drosophilia data (Assis and Bachtrog, 2013) and show that conservation of the original function in both copies is rather unlikely, the general trend being that the more frequent fates are those exhibiting a higher level of function acquisition.

2 Post-duplication fates

Several classes and sub-classes of post-duplication fates have been proposed in the literature; here we recall the main ones that we model in our framework. These fates have been chosen because they are generally agreed upon, as discussed in various surveys (see e.g. (Zhang, 2003; Hahn, 2009)); each class is assigned an acronym that we shall use in the following of the paper.

Pseudogenization (P): one copy retains its functions, while the other diverges and becomes non-functional (Ohno, 2013). Pseudogenization is believed to be

very likely, since losing one copy can repair an "accidental" duplication. In this study, we consider only a type of pseudogenization, called *compensatory drift*, in which the expression level of at least one of the duplicated genes is too low to supply the function (Birchler and Yang, 2022; Thompson et al., 2016). Note that a gene could be lost by a deletion event or by a mutation that would, e.g., inactivate its promoter. However, these fates are not considered here as we focus on gene duplication leading to paralogy in extant genomes.

Neofunctionalization (N): when one copy diverges as above, it may acquire novel functions instead of pseudogeneizing (Force et al., 1999). This is often believed to be a major mechanism of function acquisition, as neofunctionalization can use a copy of a functional gene as a template to favor adaptation (Lynch and Conery, 2000).

Double-neofunctionalization (DN): both copies acquire distinct functions that are different from the original gene (hence, the original function is not performed by any of the two copies). To our knowledge, there is no established name for this fate, although this phenomenon occurs frequently in our experiments. Double-neofunctionalization can arise when a gene is not required for survival, for instance when a copy of a duplicated gene undergoes a second duplication. In this case, both sub-copies are free to develop new functions.

Conservation (C): this process is such that neither of the duplicated copies changes, both performing the same functions as the original gene, potentially doubling its expression level. One could argue that this provides no advantage to an adapted organism (it could even be harmful due to dosage effect). However, conservation can also be advantageous when increased gene dosage is required for adaptation (Panchy et al., 2016), or when one copy needs to be kept as a "backup" (Birchler and Yang, 2022).

Subfunctionalization (SF): the copies partition the original functions and

are thus complementary and necessary to perform them (Conrad and Antonarakis, 2007). This is sometimes called duplication-degeneration-complementation (DDC) (Panchy et al., 2016). Subfunctionalization has also been associated with changes in expression patterns (Birchler and Yang, 2022), especially in cases where the copies become expressed less but, together, still produce the same amount of proteins as before. The latter is sometimes distinguished as hypofunctionalization (Veitia, 2017). In this paper, we consider both situations as mere subfunctionalization.

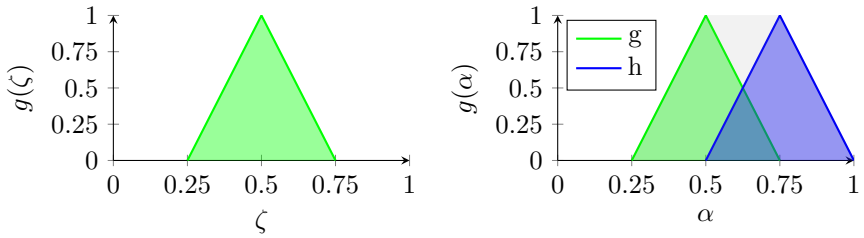
Specialization (SP): this fate occurs when the genes copies are able to perform the original functions, but *also* both develop novel functions. This differs from DN, since the original function is still performed, but also differs from SF because of the novel functions. The term was introduced in (Otto and Yong, 2002) and described as a mix of SF and N. In this work, we consider that this fate occurs as long as the original function exists (whether it is by SF or not) and both copies acquire a significant amount of new functions.

3 Methods

We first describe our theoretical model of fate classification, and then proceed to describe our experiments.

We assume the existence of a set of possible biological functions that we denote by $\mathcal{F}$. We allow any representation of functions as a set and $\mathcal{F}$ can be discrete or continuous (for instance, Gene Ontology terms, or coordinates in a multidimensional functional universe). A *gene* g expresses some functions of $\mathcal{F}$ to some degree. For this purpose, we model a gene as a (mathematical) function $g : \mathcal{F} \to \mathbb{R}$, where $g(\zeta)$ represents the activation level of function $\zeta \in \mathcal{F}$. If $g(\zeta) = 0$, then g does not contribute to performing function ζ. Importantly, notice that $g(\zeta)$ can be negative, which models the fact that g *inhibits* function

ζ. These concepts are illustrated in Figure (1.a), which shows a gene whose expression pattern has a triangular shape (note that this shape is merely for illustration, as our model applies to any shape). This gene expresses functions in the range $[0.25, 0.75]$, and the expression of each function ζ in this range is the height of the triangle at x-coordinate ζ (for instance, $g(0.5) = 1$ and $g(0.75) = 0$).



(a) A gene g.

(b) Two genes g and h.

Figure 1: An illustration of genes expressing functions in a triangle pattern.

We define the following comparative tools for two genes g and h:

- $[g+h]$ represents function addition, which can be seen as a gene described by the functional landscape that g and h accomplish together (note that they may cancel each other in case of inhibition). For each $\zeta \in \mathcal{F}$, it is defined as

$$[g + h](\zeta) = g(\zeta) + h(\zeta)$$

- $[g \cap h]$ represents function intersection and, for each $\zeta \in \mathcal{F}$, is defined as

$$[g \cap h](\zeta) = \begin{cases} \min(g(\zeta), h(\zeta)) & \text{if } g(\zeta) \geq 0, h(\zeta) \geq 0 \\ \max(g(\zeta), h(\zeta)) & \text{if } g(\zeta) < 0, h(\zeta) < 0 \\ 0 & \text{otherwise} \end{cases}$$

- for gene g, we define $\mathit{contrib}(g)$ as the total functional contribution of the gene, i.e. as the sum of absolute values of its expression levels. If $\mathcal{F}$ is discrete, we define $\mathit{contrib}(g) = \sum_{\zeta \in \mathcal{F}} |g(\zeta)|$, and if $\mathcal{F}$ is continuous, we define $\mathit{contrib}(g) = \int_{\mathcal{F}} |g(\zeta)| d\zeta$.

- $i_{g|h}$ represents the function coverage of g by h, i.e. the proportion of functions of g that can be performed by h, and is defined as

$$i_{g|h} = \frac{\mathit{contrib}([g \cap h])}{\mathit{contrib}(g)}$$

We may write $g + h$ and $g \cap h$ without brackets when no confusion can arise. Note that $[g + h] = [h + g]$ and $[g \cap h] = [h \cap g]$, but $i_{g|h}$ differs from $i_{h|g}$ if $\mathit{contrib}(g) \neq \mathit{contrib}(h)$. These notions can be visualized from Figure (1.b): $[g + h]$ can be seen as the points on the leftmost diagonal edge of the g triangle, on the top edge of the light gray area, and on the rightmost diagonal edge of the h triangle; $[g \cap h]$ can be seen as the points on the diagonal edges of the triangle formed by the overlap of the g and h triangles, which is another triangle with height $1/2$ and width $1/4$. Hence $\mathit{contrib}([g \cap h]) = ((1/2) \cdot (1/4))/2 = 1/16$. Since $\mathit{contrib}(g) = \mathit{contrib}(h) = 1/4$ we have: $i_{g|h} = i_{h|g} = (1/16)/(1/4) = 1/4$.

3.1 Classifying the fates of paralogs

Suppose that a and b are two extant paralogs and that their least common ancestor is g. For each fate described in Section 2, i.e. for each fate $X \in \{P, N, DN, C, SF, SP\}$, we quantify how much a and b appear to have undergone X, using appropriate $i_{g|h}$ proportions as defined above. The main challenge in developing a continuum between fates is to ensure that each fate has a distinguishing feature against the others. In our design, each pair of fates has a factor that contributes conversely to the two fates (while also correctly modeling them, of course). For example, N expects exactly one of $i_{a|g}$ or $i_{b|g}$ to be 1, whereas DN expects both to be 0, and values in-between have opposite effects. It was also necessary to include thresholds to model some of the fates properly, as follows:

- $\delta_\tau(x) = \max(0, \frac{x-\tau}{1-\tau})$ is a generic *threshold function* with respect to a parameter τ. It equals 0 for $x \leq \tau$, and then increases linearly from 0 to 1 in the interval $x \in [\tau, 1]$. This is useful to model fates that require a threshold.

- $\rho \in [0, 1]$ is a *pseudogene threshold*, used to determine how much functionality a copied gene must lose to be considered a pseudogene. For example, if $\rho = 0.2$, the amount of P of a gene linearly increases from 0 to 1 as its coverage of its parent drops between one fifth and 0.

- $\nu \in [0, 1]$ is a *novelty threshold* that determines how much a copy must dedicate to the parental functions to be considered as "not too new". For instance if $\nu = 0.25$, the fates C, SF require the copied genes to dedicate a quarter or more of their functions to the parental functions, and otherwise they are excluded as possible fates. Conversely, $1 - \nu$ could be interpreted as "new enough", and determines how much novelty is needed for SP.

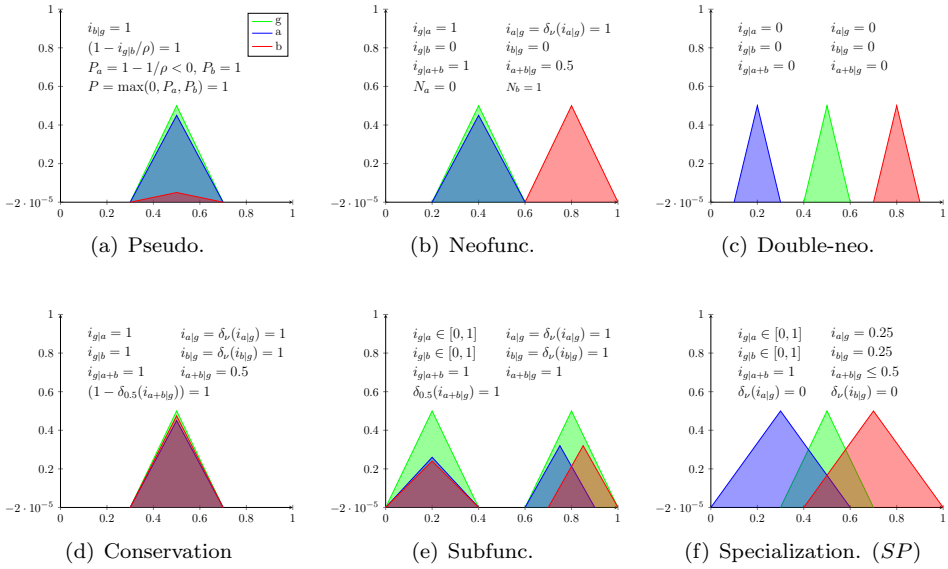
The formulas for computing the proportion of each fate are detailed in Table 1.

Fate	Formula
Pseudogenization (P)	$P_a = i_{a\|g} \cdot (1 - \frac{i_{g\|a}}{\rho})$ $P_b = i_{b\|g} \cdot (1 - \frac{i_{g\|b}}{\rho})$ $P = \max(0, P_a, P_b)$
Neofunc. (N)	$N_a = (1 - i_{a\|g}) \cdot \delta_\nu(i_{b\|g}) \cdot i_{g\|b}$ $N_b = (1 - i_{b\|g}) \cdot \delta_\nu(i_{a\|g}) \cdot i_{g\|a}$ $N = \max(N_a, N_b) \cdot (1 - P)$
Double-neo. (DN)	$DN = (1 - i_{a\|g})(1 - i_{b\|g})(1 - i_{g\|b})(1 - i_{g\|b})(1 - P)$
Conservation (C)	$C = \delta_\nu(i_{a\|g}) \cdot \delta_\nu(i_{b\|g}) \cdot i_{g\|a+b} \cdot (1 - \delta_{0.5}(i_{a+b\|g})) \cdot (1 - P)$
Subfunc. (SF)	$SF = \delta_\nu(i_{a\|g}) \cdot \delta_\nu(i_{b\|g}) \cdot i_{g\|a+b} \cdot \delta_{0.5}(i_{a+b\|g}) \cdot (1 - P)$
Specialization (SP)	$SP = i_{g\|a+b} \cdot (1 - \delta_\nu(i_{a\|g})) \cdot (1 - \delta_\nu(i_{b\|g})) \cdot (1 - P)$

Table 1: The formulas used to compute the proportion of each fate.

Using the triangular gene illustrations, Figure 2 shows that each canonical fate has an inferred proportion of 1 in our model. It can also be verified that when this occurs, the other fates have proportion 0. Also note that P and N are the only fates to use a maximum of two values. This is because there are two ways in which P can occur (either gene loses functions), and in which N can occur (either gene diverges). In the other fates (DN, C, SF, SP), the two genes behave in a similar manner instead. Although it is difficult to validate the formulas formally, we provide the rationale behind each of them:

- *Pseudogenization*: P_a should be close to 1 when a has not developed novel functions *and* has lost most of g's functions. The $i_{a|g}$ factor ensures the first condition by checking that a is covered by g. The $(1 - \frac{i_{g|a}}{\rho})$ factor implements our threshold idea for the second condition, as this factor increases linearly as a covers less functions of g, but only once the threshold ρ is crossed. The same applies to b and P_b, and P is the maximum of P_a and P_b.



(a) Pseudo. (b) Neofunc. (c) Double-neo.

(d) Conservation (e) Subfunc. (f) Specialization. (SP)

Figure 2: The canonical fates using the triangle representation (note that two possible ways in which SF can occur are shown in the same subfigure). We assume thresholds $\rho = 0.2$ (relevant for P) and $\nu = 0.25$ (mostly relevant for SP).

Note that all further fates consider the level of pseudogenization P by multiplying them by $(1 - P)$. This is because the more a gene has pseudogeneized, the less it should be considered for other fates.

- *Neofunctionalization*: N_a should be close to 1 when a acquires entirely new functions. Since a is novel, $1 - i_{a|g}$ should equal 1, and since b should only perform g, $\delta_\nu(i_{b|g})$ should be 1 (and $i_{g|b}$ should equal 1 because b covers g). The same applies to b and N_b when b neofunctionalizes.

- *Double-neo*: neither of a and b should intersect with g, and thus each of $i_{a|g}, i_{b|g}, i_{g|b}, i_{g|a}$ should be close to 0.

- *Conservation*: a and b should be identical to g, and thus a, b should be dedicated to g without "too much" novelty (i.e. $\delta_\nu(i_{a|g}), \delta_\nu(i_{b|g})$ should

be 1), and g should be covered by $[a+b]$ ($i_{g|a+b}$ should be 1). Moreover, $[a+b]$ should double each of g's functions. The $1-\delta_{0.5}(i_{g|a+b})$ factor hence expects $[a+b]$ to be covered by g by a proportion of 0.5 or less, and penalizes the fate if the coverage is higher. This factor separates C from SF.

- *Subfunctionalization*: a and b should be dedicated to performing g without too much novelty ($\delta_\nu(i_{a|g}), \delta_\nu(i_{b|g})$ should be 1), and $a+b$ should perform g together ($i_{g|a+b}$ should be 1). Unlike conservation, $[a+b]$ should be entirely covered by g since a and b have split the functions of g. The $\delta_{0.5}(i_{a+b|g})$ factor increases linearly from 0 to 1 for $i_{a+b|g} \in [0.5, 1]$, which is the opposite of conservation.

- *Specialization*: g should be performed by a and g, and thus $i_{g|a+b}$ should be 1. Moreover, a and b should both develop enough novel functions. For a, the amount of novelty is expressed as $1-i_{a|g}$. We cannot expect this term to be 1 in the SP fate, since a portion of a performs g. Using $1-\delta_\nu(i_{a|g})$ instead tolerates a to dedicate a proportion of up to ν to perform g without penalty, as long as a has enough novelty. The same holds for b.

If one considers our formulas as a probability distributions on fates, the sum of values of each fate should sum to 1 (i.e. $P+N+C+SF+SP+DN=1$). However, the six categories presented here may not cover all the possible fates of genes after a duplication. Indeed, in our experiments, we regularly observed situations where $P+N+C+SF+SP+DN<1$. Note however that we never observed situations where the sum of fate values is larger than 1 (see table 4). Since we studied thousands of duplications, we conjecture that the sum of fate values should be bounded by 1, leaving the proof as an open problem.

3.2 Computing the fate between all paralogs in a gene tree

The previous section describes how to compute the fate of a gene g and two of its paralogous descendants a and b. However, in the case of successive duplications, g may have multiple pairs of such paralogous descendants. In Algorithm 1, we describe how to compute the fate proportions between all paralogs in a gene tree G, in which leaves are extant genes and internal nodes are ancestral genes. For the purposes of our algorithm, we assume that the functions of both extant and ancestral genes are known. We also assume knowledge of a set of duplication nodes D, which can be inferred through reconciliation (Chauve and El-Mabrouk, 2009; Jacox et al., 2016). Then for each gene $g \in D$ affected by a duplication, the algorithm looks at its two child copies g_1 and g_2. It then finds the extant descendants $a_1, \ldots, a_n$ of g_1 (left leaves of g) and $b_1, \ldots, b_m$ of g_2 (right leaves of g), and calculates each fate for each triple of the form g, a_i and b_j. In our results, we report the average proportion of each fate, taken over all pairs of paralogs analyzed, as computed in Algorithm 1.

3.3 Simulations

As already mentioned, to test our method, we used simulated data generated using the aevol platform. Aevol is an *in silico* experimental evolution platform that simulates the evolution of a population or digital organisms[1]. In aevol, each organism owns a genome (double-stranded circular sequence inspired from bacterial chromosome, see Figure 3, upper part) and the model simulates transcription and translation to identify genes on the sequence. Each gene is then decoded into a $[0, 1] \rightarrow [-1, 1]$ mathematical kernel function (a "protein") and all the kernels are linearly combined to compute the phenotype (a $[0, 1] \rightarrow [0, 1]$ function – Figure 3, bottom). A population of such organisms replicate through

[1] `http://www.aevol.fr` and `http://https://gitlab.inria.fr/aevol/aevol`

Algorithm 1: Algorithm to classify duplication events. The input is a gene tree G and the set of duplication nodes D. The function $ComputeFate[X](g, a_i, b_j)$ calculates the average proportion of each fate for each triple g, a_i and b_j.

$Fates \leftarrow$ array of 6 values, initialized to 0;
$NbParalogies \leftarrow 0$;
for *each* $g \in D$ **do**
  Let g_1, g_2 be two children of g in G;
  Let $A = \{a_1, a_2, \ldots, a_n\}$ be extant descendants of g_1;
  Let $B = \{b_1, b_2, \ldots, b_m\}$ be extant descendants of g_2;
  for *each* $X \in \{P, N, DN, C, SF, SP\}$ **do**
    for *each* $a_i \in A$ **do**
      for *each* $b_j \in B$ **do**
        $Fates[X] += ComputeFate[X](g, a_i, b_j)$;
        $NbParalogies += 1$;
      end
    end
  end
end
for each $X \in \{P, N, DN, C, SF, SP\}$ **do** $Fates[X] = \frac{Fates[X]}{NbParalogies}$;

a Wright-Fisher scheme. At each generation, the fitnesses of all the organisms are computed by comparing the phenotypic function with a target function that indirectly represents the environment (see Figures 3 and 4) and, during replication, organisms may undergo various kinds of sequence mutations, including substitutions, Indels and chromosomal rearrangements (including inversions, duplications and deletions). Organisms are thus embeded into an evolutionary loop, enabling to study the relative effects of the different evolutionary forces on genome structure, genome sequence and gene repertoire.

As aevol has already been extensively described elsewhere (Knibbe, 2006; Knibbe et al., 2007b; Batut et al., 2013; Rutten et al., 2019; Liard et al., 2020), we will not describe it in more details here. Now, given our objective, there are a number of advantages of using aevol. First, the platform enables both variation of gene content and genes sequences, a mandatory property to study the fate

of duplicated genes. Second, in aevol, each gene is decoded into a mathematical function representing the genes function and the sum of all genes functions enables computing the organisms phenotype. The reproductive success (or the extinction) of an organism then depends on the adequacy of its phenotype function and the target function representing the environmental conditions. This enables a formal characterisation of genes functions, hence of the different possible fates of gene duplicates. Finally, the aevol platform has already – and successfully – be used as a benchmark to test bioinformatics methods (Biller et al., 2016). Furthermore, it has not been designed specifically to test our framework, hence providing an independent test-bed.

We now discuss our simulation framework. As briefly described above, in aevol the environment is represented by a $[0, 1] \rightarrow [0, 1]$ target function that the phenotypes must fit. We considered four different environments shown in Figure 4. We used environment (a) to generate the initial genomes, which means we let a population evolve for 1.1 million generations in this environment[2], extracted ancestor individual of final population at generation 1 million, and used it as an initial genome for further simulations. These initial genomes are called *wild-types*, and are well adapted to their environment (this "pre-evolution" step is required since evolution is heavily random in naive populations). In aevol a specific parameter ($0 < w_{max} \leq 1$) enables tuning the maximum pleiotropy in the model (the higher w_{max} the higher the pleiotropy level – $w_{max} = 1$ representing the maximum, where a gene can have an effect on all functions). As pleiotropy level is suspected to influence the fate of duplicated genes (Guillaume and Otto, 2012), we generated wild-types with four different values of $w_{max} \in$

[2]All evolutionary simulations were conducted with a population size of 1024 individuals and a mutation rate of 10^{-6} mutations per base pair per generation for each kind of mutational event. Previous experiments with the model showed that this parameter set leads to genomic structures akin to prokaryotic ones, though globally smaller (Knibbe et al., 2007a). For instance, the wild-type presented on Figure 3 has a 10,541 bp-long genome carrying 118 genes located on 50 mRNAs with a coding fraction of 77%.

$\{0.01, 0.1, 0.5, 1\}$ in environment (a). These four different wild-types enable us to test whether the pleiotropy of an organism has an impact on duplication fates. Figure 3 shows the sequence level (top) and functional level (bottom) of a wild-type evolved for 1 million generations with a minimal pleiotropy level ($w_{max} = 0.01$). Note the gene highlighted in red on the bottom-left figure. Though not active enough to reach the target, it exists in three copies on the genome, hence increasing its effect (red triangle on the bottom right). This results from two successive duplication events with fate C.

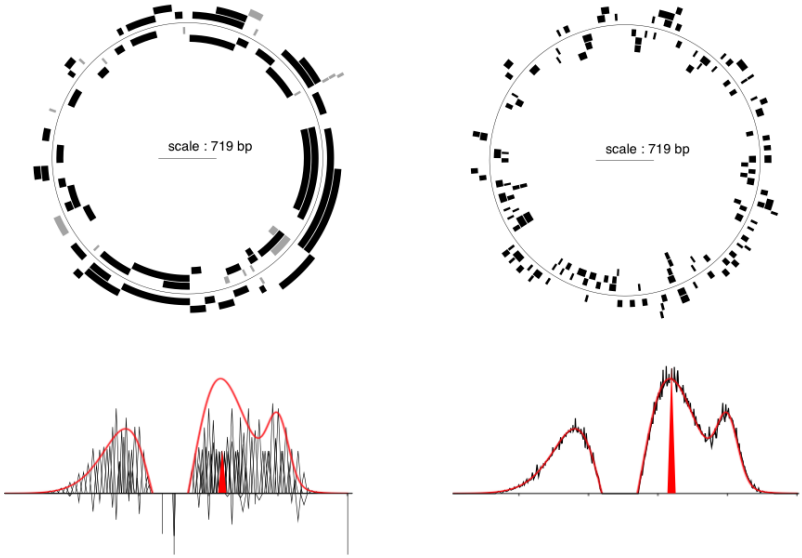


Figure 3: Overview of an aevol wild-type. Top: sequence level (genome, RNAs and genes). The double-stranded genome is represented by a circle (thin line). Black arcs represent RNAs (left) and genes (right) on each strand (grey arcs represent non-coding RNAs and non-functional genes respectively). Note the presence of polycistronic sequences. Bottom: environmental target (red curve) and functional levels (genes and phenotype, in black) with one specific function highlighted in red (see main text for details). Left: each triangle corresponds to a mathematical kernel which parameters are decoded from a gene sequence. Note the presence of function-activating/repressing genes (positive/negative triangles respectively). Right: organism's phenotype resulting from the sum of all kernels.

We used each generated wild-type as an initial genome for further 1 million generations of evolution in our four different environments. Note that, since

wild-types are already adapted to environment (a), we expect very few duplications to occur in this environment. The other three environments range from mild, medium, and heavy change with respect to the original environment; the intent of these simulations is to evaluate how individuals respond to different degrees of changes in their environment. Therefore, we expect the genomes that evolve under (d) to undergo more duplications. For each wild-type and each environment, we then performed 20 independent simulations.

Finally, we collected the most fit individuals at the end of each simulation. The extant paralogs that we analyzed were those found in their genome at the end of the process. As explained above, this procedure does not consider genes lost after duplication (either through sequence deletion or inactivation of transcription/translation initiation sequences). Thus, the pseudogenization fate here only considers extant genes whose activity has been strongly reduced[3].

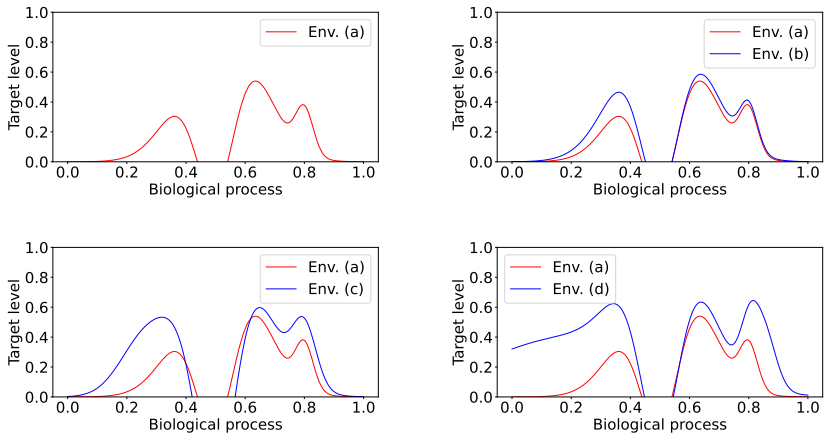


Figure 4: The four different environments used in the simulations. On the x-axis, we assume that the set of functions (biological processes) is the interval $[0, 1]$. The y-axis depicts the target level which, for each function, indicates the ideal amount of expression to survive in the environment.

[3]The source code is available at `https://github.com/r3zakalhor/Post-Duplication-Fate-Framework`.

4 Results

4.1 Fates of duplication

As explained above, starting from wild-types evolved in environment (a) with different maximum pleiotropic levels w_{max}, we simulated the evolution of 20 populations in 4 environments (ordered by increased variation compared to the environment of the wild-type) and for 1 million generations. We first verified that our phylogenies contain enough fixed duplications to enable studying the fate of duplicated genes with a reasonable precision. Table 2 shows the number of duplications per million generations observed for each environment. Recall that observed duplications are only those that result in extant paralogs, i.e. we do not consider duplications in intergenic regions, or in which a copy is lost.

	Env. (a)	**Env. (b)**	**Env. (c)**	**Env. (d)**
Gene dup. rate	3.559	15.825	40.712	55.326

Table 2: Rate of gene duplications for each environment (number of gene duplications fixed per million generations, averaged over every possible w_{max}).

	$\boldsymbol{w_{max} = 0.01}$	$\boldsymbol{w_{max} = 0.1}$	$\boldsymbol{w_{max} = 0.5}$	$\boldsymbol{w_{max} = 1}$
Gene dup. rate	53.735	28.220	17.641	15.825

Table 3: Rate of gene duplications for each pleiotropy level (number of gene duplications fixed per million generations, averaged over every environment).

Not surprisingly, the rate of fixed duplications is minimum when the organisms evolve in the constant environment (a) and it increases with the amount of change in the environments. Since each dataset comprises a million generations, the number of duplications is large enough to observe a large variety of fates. Interestingly, the number of gene duplications not only depends on the amount of environmental variation but also on the degree of pleiotropy. Indeed, Table 3 clearly shows that the lower the pleiotropy (i.e. the smaller w_{max}), the higher the number of fixed gene duplications (hence the higher the number of paralogs

at the end of the simulation). One explanation is that a smaller w_{max} implies that genes have a narrower function spectrum. Thus, having more genes may increase the chance of adding new functions, thus improving fitness.

w_{max}	P	N	DN	C	SF	SP	Total	Dup. rate
Environment (a)								
0.01	0.079	0.269	**0.357**	0.068	0.001	0.006	0.780	4.720
0.1	**0.395**	0.166	0.192	0.000	0.000	0.031	0.784	2.150
0.5	**0.239**	0.213	0.197	0.040	0.007	0.066	0.762	4.567
1	0.240	**0.434**	0.130	0.067	0.010	0.037	0.918	2.800
Environment (b)								
0.01	0.080	0.286	**0.459**	0.049	0.002	0.010	0.886	34.510
0.1	0.133	0.241	**0.343**	0.044	0.000	0.069	0.830	13.850
0.5	0.183	0.227	**0.261**	0.102	0.004	0.064	0.841	8.117
1	0.112	0.232	**0.244**	0.069	0.017	0.100	0.774	6.825
Environment (c)								
0.01	0.075	0.290	**0.465**	0.070	0.001	0.011	0.912	76.000
0.1	0.131	**0.265**	0.254	0.068	0.004	0.053	0.775	35.550
0.5	0.106	**0.293**	**0.293**	0.071	0.011	0.087	0.861	26.950
1	0.097	**0.283**	0.280	0.102	0.012	0.074	0.848	24.350
Environment (d)								
0.01	0.089	0.273	**0.482**	0.052	0.002	0.011	0.909	99.713
0.1	0.110	0.266	**0.340**	0.063	0.007	0.053	0.839	61.333
0.5	0.156	**0.266**	0.260	0.072	0.008	0.080	0.842	30.933
1	0.117	**0.306**	0.233	0.119	0.014	0.071	0.860	29.325

Table 4: Average fate proportions. Most frequent fates are boldfaced.

Table 4 show the proportions of the different fates estimated on the aevol simulations (for each wild-type we simulated 4 environments $\times$ 20 parallel repetitions evolved for 1 million generations[4]). Except in rare situations, all fates are observed and classified by our classification rules. The column "Total" reports the sum of proportions for each row. The gap between these values and 1 can be interpreted as the amount of fates that remained "unclassified". It would

[4]We note here that for some w_{max} we were able to generate and summarize statistics for several wild types: for $w_{max} = 0.01$ we have five wild types, for $w_{max} = 0.1$ one, for $w_{max} = 0.5$ three and for $w_{max} = 1$ two, leading to a total of 880 experiments.

be easy to turn our predictions into a probability distribution by normalizing them, but we prefer to emphasize the fact that paralogs underwent fates that, on average, had between 10-25% of their behavior that did not fit any of the canonical fates.

Several notable results can be observed from this table. When the organisms must adapt to a new environment (b, c and d), the most frequent fate of duplications is N or DN while P is more frequent when the organisms face the same environment as the wild-type. Moreover, while the rate of N seems independent of the mean level of pleiotropy imposed by w_{max}, we observe that the rate of DN decreases as pleiotroty increases. This emphasizes the need to differentiate both classes as we do. Note that this phenomenon is not surprising. Indeed, as pleiotropy increases, the range of functions performed by an individual gene increases, hence the probability that both duplicates lose the ancestral function and acquire a new one decreases. A most striking result is the very low percentage of SF fate. However, this result is coherent with the theoretical predictions of (Lynch and Force, 2000) and the experimental results of (Assis and Bachtrog, 2013), and probably results from the fact that SF provides no fitness advantage (since the extant function is the same as the ancestral one) but requires a transitory loss of fitness (when both copies have not yet diverged). Notably, the proportion of SF consistently increases with w_{max}. This may be explained by the fact that a higher pleiotropy level allows for alternative adaptive pathways (by adapting either genes with a high/low pleiotropy) which can compensate each others. The situation is slightly more favorable for C, especially when the environment has changed, in agreement with findings in (Assis and Bachtrog, 2013). This is probably due to the fact that the new environments may require dosage adaptation for genes function (see Figure 3 for an example such effect). In that case, duplicating a gene enables a rapid adaptation. A similar reasoning

applies to SP, which has low frequency. This confirms that the conservation of the original function in both copies is rather unlikely, the general trend being $SF < C, SP < DN, N$, sorted by increasing level of function acquisition.

4.2 Fates and time of duplication

We also evaluated the relationship between the fate of a duplication and the time at which it occurs (in terms of number of generations). Intuitively speaking, more recent duplications are expected to be biased towards Conservation, since there is less time to diverge, whereas more ancient duplications are expected to tend towards the development of new functions.

We formed bins of 100,000 generations each and, for each duplication event across all simulated wildtypes and environmental conditions, we put the duplication in the bin containing the generation it occurred in (recall that generation 0 is the most ancient and 1M the most recent). Then for each bin, we computed the average proportion of each fate within the bin (sum of fate proportion divided by number of duplications in the bin). Table 5 presents the number of duplications in each bin, and Figure 5 illustrates the relationship between time and fate.

Env. \Bins	**100K**	**200K**	**300K**	**400K**	**500K**	**600K**	**700K**	**800K**	**900K**	**1000K**
(a)	190	67	46	48	105	120	77	59	72	48
(b)	2,418	149	154	102	123	79	70	63	240	139
(c)	6,963	265	565	95	94	130	95	84	72	279
(d)	8,236	224	93	75	84	75	109	61	36	67
Total	17,807	705	858	320	406	404	351	267	420	533

Table 5: Number of duplications per generation bin, for bins of size 100K, for each environment. For instance, column 400K contains the number of duplications during generations 300K to 400K.

It is immediately apparent from Table 5 that almost all duplications occur within the first 100K generations when the environment changes (env. (b), (c), (d)). Although this may not appear as a surprise, recall that in aevol, dupli-

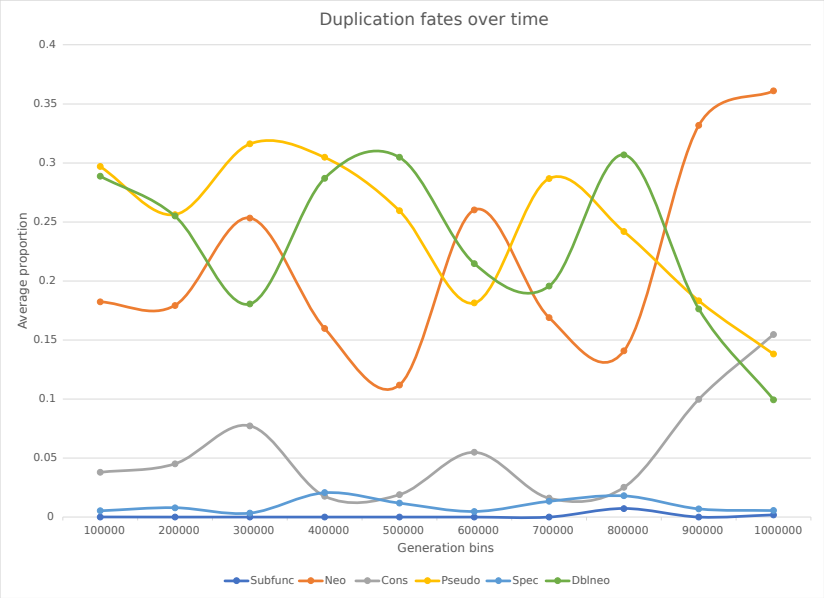


Figure 5: Average proportion of each duplication fate per generation bin.

cations are only one of the many evolutionary mechanisms that affect genome evolution (other events include substitutions, InDels, transpositions, inversions and segmental deletions). The fact that duplications are so prevalent early on therefore shows how important it is during phases of adaptation. A detailed comparison of the adaptation power of duplications against other evolutionary mechanisms is out of the scope of the current paper, but it will be interesting to perform these analyses in the future (Banse et al., 2023). In any case, there appears to be no trend in the number of duplications after 100K or 200K generations. One may arguably view the early duplications as necessary for selection, and the later ones as duplications becoming fixed by chance.

As for Figure 5, the fates DN, N, and P remain largely dominant through most generations, which is to be expected from the results of the previous section. Interestingly though, the last 200K generations introduce significant variations in the fate proportions. First, there is a sharp increase in the amount of

Conservation towards the end, going from 0.02 in the 800K bin to 0.15 in the 1000K bin, which is in line with the intuition that time is needed for divergence. It is worth noting that Conservation becomes even more frequent than Double-Neofunctionalization and Pseudogeneization, which both see a sharp decrease in the last 200K generations.

It is also noticeable that Neofunctionalization becomes the clearly dominant fate in these last generations. This may be seen as standard divergence after duplication: one copy must maintain the function and the other undergoes random drift, and since only a limited time passes, the divergent copy remains functional and observable. On the other hand, the fact that N is dominated by DN and P in more ancient duplications suggests that, given enough time, duplicated copies rather tend to eventually both diverge or to eliminate one copy. One may also notice that the C and N curves are almost parallel, and that the two fates appear to correlated.

4.3 Fates and successive duplications

We also checked whether rounds of successive duplications could affect fates. When a gene duplicates and one or both copy also duplicate later on, it is possible that a bias towards certain fates is introduced. Therefore, for each duplication g, we looked at the number of descendants of g in its gene tree (see Algorithm 1), where here the number of descendants is the number of leaves under the duplication node. For instance, g having two descendants means that no copy duplicated further, having three descendants means that one copy also duplicated, and so on. The second column of Table 6 reports the number of duplication events encountered for each number of descendants. The vast majority of duplications have only two descendants and, across all the simulations, the maximum number of descendants of a duplication is 12.

Generally speaking, we found no specific relationship between the number of descendants and fates. The numbers shown in Table 6 are distributed in a similar manner across the rows, and are also similar to the fate proportions reported in the previous section. However, we do observe a general downwards trend in the totals column. This suggests that the when successive duplications occur, the fate of the most ancestral duplication gets diluted in its descendants, making it harder to characterize. In the future, it might be beneficial to classify the fate of a duplication more "locally", that is, by looking at its descending genes until a certain point, as going too far down the gene tree may introduce interference in our analysis.

Nb descendants	Nb dups	Subfunc	Neo	Cons	Pseudo	Spec	Dblneo	Total
2	19,792	0.001	0.202	0.047	0.289	0.006	0.281	0.826
3	1,698	0.000	0.150	0.016	0.329	0.005	0.285	0.785
4	402	0.000	0.134	0.024	0.290	0.011	0.274	0.733
5	107	0.000	0.167	0.031	0.253	0.005	0.262	0.718
6	44	0.000	0.086	0.008	0.227	0.004	0.267	0.592
7	13	0.000	0.091	0.086	0.242	0.000	0.394	0.813
8	6	0.000	0.222	0.133	0.222	0.000	0.111	0.689
9	3	0.000	0.133	0.017	0.067	0.000	0.025	0.467
10	0	0.000	0.000	0.000	0.000	0.000	0.000	0.000
11	0	0.000	0.000	0.000	0.000	0.000	0.000	0.000
12	6	0.000	0.210	0.037	0.037	0.000	0.272	0.556

Table 6: Average proportion of fates per number of descendants. The second column reports the number of duplication events for each number of descendants, and the last column the sum of fate proporitions for each row.

5 Discussion

In this paper, we proposed a methodology to formally classify the fate of gene duplicates depending on the functions of the extant paralogs and of the ancestral gene. The objective is to provide the community with clear definitions as well as a mathematical toolbox to discriminate the different fates. Indeed, in the absence of such a toolbox it is almost impossible to compare experimental and/or

theoretical studies limiting the possibility of developing a global understanding of gene duplication, even though this mechanism is considered central in molecular evolution. Our framework has been extensively tested on simulated data provided by aevol, an independently designed platform. Our tests confirmed several tendencies reported in the literature (Guillaume and Otto, 2012; Assis and Bachtrog, 2013), showing the relevance of our classification. Further work will permit to study a broader set of parameters, both for the simulations and for the classification thresholds, to confirm these trends. Incidentally, our results also confirm the interest of using aevol as benchmark to test bioinformatics tools.

Our work opens several fields of research, in comparative genomics and phylogeny, in simulation and, of course, in evolutionary biology. Indeed, although the fraction of gene duplicates classified is high (> 0.75 in all situations), it also shows that further work is required to analyze the remaining fates. Also, even though these data are not reported here, we observed a small fraction of "hybrid fates" which deserve a specific study. Finally, as our methodology is based on the analysis of extant paralogs, it cannot account for the whole diversity of pseudogeneization fates. Indeed, in our results P is always lower than 20% (except in constant environment – see Table 4), which is much lower than the 80% observed in the Zebrafish (Otto and Yong, 2002). We conjecture that the difference is due to the way we selected gene duplicates in our study. Extending P class to account for the whole variety of pseudogeneization fates is an exciting direction of research. Finally, we could used real data available in published datasets such as (Gaudet et al., 2011) to further test our approach. While aevol simulations enabled testing the continuous version of our framework, other datasets could enable testing the discrete version, e.g. by classifying paralogs annotated with Gene Ontology (Zhao et al., 2020).

In this study, we used aevol to test our framework, showing that it generates data similar to real observations. This motivates us to further study gene duplications in the simulator. In particular, aevol not only provides the final organisms, but also the past individuals and the exact gene phylogeny, making it possible to know the exact fate of each gene along each branch (including gene loss). We used this information to refine our study and tested how the fate of duplicated genes evolves in time after the founding duplication event, a question that is almost impossible to study in vivo. We showed that although, on the long term, Neofunctionalization, Double-Neofunctionalization and Pseudogenization are the most probable fates, immediately after the duplication events, the dominant fates are Conservation and Neofunctionalization. Further studies could reveal which fates are more likely to open the path to others, an information that could be used to predict the evolution of specific gene branches following recent duplications. We also showed that successive duplication events rapidly blur the classification, opening questions for further refinement of the method. The model also enables "*in silico* genetic engineering". We plan to construct a series of mutants in which genes are manually duplicated and let evolve. This will open the route to a systematic study of gene duplication in the model. We could also observe the fate of duplicates in more specific settings, such as after a Whole Genome Duplication (WGD), and check whether it depends on the characteristics of the ancestral gene (e.g., on essentiality, pleiotropy or transcription level...). Another interesting line of investigation could be to understand the impact of regulation on the fates frequency and especially on subfunctionalization. While regulation was not included in the current version of aevol, an extension is under development to account for the evolution of transcription factors that will allow addressing this question.

Finally, it would also be interesting to study how specific biological dupli-

cation mechanisms, for instance unequal crossing over, tandem duplication or retrotransposition (Reams et al., 2012), are associated with fates. Such investigations would probably require to analyse not only gene functions but also gene genealogies. Combining our framework to tools such as PAINT (Gaudet et al., 2011) or PANTHER (Mi et al., 2017), that predict the functions of ancestral genes given a gene phylogeny and the functions of the extant genes, would enable us to analyse real data. We leave this for future work.

References

Ascencio, D., Diss, G., Gagnon-Arsenault, I., Dubé, A. K., DeLuna, A., and Landry, C. R. Expression attenuation as a mechanism of robustness against gene duplication. *Proceedings of the National Academy of Sciences*, 118(6): e2014345118, 2021.

Assis, R. and Bachtrog, D. Neofunctionalization of young duplicate genes in drosophila. *Proceedings of the National Academy of Sciences*, 110(43):17409–17414, 2013.

Banse, P., Luiselli, J., Parsons, D. P., Grohens, T., Foley, M., Trujillo, L., Rouzaud-Cornabas, J., Knibbe, C., and Beslon, G. Forward-in-time simulation of chromosomal rearrangements: The invisible backbone that sustains long-term adaptation. *in prep.*, 2023.

Batut, B., Parsons, D. P., Fischer, S., Beslon, G., and Knibbe, C. In silico experimental evolution: a tool to test evolutionary scenarios. In *BMC bioinformatics*, volume 14, pages 1–11. Springer, 2013.

Biller, P., Knibbe, C., Beslon, G., and Tannier, E. Comparative genomics on artificial life. In *Pursuit of the Universal: 12th Conference on Computability*

in Europe, CiE 2016, Paris, France, June 27-July 1, 2016, Proceedings 12, pages 35–44. Springer, 2016.

Birchler, J. A. and Yang, H. The multiple fates of gene duplications: deletion, hypofunctionalization, subfunctionalization, neofunctionalization, dosage balance constraints, and neutral variation. *The Plant Cell*, 2022.

Brunet, F. G., Crollius, H. R., Paris, M., Aury, J.-M., Gibert, P., Jaillon, O., Laudet, V., and Robinson-Rechavi, M. Gene loss and evolutionary rates following whole-genome duplication in teleost fishes. *Molecular biology and evolution*, 23(9):1808–1816, 2006.

Carvalho, C. M., Zhang, F., and Lupski, J. R. Genomic disorders: A window into human gene and genome evolution. *Proceedings of the National Academy of Sciences*, 107(suppl_1):1765–1771, 2010.

Chauve, C. and El-Mabrouk, N. New perspectives on gene family evolution: losses in reconciliation and a link with supertrees. In *Research in Computational Molecular Biology: 13th Annual International Conference, RECOMB 2009, Tucson, AZ, USA, May 18-21, 2009. Proceedings 13*, pages 46–58. Springer, 2009.

Conant, G. C. and Wolfe, K. H. Turning a hobby into a job: how duplicated genes find new functions. *Nature Reviews Genetics*, 9(12):938–950, 2008.

Conrad, B. and Antonarakis, S. E. Gene duplication: a drive for phenotypic diversity and cause of human disease. *Annu. Rev. Genomics Hum. Genet.*, 8: 17–35, 2007.

Darby, C. A., Stolzer, M., Ropp, P. J., Barker, D., and Durand, D. Xenolog classification. *Bioinformatics*, 33(5):640–649, 2017.

Demuth, J. P. and Hahn, M. W. The life and death of gene families. *Bioessays*, 31(1):29–39, 2009.

Diao, J., Stark, T. L., Liberles, D. A., O'Reilly, M. M., and Holland, B. R. Level-dependent qbd models for the evolution of a family of gene duplicates. *Stochastic Models*, 36(2):285–311, 2020.

Force, A., Lynch, M., Pickett, F. B., Amores, A., Yan, Y.-l., and Postlethwait, J. Preservation of duplicate genes by complementary, degenerative mutations. *Genetics*, 151(4):1531–1545, 1999.

Frénoy, A., Taddei, F., and Misevic, D. Genetic architecture promotes the evolution and maintenance of cooperation. *PLoS computational biology*, 9 (11):e1003339, 2013.

Gaudet, P., Livstone, M. S., Lewis, S. E., and Thomas, P. D. Phylogenetic-based propagation of functional annotations within the gene ontology consortium. *Briefings in bioinformatics*, 12(5):449–462, 2011.

Gu, X., Zhang, Z., and Huang, W. Rapid evolution of expression and regulatory divergences after yeast gene duplication. *Proceedings of the National Academy of Sciences*, 102(3):707–712, 2005.

Gu, Z., Rifkin, S. A., White, K. P., and Li, W.-H. Duplicate genes increase gene expression diversity within and between species. *Nature genetics*, 36(6): 577–579, 2004.

Guillaume, F. and Otto, S. P. Gene functional trade-offs and the evolution of pleiotropy. *Genetics*, 192(4):1389–1409, 2012.

Hahn, M. W. Distinguishing among evolutionary models for the maintenance of gene duplicates. *Journal of Heredity*, 100(5):605–617, 2009.

He, X. and Zhang, J. Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution. *Genetics*, 169(2):1157–1164, 2005.

Hindré, T., Knibbe, C., Beslon, G., and Schneider, D. New insights into bacterial adaptation through in vivo and in silico experimental evolution. *Nature Reviews Microbiology*, 10(5):352–365, 2012.

Huminiecki, L. and Wolfe, K. H. Divergence of spatial gene expression profiles following species-specific gene duplications in human and mouse. *Genome research*, 14(10a):1870–1879, 2004.

Jacox, E., Chauve, C., Szöllősi, G. J., Ponty, Y., and Scornavacca, C. eccetera: comprehensive gene tree-species tree reconciliation using parsimony. *Bioinformatics*, 32(13):2056–2058, 2016.

Jaillon, O., Aury, J.-M., Brunet, F., Petit, J.-L., Stange-Thomann, N., Mauceli, E., Bouneau, L., Fischer, C., Ozouf-Costaz, C., Bernot, A., et al. Genome duplication in the teleost fish tetraodon nigroviridis reveals the early vertebrate proto-karyotype. *Nature*, 431(7011):946–957, 2004.

Knibbe, C. *Structuration des génomes par sélection indirecte de la variabilité mutationnelle: une approche de modélisation et de simulation.* PhD thesis, INSA de Lyon, 2006.

Knibbe, C. What happened to my genes? insights on gene family dynamics from digital genetics experiments. In *ALIFE 14: The Fourteenth International Conference on the Synthesis and Simulation of Living Systems*, pages 33–40. MIT Press, 2014.

Knibbe, C., Coulon, A., Mazet, O., Fayard, J.-M., and Beslon, G. A long-term

evolutionary pressure on the amount of noncoding dna. *Molecular biology and evolution*, 24(10):2344–2353, 2007a.

Knibbe, C., Mazet, O., Chaudier, F., Fayard, J.-M., and Beslon, G. Evolutionary coupling between the deleteriousness of gene mutations and the amount of non-coding sequences. *Journal of Theoretical Biology*, 244(4):621–630, 2007b.

Kuzmin, E., Taylor, J. S., and Boone, C. Retention of duplicated genes in evolution. *Trends in Genetics*, 2021.

Lafond, M., Meghdari Miardan, M., and Sankoff, D. Accurate prediction of orthologs in the presence of divergence after duplication. *Bioinformatics*, 34 (13):i366–i375, 2018.

Liard, V., Parsons, D. P., Rouzaud-Cornabas, J., and Beslon, G. The complexity ratchet: Stronger than selection, stronger than evolvability, weaker than robustness. *Artificial life*, 26(1):38–57, 2020.

Lynch, M. and Conery, J. S. The evolutionary fate and consequences of duplicate genes. *science*, 290(5494):1151–1155, 2000.

Lynch, M. and Force, A. The probability of duplicate gene preservation by subfunctionalization. *Genetics*, 154(1):459–473, 2000.

Lynch, M., O'Hely, M., Walsh, B., and Force, A. The probability of preservation of a newly arisen gene duplicate. *Genetics*, 159(4):1789–1804, 2001.

Mi, H., Huang, X., Muruganujan, A., Tang, H., Mills, C., Kang, D., and Thomas, P. D. Panther version 11: expanded annotation data from gene ontology and reactome pathways, and data analysis tool enhancements. *Nucleic acids research*, 45(D1):D183–D189, 2017.

Nguyen Ba, A. N., Strome, B., Hua, J. J., Desmond, J., Gagnon-Arsenault, I., Weiss, E. L., Landry, C. R., and Moses, A. M. Detecting functional divergence

after gene duplication through evolutionary changes in posttranslational regulatory sequences. *PLoS computational biology*, 10(12):e1003977, 2014.

Ohno, S. Gene duplication and the uniqueness of vertebrate genomes circa 1970–1999. In *Seminars in cell & developmental biology*, volume 10, pages 517–522. Elsevier, 1999.

Ohno, S. *Evolution by gene duplication*. Springer Science & Business Media, 2013.

Otto, S. P. and Yong, P. The evolution of gene duplicates. *Advances in genetics*, 46:451–483, 2002.

Panchy, N., Lehti-Shiu, M., and Shiu, S.-H. Evolution of gene duplication in plants. *Plant physiology*, 171(4):2294–2316, 2016.

Reams, A. B., Kofoid, E., Kugelberg, E., and Roth, J. R. Multiple pathways of duplication formation with and without recombination (reca) in salmonella enterica. *Genetics*, 192(2):397–415, 2012.

Rutten, J. P., Hogeweg, P., and Beslon, G. Adapting the engine to the fuel: mutator populations can reduce the mutational load by reorganizing their genome structure. *BMC evolutionary biology*, 19:1–17, 2019.

Stark, T. L., Liberles, D. A., Holland, B. R., and O'Reilly, M. M. Analysis of a mechanistic markov model for gene duplicates evolving under subfunctionalization. *BMC evolutionary biology*, 17(1):1–16, 2017.

Thompson, A., Zakon, H. H., and Kirkpatrick, M. Compensatory drift and the evolutionary dynamics of dosage-sensitive duplicate genes. *Genetics*, 202(2): 765–774, 2016.

Veitia, R. A. Gene duplicates: agents of robustness or fragility? *Trends in Genetics*, 33(6):377–379, 2017.

Vosseberg, J., van Hooff, J. J., Marcet-Houben, M., van Vlimmeren, A., van Wijk, L. M., Gabaldón, T., and Snel, B. Timing the origin of eukaryotic cellular complexity with ancient duplications. *Nature ecology & evolution*, 5 (1):92–100, 2021.

Walsh, B. Population-genetic models of the fates of duplicate genes. In *Origin and Evolution of New Gene Functions*, pages 279–294. Springer, 2003.

Woods, I. G., Wilson, C., Friedlander, B., Chang, P., Reyes, D. K., Nix, R., Kelly, P. D., Chu, F., Postlethwait, J. H., and Talbot, W. S. The zebrafish gene map defines ancestral vertebrate chromosomes. *Genome research*, 15(9): 1307–1314, 2005.

Zhang, J. Evolution by gene duplication: an update. *Trends in ecology & evolution*, 18(6):292–298, 2003.

Zhao, Y., Wang, J., Chen, J., Zhang, X., Guo, M., and Yu, G. A literature review of gene function prediction by modeling gene ontology. *Frontiers in genetics*, 11:400, 2020.