A Rigorous Framework to Classify the Post-Duplication Fate of Paralogous Genes (extended version)

Reza Kalhor, Guillaume Beslon, Manuel Lafond, Celine Scornavacca

A Rigorous Framework to Classify 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]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

July 15, 2024

Abstract:

Gene duplication has a central role in evolution; still, little is known on the fates of the duplicated copies, their relative frequency, and on how environmental conditions affect them. Moreover, the lack of rigorous definitions concerning the fate of duplicated genes hinders the development of a global vision of this process. In this paper we present a new framework aiming at characterizing and formally differentiating the fate of duplicated genes. Our framework has been tested via simulations, where the evolution of populations has been simulated using aevol, an *in silico* experimental evolution platform. Our results show several patterns that confirm some of the conclusions from previous studies, while also exhibiting new tendencies; this may open up new avenues to better understand the role of duplications as a driver of evolution.

The conference version of this work was presented at RECOMB-CG 2023, accessible at: https://doi.org/10.1007/978-3-031-36911-7_1.

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 will be described in detail in the next section: pseudogenization (loss of one gene), (double) neofunctionalization (divergence in function for both/one gene), conservation (maintenance of the original functions for both genes), subfunctionalization (division of the original functions between the two copies), and specialization (division of the original functions along with the acquisition of novel ones).

Little is known about whether certain fates occur more frequently than others and how environmental conditions influence their relative occurrence. Determining the fate of paralogous genes is challenging due to two primary factors. First, the functions of their common ancestor are often unknown, impeding the ability to foresee the evolutionary development of each copy. Additionally, even with knowledge of the ancestral functions, 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. In this context, our goal is to establish a comprehensive framework for formally describing the possible fates of duplicated genes, allowing for their differentiation through the analysis of phylogenetic data. Our method involves assessing the biological functions of both the original gene and its duplicates, creating a spectrum that captures the various possible fates along a continuum.

The majority of studies addressing this subject primarily focus on theoretical aspects and put forth statistical fate models for predictive purposes. One illustrative instance is the work by Lynch et al. (Lynch and Force, 2000; Lynch et al., 2001), which conceptualizes genes as discrete sets of functions. The authors introduce a population-based model of subfunctionalization, taking into account mutation rates in regulatory regions. Notably, their findings indicate that the likelihood of subfunctionalization diminishes towards 0 with larger population sizes. 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. Utilizing data from Drosophila, they demonstrate that neofunctionalization prevails as the primary fate, followed by conservation and specialization, with a limited occurrence of subfunctionalization. In (He and Zhang, 2005), the authors use d_N/d_S ratios and expression data to distinguish subfunctionalization and neofunctionalization. They assert that dichotomous fate models fall short in elucidating the diverse functional patterns exhibited by duplicate genes. This emphasizes the necessity for the development of classification methods that take into consideration 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). Practical investigations have explored neofunctionalization, observing its occurrence through changes in both the biological processes and transcriptional expression of a duplicate. 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 accomplished by comparing the fate of three species, identifying short regulatory motifs, and statistically correlating them with observed post-translational changes. Our framework strives to generalize the methodologies established in these experimental investigations. 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.

This paper fills this gap by employing aevol to simulate the progression of individual populations. Subsequently, we utilize our framework to categorize the duplications present in the simulated data. Our tests on aevol confirm the experimental studies on drosophilia data (Assis and Bachtrog, 2013), which showed that conservation/neofunctionalization were much more likely than subfunctionalization/specialization, while at the same time exhibiting proportions that differ from ours within these two groups of fates.

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 observable 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 can occur in our experiments, albeit rarely. 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.

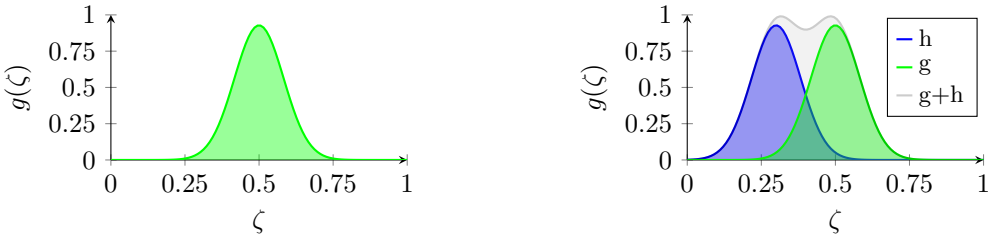
Specialization (SP): this fate occurs when the genes copies are able to perform the original functions, but *also* both develop novel functions. This differs from DN, since the original function is still performed, but also differs from SF because of the novel functions. The term was introduced in (Otto 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 Gaussian shape (note that this shape is merely for illustration, as our model applies to any shape). This gene expresses functions mainly in the range $[0.25, 0.75]$, and the expression of each function ζ in this range is the height of the shape at x-coordinate ζ (for instance in the figure on the left, $g(0.5)$ is approximately 0.9).



(a) A gene g.

(b) Two genes g and h.

Figure 1: An illustration of genes expressing functions in a Gaussian pattern. (a) shows the functions expressed by a gene. The x-axis represents the set of functions $[0, 1]$ and the y-axis the level of expression of each function. (b) shows two genes g and h and their function addition.

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. This is analogous to the area of the functional landscape

covered by g. If $\mathcal{F}$ is discrete, we define $contrib(g) = \sum_{\zeta \in \mathcal{F}} |g(\zeta)|$, and if $\mathcal{F}$ is continuous, we define $contrib(g) = \int_{\mathcal{F}} |g(\zeta)| d\zeta$.

- $i_{g|h}$ represents the function coverage of g by h, i.e. the proportion of functions of g that can be performed by h. For $contrib(g) > 0$, it is defined as

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

  For technical reasons, when $contrib(g) = 0$, we define $i_{g|h} = 1$ for any h. This corresponds to the idea that if g does nothing, any gene h can perform an empty set of functions.

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 $contrib(g) \neq contrib(h)$. These notions can be visualized in Figure (1.b): $[g + h]$ is the shape in grey that corresponds to the sum of the two Gaussians at each point; $[g \cap h]$ can be seen as the set of points formed by the overlap of the g and h shapes. Assuming that $contrib(g) = contrib(h) = 0.2$ and $contrib([g \cap h]) = 0.05$, we have $i_{g|h} = i_{h|g} = 0.05/0.2 = 0.25$.

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 $x \in [\tau, 1]$. In particular, if $x = 1$, then $\delta_\tau(x) = 1$. This is useful to model fates that require a threshold.

- $\rho \in [0, 0.5]$ is a *pseudogene threshold*, used to determine how much functionality a copied gene can preserve before starting to consider it as a pseudogene. For example, if $\rho = 0.2$, a gene is not considered as a pseudogene as long as it has not lost 4/5 of its functions, and from then the amount of P increases linearly the closer the proportion of retained functions gets to 0. We assume

that $\rho \leq 0.5$ since losing half of the functions of a gene can occur under subfunctionalization or specialization, and allowing $\rho > 0.5$ could confound P with SF or SP.

- $\nu \in [0, 1]$ 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, among other conditions, that the copied genes dedicate a quarter or more of their functions to the parental functions, and otherwise they are immediately 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.

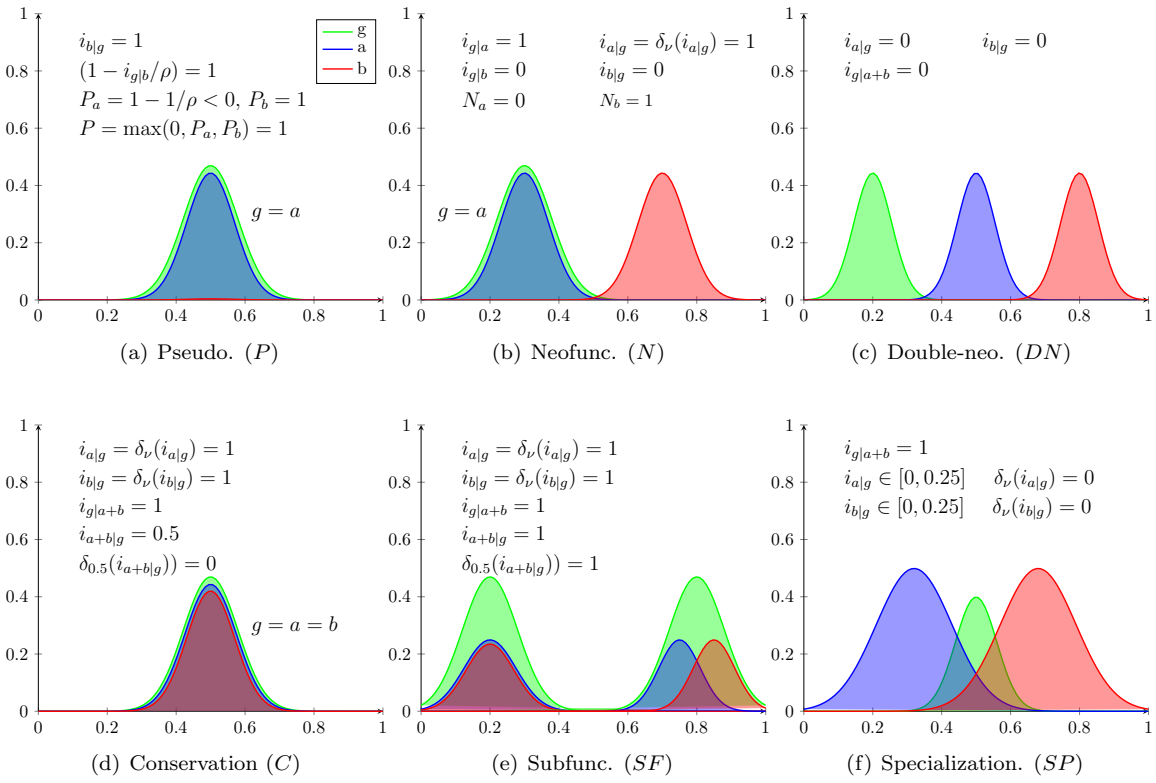
Fate	Formula
Pseudogenization (P)	$P_a = i_{a\|g} \cdot (1 - \frac{i_{g\|a}}{\rho})$ $P_b = i_{b\|g} \cdot (1 - \frac{i_{g\|b}}{\rho})$ $P = \max(0, P_a, P_b)$
Neofunc. (N)	$N_a = (1 - i_{a\|g}) \cdot \delta_\nu(i_{b\|g}) \cdot i_{g\|b}$ $N_b = (1 - i_{b\|g}) \cdot \delta_\nu(i_{a\|g}) \cdot i_{g\|a}$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

Each fate is illustrated in Figure 2, using the example of Gaussian gene functions. 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 provide an entirely formal framework for fate classification, these formulas were designed with the following two main criteria in mind:

Recognizability: when the gene copies have the exact behavior that is expected from a given fate, then our formulas assign 1 to that fate. This requires that each multiplicative factor present in a fate formula is 1.

Distinguishability: when one of the fates is clearly present and assigned 1, all the other fates are assigned 0. This requires the other fates to have at least one multiplicative factor equal to 0.



(a) Pseudo. (P) (b) Neofunc. (N) (c) Double-neo. (DN)

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

Figure 2: The canonical fates using the Gaussian representation. Note that in the case of SF and SP, $i_{g|a+b}$ is not exactly 1, but we assume that this is the case for the sake of simplicity. We assume thresholds $\rho = 0.2$ (relevant for P) and $\nu = 0.25$ (mostly relevant for SP). In (a), b has become a shape of height 0 and is considered as a pseudogene. In (e), SF can occur in two ways: a and b retain the same functions as g, but split their expression levels (left), or a and b split a portion of the functions of g (right).

For each fate, we recall the expected behavior and argue that both criteria are met by the formulas in Table 1. We also explain the rationale behind our usage of threshold functions.

- **Pseudogenization.** This fate occurs when at least one gene copy pseudogeneizes. Suppose, for instance, that b has become a pseudogene, in which case P_b should be 1 (as in Figure 2.(a)). The expected behavior for b is that it has not developed novel functions *and* has lost most of g's functions.

  Recognizability. When b does not perform novel functions outside of g, the $i_{b|g}$ factor is 1 since b is covered by g (note that, by definition, this also holds when $contrib(b) = 0$). Moreover, the factor $(1 - \frac{i_{g|b}}{\rho})$ is 1 when b does not perform any function of g, that is when the expression level of b becomes 0 as in Figure 2.(a). Hence $P_b = 1$ when both behavior occur in b.

  The same applies to a and P_a, and P is the maximum of 0, P_a and P_b, indicating the the P fate is predicted when at least one gene pseudogeneizes.

Distinguishability. Notice that every other fate includes the factor $(1 - P)$. This is because the more a gene has pseudogeneized, the less it should be considered for other fates. Hence when $P = 1$ is predicted, the other fates are assigned 0.

Although not strictly required for our criteria, note that the factor $(1 - \frac{i_{g|b}}{\rho})$ decreases linearly as b realizes more of g, and becomes 0 or less when $i_{g|b} \geq \rho$. Thus, when b performs at least a fraction of ρ of g, for example $\rho = 20\%$ of the original functions, the pseudogeneization fate for b is entirely discarded.

- **Neofunctionalization.** Suppose that a has retained the functions of g and that b has gained entirely novel functions, as in Figure 2.(b) (the case where a neofunctionalizes is symmetric). In this case N_b should be equal to 1.

  Recognizability. Since b only has novelty, none of its functions are covered by g and the factor $1 - i_{b|g}$ is 1. Moreover, and since a performs the same functions as g, they should be equal and the factors $i_{a|g}$ and $i_{g|a}$ are 1. This in turn implies that the factor $\delta_\nu(i_{a|g})$ is also 1. Hence, $N_b = 1$ in this case. Moreover, one can verify that $i_{g|a} = 1$ and $i_{b|g} = 0$ imply that $P = 0$, and the $(1 - P)$ factor is also 1. Hence, $N = N_b \cdot (1 - P) = 1$, as desired.

  Distinguishability. As already argued, $P = 0$. The fate DN is excluded since it has the factor $(1 - i_{a|g})$ whereas we assume $i_{a|g} = 1$, and C, SF are excluded since $i_{b|g} = 0$ implies $\delta_\nu(i_{b|g}) = 0$. Finally, SP is excluded since $1 - \delta_\nu(i_{a|g}) = 0$.

  We mention in passing that the factor $\delta_\nu(i_{a|g})$ in the formula for N_b, which says that a should preserve at least a fraction ν of g to even consider N_b, could not be replaced by a plain $i_{a|g}$. Indeed, in the scenario shown in Figure 2.(f), $SP = 1$ is achieved as desired. However, one can verify that using $i_{a|g}$ instead of $\delta_\nu(i_{a|g})$ in the formula for N would yield $N > 0$, preventing the distinguishability of SP.

- **Double-neo.** In this fate, the functions of a and b should be completely novel with respect to g (although a and b could have functions in common).

  Recognizability. When neither a nor b intersects with g, each of the factors $i_{a|g}, i_{b|g}, i_{g|a+b}$ is 0. Therefore, 1 minus any of these quantities, as used in the formula for DN, is 1. The factor $(1 - P)$ is 1 because $i_{a|g} = i_{b|g} = 0$.

  Distinguishability. The fate P is excluded since $P = 0$ as we just saw. The fate N is excluded since $i_{g|b} = i_{g|a} = 0$ under DN, thereby putting $N_a = N_b = N = 0$. Every other fate uses $i_{g|a+b}$, which is also 0.

- **Conservation.** Under this fate, both a and b should preserve the functions of g and, at the same time, should not have developed novelty.

  Recognizability. The factors $i_{a|g}$ and $i_{b|g}$ are 1 because their functions are the same as g and are thus covered by it. Therefore, $\delta_\nu(i_{a|g})$ and $\delta_\nu(i_{b|g})$ are also 1. The factor $i_{g|a+b}$ is 1 because g is realized by both a and b and thus also by $[a+b]$. Since $[a+b]$ doubles each of g's functions, half of $[a+b]$ must be covered by g. The $1-\delta_{0.5}(i_{a+b|g})$ factor is therefore equal to 1 when at most half of $[a+b]$ is covered by g (and decreases linearly in the interval $i_{a+b|g} \in [0.5, 1]$). One can check that the factor $(1-P)$ is 1 since $i_{g|a} = i_{g|b} = 1$.

  Distinguishability. Since $i_{a|g} = i_{b|g} = \delta_\nu(i_{a|g}) = \delta_\nu(i_{b|g}) = 1$, the fate C can be distinguished from N, DN, SP since those have a factor that is 1 minus one of these values. Importantly, C is only distinguished from SF because the latter has the factor $\delta_{0.5}(i_{a+b|g})$, which is 0 when $i_{a+b|g} = 0.5$.

- **Subfunctionalization.** In this fate, a and b must split the functions of g, and perform exactly those functions together. Thus $[a+b]$ should be equal to g (as opposed to C where $[a+b]$ is the double of g). Note that the degree of tolerable sharing is determined by ρ. For example, if $\rho = 0.2$, and a and b perform 0.8 and 0.2 of g, respectively, then no pseudogeneization is detected. However, if these proportions change to 0.99 and 0.01, then $P_b = P$ will have a much higher weight than SF (since the latter has the factor $(1-P)$).

  Recognizability. When SF occurs, a, b are covered by g and g is covered by $[a+b]$, and thus the factors $\delta_\nu(i_{a|g}), \delta_\nu(i_{b|g}), i_{g|a+b}$ are 1, as in the C fate. Also, $i_{a+b|g} = 1$ under SF, which implies $\delta_{0.5}(i_{a+b|g}) = 1$. Finally, assuming that both a and b realize at least ρ of g, $1-P = 1$ since $i_{g|a}/\rho = i_{g|b}/\rho = 1$.

  Distinguishability. The fate SF is separated from N, DN, SP for the same reasons as C, and is separated from C because of $\delta_{0.5}(i_{a+b|g}) = 1$.

- **Specialization**: In this last fate, g should be performed by a and b together, but a and b should also both develop "enough" novel functions, which is determined by the threshold ν.

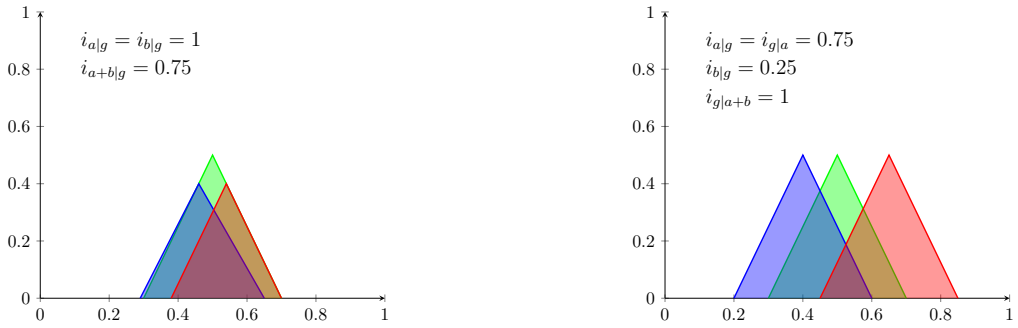
  Recognizability. The factor $i_{g|a+b}$ is 1 since a and b still realize g together. The factor $i_{a|g}$ can be seen as the proportion of non-novel functions of a. When this proportion is less than ν, a is considered to have enough novelty. We have $\delta_\nu(i_{a|g}) = 0$ when $i_{a|g} \leq \nu$, and therefore the factor $1-\delta_\nu(i_{a|g})$ is 1. The same holds for b. Finally, as in SF, assuming that a and b realize at least ρ of g, $1-P = 1$.

  Distinguishability. We just argued that $P = 0$ and it is thus excluded. We have $DN = 0$ because $i_{g|a+b} = 1$ and, since we assume that $\delta_\nu(i_{a|g}) = 0$ and $\delta_\nu(i_{b|g}) = 0$, the fates N, C, and SF will also be excluded.

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 Examples of hybrid fates

As previously mentioned, our framework allows the quantification of hybrid fates. Figure 3 illustrates two examples of such fates, which are similar to fates encountered in our experiments described later on. In these examples, we model genes as triangles, which makes the proportions easier to see and is the same representation used in our experiments.



(a) Sub. + Cons. (b) Neo. + Spec.

Figure 3: Two examples of hybrid fates. In (a) the hybrid fates are subfunctionalization and conservation while (b) shows a combination of neofunctionalization and specialization fates (see main text for details).

Figure 3.(a) illustrates a mix of SF and C. In the figure, a and b have both preserved a significant portion of g and have not developed new functions, which is the behavior of conservation, but not entirely since they also lost a portion of g. The copies also realize g together as in subfunctionalization, but this is not exactly SF since $a + b$ is much larger than g. Using threshold $\nu = 0.25$ and assuming that $i_{a+b|g} = 0.75$, we get $\delta_\nu(i_{a|g}) = \delta_\nu(i_{b|g}) = i_{g|a+b} = 1$ and $\delta_{0.5}(i_{a+b|g}) = 0.5$. In this case, we get a hybrid fate with $SF = 0.5$ and $C = 0.5$.

Figure 3.(b) displays a mix of N and SP whose sum-of-fates is less than 1. This fate is close to N since gene a maintained most of the functions of g whereas b mostly developed new functions. However,

it is not purely N because a has some new functions and b preserved some. In fact, it is also a mix of SP since the functions of g are covered by $[a + b]$, but not exactly because a is not novel enough. Plugging the numbers into our formulas using $\nu = 0.25$ yields $N = N_b = 0.375$ and $SP = 1/3$, whereas all other fates are 0.

3.3 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.

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** $Fates[X] = \frac{Fates[X]}{NbParalogies}$;

3.4 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 (a binary double-stranded circular sequence inspired from bacterial chromosome, see Figure 5, upper part) and a multi-steps Genotype-to-Phenotype map simulates transcription and translation to identify genes on the sequence, compute the phenotype that results from the interaction of the proteins encoded by these genes and ultimately compute the fitness of the organism (Figure 4). Importantly, this Genotype-to-Phenotype map is divided in two parts. (*i.*) The localization of the genes is based on the identification of signal subsequences on the genome that initiate translation (step 1 on Figure 4) and identify the mRNAs sequences. Then, on each mRNA the model searches for transcription initiation subsequences to identify the genes (step 2 on Figure 4). Hence, at the sequence level aevol mimics the structure of bacterial chromosomes, allowing for gene duplication. (*ii.*) At the functional levels (proteins and phenotype), Aevol switches to an abstract mathematical world in which biological traits are represented by a couple of values x, y with $x \in [0, 1]$ identifying the trait and $y \in [-1, 1]$ representing its inhibition (if negative) or activation (if positive) level. At this level, each gene is first transcribed into an amino-acid sequence through a simplified genetic code (step 3) and this sequence is folded to compute the mathematical function of the protein, i.e. the set of x values of this protein and the associated y values corresponding to their inhibition/activation (step 3'). For sake of simplicity, in the model all protein traits are represented by a triangle functions (see Figure 5.c), hence enabling fast computation of the protein functions from the amino-acid sequence. Finally, all triangle-proteins are linearly combined to compute the phenotype (step 4; the phenotype is a $[0, 1] \rightarrow [0, 1]$ function representing all the biological traits x resulting from the decoding of the genome – Figure 5.d). A population of such organisms replicate through 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 (step 5; see Figures 5.d and 6), the smaller the difference, the higher the fitness. Hence, the reproductive success of an organism depends on the adequacy of its phenotype function and the target function representing the environmental conditions. Finally, during replication, organisms may undergo various kinds of sequence mutations, including substitutions, Indels and chromosomal rearrangements (including inversions, duplications and deletions). Organisms are thus embedded into an evolutionary loop, enabling to study the relative effects of the different evolutionary forces on genome structure, genome sequence and gene repertoire.

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

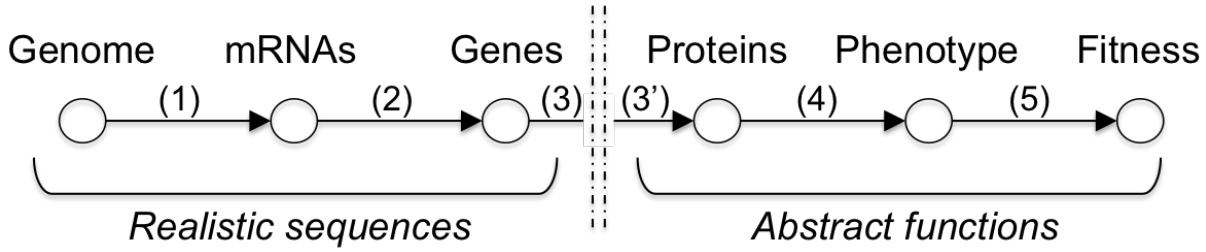


Figure 4: Overview of the Genotype-to-Phenotype map in aevol. Genomes are decoded into a phenotype through a multi-step process that first decodes the genome sequence to identify the genes (steps 1 and 2), then decodes the gene sequence to compute the protein functions (step 3 and 3'), the phenotype (step 4) and ultimately the fitness of the organism (step 5). See main text for details.

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; Banse et al., 2023), 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 the model two gene copies have exactly the same set of x, y pairs. Hence, the duplication of a given gene changes the phenotype through dosage effect (both copies contributing to the final phenotype), unless the genes sequences of one or both copies mutate. Finally, as each gene is decoded into a mathematical function, aevol enables a formal characterisation of genes functions, hence of the different possible fates of gene duplicates. Furthermore, the aevol platform has already – and successfully – be used as a benchmark to test bioinformatics methods (Biller et al., 2016) – and 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 6. We used environment (a) to generate "Wild-Type" genomes, that is initial genomes to be used in our experiments[2]. To this aim, we let several populations evolve independently for 1 million generations in this environment and extracted a single Wild-Type genome from each of each of them[3]. In aevol a specific parameter ($0 < w_{max} \leq 1$) enables tuning the maximum pleiotropy in the model (w_{max} sets the maximum range of functions to which proteins can contribute – in graphical terms, w_{max} corresponds to the maximum half-width of the base of protein triangle functions;

[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 5 has a 10,541 bp-long genome carrying 118 genes located on 50 mRNAs with a coding fraction of 77%.

[3]To choose this genome in the final population, we let the population evolve for further 100,000 generations. the wild-type genome is the genome from generation 1,000,000 that is the ancestor of the whole population at generation 1,100,000. This procedure enables extracting organisms that are well adapted to their environment (this "pre-evolution" step is required since evolution is heavily random in naive populations).

see Figure 5.c). 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 $w_{max} \in \{0.01, 0.1, 0.5, 1\}$ ($w_{max} = 1$ representing the maximum pleiotropy, where a gene can have an effect on all functions) 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 5 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.

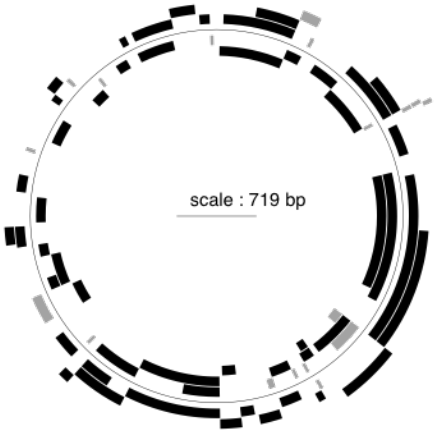
We used each generated wild-type as an initial genome for further 1 million generations of evolution in our four different environments. Note that, since wild-types are already adapted to environment (a), we expect very few duplications to occur in this environment. The other three environments range 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 (i.e. the individual which phenotype is the closest to the target function, hence which fitness is the highest). 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. The source code is available at `https://github.com/r3zakalhor/Post-Duplication-Fate-Framework`.

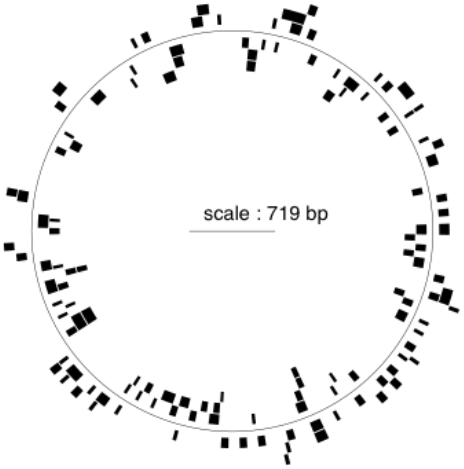
4 Results

4.1 Fates of duplication

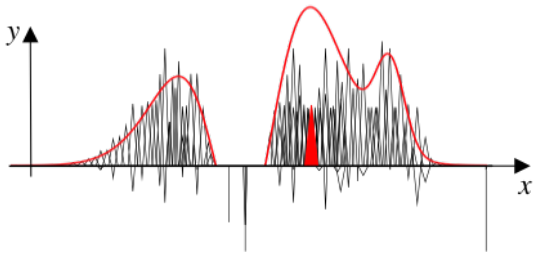
As explained above, starting from wild-types evolved in environment (a) with different maximum pleiotropic levels w_{max}, we simulated the evolution of 20 populations in 4 environments (ordered by increased variation compared to the environment of the wild-type) and for 1 million generations. We first 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



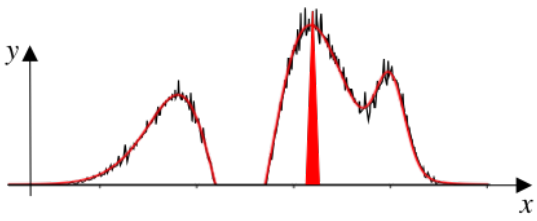
(a) Genome and mRNAs



(b) Genome and genes



(c) Protein functions



(d) Phenotypic function

Figure 5: 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 mRNAs (a) and genes (b) 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 (proteins and phenotype, in black) with one specific function highlighted in red. (c) Each triangle corresponds to a protein function which parameters are decoded from the sequence of a gene. Note the presence of function-activating/repressing proteins (positive/negative triangles respectively). (d) Organism's phenotype resulting from the sum of all protein functions. To illustrate the effect of gene duplication, one protein function has been highlighted in red on panel (c). This protein is not active enough to fit the target. However, as the corresponding gene exists in three copies on the genome, the overall effect is amplified through dosage effect (red triangle on panel d), hence increasing the contribution of this protein on the y axis without changing the set of functions it contributes to on the x axis.

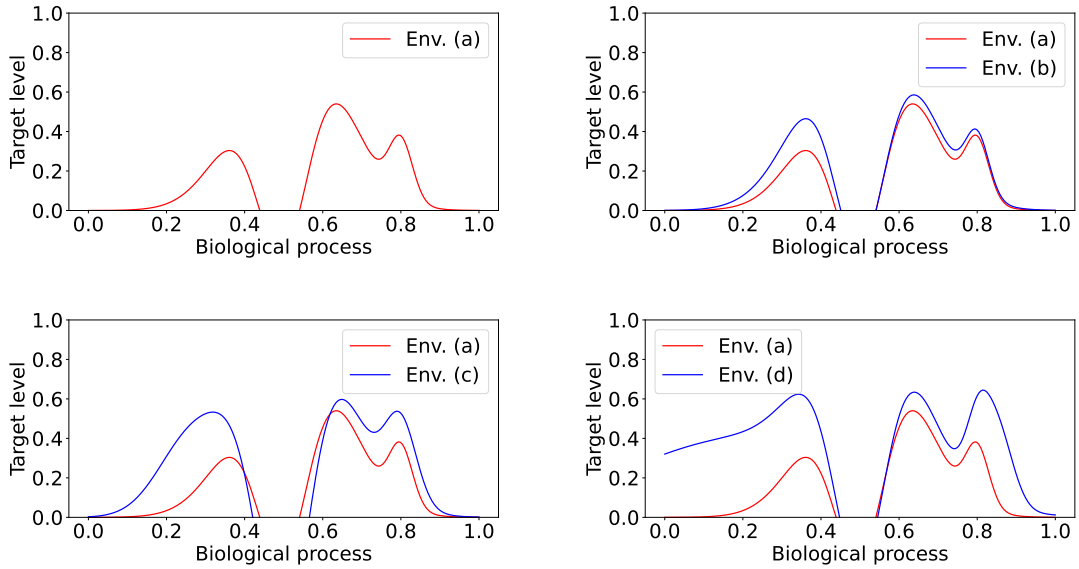


Figure 6: The four different environments used in the simulations. On the x-axis, we assume that the set of functions (biological processes) is the interval $[0, 1]$. The y-axis depicts the target level which, for each function, indicates the ideal amount of expression to survive in the environment.

environment. Recall that observed duplications are only those that result in at least one pair of 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	4.423	28.681	66.178	80.667

Table 2: Rate of observable 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	77.414	31.953	18.932	18.109

Table 3: Rate of observable 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	0.147	**0.318**	0.003	0.227	0.032	0.007	0.733	4.703
0.1	**0.333**	0.181	0.000	0.093	0.004	0.045	0.658	1.875
0.5	**0.433**	0.135	0.005	0.146	0.022	0.0.19	0.760	4.143
1	**0.304**	0.212	0.002	0.156	0.019	0.024	0.717	3.844
Environment (b)								
0.01	0.049	0.224	0.036	**0.439**	0.031	0.071	0.849	48.403
0.1	0.051	0.250	0.010	**0.498**	0.022	0.044	0.875	17.800
0.5	0.188	0.174	0.005	**0.444**	0.033	0.029	0.872	11.563
1	0.117	0.101	0.006	**0.530**	0.057	0.027	0.837	13.025
Environment (c)								
0.01	0.038	0.261	0.034	**0.420**	0.027	0.083	0.862	113.428
0.1	0.031	0.252	0.007	**0.421**	0.016	0.055	0.782	35.700
0.5	0.088	0.156	0.008	**0.607**	0.034	0.016	0.909	26.33
1	0.080	0.143	0.000	**0.547**	0.075	0.017	0.863	24.900
Environment (d)								
0.01	0.037	0.296	0.050	**0.358**	0.022	0.105	0.868	159.668
0.1	0.041	0.209	0.006	**0.475**	0.024	0.054	0.809	60.000
0.5	0.075	0.140	0.008	**0.610**	0.036	0.024	0.892	33.300
1	0.070	0.141	0.003	**0.579**	0.047	0.035	0.876	29.300

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]). The majority of the fates are 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 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. Notice that environment (a) has the lowest classification rate. This might be explained by the fact that in this setting, duplications

[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.

are less likely to contribute to fitness and may therefore undergo fates that are not as well-understood, in which case they do not correspond directly to the ones expected from the literature.

Several notable results can be observed from this table. When the organisms do not need to adapt in environment (a), P tends to be the dominant fate, even if we only detect observable pseudogenes. This is not surprising, again because duplicated genes are unlikely to contribute to fitness. The fact that the N fate dominates for $w_{max} = 0.01$ is likely because genes have limited individual effect, hence allowing their maintenance and their neofunctionalization, as discussed below. When the organisms must adapt to a new environment (b, c and d), the most frequent fate of duplications is C, followed by N. The high level of conservation may be attributed to the changes in the target levels in the fitness curves, as the new environments require dosage adaptation for genes function (see Figure 5 for an example of such effect). Indeed, one can see from Figure 6 that significant portions of the functional landscape require increased expression levels. This suggests that the genes performing these functions were duplicated and conserved for amplification, while other genes used neofunctionalization and performed the remaining extra functions required. Overall, the fates C, N, P are more frequent than SF, SP, and DN. This partially agrees with the findings of Assis and Bachtrog (2013), who found that $C, N > SF, SP$ on Drosophilia datasets. However, they found more N than C, warranting further investigation.

Moreover, we observe that the rate of N decreases as pleiotroty increases, which is not surprising. As w_{max} increases, the range of functions performed by an individual gene increases, hence the probability that one duplicate loses the ancestral function and acquires a new one decreases. A most striking result is the very low percentage of SF. 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 tends to increase 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. A similar reasoning applies to SP, which has low frequency in general, but has the inverse relationship with w_{max}. Perhaps this is because SF tends to take over the "sharing of functions" fates as pleiotropy increases, for the reasons mentioned above. Finally, DN is by far the rarest event, occurring in highest proportions when $w_{max} = 0.01$, which is expected as small pleiotropy makes it easier to diverge.

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). 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 7 illustrates the relationship between time and fate.

Env. \Bins	**100K**	**200K**	**300K**	**400K**	**500K**	**600K**	**700K**	**800K**	**900K**	**1000K**
(a)	164	67	41	45	79	119	68	60	61	50
(b)	2,207	127	145	99	94	73	67	66	209	129
(c)	5,529	218	410	97	94	110	94	77	74	246
(d)	6,570	213	90	85	92	81	107	63	49	62
Total	14,470	622	686	326	359	383	336	266	393	487

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, duplications 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 7, the fates C, N, and P remain largely dominant through most generations, which is to be expected from the results of the previous section. Notably, P is rare in the early generations since duplicates tend to be preserved for adaptation, but quickly sees a sharp increase as duplications start to introduce redundancy. Interestingly, 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 0.2 in the 800K bin to 0.55 in the 1000K bin, which is in line with the intuition that time is needed for divergence (and hence there may not be enough time for the duplicated genes to become P or N). Since

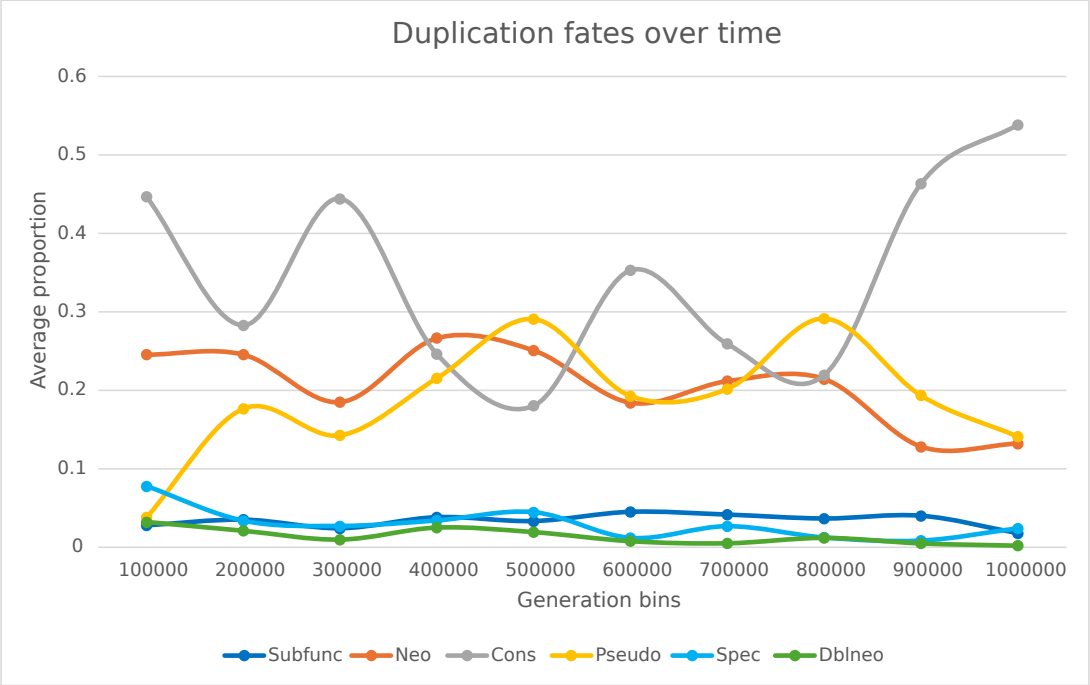


Figure 7: Average proportion of each duplication fate per generation bin.

all the predicted fates in our experiments sum to at most 1, the other fates see a decrease in these last 200K generations, which also explains why C appears to be inversely correlated with P and N.

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 16. The other columns report the average proportion of fates, for the pair of paralogs whose common ancestor has the number of descendant for the row.

Generally speaking, 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. It is worth noting that C decreases as the number of descendants increases. This is likely because when an ancestral gene

produces several pairs of paralogs, only a few of them may preserve the original function, as otherwise this would create overly high dosage effects. Therefore, even though a few pairs of paralogs may preserve the function, the number of paralogous pairs that do not tend to bring the proportion of C down. Also, DN seems to increase with the number of descendants (until 10, afterwards we do not have enough statistical power), probably because the successive duplications produce copies that are free to develop new functions. 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	14,044	0.0029	0.218	0.498	0.053	0.054	0.017	0.868
3	2,659	0.028	0.261	0.422	0.048	0.080	0.031	0.869
4	876	0.026	0.282	0.370	0.049	0.098	0.040	0.865
5	349	0.027	0.293	0.291	0.055	0.119	0.063	0.849
6	172	0.029	0.281	0.281	0.045	0.132	0.092	0.860
7	88	0.033	0.289	0.287	0.035	0.128	0.078	0.851
8	44	0.021	0.227	0.240	0.073	0.168	0.098	0.825
9	30	0.028	0.256	0.232	0.065	0.120	0.143	0.844
10	13	0.032	0.259	0.256	0.034	0.098	0.134	0.813
11	6	0.009	0.212	0.235	0.006	0.045	0.027	0.534
12	2	0.009	0.559	0.230	0.000	0.150	0.000	0.948
13	3	0.000	0.265	0.156	0.000	0.165	0.009	0.595
14	2	0.013	0.164	0.079	0.141	0.000	0.000	0.397
15	0	0.000	0.000	0.000	0.000	0.000	0.000	0.000
16	1	0.000	0.303	0.188	0.333	0.119	0.026	0.970

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 manuscript, we introduced a formal methodology for categorizing the fates of gene duplications based on the functions of existing paralogs and the ancestral gene. The aim is to offer the scientific community precise definitions and a mathematical toolkit for distinguishing between various fates. Indeed, without a comprehensive toolkit, the comparison of experimental and/or theoretical studies becomes extremely challenging, thereby restricting the potential for understanding gene duplication—a mechanism widely regarded as pivotal in molecular evolution. Our framework underwent thorough testing using

simulated data from aevol, an independently developed platform. Our tests confirmed several, but not all, 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 research opens several fields of research, spanning comparative genomics, phylogenetics, simulation, and, notably, evolutionary biology. While a substantial proportion of gene duplicates is classified (> 0.70 in all scenarios), indicating progress, it underscores the necessity for additional investigation into the unclassified fates. Furthermore, although not detailed in this report, we have identified a minor fraction of "hybrid fates" that warrant dedicated scrutiny. Lastly, it is important to acknowledge that our methodology, centered on extant paralog analysis, falls short of encompassing the complete spectrum 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 variation stems from the specific approach in choosing gene duplicates in this study. An exciting avenue for research involves expanding the P class to encompass the entire spectrum of pseudogeneization fates. Finally, we could use 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 employed aevol to evaluate our framework, demonstrating its ability to produce data resembling real-world observations. This encourages us to delve deeper into the exploration of gene duplications within the simulator. Notably, aevol not only provides the end results of organisms but also supplies information about past individuals and the precise gene phylogeny. This enables us to discern the precise trajectory of each gene along every branch, encompassing instances of 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, Conservation, Neofunctionalization and Pseudogenization are the most probable fates, immediately after the duplication events, the dominant fate is Conservation. Let us also note that it is likely that this dominance depends on the type of environmental variation as, in our experiments, the variation favors gene amplification. 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. The model also enables "*in silico* genetic engineering". We

intend to create a set of mutants by manually duplicating genes and allowing them to evolve. This approach will pave the way for a systematic examination of gene duplication within 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...). In aevol, WGD can easily be simulated by evolving wild-types with low-y target functions and then propagating these wild-types in a new environment where the target function is multiplied by two. Using this procedure, we can observe the fate of the duplicated genes immediately after the WGD but also after thousands or even millions of generations. We plan to test this exciting prospect in a near future. Another interesting avenue of research would be delving into the influence of regulation on the frequency of fates, particularly in relation to subfunctionalization. Although the current iteration of aevol does not incorporate regulation, there is an ongoing development of an extension to encompass the evolution of transcription factors. This extension aims to provide insights into how regulation might affect these aspects.

Finally, it would also be interesting to study how specific biological duplication 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.

Applying our framework to real data would require as input a set of genes and the knowledge of their functions and gene expressions. With the sequences at hand, we would need to align them and reconstruct a phylogeny from the gene family alignment, and infer duplication events via reconciliation tools (Jacox et al., 2016, for example). Then, using as input the gene phylogeny and the functions of the extant genes, the functions of ancestral genes could be predicted using tools such as PAINT (Gaudet et al., 2011) or PANTHER (Mi et al., 2017). Finally, ancestral gene expressions could be reconstructed by inferring ancestral transcriptomes, similarly to what done in (Mika et al., 2022). This would enable us to analyse real data, which we leave for future work.

Acknowledgments. The authors thank the anonymous reviewers for their helpful comments on the manuscript. We thank the ISEM platform and the Alliance Canada platform for the computing facilities.

Authorship confirmation. All of the authors contributed to every step of the research: devising the classification approach, implementing the approach, performing the simulations, analyzing the data, and writing the manuscript.

Authors' disclosure. The authors declare that they have no competing interest.

Funding statement. This work was supported by French Agence Nationale de la Recherche through the CoCoAlSeq project (ANR-19-CE45-0012), and by the Natural Sciences and Engineering Research Council of Canada - Alliance Catalyst (ALLRP 578495-22).

References

Ascencio, D., Diss, G., Gagnon-Arsenault, I., et al. Expression attenuation as a mechanism of robustness against gene duplication. *Proceedings of the National Academy of Sciences*, 118(6):e2014345118, 2021. doi: 10.1073/pnas.2014345118.

Assis, R. and Bachtrog, D. Neofunctionalization of young duplicate genes in drosophila. *Proceedings of the National Academy of Sciences*, 110(43):17409–17414, 2013. doi: 10.1073/pnas.1313759110.

Banse, P., Luiselli, J., Parsons, D. P., et al. Forward-in-time simulation of chromosomal rearrangements: The invisible backbone that sustains long-term adaptation. *Molecular Ecology*, 2023. doi: 10.1111/mec.17234. [Online] Available at: `https://onlinelibrary.wiley.com/doi/full/10.1111/mec.17234` (Accessed on 11 December 2023).

Batut, B., Parsons, D. P., Fischer, S., et al. In silico experimental evolution: a tool to test evolutionary scenarios. In *BMC bioinformatics*, volume 14, pages 1–11. Springer, 2013. doi: 10.1186/1471-2105-14-S15-S11.

Biller, P., Knibbe, C., Beslon, G., et al. 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. doi: 10.1007/978-3-319-40189-8_4.

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. doi: 10.1093/plcell/koac076.

Brunet, F. G., Crollius, H. R., Paris, M., et al. Gene loss and evolutionary rates following whole-genome duplication in teleost fishes. *Molecular biology and evolution*, 23(9):1808–1816, 2006. doi: 10.1093/molbev/msl049.

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. doi: 10.1073/pnas.0906222107.

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. doi: 10.1007/978-3-642-02008-7_4.

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. doi: 10.1038/nrg2482.

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. doi: 10.1146/annurev.genom.8.021307.110233.

Darby, C. A., Stolzer, M., Ropp, P. J., et al. Xenolog classification. *Bioinformatics*, 33(5):640–649, 2017. doi: 10.1093/bioinformatics/btw686.

Demuth, J. P. and Hahn, M. W. The life and death of gene families. *Bioessays*, 31(1):29–39, 2009. doi: 10.1002/bies.080085.

Diao, J., Stark, T. L., Liberles, D. A., et al. Level-dependent qbd models for the evolution of a family of gene duplicates. *Stochastic Models*, 36(2):285–311, 2020. doi: 10.1080/15326349.2019.1680296.

Force, A., Lynch, M., Pickett, F. B., et al. Preservation of duplicate genes by complementary, degenerative mutations. *Genetics*, 151(4):1531–1545, 1999. doi: 10.1093/genetics/151.4.1531.

Frénoy, A., Taddei, F., and Misevic, D. Genetic architecture promotes the evolution and maintenance of cooperation. *PLoS computational biology*, 9(11):e1003339, 2013. doi: 10.1371/journal.pcbi.1003339.

Gaudet, P., Livstone, M. S., Lewis, S. E., et al. Phylogenetic-based propagation of functional annotations within the gene ontology consortium. *Briefings in bioinformatics*, 12(5):449–462, 2011. doi: 10.1093/bib/bbr042.

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. doi: 10.1073/pnas.0409186102.

Gu, Z., Rifkin, S. A., White, K. P., et al. Duplicate genes increase gene expression diversity within and between species. *Nature genetics*, 36(6):577–579, 2004. doi: 10.1038/ng1355.

Guillaume, F. and Otto, S. P. Gene functional trade-offs and the evolution of pleiotropy. *Genetics*, 192 (4):1389–1409, 2012. doi: 10.1534/genetics.112.143214.

Hahn, M. W. Distinguishing among evolutionary models for the maintenance of gene duplicates. *Journal of Heredity*, 100(5):605–617, 2009. doi: 10.1093/jhered/esp047.

He, X. and Zhang, J. Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution. *Genetics*, 169(2):1157–1164, 2005. doi: 10.1534/genetics.104.037051.

Hindré, T., Knibbe, C., Beslon, G., et al. New insights into bacterial adaptation through in vivo and in silico experimental evolution. *Nature Reviews Microbiology*, 10(5):352–365, 2012. doi: 10.1038/nrmicro2750.

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. doi: 10.1101/gr.2705204.

Jacox, E., Chauve, C., Szöllősi, G. J., et al. eccetera: comprehensive gene tree-species tree reconciliation using parsimony. *Bioinformatics*, 32(13):2056–2058, 2016. doi: 10.1093/bioinformatics/btw105.

Jaillon, O., Aury, J.-M., Brunet, F., et al. Genome duplication in the teleost fish tetraodon nigroviridis reveals the early vertebrate proto-karyotype. *Nature*, 431(7011):946–957, 2004. doi: 10.1038/nature03025.

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. [Online] Available at: `https://theses.hal.science/tel-00482375/` (Accessed on 10 May 2010).

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. doi: 10.7551/978-0-262-32621-6-ch006.

Knibbe, C., Coulon, A., Mazet, O., et al. A long-term evolutionary pressure on the amount of noncoding dna. *Molecular biology and evolution*, 24(10):2344–2353, 2007a. doi: 10.1093/molbev/msm165.

Knibbe, C., Mazet, O., Chaudier, F., et al. Evolutionary coupling between the deleteriousness of gene mutations and the amount of non-coding sequences. *Journal of Theoretical Biology*, 244(4):621–630, 2007b. doi: 10.1016/j.jtbi.2006.09.005.

Kuzmin, E., Taylor, J. S., and Boone, C. Retention of duplicated genes in evolution. *Trends in Genetics*, 2021. doi: 10.1016/j.tig.2021.06.016.

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. doi: 10.1093/bioinformatics/bty 242.

Liard, V., Parsons, D. P., Rouzaud-Cornabas, J., et al. The complexity ratchet: Stronger than selection, stronger than evolvability, weaker than robustness. *Artificial life*, 26(1):38–57, 2020. doi: 10.1162/artl _a_00312.

Lynch, M. and Conery, J. S. The evolutionary fate and consequences of duplicate genes. *science*, 290 (5494):1151–1155, 2000. doi: 10.1126/science.290.5494.1151.

Lynch, M. and Force, A. The probability of duplicate gene preservation by subfunctionalization. *Genetics*, 154(1):459–473, 2000. doi: 10.1093/genetics/154.1.459.

Lynch, M., O'Hely, M., Walsh, B., et al. The probability of preservation of a newly arisen gene duplicate. *Genetics*, 159(4):1789–1804, 2001. doi: 10.1093/genetics/159.4.1789.

Mi, H., Huang, X., Muruganujan, A., et al. 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. doi: 10.1093/nar/gkw1138.

Mika, K., Whittington, C. M., McAllan, B. M., et al. Gene expression phylogenies and ancestral transcriptome reconstruction resolves major transitions in the origins of pregnancy. *eLife*, 11:e74297, jun 2022. ISSN 2050-084X. doi: 10.7554/eLife.74297.

Nguyen Ba, A. N., Strome, B., Hua, J. J., et al. Detecting functional divergence after gene duplication through evolutionary changes in posttranslational regulatory sequences. *PLoS computational biology*, 10(12):e1003977, 2014. doi: 10.1371/journal.pcbi.1003977.

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. doi: 10.1006/scdb.1999.0332.

Ohno, S. *Evolution by gene duplication*. Springer Science & Business Media, 2013. doi: 10.1007/978-3-642-86659-3.

Otto, S. P. and Yong, P. The evolution of gene duplicates. *Advances in genetics*, 46:451–483, 2002. doi: 10.1007/s12041-013-0212-8.

Panchy, N., Lehti-Shiu, M., and Shiu, S.-H. Evolution of gene duplication in plants. *Plant physiology*, 171(4):2294–2316, 2016. doi: 10.1104/pp.16.00523.

Reams, A. B., Kofoid, E., Kugelberg, E., et al. Multiple pathways of duplication formation with and without recombination (reca) in salmonella enterica. *Genetics*, 192(2):397–415, 2012. doi: 10.1534/genetics.112.142570.

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. doi: 10.1186/s12862-019-1507-z.

Stark, T. L., Liberles, D. A., Holland, B. R., et al. Analysis of a mechanistic markov model for gene duplicates evolving under subfunctionalization. *BMC evolutionary biology*, 17(1):1–16, 2017. doi: 10.1186/s12862-016-0848-0.

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. doi: 10.1534/genetics.115.178137.

Veitia, R. A. Gene duplicates: agents of robustness or fragility? *Trends in Genetics*, 33(6):377–379, 2017. doi: 10.1016/j.tig.2017.03.006.

Vosseberg, J., van Hooff, J. J., Marcet-Houben, M., et al. Timing the origin of eukaryotic cellular complexity with ancient duplications. *Nature ecology & evolution*, 5(1):92–100, 2021. doi: 10.1038/s41559-020-01320-z.

Walsh, B. Population-genetic models of the fates of duplicate genes. In *Origin and Evolution of New Gene Functions*, pages 279–294. Springer, 2003. doi: 10.1023/A:1024194802441.

Woods, I. G., Wilson, C., Friedlander, B., et al. The zebrafish gene map defines ancestral vertebrate chromosomes. *Genome research*, 15(9):1307–1314, 2005. doi: 10.1101/gr.4134305.

Zhang, J. Evolution by gene duplication: an update. *Trends in ecology & evolution*, 18(6):292–298, 2003. doi: 10.1016/S0169-5347(03)00033-8.

Zhao, Y., Wang, J., Chen, J., et al. A literature review of gene function prediction by modeling gene ontology. *Frontiers in genetics*, 11:400, 2020. doi: 10.3389/fgene.2020.00400.