Theoretical Population Biology 93 (2014) 38–51

Contents lists available at ScienceDirect

Theoretical Population Biology

journal homepage: www.elsevier.com/locate/tpb

On the retention of gene duplicates prone to dominant deleterious mutations

Giulia Malaguti, Param Priya Singh, Hervé Isambert *

Institut Curie, CNRS-UMR168, UPMC, 26 rue d'Ulm, 75005 Paris, France

ARTICLE INFO

Article history:
Received 15 October 2013
Available online 12 February 2014

Keywords:
Population genetics models
Whole-genome duplication
Small scale duplication
Dominant deleterious mutations
Purifying selection

ABSTRACT

Recent studies have shown that gene families from different functional categories have been preferentially expanded either by small scale duplication (SSD) or by whole-genome duplication (WGD). In particular, gene families prone to dominant deleterious mutations and implicated in cancers and other genetic diseases in human have been greatly expanded through two rounds of WGD dating back from early vertebrates. Here, we strengthen this intriguing observation, showing that human oncogenes involved in different primary tumors have retained many WGD duplicates compared to other human genes. In order to rationalize this evolutionary outcome, we propose a consistent population genetics model to analyze the retention of SSD and WGD duplicates taking into account their propensity to acquire dominant deleterious mutations. We solve a deterministic haploid model including initial duplicated loci, their retention through sub-functionalization or their neutral loss-of-function or deleterious gain-of-function at one locus. Extensions to diploid genotypes are presented and population size effects are analyzed using stochastic simulations. The only difference between the SSD and WGD scenarios is the initial number of individuals with duplicated loci. While SSD duplicates need to spread through the entire population from a single individual to reach fixation, WGD duplicates are *de facto* fixed in the small initial post-WGD population arising through the ploidy incompatibility between post-WGD individuals and the rest of the pre-WGD population. WGD duplicates prone to dominant deleterious mutations are then shown to be indirectly selected through purifying selection in post-WGD species, whereas SSD duplicates typically require positive selection. These results highlight the long-term evolution mechanisms behind the surprising accumulation of WGD duplicates prone to dominant deleterious mutations and are shown to be consistent with cancer genome data on the prevalence of human oncogenes with WGD duplicates.

1. Introduction

Gene duplication has long been recognized as a major source of genetic innovation in the course of evolution through the retention and divergence of specific gene duplicates arising by chance (Ohno, 1970; Holland et al., 1994; Sidow, 1996). Gene duplicates are also thought to confer some mutational robustness against loss-of-function mutations (Winzeler et al., 1999; Gu et al., 2003; Kamath et al., 2003; Gu, 2003). Conversely, however, the duplication of genes prone to dominant deleterious mutations, such as gain-of-function mutations, is expected to lead to an enhanced susceptibility to genetic diseases and, hence, be opposed by purifying selection (Furney et al., 2006; Blekhman et al., 2008; Cai et al., 2009). Yet, surprisingly, such "dangerous" gene families prone to dominant deleterious mutations have often been greatly expanded by duplication in the course of evolution, see *e.g.* Ise et al. (2000) and Esteban et al. (2001).

In particular, gene families frequently implicated in cancer and other genetic diseases in vertebrates have been greatly expanded through two rounds of whole-genome duplication (WGD) dating back from the onset of jawed vertebrates (Singh et al., 2012). By contrast, gene families lacking such a susceptibility to dominant deleterious mutations have been more typically expanded through small scale duplication (SSD) (Singh et al., 2012). More generally, gene duplicates originated from SSD or WGD events have been shown to exhibit antagonist retention patterns, with gene families expanded through WGD having typically few additional SSD genes and, vice versa, for gene families expanded mostly through SSD which exhibit few additional retained duplicates from WGD (Makino and McLysaght, 2010; Huminiecki and Heldin, 2010; Singh et al., 2012). This implies that the mode of duplication through SSD or WGD events directly impacts the selection process

* Corresponding author.
E-mail address: herve.isambert@curie.fr (H. Isambert).

http://dx.doi.org/10.1016/j.tpb.2014.01.004
0040-5809/

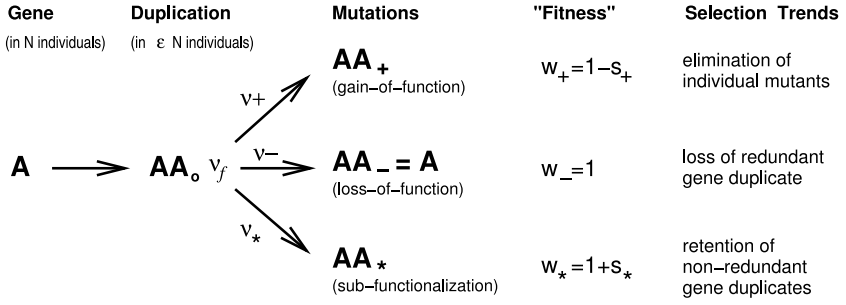

Fig. 1. *Haploid model for the retention of gene duplicates*. The model consists of two initial duplicated loci $AA_\circ$ in a haploid population with mutation rates (ν_i) towards deleterious gain-of-function mutants (AA_+ with $w_+ = 1 - s_+ < 1$), neutral loss-of-function mutants at a single locus (AA_- with $w_- = 1$) and neutral or beneficial fixed duplicates through sub-functionalization ($AA_\star$ with $w_\star = 1 + s_\star \geqslant 1$). The only difference between the two duplication scenarios is the initial fraction ϵ of individuals with duplicated loci, which is $\epsilon \simeq 1/N$ in the post-SSD population of size N, while it is $\epsilon \simeq 1$ in the post-WGD population arising through WGD-induced speciation. This WGD-induced speciation results from the ploidy incompatibility between post-WGD individuals and the rest of the pre-WGD population. See main text for model details.

of gene duplicates. Hence, their retention cannot be explained by the same *ad hoc* selection mechanism independent of the SSD or WGD modes of duplication.

These different retentions of SSD and WGD duplicates have been frequently associated to dosage balance constraints (Birchler et al., 2001; Veitia, 2002; Papp et al., 2003; Aury et al., 2006; Makino and McLysaght, 2010). However, extensive statistical analysis combining multiple properties of human genes (such as dosage balance constraints, association to cancers and genetic diseases and expression levels) have recently demonstrated (Singh et al., 2012) that the retention of WGD duplicates in vertebrates is more directly related to their susceptibility to dominant deleterious mutations than to dosage balance constraints or expression levels.

In this paper, we further strengthen this observation, showing that human oncogenes involved in different primary tumors have retained many WGD duplicates as compared to other human genes. This intriguing observation on the different retention patterns of WGD and SSD duplicates calls for a consistent population genetics model taking into account their propensity to acquire dominant deleterious mutations. To this end, we propose such a model focusing first on a simple, analytically tractable approach valid for large population sizes, before resorting to numerical simulations to analyze the consequences of stochastic fluctuations arising from finite population size. In particular, in order to analyze the retention of SSD *versus* WGD duplicates, we first use a simple deterministic model of two duplicated loci with neutral fixable genotypes in a haploid population of fixed size N and uncoupled mutation/selection dynamics. The only difference between SSD and WGD scenarios concerns the initial condition for each mode of duplication: the SSD case corresponds to a gene duplication in the genome of a *single* individual in the initial population, while the WGD case implies the genome duplication of *all* individuals in the small initial population arising through WGD. This is because WGD also induces a speciation event due to the ploidy incompatibility of the post-WGD individuals with the rest of the pre-WGD population. Although simplified, the asymptotic solutions of this deterministic population genetics model allow to capture the main evolutionary process responsible for the different retention of SSD *versus* WGD duplicates caused by dominant deleterious mutations. This haploid model is also extended into a simplified diploid model with three neutral haplotypes and one dominant deleterious haplotype. Then, to go beyond deterministic solutions for large populations, we use the formalism of one-step-process master equations and stochastic simulations to analyze the effect of finite population sizes on the retention of SSD *versus* WGD duplicates. All in all, this population genetics model supports the idea that the enhanced retention of "dangerous" WGD duplicates prone to dominant deleterious mutations is an indirect consequence of the initial speciation events triggered by WGD and the ensuing purifying selection in post-WGD species.

These results are then compared to the retention biases of SSD *versus* WGD duplicates for gene families with oncogenic properties and responsible for a broad range of primary tumors in human. Our application to genomic data will focus on the example of human oncogenes for which increasing amounts of data have recently become available from large scale cancer genome sequencing studies. Yet, unlike typical models on cancer genomics, *e.g.* Michor et al. (2004), Merlo et al. (2006), Beerenwinkel et al. (2007) and Bozic et al. (2010), our analysis of driver mutations from cancer genome data will *not* aim at modeling the *in situ* proliferation and selection of tumor cells within healthy tissues. Instead, it will concern the long-term evolution mechanisms that favored the surprising retention of WGD duplicates prone to dominant deleterious mutations in vertebrate genomes.

2. Model

We model the fixation of gene duplicates following either a SSD or a WGD event. In the following, we will first assume an haploid deterministic model to limit the number of two-locus combinations and stochastic effects to be considered. Extensions to diploid models and stochastic effects due to finite population size will then be analyzed in some details. Finally, the analytical and numerical solutions of these deterministic models will be compared to simulations of the corresponding stochastic population genetics models.

2.1. Haploid model for SSD and WGD duplicate retention

We start from the duplication event $A \rightarrow AA$ in a haploid genome, assuming that the newly duplicate gene is initially functionally redundant (Force et al., 1999; Lynch and Force, 2000a; Lynch and Conery, 2000; Lynch et al., 2001). Therefore, we assign to the initial (unstable) genotype with two redundant duplicated loci $AA_\circ$ a neutral fitness parameter $w_\circ = 1$. Then, we will consider three possible mutation-selection scenarios, corresponding to the emergence of three different phenotypes from the initial genotype $AA_\circ$, with mutation rates ν_-, $\nu_\star$, ν_+, and fitness parameters w_-, $w_\star$ and w_+, Fig. 1. Classical models suggest three alternative outcomes in the evolution of duplicate genes (Force et al., 1999; Lynch and Force, 2000a; Lynch and Conery, 2000; Lynch et al., 2001; Zhang, 2003). (i) One copy may become silenced by the accumulation of degenerative mutations and eventually become non-functionalized, while the other (fully functional) copy is retained. In our model, this corresponds to a neutral phenotype due to a loss-of-function of one of the duplicates (AA_-) with neutral fitness $w_- = 1$. (ii) Both copies may be reciprocally preserved

through the fixation of complementary loss-of-subfunction mutations, which results in a partitioning of the tasks of the ancestral gene. This means that if duplicate genes loose different subfunctions, then they must complement each other by jointly retaining the original function. In our model, this corresponds to a neutral or possibly beneficial phenotype corresponding to the retention of two non-equivalent duplicated loci through sub-functionalization ($AA_\star$) with a resulting fitness $w_\star = 1 + s_\star \geqslant 1$. (iii) One copy may acquire a novel beneficial function while the other retains the original function, but the incidence of beneficial mutations is negligible compared to the frequency of other mutations and for this reason we do not include this possibility in our model. In addition to these classical scenarios, we specifically address the case of gain-of-function mutations, that lead to an enhanced activity of the gene and are associated to a dominant deleterious phenotype. This is in particular the case of oncogenes and genes with autoinhibitory protein folds. Dominant deleterious mutations drastically reduce the fitness of the individual and correspond in our model to a deleterious phenotype due to constitutive gain-of-function mutations on one of the duplicates (AA_+) with a fitness decrement $w_+ = 1 - s_+ < 1$, Fig. 1. Sub-functionalization of the duplicated loci ($AA_\star$) implies, in principle, mutations at both loci, which can no longer individually perform all functions of the ancestral gene A (Hughes, 1994; Lynch and Force, 2000b). By contrast, loss-of-function (AA_-) and gain-of-function (AA_+) genotypes are assumed to retain a functional copy of the ancestral gene A. However, while this functional copy can mask the deleterious effect of loss-of-function mutations in AA_-, resulting in neutral fitness $w_- = 1$, it is unable to mask the deleterious effect of gain-of-function mutations in AA_+, resulting in a fitness decrement, $w_+ = 1 - s_+ < 1$. For simplicity in this haploid model, we will not distinguish on which copy the loss-of-function and gain-of-function mutations occur, assuming, in particular, that the loss-of-function mutations on either duplicate copy are equivalent to the ancestral genotype with a single gene copy, $AA_- \equiv A\_A \equiv A$.

2.2. Deterministic solutions for neutral sub-functionalization

Let us first illustrate the main theoretical results of this paper, using a simple deterministic model for the fixation of SSD *versus* WGD duplicates assuming that the two fixable genotypes, AA_- and $AA_\star$, have a neutral fitness, $w_- = 1$ and $w_\star = 1$, respectively (*i.e.* AA_+ with $w_+ < 1$ cannot be fixed in the limit of large population, $N \to \infty$). We will further assume, in this section, a population genetics model with an uncoupled mutation/selection dynamics. While somewhat simplified, this deterministic haploid model allows to capture the main evolutionary process responsible for the differential retention of SSD *versus* WGD duplicates caused by deleterious gain-of-function mutations. Besides, as we will show in the subsequent Sections 2.3 and 2.4, more advanced population genetics models, including more realistic diploid genotypes, or stochastic effects due to finite population size, exhibit in fact very similar asymptotic solutions in the limit of large populations.

In the following, we note $\phi_\circ(t), \phi_+(t), \phi_-(t)$ and $\phi_\star(t)$ the fractions of individuals in the population with the corresponding genotypes for the duplicated loci, $AA_\circ, AA_+, AA_-$ and $AA_\star$. We then write down the simplest noiseless population genetics model linking these genotypes with uncoupled mutation/selection dynamics as,

$$\begin{aligned}
d_t\phi_\circ &= (w_\circ - \bar{w})\phi_\circ - (\nu_+ + \nu_- + \nu_\star)\phi_\circ \\
d_t\phi_+ &= (w_+ - \bar{w})\phi_+ + \nu_+\phi_\circ \\
d_t\phi_- &= (w_- - \bar{w})\phi_- + \nu_-\phi_\circ \\
d_t\phi_\star &= (w_\star - \bar{w})\phi_\star + \nu_\star\phi_\circ
\end{aligned} \tag{1}$$

where $\bar{w}(t) = \sum_i w_i\phi_i(t)$ is the average fitness of the population. In particular, noting $S = \sum_i \phi_i$, one can check that $d_tS = \bar{w}(1-S)$ leads to the expected constant, $S(t) = 1$ at all time, providing that $S(t=0) = 1$ is taken as initial condition. Using $\bar{w} = \sum_i w_i\phi_i = 1 - s_+\phi_+$ in the case of neutral fixable genotypes for the duplicated loci ($w_\star = w_- = 1$), this system can be expressed as

$$\begin{aligned}
d_t\phi_\circ &= s_+\phi_+\phi_\circ - (\nu_+ + \nu_- + \nu_\star)\phi_\circ \\
d_t\phi_+ &= (s_+\phi_+ - s_+)\phi_+ + \nu_+\phi_\circ \\
d_t\phi_- &= s_+\phi_+\phi_- + \nu_-\phi_\circ \\
d_t\phi_\star &= s_+\phi_+\phi_\star + \nu_\star\phi_\circ.
\end{aligned} \tag{2}$$

The solutions for $\phi_\circ(t)$, $\phi_-(t)$ and $\phi_\star(t)$ can thus be expressed in terms of a time integral $\Phi_+(t)$ of the fraction $\phi_+(t)$ of the population with deleterious gain-of-function mutations at the duplicated loci, as

$$\begin{aligned}
\phi_\circ(t) &= \epsilon e^{-\nu_f t}\Phi_+(t) \\
\phi_-(t) &= \left(\frac{\epsilon\nu_-}{\nu_f}\left(1 - e^{-\nu_f t}\right) + 1 - \epsilon\right)\Phi_+(t) \\
\phi_\star(t) &= \frac{\epsilon\nu_\star}{\nu_f}\left(1 - e^{-\nu_f t}\right)\Phi_+(t) \\
\Phi_+(t) &= \exp\left(\int_0^t s_+\phi_+(t')dt'\right)
\end{aligned}$$

where $\nu_f = \nu_+ + \nu_- + \nu_\star$ is the total rate of mutations with functional effect (*i.e.*, gain- or loss-of-function or sub-functionalization) and $\epsilon = \phi_\circ(0)$ is the initial fraction of individuals in the population with duplicated loci, $AA_\circ$. The remaining individuals present only a single functional locus, A, which is assumed to be equivalent to the loss-of-function mutation, AA_-, at either duplicated locus, *i.e.* $\phi_-(0) = 1 - \phi_\circ(0) = 1 - \epsilon$, whereas $\phi_+(0) = 0$ and $\phi_\star(0) = 0$. As a WGD event leads to a concomitant speciation event due to the ploidy incompatibility with pre-WGD individuals, it implies that all individuals of the post-WGD population have a duplicated genome, corresponding to the case $\epsilon = 1$. By contrast, a SSD event does not typically lead to a speciation leaving a single individual with one (or a few) duplicated gene(s) in the post-SSD population corresponding to $\epsilon \simeq 1/N \ll 1$. Note that ϵ is also the expected fixation rate in absence of mutation, if all fixable genotypes are neutral, $\Pi_e = \epsilon$. Hence, using the asymptotic condition, $\phi_-(\infty) + \phi_\star(\infty) = 1$, for the fractions of individuals with the only fixable genotypes in the large population size limit, corresponding to the loss of one duplicate (AA_-) or the retention of both duplicates through sub-functionalization ($AA_\star$), we obtain,

$$\Phi_+(\infty) = \frac{\nu_f}{\nu_f - \epsilon\nu_+}$$

and thus the asymptotic fraction of sub-functionalized duplicated loci becomes,

$$\phi_\star(\infty) = \frac{\epsilon\nu_\star}{\nu_f - \epsilon\nu_+}.$$

Note, that the same result is obtained if the fitness parameters are rescaled by the average fitness, $w_i \to w_i/\bar{w}$, which only affects transient regimes but not asymptotic distributions.

For neutral fixable genotypes, AA_- and $AA_\star$ ($w_- = w_\star = 1$), $\phi_\star(\infty)$ corresponds to the expected fixation rate of $AA_\star$ in the population by coalescence, $\Pi_\star = \phi_\star(\infty)$. Thus, we obtain the following expressions for SSD duplicates with $\Pi_e^{\mathrm{SSD}} = \epsilon = 1/N \ll 1$ and WGD duplicates with $\Pi_e^{\mathrm{WGD}} = \epsilon = 1$,

$$\begin{aligned}
\Pi_\star^{\mathrm{SSD}} &\simeq \frac{\nu_\star}{\nu_f}\Pi_e^{\mathrm{SSD}} = \frac{\nu_\star}{\nu_+ + \nu_- + \nu_\star}\Pi_e^{\mathrm{SSD}} \\
\Pi_\star^{\mathrm{WGD}} &= \frac{\nu_\star}{\nu_f - \nu_+}\Pi_e^{\mathrm{WGD}} = \frac{\nu_\star}{\nu_- + \nu_\star}\Pi_e^{\mathrm{WGD}}.
\end{aligned} \tag{3}$$

Hence, the mutation rate, ν_+, leading to deleterious phenotypes with decreasing fitness ($w_+ < 1$) favors the elimination of "dangerous" duplicates after SSD events, as expected. However, the same mutation rate leading to deleterious phenotypes (ν_+) does *not* appear in the fixation rate of gene duplicates following a WGD-induced speciation event. It implies that the mechanism of purifying selection does *not* contribute to the elimination of "dangerous" duplicates in post-WGD populations following a WGD-induced speciation event ($\epsilon = 1$), unlike what happens in post-SSD populations without speciation ($\epsilon \ll 1$), as discussed in Singh et al. (2012).

Thus, from Eq. (3), we find a different fixation of duplicates through WGD and SSD events favoring, somewhat counterintuitively, the retention of WGD duplicates prone to dominant deleterious (*e.g.* gain-of-function) mutations,

$$\frac{\Pi_\star^{\mathrm{WGD}}/\Pi_e^{\mathrm{WGD}}}{\Pi_\star^{\mathrm{SSD}}/\Pi_e^{\mathrm{SSD}}} \simeq \frac{\nu_f}{\nu_f - \nu_+}. \tag{4}$$

Indeed, for genes prone to deleterious gain-of-function mutations ($\nu_+ \gtrsim \nu_- + \nu_\star$, *i.e.* $\nu_f > \nu_f - \nu+$) we find a significantly enhanced retention of duplicates through WGD as compared to SSD events ($\Pi_\star^{\mathrm{WGD}}/\Pi_e^{\mathrm{WGD}} > \Pi_\star^{\mathrm{SSD}}/\Pi_e^{\mathrm{SSD}}$). By contrast, for most genes which lack gain-of-function mutations ($\nu_+ \ll \nu_f$), we find a comparable retention of neutral duplicates through WGD and SSD events ($\Pi_\star^{\mathrm{WGD}}/\Pi_e^{\mathrm{WGD}} \simeq \Pi_\star^{\mathrm{SSD}}/\Pi_e^{\mathrm{SSD}}$). This effect of dominant deleterious mutations on the retention of WGD duplicates (Eq. (4)) is the main result of this study, which rationalizes, from a population genetics perspective, empirical evidences available from the literature, as will be shown below in the Section 3.4 analyzing the prevalence of human oncogenes with WGD duplicates from recent cancer genome data.

2.3. Extension to diploid models

The extension of the previous haploid model to a diploid model including epistatic interaction and recombination between four different alleles at each duplicated locus implies a combinatorial proliferation of two-locus diploid genotypes, such as $A_\circ A_-/A_+A_\circ$, $A_\circ A_\star/A_+A_-$, etc.

In addition to these multiplicity of diploid states, we also expect further complications due to the process of duplication-driven speciation proposed in Werth and Windham (1991) and Lynch and Force (2000a), see Discussion. Indeed, reciprocal gene loss at duplicated loci (*i.e.* $A_-A_\circ$ or $A_\circ A_-$) has been shown to lead to duplication-driven speciation due to the interbreeding barriers between individuals having lost different copies of the duplicated loci. This is because the diploid combination of haplotypes with reciprocal loss of duplicates, $A_\circ A_-/A_-A_\circ$, readily recombines to yield a double mutant haplotype, A_-A_-, and ultimately a non-functional diploid genotype A_-A_-/A_-A_-, which effectively lowers the interspecific compatibility between individuals coming from subpopulations carrying primarily the $A_\circ A_-$ or the $A_-A_\circ$ haplotype. Similar subpopulation structures are expected to arise from independent breakings of symmetry in the divergence of multiple alleles at duplicated loci, such as with the two functional haplotypes $A_-A_\circ$ and $A_\circ A_\star$ (with functional $A_\circ$, non-functional A_- and sub-functional $A_\star$) which lead, after recombination, to the non-functional diploid genotype, $A_-A_\star/A_-A_\star$.

To circumvent these complications in analyzing the retention of a single or two gene copies with multiple alleles at duplicated loci, we will in fact consider only one breaking of symmetry and reciprocal gene loss scenario below, while keeping in mind that alternative scenarios can exist and possibly co-exist as different subpopulations or species, see Discussion. This amounts to simplifying the actual two-locus four-allele diploid system into an effective one-locus four-allele diploid system, based on the four haplotypes introduced earlier, *i.e.* $AA_\circ$, AA_-, AA_+ and $AA_\star$. If we further assume, for simplicity, that there is no difference between maternal and paternal inherited haplotypes, we are left to consider only ten diploid combinations of these haplotypes, *i.e.* $AA_\circ/AA_\circ$, $AA_\circ/AA_-$, $AA_\circ/AA_+$, $AA_\circ/AA_\star$, AA_-/AA_-, AA_-/AA_+, $AA_-/AA_\star$, AA_+/AA_+, $AA_+/AA_\star$, $AA_\star/AA_\star$.

We can then study the effects of the dominant deleterious phenotype caused by the AA_+ haplotype, assuming, otherwise, a neutral fitness for all diploid combinations without AA_+. This leads to the following marginal fitness for each haplotype,

$$w_i = w_i^\circ(1-\phi_+) + w_i'\phi_+,$$

where $w_i^\circ = 1$ and $w_i' = 1 - hs_+$ for $i = \circ, -, \star$ and $w_+^\circ = 1 - hs_+$ and $w_+' = 1 - s_+$, where h is the dominance coefficient of the heterozygous diploid genotypes including one haplotype AA_+. In particular, $h = 1$ corresponds to a simple dominant deleterious mutant, while $h = 1/2$ corresponds to a co-dominant deleterious mutant with additive deleterious effects for the AA_+/AA_+ diploid genotype. This leads to the average marginal fitness of the population,

$$\bar{w} = \sum_i \phi_i w_i = 1 - 2hs_+\phi_+ + s_+(2h-1)\phi_+^2$$

and the relative marginal fitness for each haplotype,

$$w_+ - \bar{w} = -hs_+ + s_+(3h-1)\phi_+ - s_+(2h-1)\phi_+^2$$
$$w_i - \bar{w} = hs_+\phi_+ - s_+(2h-1)\phi_+^2,$$

for $i = \circ, -, \star$.

Thus, if the fraction of dominant deleterious haplotype AA_+ remains small in the population, $\phi_+ \ll 1$, as expected and confirmed by simulations (see Section 3.1), we retrieve the same population genetics system as for the haploid model studied earlier, Eqs. (1) and (2), in the case of dominant deleterious mutations ($h = 1$) or in the case of incomplete dominance ($0 < h < 1$), if the fitness decrement is rescaled as $s_+ \to s_+h^{-1}$. If we further assume that the dominance coefficient h can be approximated as the average fraction of dominant deleterious mutations, *i.e.*, $h \simeq \nu_+/\nu_f$, then Eq. (4) leads to the following retention rate of WGD duplicates with dominance coefficient h in diploid genomes,

$$\Pi_\star^{\mathrm{WGD}}(h) \simeq \frac{\Pi_\star^{\mathrm{WGD}}(0)}{1-h}. \tag{5}$$

Hence, with these simplifications, the two-locus, four-allele diploid system of duplicated loci behaves essentially like a one-locus, four-allele haploid system. This is the population genetics model that we will further consider below to study the stochastic effects in populations of finite size and coupled mutation/selection dynamics.

2.4. Extension to stochastic effects in small populations

We now consider a stochastic approach to describe the dynamics of a population of fixed size N, based on the generalization to more realistic coupled mutation/selection dynamics such as the Moran model (Moran, 1958).

We use a one-step process master equation formalism between $K \geqslant 2$ alleles (A), in the context of the one-locus, four-allele haploid system introduced in Section 2.1. Each sub-population j of size n_j ($\sum_{j=1}^K n_j = N$) has transition rates $W_{ij}(n_1, \ldots, n_K) = (n_j/N)\sum_k \beta_{ik}^{(j)} n_k$ from allele j to allele i, where n_j/N is the probability that one individual with the allele j is randomly chosen to die and $\beta_{ik}^{(j)} n_k$ is the rate at which one individual with allele k is chosen to reproduce and mutate into the allele i, given

that an individual with allele j has been chosen to die. This general expression enables to include both coupled and uncoupled mutation/selection dynamics depending on the definition of the reproduction/mutation rates $\beta_{ik}^{(j)}$.

Three main population genetics models have been studied in the literature: two models with coupled mutation/selection processes correspond to the first and second Moran models (Moran, 1958) with mutations occurring either before or after selection, respectively. The first Moran model essentially selects on the lifespan of adults rather than their reproductive success, while the second Moran model amounts to a gametic selection independent of death rate, see A. By contrast, the uncoupled mutation/selection model outlined in the Section 2.1 amounts to use, as model parameters, "average" mutation rates $\bar{\nu}_{ij} = \sum_{k\neq i} \beta_{ik}^{(j)} \phi_k$, for $j \neq i$, and "average" selection rates $\bar{w}_i = \sum_j \beta_{ii}^{(j)} \phi_j$ and $\bar{w}^{(i)} = \sum_k \beta_{kk}^{(i)} \phi_k$, see A. Uncoupled mutation/selection models have been frequently used in recent years for multiallelic systems (Eldon and Wakeley, 2006; Muirhead and Wakeley, 2009; Etheridge and Griffiths, 2009; Etheridge et al., 2010; Vogl and Clemente, 2012).

These different mutation/selection models can then be applied to study the fixation of gene duplicates following either a SSD or a WGD event. To this end, we consider the multiallelic model with the four different alleles introduced earlier in Section 2.1 ($K = 4$, Fig. 1) corresponding to the initial (unstable) duplicate state $AA_\circ$ as well as the three alleles arising through mutations from $AA_\circ$, namely, $AA_- \equiv A$, AA_+, and $AA_\star$. The rates of mutations from j to i then correspond to $\bar{\nu}_{ij} = \bar{\nu}_{i\circ(i\neq\circ)} = \nu_i$ with $i = \star, -, +$.

In fact, when all fitness parameters are neutral except for the fitness disadvantage of dominant deleterious mutants (*i.e.* $w_\circ = w_- = w_\star = 1$ and $w_+ = 1 - s_+$, where $s_+ \ll 1$), the two coupled mutation/selection models by Moran (1958) lead to very similar deterministic equation systems as the uncoupled mutation/selection model of Eqs. (1) and (2) in the large population size limit ($N \gg 1$), see B.

Extensions to adaptive selection of duplicates with $w_\star > 1$ are discussed in the Result Section 3.3.

2.5. Stochastic simulations

We performed stochastic simulations of the birth, death and mutation processes for the three population genetics models outlined above and corresponding to the one-step process master equation detailed in A. For each of the four alleles $k = \{AA_\circ, AA_+, AA_-, AA_*\}$, we keep track of a random variable $n_k(t)$ representing the number of individuals with allele k and fitness w_k at time t. We subdivide one generation into small time steps of length δt and update the frequency of each allele after every such time step.

We first consider the model with uncoupled selection and mutation, corresponding to Eqs. (1) and (2) in the deterministic limit of large population size. In each time step, the number of offspring b_k with allele k is obtained from a binomial distribution with mean $n_k w_k \delta t$. We then randomly remove a number of individuals d_k from the sub-population of allele k, so as to keep the overall population size constant, $\sum_k n'_k = \sum_k (n_k + b_k - d_k) = N$, where $n'_k = n_k + b_k - d_k \geqslant 0$ corresponds to the updated size of the sub-population k, after birth and death steps. Finally, the stochastic mutations are generated independently from the selection process for the $n'_{AA_\circ}$ individuals in the unstable duplicate allele class $AA_\circ$ with mutation probability $p'_k = \nu_k \delta t$ from allele $AA_\circ$ to allele k where ν_k is the corresponding mutation rate per generation. The sub-population sizes are then updated to $n_{AA_\circ}(t + \delta t) = n'_{AA_\circ} - \sum_k m_k$ for the $AA_\circ$ allele and to $n_k(t + \delta t) = n'_k + m_k$ for $k = AA_-, AA_+, AA_\star$, where m_k represents the number of individuals mutated from allele $AA_\circ$ to the allele k. The time step δt is typically chosen in the range of 0.01–0.1 generation.

In the case of the Moran models with coupled mutation/selection, the transition W_{ij} removes one individual from class j (*i.e.* $j \to j - 1$) and replicates one individual of class i (*i.e.* $i \to i + 1$) in the time step δt, taking into account the coupling between birth, death and mutation at the same time. At each time step the transition rates $W_{ij}(n_1, \ldots, n_K) = (n_j/N) \sum_k \beta_{ik}^{(j)} n_k$ are computed with the coefficients $\beta_{ik}^{(j)}$ from the corresponding Moran models, with either mutations before selection (model 1) or mutations after selection (model 2). The transition $j \to i$ is then chosen stochastically according to its rate W_{ij} leading to the population updates $n_j = n_j - 1$ and $n_i = n_i + 1$, and a time increment $\delta t = (\sum_{ij} W_{ij})^{-1}$ summed over all possible transitions.

In the case of the Moran model controlling death rate, we choose the death rate $\lambda_k = w_k^{-1}$, that approximates to $\lambda_k \simeq 1$ for $k = AA_\circ, AA_\star, AA_-$ and $\lambda_k \simeq 1 + s_+$ for $k = AA_+$, in the limit of small s_+, $0 < s_+ \ll 1$.

The mutation and selection parameters of the models are chosen in agreement with the available estimates in the literature (Lynch, 2010). The total mutation rate in the germline of vertebrates such as mouse or human is of the order of $1 - 4 \times 10^{-8}$ per nucleotide site per generation (Lynch, 2010). Taking an average gene length of 1000 to 1500 nt leads to an average mutation rate of $\nu_f = 4 \times 10^{-5}$ mutation per gene per generation. As the rate of sub-functionalization $\nu_\star$ is expected to be a small fraction of ν_f, we assume $\nu_\star = \nu_f/10 = 4 \times 10^{-6}$ per gene per generation, which corresponds to a fixation rate of about 10% of typical duplicates after WGD (according to Eq. (3) with $\Pi_e^{\mathrm{WGD}} = \epsilon = 1$ and $\nu_f \gg \nu_+$), in agreement with the average retention of ohnologs from each round of WGD at the origin of vertebrates, see Section 3.4.2. In addition, we assume that the local rates of gain-of-function and loss-of-function mutations vary depending on the gene local susceptibility to gain-of-function *versus* loss-of-function mutations at each position with a constant averaged sum across all genes, $\nu_+ + \nu_- = \nu_f - \nu_\star = 3.6 \times 10^{-5}$ per gene per generation. Hence, in the following, we will simply assign increasing values to ν_+, while keeping the sum $\nu_- + \nu_+$ fixed. The selective disadvantage s_d of a deleterious allele is known to be typically in the range of $s_d \simeq 10^{-3}/10^{-2}$ (Lynch, 2010), thus, the value of the selection coefficient s_+ for the dominant deleterious mutant AA_+ is chosen as $s_+ = 0.05$ to emphasize its deleterious phenotypic effect. Finally, we start either with a single individual with a SSD duplicate, leading to $\epsilon = 1/N$ for the SSD scenario, or with all individuals with WGD duplicates, leading to $\epsilon = 1$ for the WGD scenario.

3. Results

3.1. Fixation rates for neutral sub-functionalization

We first performed stochastic simulations to compute the fixation rate of gene duplicates through neutral sub-functionalization ($w_\star = 1$), in order to study the different retention of SSD *versus* WGD duplicates in the cases that, in the large population size limit, correspond to the deterministic system, Eq. (2). We analyze the probability of fixation for the allele $AA_\star$ as a function of the ratio ν_+/ν_f, which measures the "dangerousness" of the gene duplicates, that is their susceptibility to dominant deleterious mutations. Fig. 2 shows the results comparing SSD and WGD scenarios. The simulations are performed for the uncoupled mutation/selection model and a population size ranging from $N = 10^3$ (violet) to 10^5 (red). For a given ratio ν_+/ν_f, the simulated fixation rate is averaged over 10^2 to 10^4 (WGD) or 10^6 to 10^7 (SSD) fixation trajectories and the standard deviations are shown as error bars. For $Ns_+ \gg 1$ (*i.e.* $N \gg 20$, see next Section 3.2 on finite size effects), the strong fitness disadvantage s_+ prevents the fixation of the allele AA_+, and leads to an eventual competition between two

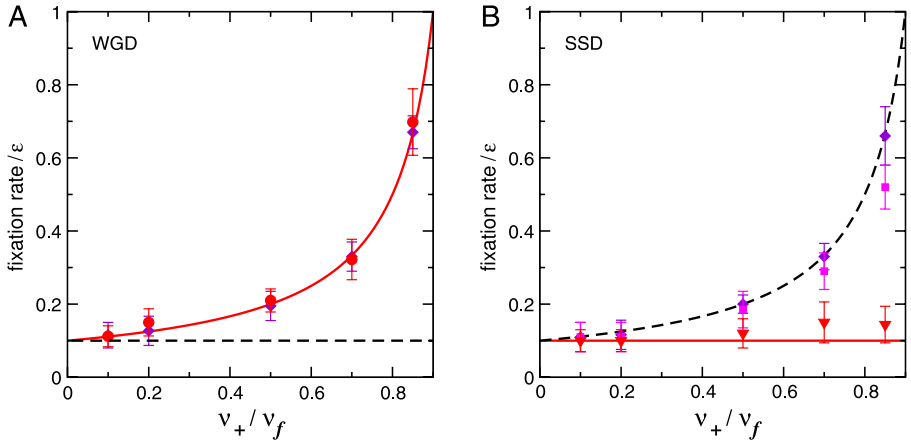

Fig. 2. *Theoretical and simulated fixation rates of WGD and SSD duplicates.* The theoretical and simulated fixation rates of gene duplicates through neutral sub-functionalization ($w_\star = 1$) are plotted for (A) WGD and (B) SSD scenarios as a function of the ratio ν_+/ν_f, which measures their susceptibility to dominant deleterious mutations. In the stochastic simulations, the uncoupled mutation/selection model is used and the different population sizes are $N = 10^3$ (diamonds, violet), 10^4 (squares, magenta) and 10^5 (triangles, red). The selection coefficient for the deleterious allele AA_+ is $s_+ = 0.05$, the total mutation rate is $\nu_f = 4 \times 10^{-5}$ per gene per generation and the rate of sub-functionalization is $\nu_\star = 4 \times 10^{-6}$ per gene per generation. For a given ratio ν_+/ν_f, the simulated fixation rate is averaged over 10^2 to 10^4 (WGD) and 10^6 to 10^7 (SSD) fixation trajectories and the standard deviations are shown as error bars. For each WGD or SSD scenario, the theoretical curves for the fixation rate are obtained from the deterministic uncoupled model, Eq. (6), and represented as a continuous red line (the dotted line is the theoretical curve for the other scenario). Finite size effects are clearly visible for the SSD scenario at small population size. For $N = 10^3$ (diamonds, violet), the simulated fixation rates for SSD essentially reduce to the corresponding fixation rates for WGD. Yet, for increasing population size ($N \geqslant 10^4$, squares, magenta), the fixation rates of SSD duplicates prone to dominant deleterious mutations ($\nu_+/\nu_f > 0.5$) become lower than for the corresponding WGD duplicates and eventually reach the SSD asymptotic limit, $\nu_\star/\nu_f$, for $N \geqslant 10^5$ (continuous red line).

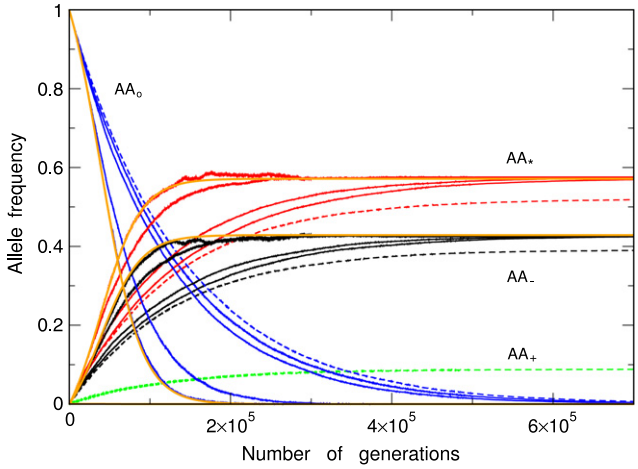

Fig. 3. *Population size effect on WGD fixation trajectories.* The effect of population size on the fixation trajectories of alleles $AA_\circ$ (blue), AA_+ (green), AA_- (black) and $AA_\star$ (red), is studied for WGD populations of increasing size, $N = 10^2$, 10^3, 10^4, 10^5 and 10^6. The selection and mutation parameters are $s_+ = 0.05$ and $\nu_+ = 3.3 \times 10^{-5}$, $\nu_\star = 4 \times 10^{-6}$ and $\nu_- = 3 \times 10^{-6}$, in order to analyze the fixation trajectories of WGD duplicates with strong susceptibility to dominant deleterious mutations ($\nu_+ \gg \nu_-$). The fixation trajectories are averaged over 10^2 to 10^4 trajectories. Finite population size affects both transient frequencies and final fixation rates for $N = 10^2$ (dotted lines), corresponding to $Ns_+ \simeq 1$, but only the transient frequencies for $N \geqslant 10^3$, reaching eventually the deterministic trajectories (orange lines) for $N = 10^6$, corresponding to $N\nu_\pm \gg s_+$, see main text. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

neutral alleles where the fixation rate $\Pi_\star$ corresponds to the allele frequency in the asymptotic limit, $\phi_\star(\infty)$, as given by Eq. (3),

$$\frac{\Pi_\star^{\mathrm{WGD}}}{\epsilon} = \frac{\nu_\star}{\nu_f}\,\frac{1}{1-\nu_+/\nu_f}$$

$$\frac{\Pi_\star^{\mathrm{SSD}}}{\epsilon} = \frac{\nu_\star}{\nu_f}. \qquad (6)$$

As the ratio $\nu_\star/\nu_f$ is kept fixed at the value 0.1, the two theoretical curves for $\Pi_\star^{\mathrm{WGD}}/\epsilon$ and $\Pi_\star^{\mathrm{SSD}}/\epsilon$ become simple functions of the ratio ν_+/ν_f and are plotted as continuous red lines in Fig. 2.

The comparison between WGD and SSD scenarios for large population size $N \geqslant 10^5$ gives interesting insights into the retention of "dangerous" duplicates prone to gain-of-function mutations. First, the retention of neutral duplicates $AA_\star$ is associated to a low fixation rate for both SSD and WGD duplicates *lacking* dominant deleterious mutations (corresponding to the region $\nu_+/\nu_f \to 0$). Conversely, for gene duplicates *prone* to dominant deleterious mutations (corresponding to $\nu_+/\nu_f \to 1 - \nu_\star/\nu_f$), the retention of neutral duplicates $AA_\star$ is clearly enhanced after WGD events for all population sizes ($N \geqslant 10^3$), while the retention of such "dangerous" SSD duplicates becomes lower than their WGD counterparts for $N \geqslant 10^4$ and reaches a limit independent of ν_+/ν_f for $N \geqslant 10^5$. These results are in agreement with the predictions of the initial simplified deterministic model for large population (Eq. (3)) and support the idea that WGD events have effectively favored the expansion of gene families prone to dominant deleterious mutations.

Note, however, that the agreement of the asymptotic allele frequency with the fixation rate only holds for large enough population size in the SSD scenario. The discrepancy at lower population sizes is due to finite size effects that allow the initial unstable duplicate $AA_\circ$ to reach fixation by drift before the mutations actually occur, hence, making the duplicates' fixation rate converge towards the WGD scenario. These finite size effects, which are the hallmark of population genetics, are analyzed in more details in the next Section 3.2.

3.2. Finite size effects

The emergence of finite size effects in the fixation rate of SSD duplicates is clearly visible on Fig. 2. Their interpretation requires, however, a detailed analysis of the consequences of stochastic noise on the evolutionary dynamics of a population of finite size N. We consider separately the WGD and SSD scenarios, below, illustrating the average fixation trajectories in Figs. 3 and 4 for duplicates with a very high susceptibility to dominant deleterious mutations, $\nu_+/\nu_f = 0.825$, to emphasize the different evolutionary scenarios of the proposed population genetics model.

In the WGD case, the effect of stochastic sampling is only visible for a very small population size ($N = 10^2$, Fig. 3), when drift can

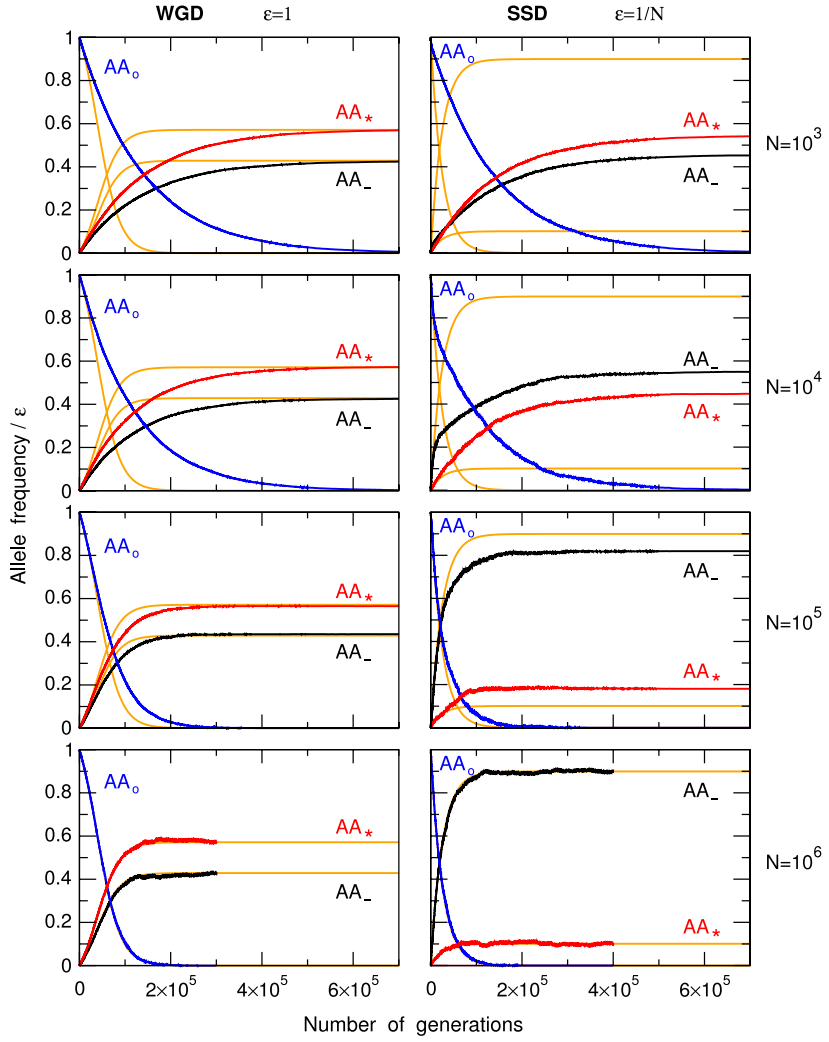

Fig. 4. *Comparison of fixation trajectories for WGD versus SSD duplicates.* The effect of population size on the fixation trajectories of alleles $AA_\circ$ (blue), AA_- (black) and $AA_\star$ (red), is compared for WGD (left) and SSD (right) scenarios. The selection and mutation parameters are $s_+ = 0.05$ and $\nu_+ = 3.3 \times 10^{-5}$, $\nu_\star = 4 \times 10^{-6}$, $\nu_- = 3 \times 10^{-6}$, as in Fig. 3. The SSD fixation trajectories are averaged over 10^6 to 10^7 trajectories. The finite size only delays the fixation trajectories of WGD duplicates for small population size before converging to the deterministic solution for $N = 10^6$ (orange line). By contrast, finite size effects affect both the trajectories and the fixation rates of SSD duplicates, which are similar to the WGD scenario for $N = 10^3$ (top), before the $AA_\star$ (red) and AA_- (black) allele trajectories invert themselves for $N = 10^4$ to eventually reach the SSD asymptotic limit (orange) for $N = 10^6$ (bottom). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

outcompete purifying selection ($Ns_+ \simeq 1$) and results in a non-negligible fixation of the deleterious allele AA_+ (green dotted line) and a simultaneous reduction of the frequencies of the other fixable alleles $AA_\star$ (red dotted line) and AA_- (black dotted line). Then, as the population size increases above $N = 10^3$, the condition $Ns_+ \gg 1$ is always satisfied, leading to the expected fixation rates of the deterministic limit, Fig. 2A (*i.e.* for a negligible fixation of the deleterious allele AA_+). Yet, we expect some additional stochastic effects on the population dynamics due to the discretization of frequencies if $\delta\phi_\pm s_+ \equiv s_+/N \gtrsim \nu_\pm$, that is $s_+ \gtrsim N\nu_\pm$. In practice, this condition delays the transient dynamics of the simulated trajectories with respect to the deterministic solution (Fig. 3). However, as N increases from 10^3 to 10^6, the large population condition is more and more verified, $s_+ \ll N\nu_\pm$, leading to average stochastic trajectories (red, blue and black lines) that eventually converge to their deterministic solutions (orange lines), for $N = 10^6$.

In the SSD case, stochastic noise affects not only the transient dynamics but also the fixation rate of SSD duplicates for a wider range of population sizes, Fig. 2B. A detailed analysis based on the comparison with the WGD case is shown in Fig. 4. In the WGD scenario with $N \geqslant 10^3$, finite size effects affect only the transient dynamics, as discussed above. By contrast, in the SSD scenario, drift caused by stochastic sampling in small population results in the spreading of the initial $AA_\circ$ duplicates to the whole population before they have the chance to mutate into other alleles, leading to a population dynamics after SSD that resembles the WGD scenario, Fig. 4 (top). This effect is evident and strong for population size $N = 10^3$, where the average simulated trajectories for SSD essentially reduce to the corresponding trajectories for WGD, after proper rescaling by $\epsilon = 1/N$. For increasing population size, this effect weakens and the fixation rates of SSD duplicates become lower than for WGD duplicates for $N \geqslant 10^4$ and eventually reach their asymptotic limit at $N \geqslant 10^5$ for SSD duplicates prone to dominant deleterious mutations, Figs. 2B and 4 (bottom).

3.3. Extension to adaptive sub-functionalization

The previous Sections 3.1 and 3.2 demonstrate that the fixation of neutral SSD duplicates by drift is at most equals to the initial fraction of SSD duplicates in the population, that is $\Pi_\star^{\text{SSD}} \leqslant \epsilon \simeq 1/N$, which is further reduced to $\Pi_\star^{\text{SSD}} \simeq \nu_\star/(\nu_f N)$ for large population, as the initial $AA_\circ$ duplicates can be lost through loss-of-function or gain-function mutations before they become fixed as $AA_\star$ through sub-functionalization, Fig. 2B.

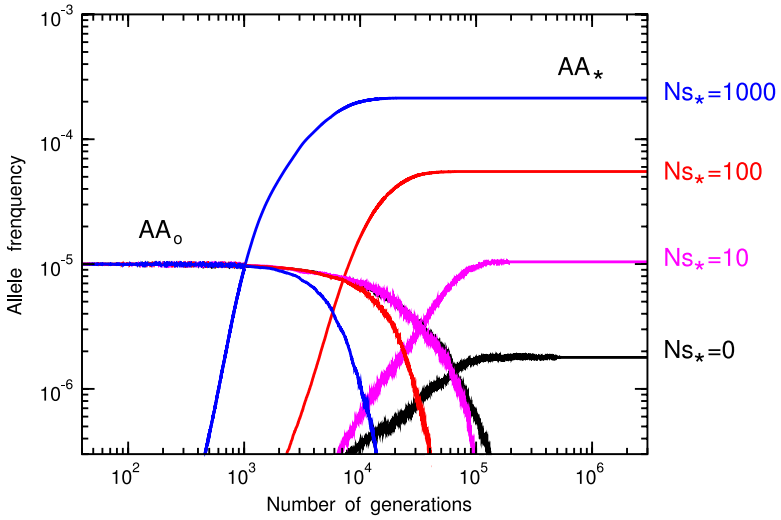

Fig. 5. *SSD fixation requires positive selection in large populations.* Lines of the same color represent the decreasing $AA_\circ$ average frequencies (left) and the corresponding increasing $AA_\star$ average frequencies (right), for increasing values of the fitness benefice ($s_\star$) of the fixed duplicates $AA_\star$ with $s_\star = 0$ (black), 10^{-4} (magenta), 10^{-3} (red) and 10^{-2} (blue). The population size is $N = 10^5$ and all other parameters are the same as in Fig. 4. The frequency of the fixed duplicates $AA_\star$ increases rapidly with a small fitness benefice ($s_\star > 0$) demonstrating that the fixation of SSD duplicates, which is inefficient by drift in large populations, is strongly enhanced under positive selection. Note that the fixation rate $\Pi_\star$ approaches the asymptotic value of the classical two-allele models ($\Pi_\star = s_\star$, Appendix C) times the fraction of mutation rates leading to sub-functionalization, *i.e.* $\Pi_\star \simeq s_\star \times \nu_\star/\nu_f = s_\star/10$, see Section 3.3. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Hence, the fixation of SSD duplicates by drift is clearly inefficient and should be quite rare in large populations (Otto and Yong, 2002; Kondrashov and Kondrashov, 2006). However, beneficial mutations are likely to be particularly important for adaptation (Fisher, 1930; Crow and Kimura, 1965; Patwa and Wahl, 2008). Indeed, it is easy to see that the fixation of SSD duplicates increases rapidly if their retention is associated even to a small fitness benefice ($s_\star > 0$) as shown on Fig. 5. A sharp rise in the average fixation trajectories is obtained for increasing values of the fitness parameter from $s_\star = 0$ (black), 10^{-4} (magenta), 10^{-3} (red) to 10^{-2} (blue). This demonstrates that the fixation of SSD duplicates is strongly enhanced under positive selection compared to the low fixation rates of neutral SSD duplicates by drift in large population. Note, in particular, that the fixation rate $\Pi_\star^{\mathrm{SSD}}$ approaches the asymptotic value of the classical two-allele models ($\Pi_\star^{\mathrm{SSD}} = s_\star$, C) times the fraction of mutation rates leading to sub-functionalization, *i.e.* $\Pi_\star^{\mathrm{SSD}} \simeq s_\star \times \nu_\star/\nu_f = s_\star/10$. This takes into account the fact that the sub-functionalized duplicates $AA_\star$ arise from the initial redundant duplicates $AA_\circ$ through a mutation rate $\nu_\star$ ten times smaller than ν_f. The slight discrepancy (increasing with increasing $s_\star$) of this estimate from the simulated $AA_\star$ fixation rate (Fig. 5) is related to the fixation time of a new beneficial mutant, $t_{\mathrm{fix}} \simeq 1/s_\star$. Indeed for increasing $s_\star$, t_{fix} becomes shorter and shorter such that no other $AA_\circ$ individuals, if present, can significantly affect the dynamics of the fixation trajectory, as they are unlikely to experience themselves sub-functionalization mutations before the first $AA_\star$ mutant spreads through the entire population by positive selection. This reduces, in practice, the apparent initial fraction of $AA_\circ$ alleles that effectively contribute to the fixation rate of $AA_\star$ through positive selection. Alternatively, positive selection might also favor the enhanced expression levels of initial SSD duplicates prior to mutations (Otto and Yong, 2002; Kondrashov and Kondrashov, 2006), leading to the classical result, $\Pi_\star^{\mathrm{SSD}} = s_\star$ (C).

These results demonstrate that the fixation of SSD duplicates typically requires positive selection in large population, while a different mechanism based on purifying selection governs the fixation of "dangerous" WGD duplicates prone to dominant deleterious mutations following WGD-induced speciation. Besides, as noted earlier (Singh et al., 2012), we expect that the population bottleneck associated with WGD-induced speciation limits the efficacy of the retention of beneficial WGD duplicates through positive selection.

3.4. Application to the prevalence of human oncogenes with WGD versus SSD duplicates

The results of these population genetics models can be applied to interpret the retention of WGD duplicates from genes prone to dominant deleterious mutations, as outlined in the Introduction. Here, we define our datasets of human oncogenes and ohnologs and illustrate this biased retention of WGD duplicates with the prevalence of human oncogenes with WGD *versus* SSD duplicates in different human primary tumors.

3.4.1. Identification of human oncogenes and ohnologs

Data on human oncogenes have recently become increasingly available thanks to the numerous cancer genome sequencing projects covering a broad range of primary tumors.

We obtained human oncogenes with mutations in different primary tumors from the Catalogue of Somatic Mutations in Cell (COSMIC) database (Forbes et al., 2011). COSMIC holds data primarily from cancer genome sequencing projects, however, it lacks the information about the dominance of mutations. Therefore, oncogene status of COSMIC genes were obtained from multiple databases including Cancer Census (Futreal et al., 2004) and SwissProt (Magrane and UniProt Consortium, 2011), and divided into two categories. First, we restricted our dataset to experimentally verified and manually curated oncogenes from Cancer Census (369) and SwissProt (223) and added a limited set of COSMIC genes, restricted to genes with at least 50 non-synonymous mutations including one recurrent non-synonymous mutation from all mutated samples in the COSMIC database (v64 release). Oncogene status of COSMIC genes were obtained either from text searches in OMIM (Hamosh et al., 2005), Ensembl (Flicek et al., 2013), Entrez gene (Maglott et al., 2011), Gene Cards (Safran et al., 2010) and Tumor Associated Genes (Chen et al., 2013), or predicted following the procedure described in Bozic et al. (2010). We also constructed a more extended

Table 1
Restricted set of oncogenes with WGD or SSD duplicates for different human primary tumors. Human oncogenes mutated in different primary tumors have significantly retained an excess of WGD duplicates as compared to the average retention of ohnologs in the whole human genome, *i.e.* 58.3% *vs* 34.7% ($P = 3.39 \times 10^{-103}$, χ^2 test). Conversely, human oncogenes are only slightly depleted in SSD duplicates as compared to the average SSD retention in the whole human genoma, *i.e.* 43.9% *vs* 48.6% ($P = 5.35 \times 10^{-5}$, χ^2 test). Data are shown for the restricted dataset, see Section 3.4.1 for details.

Human primary tumors	# of genes	# of WGD	% of WGD	WGD χ^2 *p*-value	# of SSD	% of SSD	SSD χ^2 *p*-value
Reference (All protein coding genes)	20 415	7075	34.7%	*Reference*	9916	48.6%	*Reference*
All primary tumors	1 883	1098	58.3%	3.39×10^{-103}	827	43.9%	5.35×10^{-05}
Lung	1 832	1068	58.3%	2.46×10^{-100}	811	44.3%	2.28×10^{-04}
Large intestine	1 827	1066	58.3%	1.75×10^{-100}	806	44.1%	1.38×10^{-04}
Endometrium	1 692	1005	59.4%	1.80×10^{-101}	748	44.2%	3.29×10^{-04}
Kidney	1 600	957	59.8%	3.03×10^{-99}	698	43.6%	7.52×10^{-05}
Ovary	1 551	928	59.8%	2.05×10^{-96}	680	43.8%	1.94×10^{-04}
Skin	1 528	899	58.8%	8.77×10^{-88}	699	45.7%	0.03
Prostate	1 475	870	59.0%	8.02×10^{-86}	663	44.9%	5.37×10^{-03}
Breast	1 456	867	59.5%	1.26×10^{-88}	639	43.9%	3.48×10^{-04}
Upper aerodigestive tract	1 159	697	60.1%	2.98×10^{-74}	539	46.5%	0.16
Urinary tract	1 029	602	58.5%	3.80×10^{-58}	465	45.2%	0.03
Central nervous system	996	601	60.3%	4.57×10^{-65}	435	43.7%	1.99×10^{-03}
Pancreas	993	596	60.0%	2.61×10^{-63}	431	43.4%	1.12×10^{-03}
Hematopoietic and lymphoid tissue	941	569	60.5%	3.65×10^{-62}	424	45.1%	0.03
Stomach	639	369	57.7%	1.38×10^{-34}	293	45.9%	0.17
Cervix	629	342	54.4%	2.73×10^{-25}	318	50.6%	0.32
Liver	392	243	62.0%	5.74×10^{-30}	173	44.1%	0.08
Esophagus	359	211	58.8%	7.77×10^{-22}	168	46.8%	0.50
Autonomic ganglia	119	61	51.3%	1.41×10^{-04}	62	52.1%	0.44
Biliary tract	71	49	69.0%	1.17×10^{-09}	27	38.0%	0.08
Soft tissue	53	39	73.6%	2.59×10^{-09}	12	22.6%	1.59×10^{-04}

Table 2
Extended set of oncogenes with WGD or SSD duplicates for different human primary tumors. Human oncogenes mutated in different primary tumors have significantly retained an excess of WGD duplicates as compared to the average retention of ohnologs in the whole human genome, *i.e.* 48.3% *vs* 34.7% ($P = 4.41 \times 10^{-109}$, χ^2 test). By contrast, no significant SSD bias is observed in human oncogenes as compared to the average SSD retention in the whole human genome, *i.e.* 49.3% *vs* 48.6% ($P = 0.29$, χ^2 test). Data are shown for the extended dataset, see Section 3.4.1 for details.

Human primary tumors	# of genes	# of WGD	% of WGD	WGD χ^2 *p*-value	# of SSD	% of SSD	SSD χ^2 *p*-value
Reference (All protein coding genes)	20 415	7075	34.7%	*Reference*	9916	48.6%	*Reference*
All primary tumors	5 956	2879	48.3%	4.41×10^{-109}	2934	49.3%	0.29
Lung	5 853	2827	48.3%	1.20×10^{-106}	2891	49.4%	0.21
Large intestine	5 799	2809	48.4%	8.20×10^{-108}	2850	49.1%	0.38
Endometrium	5 231	2563	49.0%	2.57×10^{-105}	2562	49.0%	0.56
Kidney	4 481	2233	49.8%	4.00×10^{-101}	2177	48.6%	0.99
Skin	4 257	2101	49.4%	2.58×10^{-90}	2119	49.8%	0.12
Ovary	4 200	2101	50.0%	2.92×10^{-97}	2052	48.9%	0.71
Prostate	3 984	1973	49.5%	1.47×10^{-86}	1980	49.7%	0.15
Breast	3 714	1903	51.2%	4.39×10^{-100}	1748	47.1%	0.07
Upper aerodigestive tract	2 860	1460	51.0%	8.65×10^{-76}	1463	51.2%	0.01
Urinary tract	2 451	1234	50.3%	6.66×10^{-60}	1176	48.0%	0.56
Pancreas	2 344	1188	50.7%	9.04×10^{-60}	1126	48.0%	0.60
Central nervous system	2 288	1195	52.2%	7.96×10^{-70}	1094	47.8%	0.47
Hematopoietic and lymphoid tissue	1 984	1049	52.9%	3.42×10^{-65}	957	48.2%	0.76
Stomach	1 407	669	47.5%	2.93×10^{-24}	728	51.7%	0.02
Cervix	1 332	636	47.7%	1.01×10^{-23}	657	49.3%	0.58
Liver	890	477	53.6%	1.63×10^{-32}	426	47.9%	0.67
Esophagus	691	347	50.2%	8.26×10^{-18}	354	51.2%	0.16
Autonomic ganglia	269	123	45.7%	1.36×10^{-04}	145	53.9%	0.08
Biliary tract	109	72	66.1%	5.63×10^{-12}	42	38.5%	0.04
Soft tissue	87	56	64.4%	5.76×10^{-09}	29	33.3%	0.004

dataset of oncogenes starting from COSMIC genes with at least 15 non-synonymous mutations and one recurrent non-synonymous mutation. The restricted dataset (Table 1) and extended dataset (Table 2) include a total of 1,883 and 5,956 oncogene candidates, respectively.

Human WGD duplicated genes or ohnologs were obtained from Makino and McLysaght (2010) who used Ensembl 52 release. These ohnolog pairs were mapped to Ensembl 70 release using BioMart, leading to a total of 7075 ohnologs. Following Singh et al. (2012), SSD duplicated genes were then identified by running an all-against-all BLASTp (Altschul et al., 1997) using human protein sequences from Ensembl 70 release. We identified the best non-self hits (E-value $< 10^{-7}$), and for all ohnolog genes, we assessed whether their best non-self hit corresponds to (one of) their ohnolog partner(s). If it is the case, the ohnolog is regarded as a non-SSD ohnolog (5653), otherwise it is considered to have been duplicated by SSD (1422). For all non-ohnolog genes, if they have a significant best hit paralog, they are considered to have experienced a SSD (8494) or else they are counted as non-SSD genes (4846).

3.4.2. Enhanced ohnolog retention in human oncogenes

Both restricted dataset (Table 1) and extended dataset (Table 2) show that human oncogenes mutated in different primary tumors

have indeed retained an excess of ohnologs dating back from the onset of jawed vertebrates. These enhanced ohnolog retentions are highly significant for both datasets as compared to the average retention of ohnologs in the whole human genome, *i.e.* 58.3% *vs* 34.7% for the restricted dataset ($P = 3.39 \times 10^{-103}$, χ^2 test, Table 1) and 48.3% *vs* 34.7% for the extended dataset ($P = 4.41 \times 10^{-109}$, χ^2 test, Table 2). Interestingly, mutated oncogenes from most primary tumors have even higher ohnolog retention biases than the average over all primary tumors, as some ohnolog oncogenes tend to exhibit driver mutations in multiple primary tumors.

By contrast, human oncogenes are only slightly depleted in SSD duplicates, as compared to the average SSD retention in the whole human genome, for the restricted dataset, *i.e.* 43.9% *vs* 48.6% ($P = 5.35 \times 10^{-5}$, χ^2 test, Table 1). No significant SSD bias is even observed on the extended dataset, *i.e.* 49.3% *vs* 48.6% ($P = 0.29$, χ^2 test, Table 2).

These results are consistent with the evolutionary model proposed in the present paper. They predict that the retention of WGD duplicates should be enhanced for genes prone to dominant deleterious mutations as for the human oncogenes considered in the above datasets, Tables 1 and 2, while the retention of SSD is predicted to be largely independent of dominant deleterious mutations, requiring instead positive selection of higher expression levels or advantageous mutations, as outlined in the previous Section 3.3 as well as in earlier studies (Otto and Yong, 2002; Kondrashov and Kondrashov, 2006).

In order to be more quantitative in comparing the available experimental data on WGD duplicates of human oncogenes with the proposed model, it is however necessary to translate the observed fraction of ohnologs, f_s, for a gene class s into an average ohnolog retention rate, p_s, over the two rounds of WGD that occurred at the onset of jawed vertebrates. This can be done through a simple mean field approximation, leading to the following expression, $p_s = 2/f_s - 1 - \sqrt{(2/f_s - 1)^2 - 1}$ (Singh et al., 2012). Hence, the observed fraction of ohnologs for human oncogenes, $f_{\text{onc}} = 58.3\%$, corresponds to an average ohnolog retention rate of $p_{\text{onc}} = 21.5\%$ at each round of WGD for the restricted dataset, while the fraction of ohnologs for the extended dataset, $f'_{\text{onc}} = 48.3\%$, corresponds to an average ohnolog retention rate, $p'_{\text{onc}} = 16.3\%$, at each round of WGD. Similarly, the reference over the whole human genome, which corresponds to the observed fraction of ohnologs, $f_{\text{ref}} = 34.7\%$, leads to an average ohnolog retention rate, $p_{\text{ref}} = 10.6\%$, at each round of WGD, as assumed in the Model Section 2.5. Thus, it implies that the observed ohnolog retention bias of human oncogenes (*i.e.* 16.3%–21.5% *vs* 10.6%) is consistent with an average degree of dominance, $h \simeq 0.35$–0.5, according to Eq. (5)], *i.e.*, 0.163–$0.215 \simeq 0.106/(1-h)$.

4. Discussion

It has long been recognized that gene duplicates located at separate loci favor the emergence of new species (Werth and Windham, 1991; Lynch and Force, 2000a). This results from a progressive incompatibility between mating partners undergoing reciprocal gene silencing of different duplicate copies, as outlined above in the Section 2.3 on the extension to diploid models. In particular, the efficiency of such speciation mechanism is expected to increase with the number of genes simultaneously duplicated in a genome and, therefore, to be most effective after WGD events in the course of evolution (Werth and Windham, 1991). In particular, such interspecific incompatibilities after WGD are likely responsible for the radiations of species that have been reported in plant genomes, such as in angiosperms at early Cretaceous some 140 MY ago (De Bodt et al., 2005), as well as in animal genomes, such as in early jawed vertebrates some 500 MY ago and subsequently in teleost fish some 300 MY ago (Kuraku and Meyer, 2010).

By contrast, the reciprocal effect of speciation on the selection of specific gene duplicates, which is the subject of the present paper, has been largely overlooked so far. This is because the fixation of a SSD duplicate is typically thought to be faster than the emergence of a new species, implying that the fixation of single gene duplicates in a population typically precedes speciation events. Yet, it is no longer the case when gene duplicates arise through WGD rather than SSD events (Innan and Kondrashov, 2010). This is because successful WGD are necessarily coupled to a concomitant speciation event, due to the ploidy mismatch between pre- and post-WGD relatives. The subsequent elimination of many WGD duplicates in post-WGD species then unfolds over tens to hundreds millions of years starting from post-WGD populations with already fixed ohnolog duplicates. In particular, this initial fixation of ohnologs is expected to enable the retention of gene duplicates that would have been normally eliminated through purifying selection following an SSD event in the genome of a single individual.

This is especially the case for SSD duplicates of genes prone to dominant deleterious mutations, such as SSD of oncogenes, which are expected to be eliminated by purifying selection, before they can be fixed in a population. Conversely, WGD duplicates prone to dominant deleterious mutations have been preferentially retained in the human genome. This was shown in Singh et al. (2012) for gene families implicated in cancer and other genetic diseases, that have been greatly expanded at the two rounds of WGD dating back from the onset of jawed vertebrates, by contrast to gene families lacking such a susceptibility to dominant deleterious mutations.

In the present study, we analyzed in more details the prevalence of human oncogenes with WGD duplicates from the available cancer genome data on a broad range of primary tumors. We performed a quantitative comparison of the model and the observed ohnolog retention bias of human oncogenes. The only adjustable parameter used to fit the data corresponds to the average degree of dominance of human oncogenes, which is estimated to be in the range of $h \simeq 0.35$–0.5. Although no large scale measurement of the degree of dominance of human oncogenes is currently available from the literature, the inferred estimate seems rather consistent with a number of independent reports on the average and variance of dominance coefficients in other organisms (Deng and Lynch, 1996; Vassilieva et al., 2000; Deng et al., 2002; Fry and Nuzhdin, 2003; Zhang et al., 2004; Phadnis and Fry, 2005; Agrawal and Whitlock, 2011). While the reported average degrees of dominance are relatively low (*e.g.*, $h \simeq 0.1$–0.2), their typical distributions appear to be quite broad across gene classes, making our estimate for human oncogenes rather expected for gene classes prone to dominant deleterious mutations (*i.e.*, $h \simeq 0.35$–0.5). But beyond this consistent fitting value for oncogenes, an interesting outcome of this analysis is to provide a theoretical rationale linking their mutational effects at vastly different time scales, from the effect of somatic mutations in tumor progression to the long-term evolution of vertebrate genomes through germline mutations and purifying selection in post-WGD species since early vertebrates.

From a broader context, the selection of gene mutants with slightly deleterious mutations has a long history starting with the nearly neutral theory devised by Ohta (1972). According to the nearly neutral theory, slightly deleterious mutations inevitably accumulate by drift in small populations, thereby, reducing the average fitness and, potentially, the population size itself. This implies that more deleterious mutations might become fixed and, in extreme cases, lead to the extinction of the population through mutational meltdown, for species with less than a few hundreds remaining individuals (Lande, 1994).

In conclusion, beyond this accumulation of slightly deleterious mutations, we propose that the specific role of WGD-induced speciation should also be taken into account to interpret the enhanced retention of the most "dangerous" WGD duplicates prone

to strongly deleterious mutations with dominant phenotypes. This suggests that not only slightly deleterious but also strongly deleterious mutations have impacted the long-term evolution and organismal complexity of vertebrates following their early two rounds of WGD.

All in all, these findings rationalize, from an evolutionary perspective, the surprising accumulation of WGD and not SSD duplicates in gene families frequently implicated in genetic disorders and cancers.

Acknowledgments

We thank Victor Mendoza for technical assistance to access the European Grid Infrastructure where most of the numerical simulations have been performed. GM is supported by a Ph.D. fellowship from the French Ministry of Research. PPS is a fellowship recipient of La Ligue Contre Le Cancer. HI acknowledges a research grant from Fondation P. G. de Gennes.

Appendix A. General stochastic models using a master equation

We present in this appendix a general approach to describe the stochastic dynamics of a population of fixed size N, based on a generic one-step process master equation for K alleles $A_1, \ldots, A_K$ ($K \geqslant 2$) governing the probability, $P(n_1, \ldots, n_K, t)$, of observing n_i individuals with allele A_i at time t (with $\sum_i n_i = N$), as

$$\frac{\partial P}{\partial t} = \sum_{i,j=1}^{K} (\mathbb{E}_i^{-1}\mathbb{E}_j^{1} - 1) W_{ij}(\{n_k\}) P(\{n_k\}, t)$$

where $\mathbb{E}^{\pm 1}$ is the "step operator" (van Kampen, 2007) such that $\mathbb{E}_i^{\pm 1} f(n_i) = f(n_i \pm 1)$ and $W_{ij}(\{n_k\})$ is the transition rate from allele j to allele i, which can be expressed in terms of the numbers of individuals with the different alleles as,

$$W_{ij}(n_1, \ldots, n_K) = \frac{n_j}{N} \sum_k \beta_{ik}^{(j)} n_k$$

where n_j/N is the probability that one individual with allele j is randomly chosen to die, while $\beta_{ik}^{(j)} n_k$ is the rate at which one individual with allele k is chosen to reproduce and mutate into allele i, given that an individual with allele j has been chosen to die. In particular, this general expression enables to include either coupled or uncoupled mutation/selection dynamics depending on the definition of the transition rates $\beta_{ik}^{(j)}$, see below.

Following the van Kampen's expansion (van Kampen, 2007), we apply the following transformation $n_i = N\phi_i + N^{\frac{1}{2}}\xi_i$ to the master equation, where $\phi_i(t)$ correspond to the noiseless deterministic solutions of the dynamics in the large population size limit $N \gg 1$, while the new variables ξ_i, which will replace n_i in the master equation, correspond to the stochastic noise in n_i for a finite size population.

Accordingly, the distribution $P(n_1, \ldots, n_K, t)$ is now written as a function of ξ_i as, $P(n_1, \ldots, n_K, t) = \Pi(\xi_1, \ldots, \xi_K, t)$. The one-step operator $\mathbb{E}_i^{\pm 1}$ changes n_i into $n_i \pm 1$ and therefore ξ_i into $\xi_i \pm N^{-1/2}$, so that

$$\mathbb{E}_i^{\pm 1} = 1 \pm N^{-1/2}\frac{\partial}{\partial \xi_i} + \frac{1}{2}N^{-1}\frac{\partial^2}{\partial^2 \xi_i} \pm \cdots$$

while the time derivative $\partial_t P(n_1, \ldots, n_K, t)$ is taken with constant n_i, leading to

$$\frac{\partial P}{\partial t} = \frac{\partial \Pi}{\partial t} - N^{\frac{1}{2}} \sum_i \frac{d\phi_i}{dt}\frac{\partial \Pi}{\partial \xi_i}.$$

Hence the master equation in the new variables ξ_i takes the form of an expansion in $N^{-1/2}$,

$$\begin{aligned}
&\frac{\partial \Pi}{\partial t} - N^{\frac{1}{2}} \sum_i \frac{d\phi_i}{dt}\frac{\partial \Pi}{\partial \xi_i} \\
&= \sum_{i,j=1}^{K} \left[N^{-\frac{1}{2}} \left(\frac{\partial}{\partial \xi_j} - \frac{\partial}{\partial \xi_i} \right) + \frac{1}{2} N^{-1} \left(\frac{\partial}{\partial \xi_j} - \frac{\partial}{\partial \xi_i} \right)^2 + \cdots \right] \\
&\quad \times \left[N\phi_j \sum_k \beta_{ik}^{(j)} \phi_k + N^{\frac{1}{2}} \left(\phi_j \sum_k \beta_{ik}^{(j)} \xi_k \right.\right. \\
&\quad \left.\left. + \xi_j \sum_k \beta_{ik}^{(j)} \phi_k \right) + \xi_j \sum_k \beta_{ik}^{(j)} \xi_k \right] \Pi.
\end{aligned}$$

The largest terms of order $N^{\frac{1}{2}}$ cancel each other out if $\phi_i(t)$ are taken as the solutions of the deterministic equations,

$$\frac{d\phi_i}{dt} = \phi_i \sum_k (\beta_{ii}^{(k)} - \beta_{kk}^{(i)})\phi_k - \phi_i \sum_{l;k\neq l} \beta_{lk}^{(i)} \phi_k + \sum_{j,k\neq i} \phi_j \beta_{ik}^{(j)} \phi_k$$

which leads to the following deterministic equations in the three main population genetics models described in the literature,

- 1- The first Moran model (Moran, 1958) of coupled mutation/selection processes with mutations occurring *before* selection, which is assumed to control death rates. This is a selection on the lifespan of adults rather than their reproductive success,

  $$\beta_{ik}^{(j)} = \lambda_j \nu_{ik}, \quad \text{for } k \neq i,$$
  $$\beta_{ii}^{(j)} = \lambda_j \left(1 - \sum_{l\neq i} \nu_{li} \right),$$
  $$\frac{d\phi_i}{dt} = \phi_i \left(\sum_k \lambda_k \phi_k - \lambda_i \right) - \phi_i \sum_{l\neq i;k} \nu_{li}\lambda_k \phi_k + \sum_{j,k\neq i} \phi_j \lambda_j \nu_{ik} \phi_k$$

  where ν_{ik} corresponds to the probability to experience a mutation from k to i at the time scale of death rate λ_j.
- 2- The second Moran model (Moran, 1958) of coupled mutation/selection processes with mutations occurring *after* selection, which is assumed to control birth rates. This is a gametic selection with a death independent rate $\beta_{ik}^{(j)} \equiv \beta_{ik}$,

  $$\beta_{ik}^{(j)} \equiv \beta_{ik} = \nu_{ik} w_k, \quad \text{for } k \neq i$$
  $$\beta_{ii}^{(j)} \equiv \beta_{ii} = \left(1 - \sum_{l\neq i} \nu_{li} \right) w_i,$$
  $$\frac{d\phi_i}{dt} = \phi_i \left(w_i - \sum_k w_k \phi_k \right) - \phi_i w_i \sum_{l\neq i} \nu_{li} + \sum_{k\neq i} \nu_{ik} w_k \phi_k$$

  where ν_{ik} corresponds to the probability to experience a mutation from k to i at the time scale of birth rate w_k.
- 3- The case of uncoupled mutation/selection outlined in the first section which amounts to use "average" mutation rates $\bar{\nu}_{ij}$ and "average" selection rates $\bar{w}_i$ and $\bar{w}^{(i)}$ as model parameters, as frequently used in recent years (Eldon and Wakeley, 2006; Muirhead and Wakeley, 2009; Etheridge and Griffiths, 2009; Etheridge et al., 2010; Vogl and Clemente, 2012),

  $$\bar{\nu}_{ij} = \sum_{k\neq i} \beta_{ik}^{(j)} \phi_k, \quad \text{for } j \neq i$$
  $$\bar{\nu}_{ii} = 0,$$
  $$\bar{w}_i = \sum_j \beta_{ii}^{(j)} \phi_j,$$

$$\bar{w}^{(i)} = \sum_k \beta_{kk}^{(i)} \phi_k$$

$$\frac{d\phi_i}{dt} = \phi_i(\bar{w}_i - \bar{w}^{(i)}) - \phi_i \sum_l \bar{\nu}_{li} + \sum_j \bar{\nu}_{ij}\phi_j.$$

Appendix B. Four-allele models of SSD *versus* WGD retention

In this appendix, we apply the three mutation/selection models defined in A to study the fixation of gene duplicates following either a SSD or a WGD event. We consider the four different genotypes described in the main text: the initial (unstable) duplicates, $AA_\circ$, and the three alleles arising by mutation from $AA_\circ$, *i.e.* $AA_- \equiv A$, AA_+ and $AA_\star$. The mutations with functional effect are therefore occurring from allele $j = \circ$ to $i = +, -, \star$ with probabilities (or rates) $\nu_{ij} = \nu_{i\circ(i\neq\circ)} = \nu_i$ for $i = +, -, \star$. For the first Moran model where mutations occur before selection, the deterministic system of equations becomes

$$\begin{aligned}
d_t\phi_\circ &= \phi_\circ(\bar{\lambda} - \lambda_\circ) - \nu_f\bar{\lambda}\phi_\circ \\
d_t\phi_- &= \phi_-(\bar{\lambda} - \lambda_-) + \nu_-\bar{\lambda}\phi_\circ \\
d_t\phi_+ &= \phi_+(\bar{\lambda} - \lambda_+) + \nu_+\bar{\lambda}\phi_\circ \\
d_t\phi_\star &= \phi_\star(\bar{\lambda} - \lambda_\star) + \nu_\star\bar{\lambda}\phi_\circ,
\end{aligned}$$

where $\bar{\lambda} = \sum_k \lambda_k\phi_k$. For the second Moran model where mutations occur after selection,

$$\begin{aligned}
d_t\phi_\circ &= \phi_\circ(w_\circ - \bar{w}) - \nu_f w_\circ\phi_\circ \\
d_t\phi_- &= \phi_-(w_- - \bar{w}) + \nu_- w_\circ\phi_\circ \\
d_t\phi_+ &= \phi_+(w_+ - \bar{w}) + \nu_+ w_\circ\phi_\circ \\
d_t\phi_\star &= \phi_\star(w_\star - \bar{w}) + \nu_\star w_\circ\phi_\circ,
\end{aligned}$$

where $\bar{w} = \sum_k w_k\phi_k$. For the case of uncoupled selection/mutation we retrieve the initial uncoupled mutation/selection dynamics of Eq. (1),

$$\begin{aligned}
d_t\phi_\circ &= \phi_\circ(w_\circ - \bar{w}) - \nu_f\phi_\circ \\
d_t\phi_- &= \phi_-(w_- - \bar{w}) + \nu_-\phi_\circ \\
d_t\phi_+ &= \phi_+(w_+ - \bar{w}) + \nu_+\phi_\circ \\
d_t\phi_\star &= \phi_\star(w_\star - \bar{w}) + \nu_\star\phi_\circ.
\end{aligned}$$

The structure of these equations is the same for all models and the differences come from the specific choices of the parameters in the transition rates. The two Moran models can be compared using $\lambda_k = w_k^{-1}$.

In the limit of small fitness decrement caused by deleterious mutations, $s_+ \ll 1$, all three models lead to the same approximate equation system between the four alleles and, therefore, to very close deterministic solutions. However, the stochastic effects encompassed in the full distribution, solution of the master equation, are not accessible analytically in the case of four alleles. Yet, the corresponding population genetics system with only two alleles can be solved exactly, as first shown in Moran (1958), and can bring insights on the competition between the two main fixable alleles of our four-allele system (*i.e.* AA_- and $AA_\star$) as shown in C, below.

Appendix C. Exact results for two-allele stochastic models

We consider the continuous time, one-step process master equation for death–birth and mutation stochastic processes between only two alleles *A* and *a* in a haploid population of fixed size N, with n *A*-individuals and $(N - n)$ *a*-individuals. This equation does not include any approximation, unlike the diffusion equation that is valid in the limit of large populations and small selection, and allows to obtain an exact analytical solution in terms of hypergeometric functions for any values of the model parameters. Moreover, we will show below that it is possible to retrieve classical results of the Wright–Fisher model (Ewens, 1979) as approximations.

The one-step transition rates correspond to the probability density for the system to change its number of individuals with allele *A* from n to $n + 1$ or $n - 1$ during an infinitesimal time dt,

$$\begin{aligned}
W(n \to n + 1) &= W^+(n) \\
W(n \to n - 1) &= W^-(n)
\end{aligned}$$

while $W(n \to n \pm k) = 0$ if $|k| > 1$. The master equation governing the probability, $P(n, t)$, of observing n individuals with allele *A* at time t, is given by

$$\begin{aligned}
\partial_t P(n,t) &= (\mathbb{E}^{-1} - 1)W^+(n)P(n,t) + (\mathbb{E}^1 - 1)W^-(n)P(n,t) \\
&= W^+(n-1)P(n-1,t) - W^+(n)P(n,t) \\
&\quad + W^-(n+1)P(n+1,t) - W^-(n)P(n,t)
\end{aligned}$$

where $\mathbb{E}^{\pm 1}$ is the one-step operator. Using the following transition rates, $W^\pm(n)$, for the three models outlined above, we obtain,

- 1- For the first Moran model (Moran, 1958) of coupled mutation/selection with mutations occurring before selection which controls death rates,

  $$W^+(n) = \mu\,\frac{(N-n)}{N}\,[n(1-\nu_1) + (N-n)\nu_2]$$
  $$W^-(n) = \mu\,\frac{n}{N}\,\frac{1}{(1+s)}\,[(N-n)(1-\nu_2) + n\nu_1]$$

  where μ is the equal birth–death rate of each allele and ν_1 [resp. ν_2] the mutation probability from allele *A* to *a* [resp. from *a* to *A*]. The factor $1/(1 + s)$ implies a reduced ($s > 0$) or enhanced ($s < 0$) death rate of allele *A*.
- 2- For the second Moran model (Moran, 1958) of coupled mutation/selection with mutations occurring after selection which controls birth rates,

  $$W^+(n) = \mu\,\frac{(N-n)}{N}\,[n(1+s)(1-\nu_1) + (N-n)\nu_2]$$
  $$W^-(n) = \mu\,\frac{n}{N}\,[(N-n)(1-\nu_2) + n(1+s)\nu_1]$$

  where $1 + s$ is now the gain ($s > 0$) or loss ($s < 0$) of reproductive success of allele *A*.
- 3- For the uncoupled mutation/selection model with averaged transition parameters outlined above,

  $$W^+(n) = \mu\,(N-n)\frac{n}{N}(1+s) + (N-n)u_2$$
  $$W^-(n) = \mu\,n\frac{(N-n)}{N} + nu_1$$

  where u_1 (resp. u_2) is the mutation rate from allele *A* to *a* (resp. from allele *a* to *A*).

Introducing the rescaled mutation rates $\nu_1 = u_1/\mu$ and $\nu_2 = u_2/\mu$ for the uncoupled mutation/selection model leads to a common form for all three models,

$$W^+(n) = \mu\,\frac{(N-n)}{N}\left(nA^+ + B^+\right)$$

$$W^-(n) = \mu\,\frac{n}{N}\left(nA^- + B^-\right)$$

where $A^+ = 1 - \nu_1 - \nu_2, B^+ = N\nu_2, A^- = -(1 - \nu_1 - \nu_2)/(1 + s), B^- = N(1 - \nu_2)/(1 + s)$, for the first Moran model; $A^+ = (1 - \nu_1)(1 + s) - \nu_2, B^+ = N\nu_2, A^- = -(1 - \nu_1(1 + s) - \nu_2), B^- = N(1 - \nu_2)$, for the second Moran model and $A^+ = 1 + s, B^+ = N\nu_2, A^- = -1, B^- = N(1 + \nu_1)$, for the uncoupled

mutation/selection model. The corresponding master equation can then be solved by introducing the generating function,

$$\phi(z,t) = \sum_n z^n P(n,t),$$

which leads to the following differential equation (using the boundary conditions $W^+(N) = W^-(0) = 0$),

$$\begin{aligned}\partial_t \phi(z,t) &= (z-1)\sum_n W^+(n) z^n P(n,t) \\ &\quad + (z^{-1}-1)\sum_n W^-(n) z^n P(n,t)\end{aligned}$$

$$\begin{aligned}\frac{N}{\mu}\partial_t\phi &= (z-1)\left(A^+ z\partial_z(N\phi - z\partial_z\phi) + B^+(N\phi - z\partial_z\phi)\right) \\ &\quad + (z^{-1}-1)\left(A^- z\partial_z(z\partial_z\phi) + B^- z\partial_z\phi\right) \\ &= (z-1)\Big[A^+\left((N-1)z\partial_z\phi - z^2\partial_z^2\phi\right) \\ &\quad + B^+(N\phi - z\partial_z\phi) - A^-\left(\partial_z\phi + z\partial_z^2\phi\right) - B^-\partial_z\phi\Big] \\ &= (z-1)\Big[\left(-z^2A^+ - zA^-\right)\partial_z^2\phi \\ &\quad + \left((A^+(N-1) - B^+)z - A^- - B^-\right)\partial_z\phi + B^+N\phi\Big].\end{aligned}$$

The stationary solutions correspond to the following homogeneous second order ordinary differential equation

$$\left(-z^2A^+ - zA^-\right)\partial_z^2\phi + \left[(A^+(N-1) - B^+)z - A^- - B^-\right]\partial_z\phi + B^+N\phi = 0$$

which can be transformed into the hypergeometric differential equation through the rescaling $z \to -zA^-/A^+$, see Abramowitz and Stegun (1964),

$$z(z-1)\partial_z^2\phi + ((\alpha+\beta+1)z - \gamma)\,\partial_z\phi + \alpha\beta\phi = 0$$

where $\alpha = -N$, $\beta = B^+/A^+$, $\gamma = 1 + B^-/A^-$. The only acceptable solution is a polynomial of finite degree N corresponding to the following hypergeometric function (as $\alpha = -N$ is a negative integer),

$$\phi_s(z) = 1 + \sum_{n=1}^{N} \frac{(\alpha)_n(\beta)_n}{(\gamma)_n}\frac{(1+s)^n}{n!}z^n$$

where $(u)_n$ is the Pochhammer symbol, $(u)_n = u(u+1)\cdots(u+n-1) = \Gamma(u+n)/\Gamma(u)$. These stationary solutions can be rewritten, using the Γ function, as,

$$\phi_s(z) = \sum_{n=0}^{N}\binom{N}{n}\frac{\Gamma(\delta_1 - n)\Gamma(\delta_2 + n)}{\Gamma(\delta_1)\Gamma(\delta_2)}(1+s)^n z^n$$

where

$$\delta_1 = 1 - \gamma = -B^-/A^- = N(1+\nu_1) > N$$

$$\delta_2 = \beta = B^+/A^+ = N\nu_2/(1+s)$$

for the parameters of the uncoupled mutation/selection model above. This leads to the exact stationary distribution, $P_s(n)$, of individuals with allele A, for all $n = 0, \ldots, N$ and δ_1 and δ_2 expressions valid for any population size N, fitness increment s and mutation rates ν_1 and ν_2,

$$P_s(n) = \binom{N}{n}(1+s)^n\frac{\Gamma(\delta_1 - n)\Gamma(\delta_2 + n)}{\Gamma(\delta_1)\Gamma(\delta_2)}. \tag{7}$$

This expression holds, however, only if the arguments of the Γ functions are different from zero (*i.e.* $\delta_1, \delta_2 \neq 0, \delta_1 - n \neq 0, \delta_2 + n \neq 0$). This means that the Moran model in absence of mutations cannot be derived as the limit of this exact stationary distribution for $\nu_1, \nu_2 \to 0$. To retrieve this case, the partial differential equation for the probability generating function has to be directly rewritten for the suitable transition rates $W^{\pm}(n, \nu_1 = 0, \nu_2 = 0)$, which are equivalent for all three models (up to a rescaling of time scale for the first Moran model). The stationary solution has the form, $\phi_s(z) = \Pi_N z^N + \Pi_0$, and can be solved, following Houchmandzadeh and Vallade (2010), leading to $\Pi_N = (1-\sigma^{n_0})/(1-\sigma^N)$, $\Pi_0 = (\sigma^{n_0} - \sigma^N)/(1-\sigma^N)$, where $\sigma = 1/(1+s)$ and n_0 is the initial number of A-individuals in the population. In particular, the well-known result for the fixation probability in an haploid populations is readily retrieved as an approximation for a small selection coefficient s, noting that $\sigma^N = e^{n\log\sigma} \simeq e^{-ns}$,

$$\Pi_N = \frac{1 - e^{-sn_0}}{1 - e^{-sN}}.$$

In particular, the probability of fixation of a new arisen mutant ($n_0 = 1$) in a large population reduces to $\Pi_N \simeq s$.

Using the exact solution of Eq. (7) and the Stirling's approximation for large factorials ($\Gamma(z+1) = z! \simeq e^{z\ln z - z}$), in addition to low fitness gain $s \ll 1$ and mutation rates $\nu_1, \nu_2 \ll 1$, then leads to the approximation,

$$P_s(p) \simeq (1+s)^{Np}(1-p)^{N\nu_1 - 1}\, p^{N\nu_2 - 1},$$

where $p = n/N$. This allows to recover the well-known approximate solution of the Wright–Fisher model (Ewens, 1979) for a diploid population ($N = 2N_e$) with non-overlapping generation,

$$P_s(p) \propto \bar{w}^{2N_e} p^{4N_e\nu_2' - 1}(1-p)^{4N_e\nu_1' - 1}$$

with $\bar{w} = 1 + ps$ and $2\nu_i' = \nu_i$ due to the difference in the distribution of offspring between the overlapping and non-overlapping generation models. Note that this factor 2 can be readily recovered in the uncoupled model assigning μ as the equal birth-or-death rate of each allele per generation and thus, $\mu/2$, as the rate of *a*-death-*A*-birth process in $W^+(n)$ and, similarly, as the converse rate of *A*-death-*a*-birth process in $W^-(n)$.

References

Abramowitz, M., Stegun, I., 1964. Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables. Dover.

Agrawal, A.F., Whitlock, M.C., 2011. Inferences about the distribution of dominance drawn from yeast gene knockout data. Genetics 187, 553–566.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., et al., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25, 3389–3402.

Aury, J., Jaillon, O., Duret, L., B, Noel, Jubin, C., et al., 2006. Global trends of whole-genome duplications revealed by the ciliate paramecium tetraurelia. Nature 444, 171–178.

Beerenwinkel, N., Antal, T., Dingli, D., Traulsen, A., Kinzler, K.W., Velculescu, V.E., Vogelstein, B., Nowak, M.A., 2007. Genetic progression and the waiting time to cancer. PLoS Comput. Biol. 3, e225.

Birchler, J.A., Bhadra, U., Bhadra, M.P., Auger, D.L., 2001. Dosage-dependent gene regulation in multicellular eukaryotes: implications for dosage compensation, aneuploid syndromes, and quantitative traits. Dev. Biol. 234, 275–288.

Blekhman, R., Man, O., Herrmann, L., Boyko, A.R., Indap, A., et al., 2008. Natural selection on genes that underlie human disease susceptibility. Curr. Biol. 18, 883–889.

Bozic, I., Antal, T., Ohtsuki, H., Carter, H., Kim, D., Chen, S., Karchin, R., Kinzler, K.W., Vogelstein, B., Nowak, M.A., 2010. Accumulation of driver and passenger mutations during tumor progression. Proc. Natl. Acad. Sci. 107, 18545–18550.

Cai, J.J., Borenstein, E., Chen, R., Petrov, D.A., 2009. Similarly strong purifying selection acts on human disease genes of all evolutionary ages. Genome Biol. Evol. 1, 131–144.

Chen, J.S., Hung, W.S., Chan, H.H., Tsai, S.J., Sun, H.S., 2013. In silico identification of oncogenic potential of fyn-related kinase in hepatocellular carcinoma. Bioinformatics 29, 420–427.

Crow, J.F., Kimura, M., 1965. Evolution in sexual and asexual populations. Amer. Nat. 439–450.

De Bodt, S., Maere, S., Van de Peer, Y., 2005. Genome duplication and the origin of angiosperms. Trends Ecol. Evolut. (Amst.) 20, 591–597.

Deng, H.W., Gao, G., Li, J.L., 2002. Estimation of deleterious genomic mutation parameters in natural populations by accounting for variable mutation effects across loci. Genetics 162, 1487–1500.

Deng, H.W., Lynch, M., 1996. Estimation of deleterious-mutation parameters in natural populations. Genetics 144, 349–360.

Eldon, B., Wakeley, J., 2006. Coalescent processes when the distribution of offspring number among individuals is highly skewed. Genetics 172, 2621–2633.
Esteban, L.M., Vicario-Abejon, C., Fernandez-Salguero, P., Fernandez-Medarde, A., Swaminathan, N., et al., 2001. Targeted genomic disruption of H-ras and N-ras, individually or in combination, reveals the dispensability of both loci for mouse growth and development. Mol. Cell. Biol. 21, 1444–1452.
Etheridge, A., Griffiths, R., 2009. A coalescent dual process in a moran model with genic selection. Theor. Popul. Biol. 75, 320–330.
Etheridge, A.M., Griffiths, R.C., Taylor, J.E., 2010. A coalescent dual process in a moran model with genic selection, and the lambda coalescent limit. Theor. Popul. Biol. 78, 77–92.
Ewens, W.J., 1979. Mathematical Population Genetics. Springer-Verlag, New York.
Fisher, R.A., 1930. The Genetical Theory of Natural Selection. Clarendon Press, Oxford.
Flicek, P., Ahmed, I., Amode, M.R., Barrell, D., Beal, K., et al., 2013. Ensembl 2013. Nucleic Acids Res. 41, 48–55.
Forbes, S.A., Bindal, N., Bamford, S., Cole, C., Kok, C.Y., et al., 2011. COSMIC: mining complete cancer genomes in the catalogue of somatic mutations in cancer. Nucleic Acids Res. 39, D945–D950.
Force, A., Lynch, M., Pickett, F.B., Amores, A., Yan, Y.l., Postlethwait, J., 1999. Preservation of duplicate genes by complementary, degenerative mutations. Genetics 151, 1531–1545.
Fry, J.D., Nuzhdin, S.V., 2003. Dominance of mutations affecting viability in Drosophila melanogaster. Genetics 163, 1357–1364.
Furney, S.J., Alba, M.M., Lopez-Bigas, N., 2006. Differences in the evolutionary history of disease genes affected by dominant or recessive mutations. BMC Genomics 7, 165.
Futreal, P.A., Coin, L., Marshall, M., Down, T., Hubbard, T., et al., 2004. A census of human cancer genes. Nat. Rev. Cancer 4, 177–183.
Gu, X., 2003. Evolution of duplicate genes versus genetic robustness against null mutations. Trends Genet. 19, 354–356.
Gu, Z., Steinmetz, L.M., Gu, X., Scharfe, C., Davis, R.W., Li, W.H., 2003. Role of duplicate genes in genetic robustness against null mutations. Nature 421, 63–66.
Hamosh, A., Scott, A.F., Amberger, J.S., Bocchini, C.A., McKusick, V.A., 2005. Online mendelian inheritance in man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res. 33, D514–517.
Holland, P.W., Garcia-Fernandez, J., Williams, N.A., Sidow, A., 1994. Gene duplications and the origins of vertebrate development. Dev. Suppl. 125–133.
Houchmandzadeh, B., Vallade, M., 2010. Alternative to the diffusion equation in population genetics. Phys. Rev. E 82, 051913.
Hughes, A.L., 1994. The evolution of functionally novel proteins after gene duplication. Proc. Biol. Sci. 256, 119–124.
Huminiecki, L., Heldin, C.H., 2010. 2r and remodeling of vertebrate signal transduction engine. BMC Biology 8, 146.
Innan, H., Kondrashov, F., 2010. The evolution of gene duplications: classifying and distinguishing between models. Nature Rev. Genet. 11, 97–108.
Ise, K., Nakamura, K., Nakao, K., Shimizu, S., Harada, H., Ichise, T., Miyoshi, J., Gondo, Y., Ishikawa, T., Aiba, A., et al., 2000. Targeted deletion of the h-ras gene decreases tumor formation in mouse skin carcinogenesis. Oncogene 19, 2951.
Kamath, R.S., Fraser, A.G., Dong, Y., Poulin, G., Durbin, R., et al., 2003. Systematic functional analysis of the Caenorhabditis elegans genome using RNAi. Nature 421, 231–237.
Kondrashov, F.A., Kondrashov, A.S., 2006. Role of selection in fixation of gene duplications. J. Theoret. Biol. 239, 141–151.
Kuraku, S., Meyer, A., 2010. Whole genome duplications and the radiation of vertebrates. In: Dittmar, K., Liberles, D. (Eds.), Evolution After Gene Duplication. John Wiley & Sons, Inc., pp. 299–311.
Lande, R., 1994. Risk of population extinction from fixation of new deleterious mutations. Evolution 48, 1460–1469.
Lynch, M., 2010. Evolution of the mutation rate. Trends Genet. 26, 345–352.
Lynch, M., Conery, J.S., 2000. The evolutionary fate and consequences of duplicate genes. Science 290, 1151–1155.
Lynch, M., Force, A., 2000a. Gene duplication and the origin of interspecific genomic incompatibility. Amer. Nat. 156, 590–605.
Lynch, M., Force, A., 2000b. The probability of duplicate gene preservation by subfunctionalization. Genetics 154, 459–473.
Lynch, M., O'Hely, M., Walsh, B., Force, A., 2001. The probability of preservation of a newly arisen gene duplicate. Genetics 159, 1789–1804.
Maglott, D., Ostell, J., Pruitt, K.D., Tatusova, T., 2011. Entrez gene: gene-centered information at NCBI. Nucleic Acids Res. 39, D52–D57.
Magrane, M., UniProt Consortium „ 2011. UniProt Knowledgebase: a hub of integrated protein data. Database (Oxford) 2011, bar009.
Makino, T., McLysaght, A., 2010. Ohnologs in the human genome are dosage balanced and frequently associated with disease. Proc. Natl. Acad. Sci. 107, 9270–9274.
Merlo, L.M., Pepper, J.W., Reid, B.J., Maley, C.C., 2006. Cancer as an evolutionary and ecological process. Nat. Rev. Cancer 6, 924–935.
Michor, F., Iwasa, Y., Nowak, M.A., 2004. Dynamics of cancer progression. Nat. Rev, Cancer 4, 197–205.
Moran, P.A.P., 1958. The effect of selection in a haploid genetic population. Proc. Camb. Phil. Soc 54, 463–467.
Muirhead, C.A., Wakeley, J., 2009. Modeling multiallelic selection using a moran model. Genetics 182, 1141–1157.
Ohno, S., 1970. Evolution by Gene Duplication. Springer.
Ohta, T., 1972. Population size and rate of evolution. J. Mol. Evol. 1, 305–314.
Otto, S.P., Yong, P., 2002. The evolution of gene duplicates. Adv. Genet. 46, 451–483.
Papp, B., Pál, C., Hurst, L.D., 2003. Dosage sensitivity and the evolution of gene families in yeast. Nature 424, 194–197.
Patwa, Z., Wahl, L., 2008. The fixation probability of beneficial mutations. J. R. Soc. Interface 5, 1279–1289.
Phadnis, N., Fry, J.D., 2005. Widespread correlations between dominance and homozygous effects of mutations: implications for theories of dominance. Genetics 171, 385–392.
Safran, M., Dalah, I., Alexander, J., Rosen, N., Iny Stein, T., et al. 2010. GeneCards Version 3: the human gene integrator. Database (Oxford) 2010, baq020.
Sidow, A., 1996. Gen(om)e duplications in the evolution of early vertebrates. Curr. Opin. Genetics Dev. 6, 715–722.
Singh, P.P., Affeldt, S., Cascone, I., Selimoglu, R., Camonis, J., Isambert, H., 2012. On the expansion of dangerous gene repertoires by whole-genome duplications in early vertebrates. Cell Rep. 2, 1387–1398.
van Kampen, N., 2007. Stochastic Processes in Physics and Chemistry, third ed. North-Holland Personal Library.
Vassilieva, L.L., Hook, A.M., Lynch, M., 2000. The fitness effects of spontaneous mutations in Caenorhabditis elegans. Evolution 54, 1234–1246.
Veitia, R.A., 2002. Exploring the etiology of haploinsufficiency. Bioessays 24, 175–184.
Vogl, C., Clemente, F., 2012. The allele-frequency spectrum in a decoupled moran model with mutation, drift, and directional selection, assuming small mutation rates. Theor. Popul. Biol. 81, 197–209.
Werth, C.R., Windham, M.D., 1991. A model for divergent, allopatric speciation of polyploid pteridophytes resulting from silencing of duplicate-gene expression. Amer. Nat. 515–526.
Winzeler, E.A., Shoemaker, D.D., Astromoff, A., Liang, H., Anderson, K., et al., 1999. Functional characterization of the *S*. cerevisiae genome by gene deletion and parallel analysis. Science 285, 901–906.
Zhang, J., 2003. Evolution by gene duplication: an update. Trends Ecol. Evol. 18, 292–298.
Zhang, X.S., Wang, J., Hill, W.G., 2004. Influence of dominance, leptokurtosis and pleiotropy of deleterious mutations on quantitative genetic variation at mutation-selection balance. Genetics 166, 597–610.