Genetica
DOI 10.1007/s10709-008-9311-5

Preservation of duplicate genes by originalization

Cheng Xue · Yunxin Fu

Received: 13 November 2007 / Accepted: 2 August 2008

Abstract Neofunctionalization, subfunctionalization and increasing gene dosage were proposed to be the possible ways to explain duplicate-gene preservation in previous studies. However, in some natural populations, such as yeast *Saccharomyces cerevisiae*, a considerable proportion of the duplicate genes originated from ancient whole genomic duplication (WGD) is preserved till now, which cannot be sufficiently explained by these mechanisms. In this article, we present another possible way to explain this conundrum—originalization, by which duplicate genes are both preserved intact at a high frequency in the population under only purifying selection. With approximate equal rates of mutation at the two duplicated loci, analytical, numerical and simulation results consistently show that the mean time to nonfunctionalization for unlinked haploinsufficient gene duplication might become markedly prolonged, which results from originalization. These theoretical results imply that originalization might be an alternative effective and temporary way of preserving duplicate genes.

C. Xue
College of Life Sciences, University of Sun Yet-Sen, Guangzhou, China

C. Xue (✉)
GuangDong Institute for Monitoring Laboratory Animals, 105 Road Xingang West, Guangzhou 510260, China
e-mail: lflf27@yahoo.com.cn

Y. Fu
Laboratory for Conservation and Utilization of Bio-Resources, Yunnan University, Yunnan, China
e-mail: Yunxin.fu@uth.tmc.edu

Y. Fu
Human Genetics Center, School of Public Health, University of Texas at Houston, P.O. Box 20186, Houston, TX 77225, USA

Keywords Gene duplication · Originalization · Recombination · Selection · Preservation

The evolutionary mechanism for gene duplication has been argued for many years (Ohno 1970; Walsh 2003; Lynch and Katju 2004; Li et al. 2005). Ohno (1970) proposed that nonfunctionalization, by which one of the duplicate genes loses the function of the ancient gene while another maintains the function (this model is also called the classical model), might be the evolutionary fate for most duplicate genes because advantageous mutations are rare compared with degenerative mutations. By contrast, Force et al. (1999) and Lynch and Force (2000) argued that this might not be the case for a considerable proportion of gene duplication, because many of the duplicate genes originating from ancient whole genomic duplications (WGD) have been observed to be retained in some genomes, such as tetraploid fish (Ferris and Whitt 1979), *Xenopus Laevis* (Hughes and Hughes 1993), maize, and yeast *Saccharomyces cerevisiae* (Wolfe and Shields 1997). Force et al. (1999) proposed that subfunctionalization, by which the functions of the ancestral gene are partitioned between descendent duplicate genes so that both descendent duplicate genes are preserved by selection, is an effective way for duplicate-gene preservation in the genomes. This was named the Duplication–Degeneration–Complementation (DDC) model.

Recently the DDC model was also challenged when it is applied to explain the genomic data in large-population-size organisms, such as yeast *S. cerevisiae*. Because they usually have large effective population size (Lynch and Conery 2004) and no cell differentiation as single cell organisms, the probability of subfunctionalization for duplicate genes might be very low (Li et al. 2005; Lynch and Force 2000). In fact, a proportion (~10%) of duplicate genes originated from an

ancient WGD are maintained in yeast *S. cerevisiae* genome (Wolfe and Shields 1997; Byrne and Wolfe 2005).

Neofunctionalization (which means that a novel gene is fixed at one of the duplicated loci while another copy maintains the ancient function) has been well documented and reviewed as an important way of providing fundamental evolutionary materials (Ohno 1970; Long et al. 2003; Li et al. 2005). Since beneficial mutation rate is usually much lower than deleterious mutation rate, the evolutionary fates of duplicate genes almost are not neofunctionalization but nonfunctionalization under the classical model (Force et al. 1999), which also cannot explain the previously described genomic observations on duplicate genes.

Another hypothesis was proposed recently for this paradox that both duplicate genes might be preserved from the initiation of gene duplication by positive selection resulting from increasing gene dosage of one of the duplicate genes, and then ancillary and lowly expressed function are amplified to become highly expressed such that both copies are preserved permanently (Hooper and Berg 2003; Moore and Purugganan 2003; Clement et al. 2006). This process was named amplification. The main difference between amplification and subfunctionalization is that subfunctionalization is a relaxed, nearly neutral process, while from the initiation of gene duplication positive selection is involved in amplification.

He and Zhang (2005) also suggested that after duplication events duplicate genes might experience a relaxed and nearly neutral evolution by subfunctionalization for a transient period, followed by a prolonged neofunctionalziation. Since subfunctionalization might not be favored in large populations, it seems that there might be an alternative mechanism for retaining duplicate genes before neofunctionalization in large populations. Takahata and Maruyama (1979) reported that mean time to nonfunctionalization (T) for unlinked haploinsufficient gene duplication might be much prolonged in a large population ($N\mu = 10$, where N is the effective population size and μ is the degenerative mutation rate at both duplicated loci), but they gave no further explanation on this observation. In this context, advantageous mutations are extensively expected to be more likely to occur in the population during this prolonged nonfunctionalization period, which might facilitate neofunctionalization. However, at present no such theoretical evidence supports this hypothesis. In this article, we try to explore the mechanism underlying the evolution of duplicate genes and provide theoretical evidences and explanations for prolonged T for unlinked haploinsufficient gene duplication, by analytical, numerical and simulation approaches.

Additionally, previous theoretical studies on gene duplication (Takahata and Maruyama 1979; Li 1980; Lynch and Force 2000) mostly assumed that degenerative mutation rates at both duplicated loci are the same. However, no evidence supports or rejects this assumption. So in this article this assumption is relaxed. Haplosufficient (commonly called double null recessive, DNR) and haploinsufficient (also called partial dominant, HI) selective models are selected, which are described in detail below. In this article, we mainly focus on the evolution of duplicate genes originating from polyploidization, such as WGD, so some effects of genetic forces on small segmental duplication are not discussed, such as unequal crossing over, retroposition and gene conversion, etc.

Theoretical results consistently show that with approximate equal mutation rates for the two duplicated loci, T for unlinked haploinsufficient gene duplications becomes prolonged strikingly even in a modest population (roughly $N\mu > 0.5$), which is attributable to the high frequency of the original genes (possessing the same function as ancient gene did before duplication) at both duplicated loci. This process is named originalization, by which a high proportion of original (or wild-type) duplicate genes are preserved intact in the population for a prolonged time to nonfunctionalization under purifying selection, even with no positive selection.

Analytical analysis

Assume in a diploid, random mating population, after a WGD event there are two duplicated loci in the genome. To simplify the representation of chromosomal haplotypes, assume that the duplicated loci are on the same chromosome; a character '0' denotes the wild-type allele and '1' the mutant allele at a locus. Therefore for duplicated loci there are four types of chromosomal haplotypes, namely, "00", "01", "10", and "11", respectively.

Under the DNR (or haplosufficient) selective model, assume that the double null recessive is lethal, for example, individuals with two chromosome genotypes being both "11", are dead. Under the HI (haploinsufficient) selective model, individuals having one or no original allele at both duplicated loci are not viable, for example, individuals with two chromosomal genotypes being "10" and "11", are dead.

Let x_0, x_1, x_2, x_3 be the frequencies of chromosomal haplotypes, "00", "01", "10" and "11", respectively. Fitness of individuals with various genotypes under the DNR and HI selective models are shown in Table 1. The

Table 1 Fitness of individuals with various genotypes under the classical models[a]

Chromosomal haplotype	"00"	"01"	"10"	"11"
"00"	1	1	1	1
"01"	1	1	1	$1 - s_1$
"10"	1	1	1	$1 - s_1$
"11"	1	$1 - s_1$	$1 - s_1$	0

[a] $s_1 = 0$ under the DNR selective model; $s_1 = 1$ under the HI selective model

differential changes of chromosomal haplotype frequencies resulting from effects of recombination, mutation and selection at every generation, are given by

$$\begin{aligned}
x_0' &= x_0x_3^2 + rx_1x_2 - rx_0x_3 + 2s_1x_0x_1x_3 + 2s_1x_0x_2x_3 \\
&\quad - (\mu_1+\mu_2)x_0 \\
x_1' &= x_1x_3^2 - rx_1x_2 + rx_0x_3 - s_1x_1x_3 + 2s_1x_1^2x_3 + 2s_1x_1x_2x_3 \\
&\quad + \mu_1x_0 - \mu_2x_1 \\
x_2' &= x_2x_3^2 - rx_1x_2 + rx_0x_3 - s_1x_2x_3 + 2s_1x_1x_2x_3 + 2s_1x_2^2x_3 \\
&\quad + \mu_2x_0 - \mu_1x_2 \\
x_3' &= x_3^3 + rx_1x_2 - rx_0x_3 - s_1x_1x_3 - s_1x_2x_3 - x_3^2 + 2s_1x_1x_3^2 \\
&\quad + 2s_1x_2x_3^2 + \mu_2x_1 + \mu_1x_2
\end{aligned} \tag{1}$$

where r is the recombination rate between two duplicated loci, $r = 0$ for linked duplicated loci and $r = 0.5$ for unlinked; μ_1 and μ_2 are mutation rates at both duplicate loci, respectively. Under the DNR selective model, $s_1 = 0$; under the HI selective model, $s_1 = 1$. Equation 1 is derived briefly in Appendix.

Now consider the dynamic change of these four allele frequencies in the finite population at every generation. Two approaches were proposed in previous studies to estimate T and dynamic changes of allele frequencies at one locus and two loci, Mendelian Markov process (Khazanie and McKean 1966a, b) and diffusion approximation (Kimura 1955a, b; Kimura and King 1979; Watterson 1983; Wang and Rannala 2004). Here traditional stochastic process (Mendelian Markov process) is used to observe T and dynamic changes of chromosomal haplotype frequencies during the evolution of gene duplication.

Dynamic change of the chromosomal haplotype frequencies for gene duplication is a typical Markov process (Rice 2004). Let A_t be a vector of frequencies of the possible states at generation t. At each state, another vector, $\langle n_0, n_1, n_2, n_3\rangle$, where $n_0 + n_1 + n_2 + n_3 = 2N$, is used for the count numbers of chromosomal haplotypes in the population, "00", "01", "10" and "11", respectively. Therefore the frequencies of chromosomal haplotypes, respectively at a state is given by

$$x_i = n_i/(2N) \quad (i = 0, 1, 2, \text{ and } 3). \tag{2}$$

In a diploid population with population size N, for gene duplication, the number of possible states can be calculated by

$$Q_k = \left(\sum_{i_{k-1}=1}^{2N} \cdots \left(\sum_{i_3=1}^{i_4}\left(\sum_{i_2=1}^{i_3}\left(\sum_{i_1=1}^{i_2} 1\right)\right)\right)\right) + 1 \tag{3}$$

where k is the number of distinct chromosomal haplotypes and in the situation of gene duplication considered above, $k = 4$. Therefore, the number of possible states is also given by

$$Q_4 = N(4N^2 + 11)/3 + 4N^2 + 1. \tag{4}$$

Apparently, Q_4 is proportional to N^3. If N is large, the process of calculating will be inevitably very time-consuming. So the cases of smaller N ($N = 10$ and 20), higher mutation rates ($\mu_1 = 0.01$, 0.03, 0.05, and 0.1) and symmetry of mutation rates on duplicated loci ($\alpha = \mu_2/\mu_1 = 1$), as examples, are selected for estimating T and dynamic changes of chromosomal haplotype frequencies during the nonfunctionalization of gene duplication.

Let P be a $Q_4 \times Q_4$ transition matrix, in which each element of P, P_{ij}, is the probability that the i-th state at a generation becomes the j-th state at the next generation. Under the Wright-Fisher model, P_{ij} is a multinomial variable. Let $n_{k[i]}$ and $x_{k[i]}$ be the number of occurrences and frequency, respectively, of chromosomal haplotype k ($k = 0,1,2,3$) at the i-th state.

For calculating each element in P, first, $x_{0[i]}$, $x_{1[i]}$, $x_{2[i]}$, $x_{3[i]}$ is changed respectively according to Eq. 1, as results of recombination, mutation and selection processes; then P_{ij} is given by

$$P_{ij} = \frac{(2N)!}{n_{0[j]}!n_{1[j]}!n_{2[j]}!n_{3[j]}!}\left(x_{0[i]}^{n_{0[j]}}\right)\left(x_{1[i]}^{n_{1[j]}}\right)\left(x_{2[i]}^{n_{2[j]}}\right)\left(x_{3[i]}^{n_{3[j]}}\right). \tag{5}$$

Now let us consider the calculation of A_t at every generation. Initially, all chromosomal genotypes in the population are original chromosomal haplotype, "00", so at A_0, the frequency of the state $\langle 2N, 0, 0, 0\rangle$ is 1 while those of other states are 0. A_{t+1} is given recursively by

$$A_{t+1} = A_tP. \tag{6}$$

In fact, values of elements in A_t are the cumulative frequencies of states at generation t. Let C_t be the cumulative probability of nonfunctionalization at generation t and it is equal to the sum of the frequencies of states with $n_0 = 0$ (at nonfunctionalization the frequency of "00" must be 0, so x_0 are observed as a proxy of nonfunctionalization), while the frequency of the original chromosomal haplotype is calculated by $1 - C_t$. Therefore, the density probability of nonfunctionalization, f_{t+1}, is the change of C_t, given by

$$f_{t+1} = C_{t+1} - C_t. \tag{7}$$

Finally, when the gap between C_t and 1 is small enough, the final generation time, t_f, is obtained, which means beyond this time the nonfunctionalization is almost reached. Mean time to nonfunctionalization, T, can be calculated by

$$T = \sum(tf_t) \quad (t = 1, 2, \ldots, t_f). \tag{8}$$

At the same time, variance of time to nonfunctionalization, σ^2 can also be estimated

$$\sigma^2 = \sum \left[(t - T)^2 f_t \right] \quad (t = 1, 2, \ldots, t_f). \tag{9}$$

The analytical results are shown in Table 2. From these results, estimations are almost the same as simulation results (simulation methods are described below) and so are the distributions of time to nonfunctionalization of gene duplication (data not shown), which indicate that both the simulation and analytical results are reliable. When $N(\mu_1 + \mu_2)$ is larger (and even N is small), T for unlinked duplication under the HI selective model is much larger than those for other duplications. Dynamic change of original chromosomal haplotype frequency, x_0, is coupled with that of nonfunctionalization probability (Fig. 1). Because at nonfunctionalization of gene duplication x_0 must be 0, the observed prolongation of T for unlinked duplication under the HI selective model results from higher x_0 in the population. For unlinked gene duplication under the HI selective model, higher x_0 in the population means that original genes are preserved at both duplicated loci. Thus, we call this mathematical process of preservation of gene duplication without considering positive selection (much different from amplification)—originalization.

Table 2 Comparisons of analytical and simulation results on mean time to nonfunctionalization for gene duplication (in units of N generations, T/N) with symmetry of mutation rates at duplicated loci ($\mu_1 = \mu_2$)[a]

N	r	μ_1	DNR_ANA	DNR_SIM	HI_ANA	HI_SIM
10	0	0.01	8.78 (5.93)	8.78 (5.93)	9.71 (6.77)	9.72 (6.78)
		0.03	4.68 (2.61)	4.68 (2.61)	5.78 (3.46)	5.79 (3.46)
		0.05	3.62 (1.95)	3.62 (1.95)	4.78 (2.77)	4.78 (2.77)
		0.1	2.57 (1.42)	2.56 (1.42)	3.72 (2.17)	3.72 (2.17)
	0.5	0.01	9.43 (6.73)	9.44 (6.75)	12.2 (9.57)	12.3 (9.60)
		0.03	5.30 (3.42)	5.30 (3.43)	8.83 (6.89)	8.83 (6.92)
		0.05	4.15 (2.64)	4.16 (2.64)	8.05 (6.43)	8.05 (6.43)
		0.1	2.93 (1.87)	2.93 (1.86)	7.00 (5.78)	7.00 (5.77)
20	0	0.01	6.04 (3.53)	6.04 (3.53)	7.09 (4.38)	7.09 (4.38)
		0.03	3.64 (1.97)	3.64 (1.97)	4.80 (2.70)	4.80 (2.71)
		0.05	2.97 (1.67)	2.96 (1.67)	4.17 (2.34)	4.17 (2.35)
		0.1	2.26 (1.37)	2.26 (1.37)	3.45 (2.01)	3.45 (2.01)
	0.5	0.01	7.18 (4.93)	7.17 (4.92)	11.82 (9.52)	11.82 (9.53)
		0.03	4.58 (3.08)	4.58 (3.09)	11.75 (10.02)	11.75 (10.02)
		0.05	3.73 (2.52)	3.74 (2.52)	12.87 (11.38)	12.87 (11.41)
		0.1	2.75 (1.87)	2.75 (1.88)	15.10 (13.96)	15.14 (13.98)

[a] Values in the table are in units of N generation (T/N) and values in parentheses are standard deviations (σ). DNR_ANA and HI_ANA are analytical results of mean time to nonfunctionalization, T, and variances under the DNR and HI selective models, respectively, which are calculated by mathematical formulas described in the text; DNR_SIM and HI_SIM are simulation results of T under the DNR and HI selective models, respectively, and simulation repeats 10^6 times

This is an interesting observation, but in this section population size is very small ($N = 10$ and 20). Now let us consider it in a very large population.

Numerical analysis

Assume that the population is so large ($N\mu \gg 10$) that the effect of genetic drift is small enough to be ignored. Equation 1 can be treated as a group of ordinary differential equations (ODEs). Thus, numerical solutions of the ODEs have been obtained by the Runge–Kutta method (Kincaid and Cheney 2002), with initial conditions $x_0 = 1$, $x_1 = x_2 = x_3 = 0$, given some other appropriate conditions, such as recombination rate ($r = 0$ for linked duplication or $r = 0.5$ for unlinked duplication) and selective model (DNR or HI). Thus, dynamic changes of the chromosomal haplotype frequencies with time (generation) have been observed with $\mu_1 = 10^{-3}$ and different values of α ($\alpha = \mu_2/\mu_1$). As mentioned above, x_0 for unlinked gene duplication under the HI selective model is likely to be kept at high frequency so that original genes are preserved at both duplicated loci, which is called originalization. Now let us consider why and how they are preserved. Dynamic changes of x_0 with different genetics parameters are shown in Figs. 2 and 3 and several features are apparent as follows.

First, for linked duplication ($r = 0$), x_0 usually decreases nearly exponentially down to 0 and changes of x_0 with the same α are similar under the different selection models (see Figs. 2a, 3a). x_0 under the HI selective model at the tail in Fig. 3a is slightly higher than those under the DNR selective models in Fig. 2a, which suggest that in the larger population, T for linked duplication under the HI selective model might be slightly larger than those under the DNR selective models. These suggestions are observed in simulation (see above simulation results in Table 2), which indicates that x_0 is an appropriate and approximate proxy of T.

Second, for unlinked duplication ($r = 0.5$), x_0 usually decreases quickly to an equilibrium in case of $\alpha = 1$. If α is not equal to 1, x_0 under the DNR selective models decreases gradually to zero (see the curves of $\alpha = 0.5$ and 0.8 in Fig. 2b). However, under the HI selective model, when α is approaching 1 or only slightly deviates from 1, x_0 still decreases to the equilibrium (see the curve of $\alpha = 0.8$ in Fig. 3b) or decreases very slowly (see the curve of $\alpha = 0.5$, Fig. 3b). Although when $\alpha = 1$, x_0 for unlinked gene duplication under the DNR selective model also reach an equilibrium, slight asymmetry of mutation rates on the duplicated loci ($\alpha \neq 1$, but the gap between α and 1 is small) will break this equilibrium (see the curve of $\alpha = 0.8$ in Fig. 2b), while under HI selective model this doesn't happen even with slight asymmetry of mutation rates (see

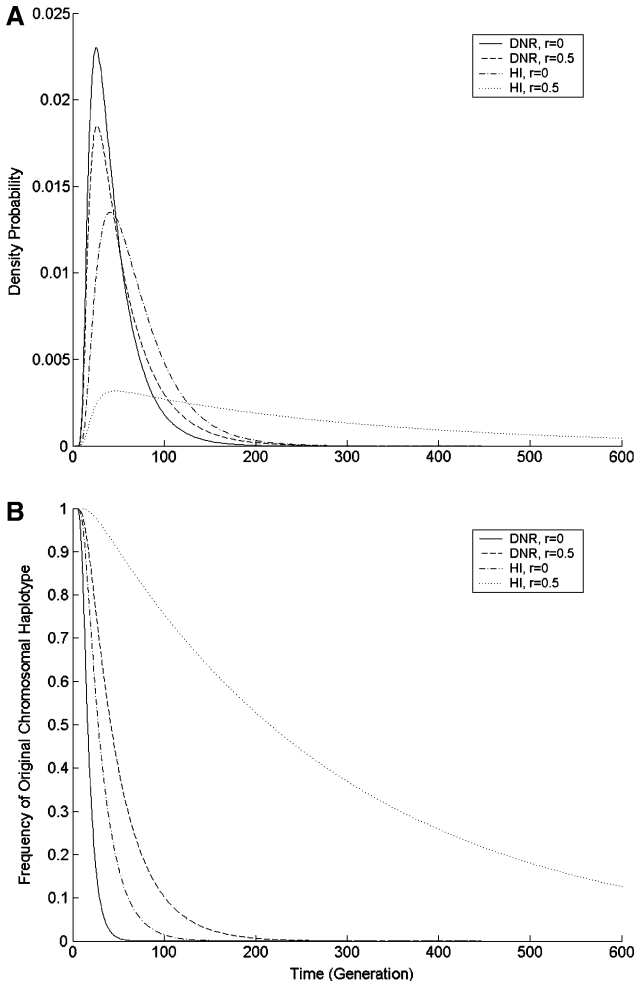

Fig. 1 Dynamic changes of density probability of nonfunctionalization (**a**) and frequency of the original chromosomal haplotype, x_0 (**b**), of linked ($r = 0$) and unlinked ($r = 0.5$) gene duplication under the DNR, and HI selective models, according to analytical results, where population size $N = 20$, and mutation rates $\mu_1 = \mu_2 = 0.1$

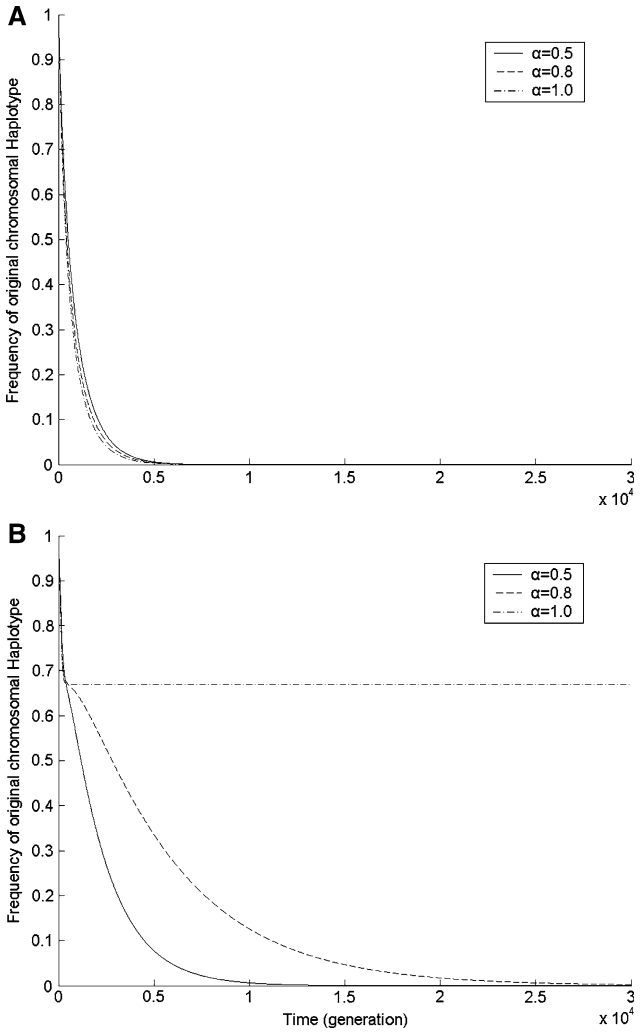

Fig. 2 Dynamic changes of original chromosomal haplotype frequency (x_0) for linked (recombination rate $r = 0$ in (**a**)) and unlinked ($r = 0.5$ in (**b**)) gene duplications under the DNR selective model, according to the numerical results of Eq. 1, where mutation rate at locus one is $\mu_1 = 10^{-3}$, and $\alpha = \mu_2/\mu_1$

the curve of $\alpha = 0.8$ in Fig. 3b). Therefore, the rigid, flexible and stable quasi-equilibrium of x_0 for unlinked duplication under the HI selective model is the main reason that T appears to be much prolonged in the analytical results above. Additionally, originalization (in which x_0 is kept higher for a longer time in the population) appears during the unlinked gene duplication not only under the HI selective model, but also under the DNR selective model (see the curves of $\alpha = 1.0$ in Figs. 2b, 3b). But originalization is more likely to appear during the evolution of unlinked gene duplication under the HI selective model than under the DNR selective model because quasi-equilibrium of x_0 is more flexible and stable under the HI selective model (see the curves of $\alpha = 0.8$ in Figs. 2b, 3b).

Finally, it usually takes much more time for x_0 for unlinked gene duplication to decrease to 0 than that for linked. When $\alpha = 1$, x_0 for unlinked gene duplication will decrease quickly to an equilibrium while that for linked will decrease exponentially to 0 (see Figs. 2, 3). These observations can explain that in simulation T for unlinked duplication is usually larger than that for linked duplication (see simulation results above; Li 1980; Lynch and Force 2000), because of higher x_0 in the population. Simply stated, high recombination (for example $r = 0.5$) forces the loss of "10","01" and "11" under purifying selection (data not shown), thus "00" is preserved at a high frequency in the population. This is the main reason that recombination between duplicated loci provokes the prolongation of T.

In above analytical section, we have shown some simulation results with genetic parameters—very small population size ($N = 10$ and 20) and very high mutation rate ($\mu_1 = \mu_2 = 0.01$–0.1); in this section, we showed the numerical results in a very large population. However,

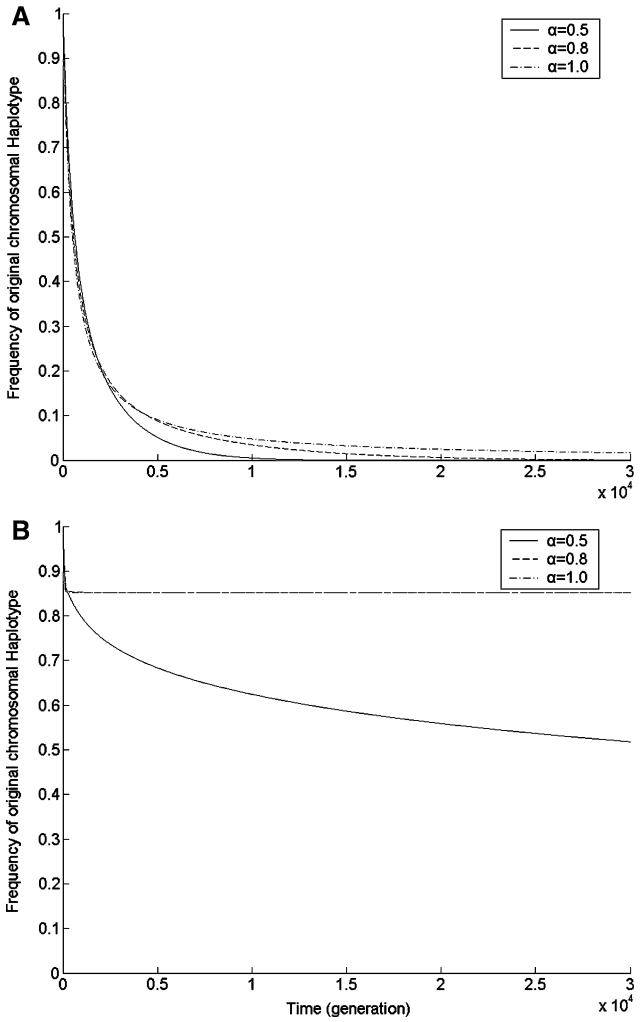

Fig. 3 Dynamic changes of original chromosomal haplotype frequency (x_0) for linked (recombination rate $r = 0$ in (**a**)) and unlinked ($r = 0.5$ in (**b**)) gene duplications under the HI selective model, according to numerical results of Eq. 1, where the mutation rate at locus one is $\mu_1 = 10^{-3}$ and $\alpha = \mu_2/\mu_1$. Curves of $\alpha = 0.8$ and $\alpha = 1.0$ in subplot (**b**) are coincident

natural population size is not extremely small or large, so further theoretical study by large-scale simulation with more realistic genetic parameters is needed to reinforce our above conclusions.

Simulation method and results

In large-scale simulations, the gamete-based algorithm described by Lynch and Force (2000) is used, and frequencies of chromosomal haplotypes are kept tract of at every generation.

In simulation, effects of recombination, mutation, selection and genetic drift are considered, as described by Lynch and Force (2000). Here to simplify and clarify the mathematical description of the simulation process, Lynch and Force's gamete-based simulation algorithm is revised. At every generation, effects of recombination, mutation and selection are calculated directly according to mathematical formulas in Eq. 1, which is different from that in previous algorithms (Lynch and Force 2000).

For the effect of genetic drift, actual chromosomal haplotype frequencies at the next generation are generated by multinomial random sampling. Therefore, the whole revised simulation algorithm is outlined:

Initially, let $x_0 = 1$, $x_1 = x_2 = x_3 = 0$ in the gene pool,

1. Chromosomal haplotype frequencies in the gene pool are differentially changed according to Eq. 1.
2. Sample the given number ($2N$) of gametes multinomially and randomly, and then calculate the actual chromosomal haplotype frequencies in the gene pool at the next generation.
3. Repeat 1, go to the next generation until the final nonfunctionalization of duplicate genes is reached.

The simulation with this algorithm requires very little memory and runs faster than Lynch's algorithm (source code kindly provided by Dr. Lynch). The results from the revised gamete-based algorithm are very similar to those from the gamete-based algorithms previously described (Lynch and Force 2000). The source code of the simulation program is available on request.

T is focused under the classical model, given some conditions, such as mutation rates on two duplicated loci, linkage and selective model, among others. For comparing simulation results with related those in previous studies and above numerical results, let the mutation rate on one of the duplicated loci be $\mu_1 = 10^{-3}$ and the ratio of two mutation rates at the duplicated loci be α, so mutation rate at another locus, μ_2, is equal to $\alpha\mu_1$. The simulation results are shown in Table 3. For a more realistic biological relevance, simulation results in the case of $\mu_1 = 10^{-6}$ and $\mu_2 = 10^{-6}$ are shown in Fig. 4. We may obtain some insights into the evolution of gene duplication as follows.

First, when N is small, with the increase of α, T is reduced (in the cases of $N = 10^1$ in Table 3), but this trend is reversed for large N (in the cases of $N = 10^3$ and 10^4 in Table 3). In a small population ($N\mu \leq 0.01$), the evolution of the duplicate genes is theoretically neutral, so T is equal to about $1/(\mu_1 + \mu_2)$ (Li 1980; Lynch and Force 2000), nearly regardless of population size. Therefore, when α is larger, T is shortened (see the cases of $N = 10^1$ in Table 3). In the larger population (see the cases of $N = 10^3$ and 10^4 in Table 3), purifying selection becomes stronger with a larger α, which results in a larger T.

Second, with symmetry of mutation rates on the duplicated loci ($\alpha = 1$), an approximation of T for unlinked duplication under the DNR selective model was given by Watterson (1983)

$$T \approx N[\log(2N) - \psi(2N\mu_1)] \quad (10)$$

where $\psi(\cdot)$ is the digamma function. Watterson stated that this formula was applicable for large N and Lynch and Force (2000) argued that it could provide a good approximation for T for unlinked duplication in the full range of N. However, further subtle comparisons show that T in simulation are slightly larger than the predictions from Eq. 10 (more data not shown), which indicate that this approximation somewhat underestimates T for unlinked duplication under the DNR selective model (see Table 3).

Third, when N is larger (for example, $N = 10^3$ and 10^4), T for unlinked duplication is usually larger than that for linked under either the DNR or HI selective models (see Table 3). This observation of T under the DNR selective model is consistent with previous results (Li 1980; Lynch and Force 2000), but it was not explained clearly. Our numerical results have explained it by originalization in above sections. On the other hand, for unlinked duplication, T under the HI selective model are larger than those under the DNR selective models and when $N(\mu_1 + \mu_2)$ and α both are approaching 1, T under the HI selective model becomes surprisingly large, which is also consistent with above theoretical results (because of originalization). In some cases ($N = 10^4$ and $\alpha = 0.8$, 1.0), values of T for unlinked ($r = 0.5$) gene duplication under the HI selective model are too large to reach nonfunctionalization, so they are written as the symbol of infinity in Table 3.

Fourth, more realistically, assume $\mu_1 = \mu_2 = 10^{-6}$. Very large T also appears for unlinked haploinsufficient gene duplication (see Fig. 4). Even in the modest population ($N > 500{,}000$ or $N\mu_1 > 0.5$), it is also very difficult to obtain T for unlinked gene duplication under the HI selective selection in simulation, because T is too large (over 10^{10} generations or 10^4 N generations), which is consistent with the above simulation observations for unlinked HI gene duplication in the cases of $\mu_1 = \mu_2 = 10^{-3}$ and $N = 10^4$ in Table 3. This indicates that the marked prolongation of T is common in a modest population (roughly $N\mu > 0.5$) during the evolution of haploinsufficient unlinked gene duplication, with realistically small mutation rates.

Although Takahata and Maruyama (1979) observed by simulation, T for unlinked duplication is larger than that for

Table 3 Simulation results of mean time to nonfunctionalization of gene duplication (in units of N generations, T/N) under the classical model[a]

N	r	α	DNR	HI	LI[b]	WAT[c]
10^1	0	0.5	53.3 (50.4)	55.1 (53.7)		
		0.8	52.3 (53.3)	53.5 (57.1)		
		1.0	51.4 (49.2)	52.7 (51.4)		
	0.5	0.5	58.5 (49.7)	57.7 (51.6)		
		0.8	56.9 (53.9)	55.7 (55.4)		
		1.0	53.7 (52.2)	54.4 (51.4)		53.4
10^2	0	0.5	10.8 (7.6)	11.6 (8.4)		
		0.8	9.7 (6.7)	10.7 (7.5)		
		1.0	9.1 (6.5)	10.1 (6.9)	9.2 (6.0)	
	0.5	0.5	12.7 (9.8)	17.2 (14.2)		
		0.8	12.0 (9.2)	18.4 (15.6)		
		1.0	11.5 (8.8)	18.4 (15.6)	12.3 (9.4)	10.9
10^3	0	0.5	3.35 (1.88)	3.94 (2.19)		
		0.8	3.63 (2.05)	4.48 (2.46)		
		1.0	3.64 (2.09)	4.52 (2.47)	5.0 (3.0)	
	0.5	0.5	4.00 (2.30)	9.45 (7.19)		
		0.8	6.49 (4.63)	136.8 (134.9)		
		1.0	7.91 (6.08)	967.1 (932.6)	12.3 (9.4)	7.2
10^4	0	0.5	0.74 (0.27)	0.92 (0.32)		
		0.8	1.48 (0.81)	1.80 (0.85)		
		1.0	2.81 (1.98)	3.31 (2.06)	4.1 (2.6)	
	0.5	0.5	0.74 (0.24)	5.77 (4.19)		
		0.8	1.47 (0.60)	∞		
		1.0	7.7 (5.8)	∞	8.1 (6.1)	6.9

[a] Values in the table are in units of N generation

[b] LI are Li's simulation results (1980)

[c] WAT are the predictions from Eq. 2 (Watterson 1983)

Parentheses are standard deviation. Degenerative mutation rate at one of the duplicated loci is $\mu_1 = 10^{-3}$ and that on another is $\mu_2 = \alpha\mu_1$, where α is equal to μ_2/μ_1. DNR and HI are simulation results under the DNR and HI selective models, respectively. Simulation repeats 10^5 times

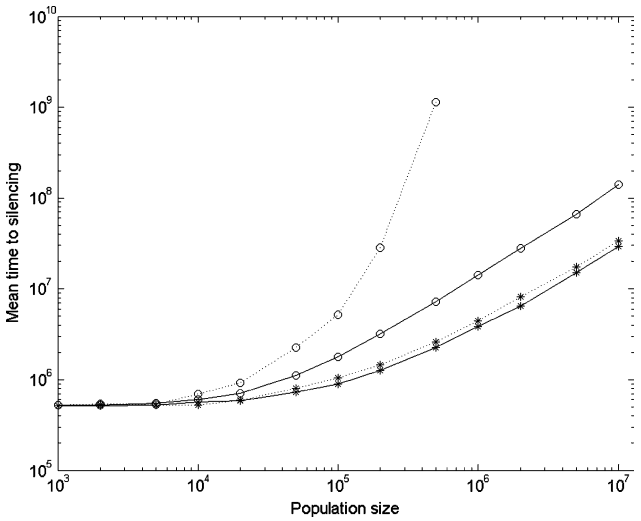

Fig. 4 Simulation results of mean time to nonfunctionalization of gene duplication under the classical model, and different selective models, where $\mu_1 = 10^{-6}$, and $\alpha = 1$. Star and circle spots are simulation results of linked ($r = 0$) and unlinked ($r = 0.5$) gene duplication, respectively. Solid and dotted lines are simulation results under the DNR and HI selective models, respectively. Simulation repeats 10^3 times

linked under the HI selective model in a large population ($N\mu_1 = 10$ and $\mu_1 = \mu_2$), the magnitude of the difference they observed is much smaller than our observation. Particularly we observed markedly prolonged T for unlinked haploinsufficient duplication even in a modest population (roughly $N\mu > 0.5$) (Table 3; Fig. 4) and with slight asymmetry of mutation rates (for example, $T = 136.8\ N$ generations in the case of $N = 10^3$ and $\alpha = 0.8$ in Table 3).

Discussion

In this study, our theoretical analyses consistently indicate that originalization might be an effective way of duplicate-gene preservation in the evolution of unlinked haploinsufficient gene duplication even only under purifying selection, especially in large-population-size organisms. This might be helpful to explain the observations that the high portion of duplicate genes originated from the ancient WGD are retained in single cell organisms, such as yeast *S. cerevisiae* genome (Wolfe and Shields 1997). Through originalization, time to nonfunctionalization of haploinsufficient gene duplication might be greatly prolonged and original genes at both duplicated loci might be maintained directly in the population. Original functional genes are more likely to accept advantageous mutations than nonfunctional genes, and during the prolonged road to nonfunctionalization advantageous genes are more likely to hit in the population, so originalization might facilitate neofunctionalization of gene duplication, by which duplicate genes are permanently preserved in the genome by selection.

Kondrashov and Koonin (2004) observed that haploinsufficient genes usually had more paralogs than haplosufficient genes. Papp et al. (2003) and Kondrashov and Koonin (2004) reported that in yeast, after ancient WGD events, haplosufficient duplicate genes were preferentially lost, while haploinsufficient duplicate genes were preferentially preserved. Our theoretical results are applicable to directly explain these two previous observations. For haplosufficient genes (under the DNR selective model), T for linked and unlinked duplications is relatively short and not sufficiently long for the original allele to wait for the arrival of advantageous mutations in the population. The quasi-equilibrium of the original chromosomal haplotype frequencies (x_0) is subject to crash because of the slight asymmetry of mutation rates (see $\alpha = 0.5$ and 0.8 in Fig. 2b). These consequences might accommodate little for advantageous genes and result in that haplosufficient duplicate genes are usually nonfunctionalization and then lost in the genome, so haplosufficient duplicate genes have fewer paralogs and are preferentially lost. By contrast, for unlinked haploinsufficient genes, duplicate genes might be preserved intact directly by originalization. On the other hand, both prolonged nonfunctionalization time and high frequency of the original gene facilitate the arrival of various advantageous mutations in the population, which results in that these various advantageous mutations are buffered without strong positive selections. However, changing environments on local subdivided populations might provide strong positive selections under which different novel genes might be fixed in different subpopulations and then gene duplicates are sequentially or accompanyingly preserved by various neofunctionalizations (unpublished materials). Therefore, the family size of haploinsufficient genes is usually larger than that of the haplosufficient genes and haploinsufficient gene duplicates are preferentially preserved.

In conventional understanding of the relationship between dominance and gene duplication (see Box 1 in Kondrashov and Koonin 2004) when the mutant allele is recessive to wild-type at the ancient locus before duplication events, fitness of the double null recessive for duplicate genes is much lower, while those of other genotypes are similar, and close to the maximum value (fitness of all wild types at duplicated loci). To simplify the mathematical description of this selection model, the DNR model was usually assumed in most theoretical studies of gene duplication (Li 1980; Watterson 1983; Lynch and Force 2000), in which the double null recessive is lethal (usually relative fitness is 0) while individuals with other genotypes have the same fitness (usually relative fitness is 1). This assumption is helpful in mathematical deriving since complex positive selection for dosage requirement is ignored. So this model can be considered as an "ideal" or preliminary model in the theoretical study. When the mutant is dominant to the wild-type allele at the ancient locus, fitness of the genotypes with no or only one wild-type allele at the duplicated loci is quite low while others are quite high, close to the maximum value (Kondrashov and Koonin 2004). Similarly, the HI selective model described above has also been treated as an "ideal" theoretical model in previous theoretical studies (Takahata and Maruyama 1979; Lynch and Force 2000). In fact, when some other complex assumptions are involved in our related studies (for example, under the HI selective model, at most one wild-type allele at the duplicated loci decreases the fitness, but is not lethal; or double null recessive is lethal, and individuals with only one wild-type allele have a low relative fitness), originalization still might appear in the evolution of unlinked gene duplication (data not shown). In experiential deletion and genomic data analysis, the prevalence of haploinsufficient genes in yeast and humans, etc., has been proved (Kondrashov and Koonin 2004; Deutschbauer et al. 2005). Therefore, DNR and HI selective models discussed in this article are possible both in theory and in reality. Of course, if the evolutionary mechanism under these "ideal" and preliminary models were well

understood, it would not be difficult for us to consider the evolution of gene duplication under more complex and realistic conditions.

Therefore, originalization might be another effective, mediated and temporary way of preserving duplicate genes, which is vastly different from other permanent-preservation methods, such as neofunctionalization and subfunctionalization.

Acknowledgments This work is partly supported by funds from 973 project (2003CB415102) and we acknowledge help from Center of High Performance Computation, Yunnan University. We thank anonymous reviewers for many valuable comments and also thank Drs. Huatao Deng, Tianhong Xu, Shuqun Liu, Yang Shen, Xianda Lu, Ren Huang, Suhua Shi, Lianghu Qu, Yupeng Cun and Michael Lynch for their assistance and Sara Barton for editorial review. The junior author also graciously acknowledges his fellowships from Guang-Dong Institute for Monitoring Laboratory Animals and Tarim Agricultural University.

Appendix

Let x_0, x_1, x_2, x_3 be the frequencies of chromosomal haplotypes, "00", "01", "10" and "11", respectively; and r be the recombination rate; mutation rates at duplicated loci are μ_1 and μ_2; for the DNR selective model, $s_1 = 0$; for the HI model, $s_1 = 1$. Fitness of individuals with various genotypes under the DNR and HI selective models are shown in Table 1. At every generation, mean population fitness and differential changes of chromosomal haplotype frequencies are given by

$$w = 1 - x_3^2 - 2s_1x_1x_3 - 2s_1x_2x_3 \tag{A1}$$

$$\begin{aligned} x_0' &= [x_0x_2/2 + x_1x_0/2 + rx_2x_1/2 + x_2x_0/2 + rx_1x_2/2 \\ &\quad + x_0x_1/2 + (1-r)x_0x_3/2 + (1-r)x_3x_0/2 + x_0^2]/w - x_0 \\ &\quad - (\mu_1 + \mu_2)x_0 \\ &= \left(x_0x_3^2 + rx_1x_2 - rx_0x_3 + 2s_1x_0x_1x_3 + 2s_1x_0x_2x_3\right)/w \\ &\quad - (\mu_1 + \mu_2)x_0 \end{aligned}$$

$$\begin{aligned} x_1' &= [(1-s_1)x_3x_1/2 + x_1x_0/2 + (1-r)x_2x_1/2 \\ &\quad + (1-r)x_1x_2/2 + x_0x_1/2 + rx_0x_3/2 \\ &\quad + (1-s_1)x_1x_3/2 + rx_3x_0/2 + x_1^2]/w - x_1 + \mu_1x_0 - \mu_2x_1 \\ &= \left(x_1x_3^2 - rx_1x_2 + rx_0x_3 - s_1x_1x_3 + 2s_1x_1^2x_3 + 2s_1x_1x_2x_3\right)/w \\ &\quad + \mu_1x_0 - \mu_2x_1 \end{aligned}$$

$$\begin{aligned} x_2' &= [x_0x_2/2 + (1-r)x_2x_1/2 + x_2x_0/2 + (1-s_1)x_2x_3/2 \\ &\quad + (1-r)x_1x_2/2 + (1-s_1)x_3x_2/2 + rx_0x_3/2 + x_2^2 \\ &\quad + rx_3x_0/2]/w - x_2 + \mu_2x_0 - \mu_1x_2 \\ &= \left(x_2x_3^2 - rx_1x_2 + rx_0x_3 - s_1x_2x_3 + 2s_1x_1x_2x_3 + 2s_1x_2^2x_3\right)/w \\ &\quad + \mu_2x_0 - \mu_1x_2 \end{aligned}$$

$$\begin{aligned} x_3' &= [(1-s_1)x_3x_1/2 + rx_2x_1/2 + (1-s_1)x_2x_3/2 \\ &\quad + rx_1x_2/2 + (1-s_1)x_3x_2/2 + (1-r)x_0x_3/2 \\ &\quad + (1-s_1)x_1x_3/2 + (1-r)x_3x_0/2]/w - x_3 \\ &\quad + \mu_2x_1 + \mu_1x_2 = \left(rx_1x_2 - rx_0x_3 - s_1x_1x_3 - s_1x_2x_3 - x_3^2\right. \\ &\quad \left. + 2s_1x_1x_3^2 + 2s_1x_2x_3^2 + x_3^3\right)/w + \mu_2x_1 + \mu_1x_2. \end{aligned} \tag{A2}$$

Because $w \approx 1$, Eq. A2 can be approximately given by

$$\begin{aligned} x_0' &\approx x_0x_3^2 + rx_1x_2 - rx_0x_3 + 2s_1x_0x_1x_3 + 2s_1x_0x_2x_3 \\ &\quad - (\mu_1 + \mu_2)x_0 \\ x_1' &\approx x_1x_3^2 - rx_1x_2 + rx_0x_3 - s_1x_1x_3 + 2s_1x_1^2x_3 + 2s_1x_1x_2x_3 \\ &\quad + \mu_1x_0 - \mu_2x_1 \\ x_2' &\approx x_2x_3^2 - rx_1x_2 + rx_0x_3 - s_1x_2x_3 + 2s_1x_1x_2x_3 + 2s_1x_2^2x_3 \\ &\quad + \mu_2x_0 - \mu_1x_2 \\ x_3' &\approx rx_1x_2 - rx_0x_3 - s_1x_1x_3 - s_1x_2x_3 - x_3^2 + 2s_1x_1x_3^2 \\ &\quad + 2s_1x_2x_3^2 + x_3^3 + \mu_2x_1 + \mu_1x_2. \end{aligned}$$

Thus, Eq. 1 in the text is obtained.

References

Byrne KP, Wolfe KH (2005) The yeast gene order browser: combing curated homology and syntenic context reveals gene fate in polyploid species. Genome Res 15:1456–1461. doi:10.1101/gr.3672305

Clement Y, Tavares R, Marais GAB (2006) Does lack of recombination enhance asymmetric evolution among duplicate genes? Insight from the *Drosophila melanogaster* genome. Gene 385:89–95. doi:10.1016/j.gene.2006.05.032

Deutschbauer AM, Jaramillo DF, Proctor M et al (2005) Mechanisms of haploinsufficiency revealed by genome-wide profiling in yeast. Genetics 169:1915–1925. doi:10.1534/genetics.104.036871

Ferris SD, Whitt GS (1979) Evolution of the differential regulation of duplicate genes after polyploidization. J Mol Evol 12:267–317. doi:10.1007/BF01732026

Force A, Lynch M, Pickett FB et al (1999) Preservation of duplicate genes by complementary, degenerative mutation. Genetics 151:1531–1545

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

Hooper SD, Berg OG (2003) On the nature of gene innovation: duplication patterns in microbial genomes. Mol Biol Evol 20:945–954. doi:10.1093/molbev/msg101

Hughes MK, Hughes AL (1993) Evolution of duplicate genes in a tetraploid animal, *Xenopus laevis*. Mol Biol Evol 10:1360–1369

Khazanie RG, McKean HE (1966a) A Mendelian Markov process with binomial transition probabilities. Biometrika 53:37–48

Khazanie RG, McKean HE (1966b) A Mendelian Markov process with multinomial transition probabilities. J Appl Probab 3:353–364. doi:10.2307/3212124

Kimura M (1955a) Solution of a process of random genetic drift with a continuous model. Proc Natl Acad Sci USA 41:144–150. doi:10.1073/pnas.41.3.144

Kimura M (1955b) Stochastic processes and distribution of gene frequencies under the natural selection. Cold Spring Harb Symp Quant Biol 20:33–53

Kimura M, King JL (1979) Fixation of a deleterious allele at one of two "duplicate" loci by mutation pressure and random drift. Proc Natl Acad Sci USA 76:2858–2861. doi:10.1073/pnas.76.6.2858

Kincaid D, Cheney W (2002) Numerical analysis: mathematics of scientific computing, 3rd edn. Brooks/Cole Publication Co, Pacific Grove

Kondrashov FA, Koonin EV (2004) A common framework for understanding the origin of genetic dominance and evolutionary fates of gene duplicates. Trends Genet 20:287–291. doi:10.1016/j.tig.2004.05.001

Li W-H (1980) Rate of gene silencing at duplicate loci: a theoretical study and interpretation of data from tetraploid fishes. Genetics 95:237–258

Li W-H, Yang J, Gu X (2005) Expression divergence between duplicate genes. Trends Genet 21:602–607

Long M-Y, Betran M, Thornton K et al (2003) The origin of new genes: glimpses from the young and old. Nat Rev Genet 4:865–875. doi:10.1038/nrg1204

Lynch M, Conery JS (2004) The origins of genomic complexity. Science 302:1401–1404. doi:10.1126/science.1089370

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

Lynch M, Katju V (2004) The altered evolutionary trajectories of gene duplicates. Trends Genet 20(11):544–549. doi:10.1016/j.tig.2004.09.001

Moore RC, Purugganan MD (2003) The early stages of duplicate gene evolution. Proc Natl Acad Sci USA 100:15682–15687. doi:10.1073/pnas.2535513100

Ohno S (1970) Evolution by gene duplication. Springer-Verlag, New York

Papp B, Pail C, Hurst LD (2003) Dosage sensitivity and the evolution of gene families in yeast. Nature 424:194–197. doi:10.1038/nature01771

Rice SH (2004) Evolutionary theory: mathematical and conceptual foundations. Sinauer Associates Inc, Sunderland

Takahata N, Maruyama T (1979) Polymorphism and loss of duplicate gene expression: a theoretical study with application to the tetraploid fish. Proc Natl Acad Sci USA 76:4521–4525. doi:10.1073/pnas.76.9.4521

Walsh JB (2003) Population-genetic models of the fates of duplicate genes. Genetica 118:279–294. doi:10.1023/A:1024194802441

Wang Y, Rannala B (2004) A novel solution for the time-dependent probability of gene fixation or loss under natural selection. Genetics 168:1081–1084. doi:10.1534/genetics.104.027797

Watterson GA (1983) On the time for gene silencing at supplicate loci. Genetics 105:745–766

Wolfe KH, Shields DC (1997) Molecular evidence for an ancient duplication of the entire yeast genome. Nature 387:708–713. doi:10.1038/42711