Genetica
DOI 10.1007/s10709-008-9319-x

Mean time to resolution of gene duplication

Cheng Xue · Yunxin Fu

Received: 13 November 2007 / Accepted: 18 August 2008

Abstract The mean time to resolution of gene duplication (T_r) is studied in this paper under the double null recessive (DNR) and haplo-insufficient (HI) models within the same analytical and simulation framework. We show that when population size is not too small (more precisely $N\mu > 0.1$), T_r for unlinked duplication is usually larger than that for linked and T_r for unlinked duplication under the HI model might be greatly prolonged, which were consistent with previous observations. Furthermore, by analytical approach we here indicate the primary underlying mechanism is that the frequency of the original (or wild-type) chromosomal haplotype of the linked duplication decreases nearly exponential to zero with time while that of the unlinked decreases quickly to an quasi-equilibrium; and this phenomenon is particularly profound under the HI model, because the quasi-equilibrium frequency of the original chromosomal haplotype (x_0) under the HI model is higher than that under the DNR model. These results suggest that recombination and HI model might jointly contribute to the marked prolongation of T_r even in a modest population. The prolonged T_r and higher quasi-equilibrium frequency of the original allele at both duplicated loci might have offered more opportunities for the emergence of novel genes.

Keywords Mean time · Resolution · Duplication · Linkage · Selection

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

Y. Fu
Human Genetics Center, School of Public Health, University of Texas at Houston, Houston, TX, USA

Introduction

Evolution through gene duplication is widely recognized as an important mechanism of molecular evolution as it provides materials for evolutionary novelty. The classical model explaining the evolutionary fate of duplicated gene postulates that gene duplication evolves by nonfunctionalization or neofunctionalization for one copy of the duplicates (Ohno 1970; Moore and Purugganan 2003; Walsh 2003; Lynch and Katju 2004). The former refers to the situation in which one of the duplicate genes becomes nonfunctional while the other maintains the original function, while the latter to the situation in which one of the duplicate genes acquires a new function and both gene copies survive.

The classical model predicts that only a very small fraction of gene duplicates will survive through neofunctionalization. This model has been recently challenged since the proportion of duplicate genes preserved in several genomes is much higher than expected under the classical model (Force et al. 1999; Lynch and Force 2000). Recognizing that many genes posses multiple functions, often through utilization of multiple regulatory regions, Force et al. (1999) proposed an alternative model, named Duplicate Degenerate and Complementary (DDC), to explain the fate of duplicate genes. The DDC model postulates that gene duplicates can be preserved through

partitioning the functions of the original gene among the duplicates so that both copies are retained complementarily. The mean time to resolution of this process is a relatively short period of time (a few millions years), and the probability of subfunctionalization is higher when duplicate genes have more subfunctions (Force et al. 1999; Lynch and Conery 2000). If this is correct, there is still a problem that is, in single cell organism, such as yeast *Saccharomyces cerevisiae*, population size is much large, which might protect duplicate genes from retaining in the genome by subfunctionalization (Lynch and Force 2000). However, in yeast, at present $\sim$10% of duplicate genes come from genomic duplication occurring $\sim$100 millions years ago (Wolfe and Shields 1997). Apparently, subfunctionalization is not a reasonable explanation for the genomic observations in yeast. Therefore, further theoretical researches on the evolution of gene duplication are needed to accommodate more genomic data observed.

For either of the two models, the fate of the gene duplicate has been studied by two related approaches. One is to analyze the probability of the preservation of gene duplication. The other is to examine the mean time, T_r, to the resolution of gene duplication. According to the definition of T_r described by Lynch and Force (2000), it is the time until the fate of the duplication is completely determined, assuming duplicate genes originating from large-scale gene duplication (such as whole genomic duplication). T_r is correlated to the rate of nonfunctionalization and subfunctionalization for gene duplication and reflects an essential facet of the evolutionary mechanism of gene duplication.

Li (1980) studied the properties of T_r under the classical model assuming the selective model to be double null recessive (DNR). When $N\mu < 0.01$ where N is the effective population size and μ is the null mutation rate of one of the gene duplicates, the two loci of gene duplication behave as neutral and T_r is about $1/(2\mu)$, regardless of the degree of linkage between the two loci. When $N\mu$ is not much larger than 1, T_r for linked and unlinked duplication are of the same order of magnitude. Watterson (1983) obtained an approximation for T_r for unlinked duplicates under the DNR model as

$$T_r \approx N[\log(2Ns) - \psi(2N\mu)] \quad (1)$$

where ψ is the digamma function and s is the purifying selection coefficient. When the double null recessive is lethal which corresponds to $s = 1$. Equation 1 becomes

$$T_r \approx N[\log(2N) - \psi(2N\mu)] \quad (2)$$

because $\psi(x)$ can be approximated by $\psi(x) \approx \log(x) - 1/x \approx \log(x)$, where $x >> 1$. Therefore T_r is approximately equal to $[\log(1/\mu)]N$ for a large population, which indicates that T_r for unlinked loci is asymptotically linear to N.

Maruyama and Takahata (1981) observed that under the DNR model T_r for unlinked duplication is much larger than that of linked in a large population. Lynch and Force (2000) examined the predictions of Eq. 2 by simulation under both the classical and DDC model and concluded that the predictions agree well with the simulation results for unlinked loci. These simulation results indicate that in a large population, T_r for both linked and unlinked duplication are approximately linear to N and T_r for the unlinked duplication is much larger than that for the linked duplication.

The aforementioned studies on T_r are all under the DNR selective model. Takahata and Maruyama (1979) studied the behavior of T_r under the Haplo-Insufficient (HI) model (also known as partial dominance) by simulation. The HI is a stronger selective model than the DNR and an individual under the HI model is viable only when there are at least two functional alleles at both duplicated loci. They showed that when $N\mu$ is large ($N\mu = 10$), T_r for unlinked duplication is much longer than that for linked under the HI model, similar to the case of the DNR model. However, a shortcoming in these studies is that these authors give no explanation on the prolongation of T_r for unlinked duplication analytically, which might help us greatly in understanding the evolutionary mechanism underlying gene duplication.

In this article, we investigate this phenomenon systematically under the classical and DDC models, and derive its proper explanation under the same theoretical framework. Both numerical analysis and simulation were used to ensure that results are consistent. In a finite population, the frequency of the original (or wild-type) chromosomal haplotype of the linked duplication decreases nearly exponential to zero with time while that of the unlinked duplication decreases quickly to a quasi-equilibrium (which means a frequency that appears to be stable) and the quasi-equilibrium frequency of the original chromosomal haplotype under the HI model is higher than that under the DNR model in a finite population.

Through this article, duplicate genes originated from ancient whole genomic duplicated are considered just like previous theoretical studies (Li 1980; Lynch and Force 2000), some genetic forces acting on small segmental duplication, such as gene conversion and unequal crossing over, are ignored.

Simulation methods and presentation of genotypes

One of our main approaches to studying T_r is computer simulation. The essence of the simulation is to follow the frequencies of various alleles in a population generation by generation until a resolution is reached. Simulation of this

type is known as perspective simulation and has been widely used in population genetics studies. In general, there are two ways to keep track of alleles in an evolving population. One is to record the genotype of each individual and the other is to record only the frequencies of alleles. The former is known as individual-based simulation and the latter as gamete-based simulation. Individual- and gamete-based simulation algorithm have been described in detail by Lynch and Force (2000). In our study, our first choice is individual-based simulation whenever it is feasible.

To simplify the description of alleles, we use chromosomal haplotype—a string of letters 0 and 1 to represent various genotypes of individuals. Assumed duplicate genes are on the same chromosome, each of which is represented by a letter with 0 and 1 meaning the original and mutant allele, respectively under the classical model considering a gene with only one function. Under the DDC model, assume a gene has 2 regulatory elements with two functions denoted as "000", in which the first two letters are for regulatory regions and that last for the coding region, so a typical chromosomal haplotype is denoted as "000000" for two sequential duplicate genes.

Assume functions considered are essential. Under the DNR selective model, the double null recessive is lethal, so under the classical model, individuals with both chromosomes being "11", are not viable, because all loci are occupied by null alleles; under the DDC model, individuals with one chromosome being "100100" and other one being "100100", are dead, because the first subfunction are lost completely on the duplicated loci. Under the HI selective model, individuals having one or none wild-type allele are not viable. For example, under the classical model, individuals with two chromosomes being "11" and "10", are also not viable, because they only have one wild-type allele on the duplicated loci; under the DDC model, individuals with two chromosomes "100100" and "100000", are also dead, because for the first subfunction, they only have one wild-type allele on the duplicated loci. Therefore, the HI model is the simplest and most representative in the dosage-requirement models.

Simulation results of the mean time to resolution under the classical and DDC model

Our simulation results about T_r under the classical model are summarized in Table 1. In particular, we compared observations under the DNR and HI selective models with those in previous studies. Several features of the evolution of gene duplication under the classical model are apparent from this Table. First, for tightly linked duplication (recombination rate $r = 0$ and $r = 10^{-3}$), our simulation results are close to Lynch's (Lynch and Force 2000) and are somewhat smaller than Li's (Li 1980) under the DNR selective model. T_r/N from our simulations in large-size populations ($N\mu \geq 10$) are quite similar for both selective models and only fluctuate in a very narrow range ($\sim$2.2–2.9 N generations). This indicates that in a large population T_r for linked gene duplication mostly depends on N.

Second, under the DNR selective model, for unlinked loci, T_r from our simulation are very close to Watterson's (1983) theoretical predictions and similar to Lynch's (Lynch and Force 2000) in the cases of $\mu = 10^{-5}$ (only simulation results with this parameter were shown in Lynch and Force's paper). In a large population, T_r in simulation is asymptotically linear to N, just as expected by Eq. 3. For example, when $N = 10^6$ and $\mu = 10^{-5}$, T_r in simulation is 11.7 N, which is very close to the prediction of $[\log(1/\mu)]N \approx 11.5$ N from Eq. 3 and Lynch's conclusion about 10 N generations (Lynch and Force 2000). These indicate that Eq. 2 can provide a good approximation of T_r for unlinked gene duplication under the DNR selective model.

Third, under both the DNR and HI selective models, T_r for unlinked duplication is usually larger than that for linked duplication, which is consistent with the observations in the previous studies (Li 1980; Lynch and Force 2000). Li (1980) reported that under the DNR model, T_r for unlinked and linked duplication are on the same order of magnitude. Our simulation results support this conclusion under the DNR model. However under the HI selective model, it is no longer the case. When $N\mu$ increases from 0.1 to 1, the difference between the T_r for unlinked duplication and the T_r for the linked duplication increases dramatically under the HI model (see Table 1). When $N\mu > 1$, the simulation becomes increasingly slow to get a resolution, so few resolutions can be obtained, and it is reported as infinity in Table 1. This indicates that T_r for unlinked duplication under the HI model is much longer than that for linked even when the population size is not too large ($0.1 < N\mu < 1$).

Finally, T_r with a low recombination rate ($r = 0.001$) has also been observed. Most results are close to Li's observations (Li 1980) when $N\mu < 0.1$, but T_r for duplication with low recombination rate ($r = 0.001$) is much larger apparently than that of linked duplication ($r = 0.0$) when this condition is violated. This indicates that even very small recombination can affect the evolution of gene duplication in a larger population.

Under the DDC model, for linked loci, it can be observed that T_r with various sets of genetic parameters are all approximately in a very narrow range of N generations when roughly $N\mu_c \geq 10$ (see Fig. 1), regardless of mutation rates and gene structure (data not shown). T_r for linked duplication is usually smaller than that for unlinked duplication under either the DNR or HI model (see Fig. 1)

Table 1 Mean times (T_r/N) to resolution of gene duplication under the classical model[a]

μ	N	r	DNR model				HI model
			T_r /N	Li[b]	GAW[c]	Lynch[d]	T_r /N
10^{-3}	10^2	0.5	11.5 (8.2)	12.3 (9.4)	10.6		18.2 (15.2)
		10^{-3}	9.2 (5.8)	9.2 (6.0)			10.2 (6.9)
		0	9.1 (6.0)				10.0 (6.9)
	10^3	0.5	7.7 (5.9)*	8.6 (6.2)	7.3		993.1 (987.6)*
		10^{-3}	4.1 (2.5)	5.0 (3.0)			6.7 (5.1)
		0	3.6 (2.1)*				4.4 (2.3)*
	10^4	0.5	7.4 (5.6)*	8.1 (6.1)	6.9		∞
		10^{-3}	3.6 (2.6)	4.1 (2.6)			3869.0 (2775.0)
		0	2.8 (2.0)*				3.0 (2.0)*
10^{-4}	10^3	0.5	13.3 (10.5)	14.3 (11.6)	12.9		25.3 (22.3)
		10^{-3}	9.9 (6.7)	10.3 (7.3)			11.7 (8.5)
		0	9.2 (6.3)				9.8 (6.43)
	10^4	0.5	9.7 (7.5)*	12.7 (8.3)	9.5		4640 (2808)
		10^{-3}	6.2 (4.3)	8.1 (5.5)			31.3 (28.8)
		0	3.6 (2.0)*				4.6 (2.5)*
	10^5	0.5	9.6 (7.4)*	9.7 (7.6)	9.2		∞
		10^{-3}	5.8 (4.4)	5.8 (4.4)			∞
		0	2.9 (2.0)*				3.2 (1.8)*
10^{-5}	10^4	0.5	15.2 (12.6)*	16.3 (13.3)	15.2	$\approx$10.3	35.9 (33.0)
		10^{-3}	13.1 (10.2)	13.4 (10.4)			17.1 (14.3)
		0	9.0 (6.0)*			$\approx$8	10.3 (7.1)*
	10^5	0.5	12.0 (9.60)	11.6 (9.3)	11.8	$\approx$10	∞
		10^{-3}	9.2 (7.3)	9.4 (7.0)			48.9 (28.1)
		0	3.7 (2.1)			$\approx$4	4.4 (2.3)
	10^6	0.5	11.7 (9.4)	10.8 (8.7)	11.5	$\approx$10	∞
		10^{-3}	5.3 (2.4)	9.3 (7.0)			∞
		0	2.9 (2.0)			$\approx$2.5	3.1 (1.8)

[a] Numbers in parentheses are standard deviations. Units of time are N generations and values in parentheses are standard deviations; values without an asterisk (*) are from simulations repeat 5,000 times, while with an asterisk (*) are from 10^6 replicates. "∞" means time to resolution is too large to be reached within current computation capacity

[b] Li's values are T_r for Li's simulation results (Li 1980) and the recombination rate for tight linked loci in Li's paper is 10^{-3}

[c] GAW's values are predictions from Eq. 2

[d] Lynch's values are read from the figure directly, so they are not exact (Lynch and Force 2000)

in a larger population, just as observed above under the classical model.

The effects of selective models on T_r are observed, which show that T_r for linked duplication under the DNR and HI models are very close to each other, while that of unlinked duplication becomes apparently distinct when the population size is larger (roughly $N\mu_c > 0.1$) and T_r for unlinked duplication becomes much larger than that for linked under the HI model, when population size is modest ($0.1 < N\mu_c < 1$) (Fig. 1). This is also consistent with the above observations under the classical model.

Therefore, several conclusions can be drawn from the simulation results: (1) T_r for linked duplication fluctuates only in a very narrow range in a large population ($N\mu > 10$); (2) in a larger population (roughly $N\mu > 0.1$), T_r for unlinked duplication is usually larger than that of linked duplication; (3) T_r for unlinked duplication is markedly larger than that for linked under the HI model, even when population size is modest ($0.1 < N\mu < 1$). Although Takahata and Maruyama (1979) observed by simulation T_r for unlinked duplication is larger than that for linked under the HI model (or partial dominance) in a large population ($N\mu = 10$), the magnitude of difference they observed is much smaller than our observation. Particularly we found that T_r for unlinked duplicates can become considerably larger even in a modest population ($0.1 < N\mu < 1$). Because

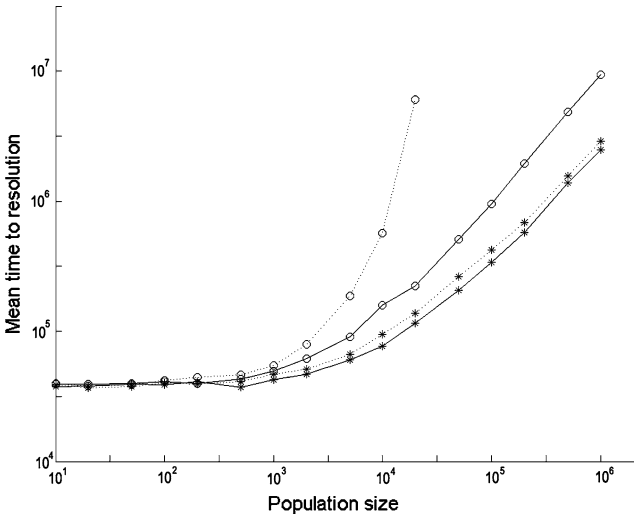

Fig. 1 Mean time to resolution of gene duplication under the DDC model, where $\mu_c = 10^{-5}$, $\mu_r = 10^{-5}$, and the number of regulatory elements $z = 2$. Star spots are simulation results of linked duplications, while circle spots are of unlinked duplications. Solid and dotted lines are simulation results under the DNR and HI selective model, respectively

of these observations, we further carried out numerical analysis and examined in more detail the dynamic changes in chromosomal haplotype frequencies.

Numerical analysis

Consider the classical model first. Assume that the population is a large random mating, diploid population. Consider a pair of duplicated loci on the same chromosome named locus 1 and 2, respectively. Frequencies of the chromosomal haplotype "00", "01", "10" and "11" are denoted as x_0, x_1, x_2 and x_3, respectively. Fitnesses of individual genotypes are shown in Table 2. Because $x_0 + x_1 + x_2 + x_3 = 1$, only 3 of 4 frequencies are independent. Therefore we will focus on the first three frequencies.

Assume $s_1 = 1$ and $s_2 = 0$ under the DNR model; $s_1 = 1$ and $s_2 = 1$ under the HI model, then total fitness

Table 2 Fitness of individual genotypes under the classical models[a]

Chromosome haplotypes	"00"	"01"	"10"	"11"
"00"	1	1	1	1
"01"	1	1	1	$1 - s_2$
"10"	1	1	1	$1 - s_2$
"11"	1	$1 - s_2$	$1 - s_2$	$1 - s_1$

[a] $s_1 = 1$ and $s_2 = 0$ under the DNR selective model while $s_1 = 1$ and $s_2 = 1$ under the HI selective model

and changes of frequencies in a generation are given by a group of ordinary differential equations (ODEs),

$$\begin{aligned} W &= 1 - 2s_2x_1x_3 - 2s_2x_1x_3 - s_1x_3^2 \\ dx_0/dt &= (x_0 - rD)/W - x_0 - 2\mu x_0; \\ dx_1/dt &= (x_1 + rD - s_2x_1x_3)/W - x_1 + (x_0 - x_1)\mu \\ dx_2/dt &= (x_2 + rD - s_2x_2x_3)/W - x_2 + (x_0 - x_2)\mu \end{aligned} \tag{3}$$

where t is time (generation), r is the recombination rate and D is the linkage disequilibrium, which is equal to $x_0x_3 - x_1x_2$.

Under the DNR model,

$$\begin{aligned} W &= 1 - s_1x_3^2 \\ dx_0/dt &= \left(x_0x_3^2 - rD\right)/W - 2x_0\mu \\ dx_1/dt &= \left(x_1x_3^2 + rD\right)/W + (x_0 - x_1)\mu \\ dx_2/dt &= \left(x_2x_3^2 + rD\right)/W + (x_0 - x_2)\mu. \end{aligned} \tag{4}$$

Equation 4 is essentially the same as Li (1980). The initial conditions for these equations are $x_{1(t\,=\,0)} = 1$, $x_{2(t\,=\,0)} = 0$ and $x_{3(t\,=\,0)} = 0$. We obtained numerical solutions to these equations by the Runge-Kutta method (Kincaid and Cheney 2002), given proper initial conditions and models, for example, selective models and recombination rate. We shall present only the results of the analysis while technical detail of the analysis is available upon request.

Most previous mathematical studies on gene duplication (Watterson 1983; Kumara and King, 1979; Takahata and Maruyama 1979) used a different method to represent the genotypes. They assumed the original allele on locus 1 is A, while the mutant allele is a, and the original allele on locus 2 is B, while the mutant allele is b. Thus, AABB is homozygous individual, and AaBb is heterozygous. The main discrepancy of this representation and ours above with '0' and '1' is that "01" is for chromosomal direction and Aa is for locus direction. It is convenient and simplified in mathematical derivation by use of the locus-direction representation without considering the effect of linkage. However, when considering the effect of linkage in this study (see above), it is necessary to use the chromosome-direction representation. Because throughout this article, we assume mutation rates on both loci are the same and selective models are the DNR and HI models, in fact, frequencies of alleles with these two-direction representations are symmetrically equal, for example, homozygous frequencies of AA and BB are equal to that of "00", and heterozygosity, frequencies of Aa and Bb are equal to the frequencies of "10" and "01", etc.

Kimura and King (1979) observed that at the mutation-selection balance for unlinked gene duplication, under the classical and DNR models, the frequencies of p and q (p, q are frequencies of the null alleles at the two duplicate loci,

a and b, respectively) will reach an equilibrium in order that $p^2q^2 = \mu/s$. Therefore, the balance frequency of the double null allele, aa or "11", for unlinked duplication under the DNR model is expected to be $\sqrt{\mu/s} \approx 0.03162$ when μ is 10^{-3} and $s = 1$. In our analysis results, x_3, the frequency of "11", is the same as this expectation. Because mutation rates on two duplicated loci are assumed to be equal, p and q are the same at the selection-mutation balance. The frequency of the null allele at a locus, p, is expected to be $p = \sqrt{pq} = \sqrt{x_3} \approx 0.1778$, which is also consistent with our numerical results (data not shown).

Because at the selection-mutation balance heterozygosity for unlinked duplication under the DNR and HI models is approximately equal to x_1 or x_2 in a large population, it is also observed as a function of mutation rate (see Fig. 2). Numerical results show that the equilibrium frequency of x_0 under the HI model is larger than that under the DNR model, while equilibrium frequency of x_1 or x_2 under the HI model is smaller than that under the DNR model; and the equilibrium frequency of x_0 decreases when mutation rate is larger. These indicate that heterozygosity for unlinked duplication increases when mutation rate is larger, and is usually larger under the DNR model than that under the HI model.

In addition, x_0 of unlinked duplication decreases quickly to a mutation-selection balance with time (generation) while that of linked duplication decreases nearly exponential to zero, which result in the larger T_r for unlinked duplication in a finite population. In another way, x_0 of unlinked duplication at the mutation-selection balance under the HI model is higher than that under the DNR model (see Figs. 2 and 3), which might result from different selection pressure.

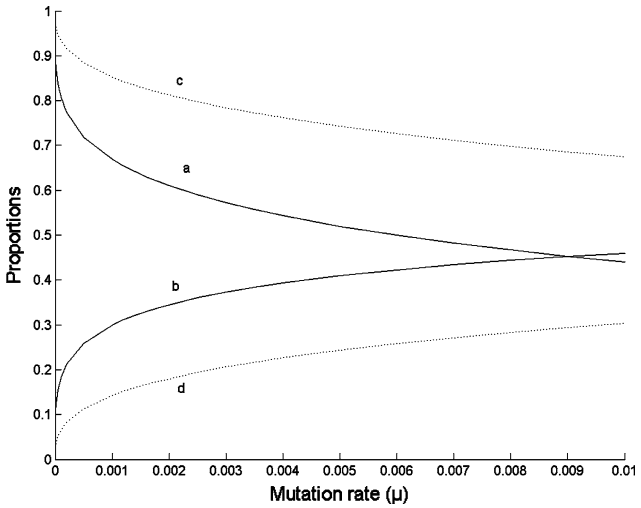

Fig. 2 Equilibrium frequencies of alleles for unlinked duplication at the mutation-selection balance in a large population under the classical model, as a function of mutation rate. Curves a and b represent numerical results for x_0 and $x_1 + x_2$ under the DNR model, respectively; c and d represent for x_0 and $x_1 + x_2$ under the HI model, respectively

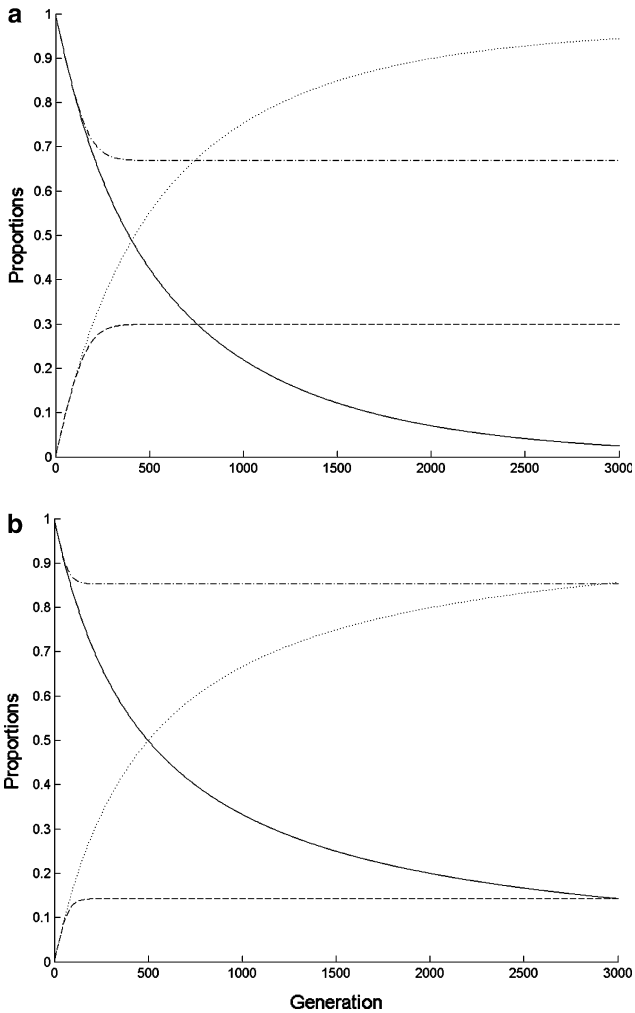

Fig. 3 Dynamical changes of allele frequencies of gene duplication with time from numerical analysis under the DNR and HI model, where $\mu = 10^{-3}$. Curves in subplot a are results under the DNR model, and in subplot b are under the HI model; solid and dash-dotted curves are x_0 of linked and unlinked loci, respectively; dotted and dashed curves are $x_1 + x_2$ of linked and unlinked loci, respectively

Because at resolution x_0 should be 0, for unlinked duplication, the equilibrium proportion of x_0 resulting from free recombination and higher equilibrium proportion of x_0 from the HI model might contribute jointly to the obvious prolongation of T_r observed in the larger population in the above simulation under the HI model.

Numerical analysis is also carried out under the DDC model. However, since the expressions of the ODEs are too lengthy, they will not be shown here. Numerical results show that the behavior of chromosomal haplotype frequencies at every generation is consistent with that under the classical model, especially the frequency x_0 of the original chromosomal haplotype "000000".

To examine whether the above numerical analysis is correct, we keep track of dynamical changes of the frequencies of chromosomal haplotypes in our simulation.

Indeed simulation results agree quite well with the numerical analysis (data not shown).

Discussion

There are two noteworthy conclusions in this research. First, T_r for unlinked duplication is usually larger than that of linked duplication when the population size is not small ($N\mu > 0.1$). Second, T_r for unlinked duplication is much larger than that for linked under the HI model even when the population size is modestly large ($0.1 < N\mu < 1$).

We also show a reasonable explanation underlying these two observations. The evolutionary trajectories of unlinked and linked duplications are shown to be quite different. In a larger finite population, the frequency of the original chromosomal haplotype for linked duplications diminishes quickly with time, while that for unlinked duplication is kept high in the population and fluctuates around the quasi-equilibrium due to mutation-selection balance (see Fig. 3). On one hand, these dynamic features result in a shorter T_r for linked than unlinked duplication. On the other hand, they might also provide more opportunities for accumulating advantageous mutations on the way to resolution of unlinked duplication than linked duplication. This suggests that recombination facilitate the emergence of novel genes. Since T_r of linked and unlinked duplications are of the same order of magnitude (Li 1980), the prolongation of T_r for unlinked gene duplication might primarily result from a higher proportion of the original allele at both unlinked duplicated loci under the classical selective model.

Data from tetraploid fish, for example, in catastomid fish, showed that the rate of resolution of gene duplication is quite slow and polymorphism is also quite small, which is not consistent with the expectation of the DNR model (Takahata and Maruyama 1979). However, this phenomenon can be explained easily under the HI model. Under the HI model, T_r for gene duplication resulting from tetraploidization might be considerably long, so the rate of resolution of gene duplication is quite slow.

In catastomid fish, about 35–65% of the genomic duplicates resulting from tetraploidization about 50 million years ago, are fixed and lost, but frequencies of the unfixed null alleles in the population is small (Ferris and Whitt 1977). In the light of our numerical analysis, if mutation rate to null is 10^{-5}, the quasi-equilibrium frequency of heterozygosity is about 0.106 under the DNR model and 0.030 under the HI model (data not shown). In a finite population, heterozygosity fluctuates around this quasi-equilibrium with the mean equal to d_{x1}/d_t (shown in Eq. 3) and the variance equal to $x_1(1-x_1)/(2\,N)$ of changes (Kimura and King 1979; Tajima 1990). Apparently, heterozygosity under the DNR model fluctuates in a larger range, which support the statement that observation of small heterozygosity in catastomid fish might not be explained under the DNR model (Takahata and Maruyama 1979). However, it can be explained under the HI model because of much lower heterozygosity and shorten fluctuating range in the numerical analysis.

Haplo-insufficient genes usually have more paralogs in the population than that of haplo-sufficient genes (Kondrashov and Koonin 2004). Because the condition of $N\mu > 0.1$ is not difficult to meet, T_r for unlinked duplication under the HI model is likely to be much longer than that under the DNR model. During the prolonged voyage of evolution in haplo-insufficient duplicate genes, there are more opportunities for advantageous mutation to arise in the population and each may be preserved in a low frequency. Without very strong selection, these advantageous mutations are difficult to be fixed in the large population. However, fluctuating environments from time to time intensify the selection, which might result in the fixation of some of these advantageous mutations, particularly when the population is subdivided. The accumulation of such divergent mutations might facilitate speciation, because speciation is a differential process at the genic level with differential adaptations (Wu 2001). Reversibly, speciation accelerates divergence of gene duplicates furthermore resulting in more prologs. Thus more prologs are observed in haplo-insufficient gene families than in haplo-sufficient gene families.

Acknowledgements This work is partly supported by fund from 973 project (2003CB415102), and we thank the help from Center of Performance Computation of Yunnan University. We also thank Drs. Shuqun Liu, Yang Shen, Xianda Lu, Ren Huang, Suhua Shi, Lianghu Qu, and M. Lynch for their helps and Sara Barton for editorial assistance. The junior author also graciously acknowledges his fellowships from GuangDong Institute for Monitoring Laboratory Animals and Tarim Agricultural University.

References

Ferris SD, Whitt GS (1977) Loss of duplicate gene expression after polyploidisation. Nature 265:258–260. doi:10.1038/265258a0

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

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 Pub. 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

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

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

Maruyama T, Takahata N (1981) Numerical studies of the frequency trajectories in the process of null genes at duplicated loci. Heredity 46:49–57. doi:10.1038/hdy.1981.5

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

Tajima F (1990) Relationship between DNA polymorphism and fixation time. Genetics 125:447–454

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

Watterson GA (1983) On the time for gene silencing at duplicate 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

Wu C-I (2001) The genic view of the process of speciation. J Evol Biol 14:851–865. doi:10.1046/j.1420-9101.2001.00335.x