BMC Genetics

RESEARCH ARTICLE Open Access

Recombination facilitates neofunctionalization of duplicate genes via originalization

Cheng Xue*[1,2], Ren Huang[1,2], Shu-Qun Liu[3] and Yun-Xin Fu[3,4]

Abstract

Background: Recently originalization was proposed to be an effective way of duplicate-gene preservation, in which recombination provokes the high frequency of original (or wild-type) allele on both duplicated loci. Because the high frequency of wild-type allele might drive the arising and accumulating of advantageous mutation, it is hypothesized that recombination might enlarge the probability of neofunctionalization (P_{neo}) of duplicate genes. In this article this hypothesis has been tested theoretically.

Results: Results show that through originalization recombination might not only shorten mean time to neofunctionalizaiton, but also enlarge P_{neo}.

Conclusions: Therefore, recombination might facilitate neofunctionalization via originalization. Several extensive applications of these results on genomic evolution have been discussed: 1. Time to nonfunctionalization can be much longer than a few million generations expected before; 2. Homogenization on duplicated loci results from not only gene conversion, but also originalization; 3. Although the rate of advantageous mutation is much small compared with that of degenerative mutation, P_{neo} cannot be expected to be small.

Background

Gene duplication is the most common way of evolving new genes [1-4], but it is still argued how new genes evolve from duplicate genes in detail [5-7]. Ohno (1970) proposed that new genes might be fixed at one of duplicated loci by genetic drift, which was called neofunctionalization. Because degenerative mutations might also be fixed on the duplicated loci (called nonfunctionalization) and the occurring rate of degenerative mutation is usually much larger than that of advantageous mutation, the evolutionary fate of most duplicate genes is nonfunctionalization [8]. However, it has been observed that many duplicate genes are retained in some genomes, such as in tetraploid fish [9], *Xenopus Laevis* [10], and yeast *Saccharomyces cerevisiae* [4,11,12]. So it is necessary to explain these observations reasonably.

Assuming double null recessive selection and unlinked duplicated loci, Walsh (1995 and 2003) modeled the state of the population as a three-state (wild-type, degenerative and advantageous alleles) Markov chain, and thus calculated the probability (P_{neo}) that the advantageous allele will fix before the nonfunctional allele does [13,14]. Under weak positive selection (roughly $Ns << 1$), P_{neo} was given by

$$P_{neo} = \{1 + [1 - EXP(-4Ns)] / (4Ns\rho)\}^{-1} \quad (1)$$

where EXP is the exponential function, ρ is the ratio of advantageous mutation rate (μ_{neo}) to degenerative mutation rate (μ_{non}), N is effective population size, and s is positive selection coefficient. Under strong positive selection, this formula is corrected,

$$P_{neo} = 1 - EXP(-8N^2 s\mu_{neo}) / (1 + 4Ns\rho) \quad (2)$$

And Walsh (2003) also suggested that recombination might enlarge P_{neo}, but he neither provided theoretical evidences, nor gave further explanation or hypothesis [14]. Recently Xue and Fu observed a mathematical process that we named originalization during the evolution of gene duplication under recombination, which can explain this suggestion [15]. During originalization, under purifying selection recombination results in the

* Correspondence: lflf27@yahoo.com.cn
[1] GuangDong Institute for Monitoring Laboratory Animals, Guangzhou, China
Full list of author information is available at the end of the article

higher frequency of the original allele on both duplicated loci, so mean time to nonfunctionalization (T_{non}) is prolonged. And it was hypothesized that prolonged T_{non} and high frequencies of the wild-type allele might confer the arising and accumulating of advantageous alleles in the population, so that P_{neo} might become larger [15-17].

In this article, we will test the hypothesis of enlarged P_{neo} for unlinked gene duplication by originalization, and explore the underlying mechanism. Our results show that under stronger positive selection (Roughly Ns > 0.5) and in larger populations (Roughly N μ_{non} > 0.1) recombination not only enlarges P_{neo}, but also shortens mean time to neofunctionalization of duplicate genes (T_{neo}). Therefore, through originalization recombination facilitates neofunctionalization of duplicate genes.

Results

Assumptions and notations

Assume that the duplicate genes originated from polyploidization, such as ancient whole genomic duplication (WGD), so that the effects of some genetic forces on small segmental duplications, such as unequal crossing over and gene conversion, are ignored, as assumed in previous theoretical studies on neofunctionalization of duplicate genes [13,14].

Assume in a random mating, diploid population, chromosomal haplotype is used to represent various genotypes of individuals [15,16]. Considering advantageous and degenerative mutations, there are three types of alleles at one of duplicated loci: wild-type allele (denoted as a character '0'), degenerative allele (denoted as a character '1'), and advantageous allele (denoted as a character '2'). In this way, there are nine possible types of chromosomal haplotypes in the population, namely, "00", "01", "02", "10", "11", "12", "20", "21" and "22", respectively.

We use the DNR (double null recessive or haplosufficient) and haploinsufficient (HI) selective models presented in our previous studies [15,16]. Under the DNR selective model, individuals with no wild-type allele at both of duplicated loci are invalid (relative fitness is 0), for example, individuals with chromosomal haplotypes "11" and "11", or "12" and "22". Under the HI selective model individuals with at least two copy of wild-type alleles on duplicated loci are valid. Assume mutation rates are the same on the duplicated loci; Transition (or mutation) from original allele to degenerative or advantageous allele is irreversible; Mutations from degenerative to advantageous and from advantageous to degenerative are ignored.

Under these assumptions, we report mean time to neofunctionalization (T_{neo}) under the model only involving neofunctionalization and P_{neo} under the model involving neofunctionalization and nonfunctionalization (details of the models are shown below).

Mean time to neofunctionalization for gene duplication

Model

Let's consider a very simple model only involving neofunctionalization for duplicate genes at first. In this model, there are only four types of possible chromosomal haplotypes in the population, "00", "02", "20" and "22", and their frequencies in the population are denoted as x_0, x_1, x_2 and x_3 respectively. Because $x_0+x_1+x_2+x_3 = 1$, three of these four frequencies are independent and x_0, x_1, x_2 are focused. Assume advantageous mutations are additive with fitness 1+ks for k advantageous allele(s) totally at duplicated loci. Fitnesses of individuals with various genotypes are shown in Table 1. Thus, without considering genetic drift (i.e. in an infinite population), differential changes of chromosomal haplotype frequencies at every generation, are given by a group of ordinary differential equations
(ODEs),

$$w = 1 - x_3^2 + 2 s x_0 x_1 + 2 s x_1^2 + 2 s x_0 x_2 + 4 s x_1 x_2 + 2 s x_2^2 + 4 s x_0 x_3 + 6 s x_1 x_3 - 2 s_1 x_1 x_3 - 6 s s_1 x_1 x_3 + 6 s x_2 x_3 - 2 s_1 x_2 x_3 - 6 s s_1 x_2 x_3$$

$$x_0' = \left(\begin{array}{l} x_0 + s x_0 x_1 + s x_0 x_2 + r x_1 x_2 + 2 r s x_1 x_2 - r x_0 x_3 + \\ 2 s x_0 x_3 - 2 r s x_0 x_3 \end{array} \right) / w - x_0 - 2 x_0 \mu_{neo}$$

$$x_1' = \left(\begin{array}{l} x_1 + s x_0 x_1 + 2 s x_1^2 - r x_1 x_2 + 2 s x_1 x_2 - 2 r s x_1 x_2 + r x_0 x_3 + \\ 2 r s x_0 x_3 + 3 s x_1 x_3 - s_1 x_1 x_3 - 3 s s_1 x_1 x_3 \end{array} \right) / w - x_1 + x_0 \mu_{neo} - x_1 \mu_{neo}$$

$$x_2' = \left(\begin{array}{l} x_2 + s x_0 x_2 - r x_1 x_2 + 2 s x_1 x_2 - 2 r s x_1 x_2 + 2 s x_2^2 + r x_0 x_3 + \\ 2 r s x_0 x_3 + 3 s x_2 x_3 - s_1 x_2 x_3 - 3 s s_1 x_2 x_3 \end{array} \right) / w - x_2 + x_0 \mu_{neo} - x_2 \mu_{neo}$$

Table 1: Fitnesses of individual genotypes for neofunctionalizaion of gene duplication *

Chromosomal Haplotypes	"00"	"02"	"20"	"22"
"00"	1	1+s	1+s	1+2s
"02"	1+s	1+2s	1+2s	$(1-s_1)(1+3s)$
"20"	1+s	1+2s	1+2s	$(1-s_1)(1+3s)$
"22"	1+2s	$(1-s_1)(1+3s)$	$(1-s_1)(1+3s)$	0

* s is positive selection coefficient. Under the DNR selective model, $s_1 = 0$, while under the HI selective model, $s_1 = 1$.

where w is mean population fitness; r is the recombination rate between two duplicated loci; μ_{neo} is the rate of advantageous mutation; under the DNR selective model, $s_1 = 0$, while $s_1 = 1$ under the HI selective model.

Based on these ODEs, given $\mu_{neo} = 10^{-6}$, dynamic changes of chromosomal haplotype frequencies were numerically obtained by the Runge-Kutta method [18] given initial conditions $x_0 = 1$, and $x_1 = x_2 = 0$; with considering genetic drift (i.e. in an finite population) simulations were also carried out to test the numerical results.

Numerical results

In an infinite population dynamic changes of chromosome haplotypes under strong positive selection (s = 0.01) are shown in Figure 1. For linked gene duplication, the frequency of original chromosomal haplotype, x_0, decreases nearly exponentially down to 0; x_1 and x_2 increase continually up to ~0.5. However, for unlinked gene duplication, the behaviors of chromosomal haplotype frequencies are more interesting. Initially, x_0 decreases to an equilibrium and then is kept at a high level while x_1 and x_2 increase also to equilibrium. This equilibrium is kept for a period of time, and then it crashes suddenly, in which x_0 drops down to very low (close to 0) suddenly, and so does one of x_1 and x_2 while another increases up to ~1 (see Figure 1). At neofunctionalization, x_1 or x_2 are equal to 1, so these numerical results suggest that in finite and large populations T_{neo} for unlinked duplicate genes might be shorter than that for

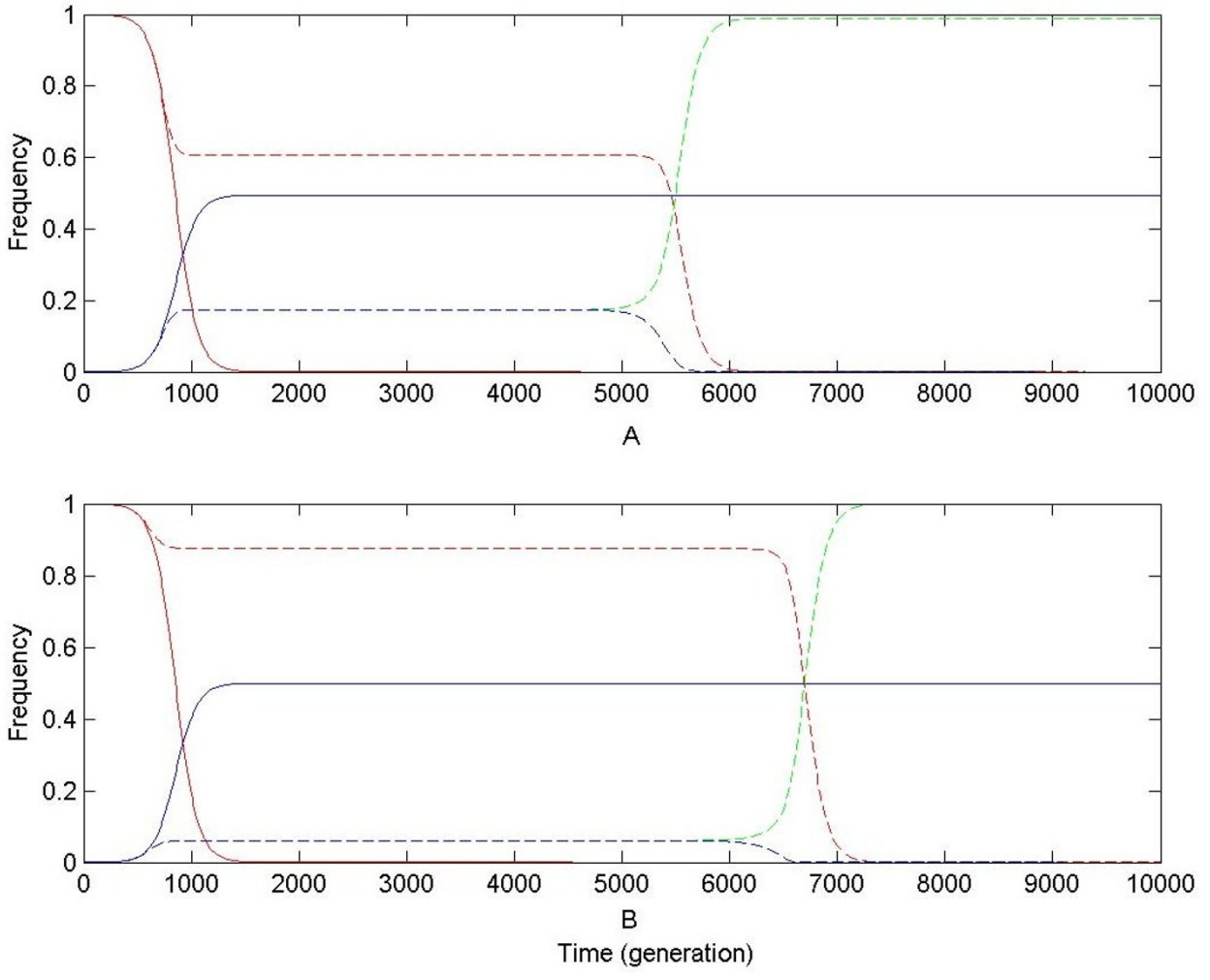

Figure 1 Dynamic changes of chromosomal haplotype frequencies for gene duplication during neofunctionalization under strong positive selection. Assume s = 0.01 and $\mu_{neo} = 10^{-6}$. In subplot **A**, are numerical results under the DNR selective model; in subplot **B**, numerical results under the HI selective model. Solid and dashed curves are numerical results for linked and unlinked gene duplication, respectively. Red, green and blue curves are numerical results for frequencies of chromosomal haplotypes "00", "02" and "20", corresponding to x_0, x_1 and x_2, respectively. In subplots A and B, for linked gene duplication, curves of x_1 and x_2 are completely coincident.

linked. Under recombination high x_0 in the population was named originalization [15], which descirbes the main difference between evolutionary trajectories of unlinked and linked gene duplications (see Figure 1; also see Ref. [15] and [16]). Therefore, these observations suggest that by originalization, under strong positive selection recombination contribute to shortened T_{neo} for unlinked gene duplication.

Simulation results

To examine this prediction of shortened T_{neo} for unlinked duplicate genes in large populations, simulation results in a larger population ($N\mu_{neo} = 0.2$) are shown in Figure 2. Of course, similar results are obtained in other larger populations ($N\mu_{neo} > 0.2$) (not shown). However, even when $N\mu_{neo} = 0.2$, the results sufficiently indicate that T_{neo} for unlinked duplicate genes is shortened when positive selection is strong (see Figure 2).

If s is small enough (or close to 0), the evolutionary behavior of an advantageous mutation is similar to that of a nearly neutral mutation [19]. Therefore, in simulation, when s is small (for example, $s = 10^{-7}$ in Figure 2) and population size is not small (roughly $N\mu_{neo} > 0.1$), T_{neo} for unlinked gene duplication is larger than that for linked under the either DNR or HI selective model; and T_{neo} for unlinked gene duplication becomes greatly prolonged under the HI selective model (see Figure 2). These observations are very consistent with those of degenerative mutations in previous studies [15,16,20-23]. When s is large (for example, $s = 0.01$), T_{neo} for unlinked gene duplication is much shortened and smaller than that for linked (see Figure 2), which is in agreement with above numerical results.

In our previous studies [15,16], we observed that under recombination T_{non} can be prolonged in a larger popula-

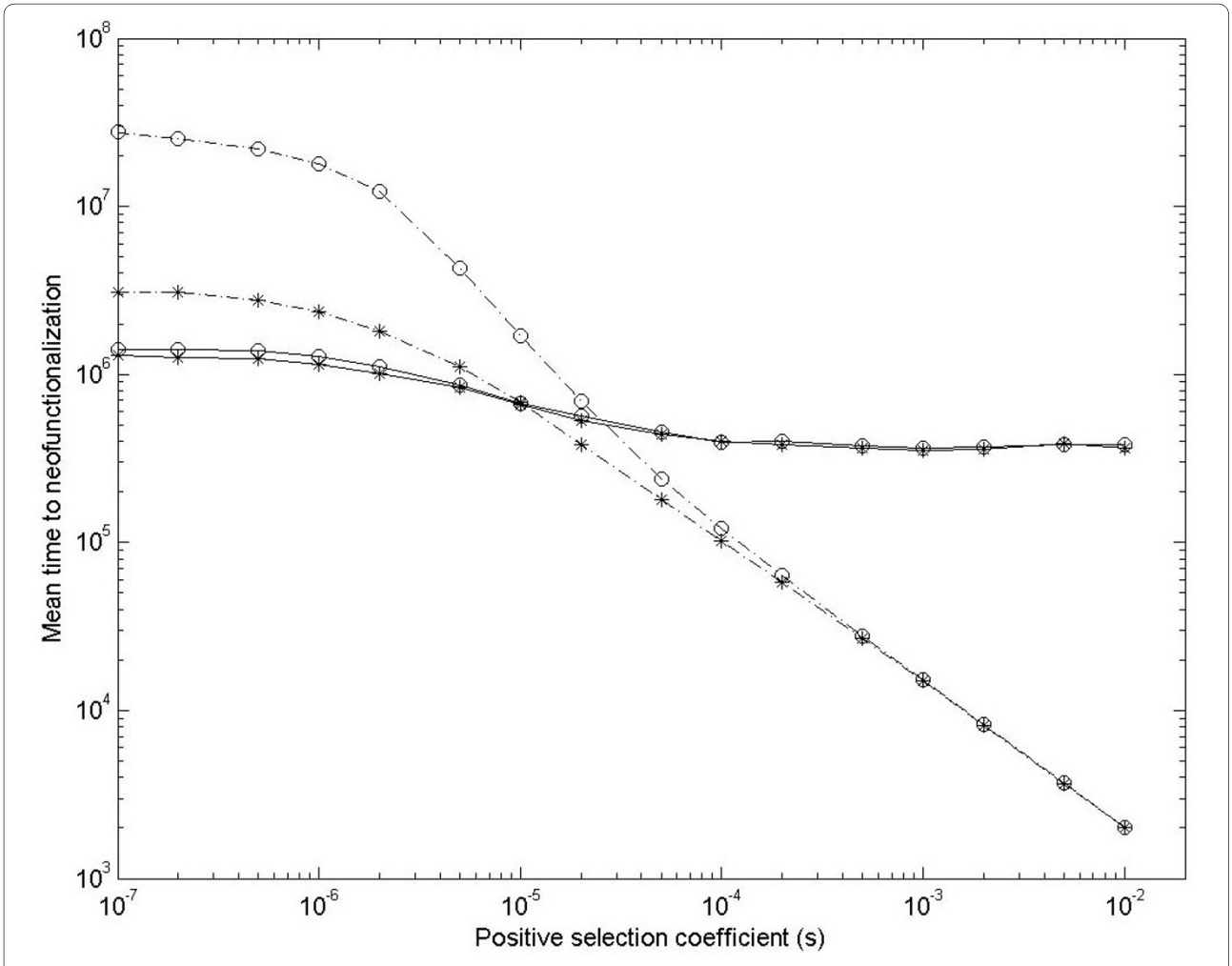

Figure 2 Simulation results for mean time to neofunctionalization of gene duplication with positive selection coefficient. Assume N = 200000 and $\mu_{neo} = 10^{-6}$. Star and circle spots are simulation results under the DNR and HI selective models, respectively. Solid and dash-dot lines are simulation results for linked and unlinked gene duplication, respectively. Simulation repeats 3000 times.

Table 2: Fitnesses of individual genotypes for resolution (neofunctionalizaion and nonfunctionalization) of gene duplication*

chromosomal haplotypes	"00"	"01"	"02"	"10"	"11"	"12"	"20"	"21"	"22"
"00"	1	1	1+s	1	1	1+s	1+s	1+s	1+2s
"01"	1	1	1+s	1	$1-s_1$	$(1-s_1)(1+s)$	1+s	$(1-s_{1)}(1+s)$	$(1-s_{1)}(1+2s)$
"02"	1+s	1+s	1+2s	1+s	$(1-s_1)(1+s)$	$(1-s_1)(1+2s)$	1+2s	$(1-s_1)(1+2s)$	$(1-s_1)(1+3s)$
"10"	1	1	1+s	1	$1-s_1$	$(1-s_1)(1+s)$	1+s	$(1-s_1)(1+s)$	$(1-s_1)(1+2s)$
"11"	1	$1-s_1$	$(1-s_1)(1+s)$	$1-s_1$	0	0	$(1-s_1)(1+s)$	0	0
"12"	1+s	$(1-s_1)(1+s)$	$(1-s_1)(1+2s)$	$(1-s_1)(1+s)$	0	0	$(1-s_1)(1+2s)$	0	0
"20"	1+s	1+s	1+2s	1+s	$(1-s_1)(1+s)$	$(1-s_1)(1+2s)$	1+2s	$(1-s_1)(1+2s)$	$(1-s_1)(1+3s)$
"21"	1+s	$(1-s_1)(1+s)$	$(1-s_1)(1+2s)$	$(1-s_1)(1+s)$	0	0	$(1-s_1)(1+2s)$	0	0
"22"	1+2s	$(1-s_1)(1+2s)$	$(1-s_1)(1+3s)$	$(1-s_1)(1+2s)$	0	0	$(1-s_1)(1+3s)$	0	0

* s is positive selection coefficient. Under the DNR selective model, $s_1 = 0$, while under the HI selective model, $s_1 = 1$.

tion (roughly N μ_{non} > 0.1); and x_0 is kept higher in the population. So prolonged T_{non}, shortened T_{neo} and high x_0 might jointly result in larger P_{neo} for unlinked gene duplication. In order to validate this prediction, direct observations of P_{neo} are also carried out.

Probability of neofunctionalization for gene duplication

Model

Now consider a model involving neofunctionalziation and nonfunctionalization. In the gene pool, there are nine possible chromosomal haplotyes in the population, "00", "01", "02", "10", "11", "12", "20", "21", "22", whose frequencies are denoted as y_0, y_1, y_2, y_3, y_4, y_5, y_6, y_7, y_8, respectively. Fitnesses of individuals with various genotypes are shown in Table 2. Under these conditions, in an infinite population another group of ODEs, just like Equation 3, have been obtained. Their expressions are too lengthy, so they are provided in Appendix. Numerical and simulation methods are the same as those in the above section. Numerical and simulation results were also obtained with the rate of degenerative mutation (μ_{non}) = 10^{-4} and that of advantageous (μ_{neo}) = 10^{-6}. Initially let $y_0 = 1$, and $y_1 = y_2 = y_3 = y_4 = y_5 = y_6 = y_7 = y_8 = 0$.

Numerical results

Numerical results are shown in Figure 3 and 4. P_{neo} can be approximately expressed as $y_2+y_5+y_8$ or $y_6+y_7+y_8$, and the

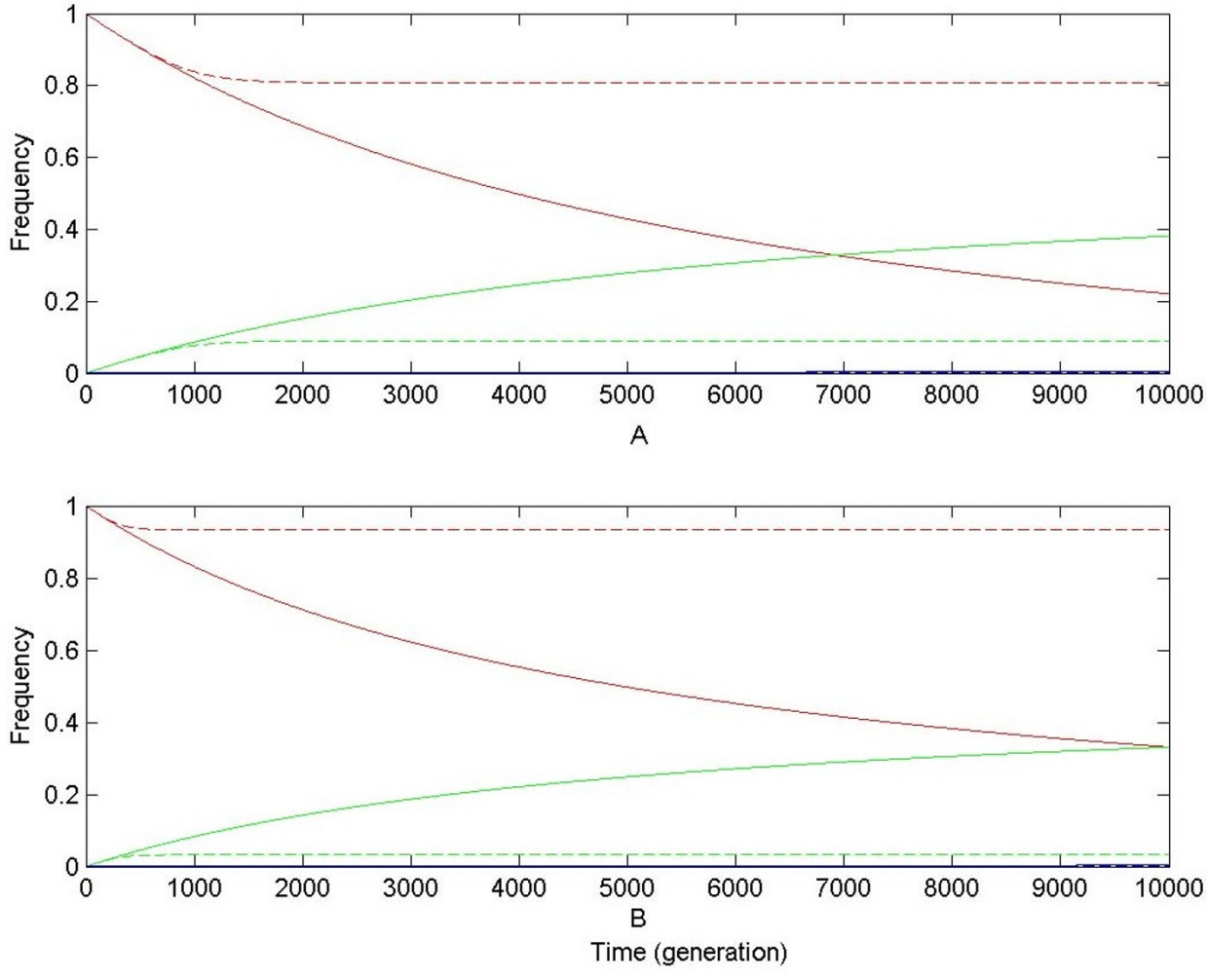

Figure 3 Dynamic changes of chromosomal haplotype frequencies for gene duplication during resolution (neofunctionalization and nonfunctionalization) under slight positive selection. Assume $\mu_{neo} = 10^{-6}$, $\mu_{non} = 10^{-4}$, and $s = 10^{-6}$. In subplot **A**, numerical results are obtained under the DNR selective model; in subplot **B**, numerical results under the HI selective model. Solid and dashed curves are numerical results for linked and unlinked gene duplication, respectively. Red, gree and blue curves are numerical results for frequencies of chromosomal haplotypes "00", "01" (or "10") and "02" (or "20"), corresponding to y_0, y_1 and y_2, respectively. In subplots A and B, for linked gene duplication, curves of y_2 are nearly coincident with x-axis.

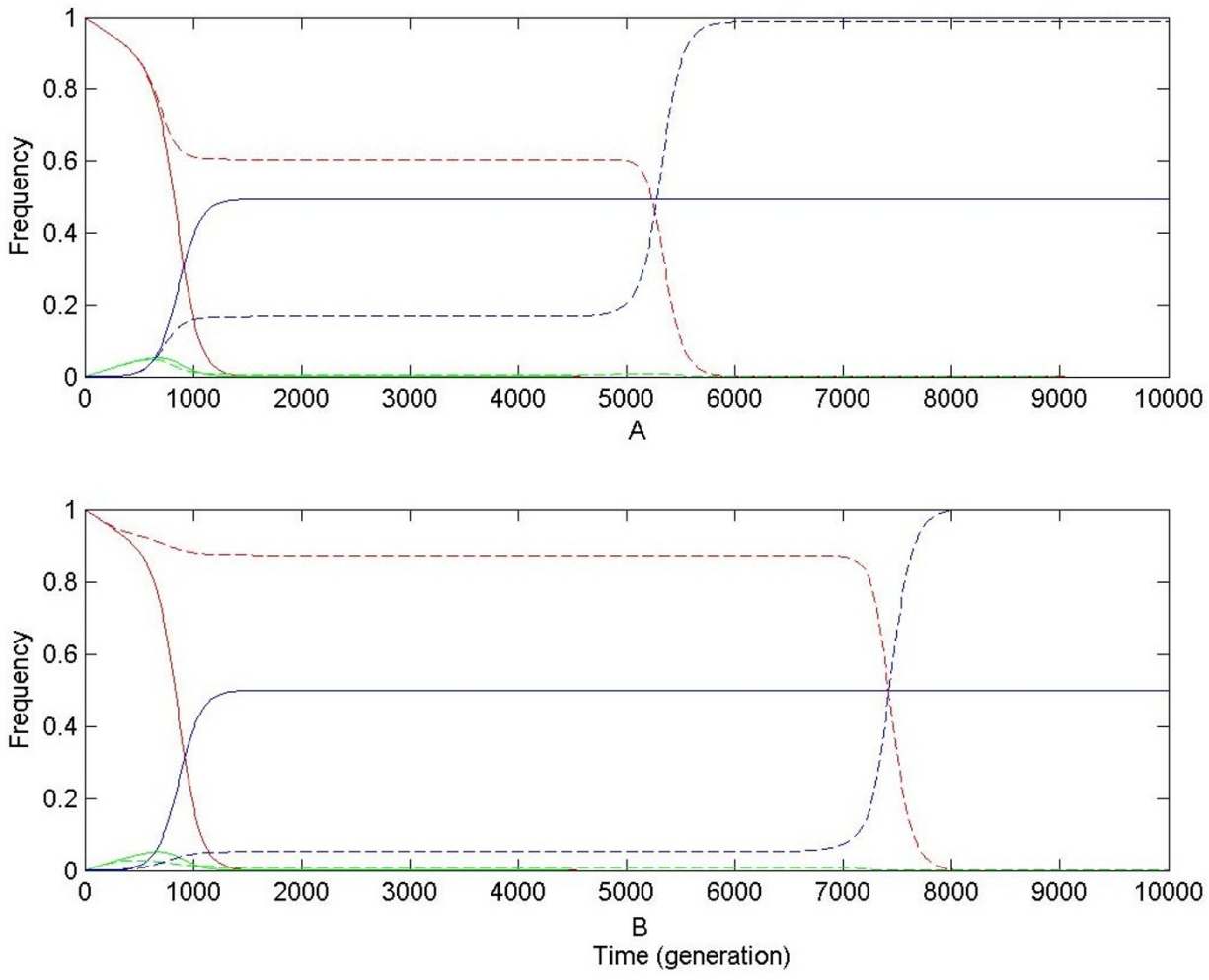

Figure 4 Dynamic changes of chromosomal haplotype frequencies for gene duplication during resolution (neofunctionalization and non-functionalization) under strong positive selection. Assume $\mu_{neo} = 10^{-6}$, $\mu_{non} = 10^{-4}$, and s = 0.01. In subplot **A**, numerical results are obtained under the DNR selective model; in subplot **B**, numerical results under the HI selective model. Solid and dashed curves are numerical results for linked and unlinked gene duplication, respectively. Red, green and blue curves are numerical results for frequencies of chromosomal haplotypes "00", "01" (or "10") and "02" (or "20"), corresponding to y_0, y_1 (or y_4) and y_2 (or y_6), respectively. In subplots A and B, for linked gene duplication, curves of y_1 are nearly coincident with x-axis.

probability of nonfunctionalization as $y_1+y_4+y_7$ or $y_3+y_4+y_5$. Because under the DNR and HI selective model described above, y_4, y_5, y_7 and y_8 are quite small and close to 0, P_{neo} is approximately equal to y_2 or y_6, and the probability of nonfunctionalization is approximately equal to y_1 or y_3. So only dynamic changes of y_2 and y_1 are shown in numerical results as the proxies for the probabilities of neofunctionalization and nonfunctionalization, respectively, and y_0 is treated as a proxy of non-resolution (or originalization) [15].

When positive selection is slight ($s = 10^{-6}$), for unlinked gene duplication, an equilibrium is quickly reached for y_0, y_1, and low-level y_2, while for linked duplication, y_0 continually decrease with increasing y_1 and very low (close to 0) y_2 (see Figure 3). These indicate that under weak positive selection high frequency of original allele and low frequency of advantageous alleles are both buffered on unlinked duplicate loci in the population.

When positive selection is strong (s = 0.01), for linked duplication, y_0 decreases exponentially down to be very low (close to 0); and y_2 increase continually up to ~0.5. However, for unlinked gene duplication, y_0 is only kept high for a period of time and then crashes while y_2 increases suddenly up to be very high (~1) (see Figure 4), which is very similar to observations in Figure 1. These results, combined with results in the above section and in our previous studies, including high y_0 and sudden increase of advantageous allele frequency at one of duplicated loci in the population (see Figure 4), prolonged T_{non} [15,16,20-23] and shortened T_{neo} (see Figure 2), jointly suggest an increase of P_{neo} for unlinked gene duplication in finite populations.

Simulation results

In finite populations, there are several features in simulation results of P_{neo}. First, under strong positive selection, when N is small (roughly N μ_{non} < 0.1), P_{neo} for unlinked gene duplication under both DNR and HI selective models are all close (see Table 3), and similar to Walsh's prediction - μ_{neo}/μ_{non} [13,14]. However, when N is larger (roughly N μ_{non}) > 0.1), both predictions from Equation 1 and 2 are different from our observations under the DNR selective model in simulation (see Table 3).

Second, in a given larger population (N μ_{non} = 0.5), simulation results of P_{neo} with positive selection coefficient (s) are shown in Table 4. If s is small (roughly Ns ≤ 0.1), P_{neo} for unlinked gene duplication under the DNR selective model are also close to Walsh's prediction - μ_{neo}/μ_{non} [14]. If s becomes larger (roughly Ns > 0.5), P_{neo} becomes different from expectations from Equation 1 and 2; and P_{neo} for unlinked gene duplication is larger than that for linked under both the DNR and HI selective models (see Figure 4). Therefore, these observations indicate that Equation 1 and 2 don't provide good approximations of P_{neo} for unlinked gene duplication under stronger positive selection; and free recombination (r = 0.5) enlarges P_{neo}, which is quite consistent with observations of P_{neo} in Table 3, in addition to numerical expectations and suggestions in our previous studies [15].

Third, these observations of P_{neo} were obtained under two extreme conditions: linked (r = 0) and unlinked (r = 0.5). However, in most real cases 0 < r < 0.5, so P_{neo} with these conditions are also simulated, and results are shown in Figure 5. Simulation results clearly show that as r is larger, P_{neo} becomes larger under both DNR and HI selective models. This reinforces our conclusion that recombination enlarges P_{neo} under strong selection.

Discussion and Conclusions

One might argue that these parameters used in above analyses are not realistic enough, for example $\mu_{neo} = 10^{-6}$, or $\mu_{non} = 10^{-4}$ and $\mu_{neo} = 10^{-6}$. They also can be changed into other more realistic values, for example $\mu_{non} = 10^{-6}$, and $\mu_{neo} = 10^{-9}$ [13,14,23], but these changes do not influence conclusions obtained above except for much prolonged time for calculations.

The sudden crash of the balance of chromosomal haplotype frequencies for unlinked gene duplication in numerical results shown in Figure 1 and 4 might be criticized to result from numerical tolerance. But P_{neo} and dynamic changes of genotypes observed directly in simulation are quite consistent with predictions from numerical results. In our previous studies, it has been observed that high x_0 at the equilibrium can be broken by genetic drift in finite populations [15,16]. In this study this balance can also be broken by strong positive selection.

According to our theoretical results presented in this study and previous studies, several views on the evolution of gene duplication should be revised and reconsidered.

T_{non} might be usually much longer than a few million generations in natural populations

It was commonly considered that for gene duplication, mean time to nonfunctionalization is a few million generations or less (assume degenerative mutation rate is ~10^{-6}) [23]. In light of our results, this view should be revised. Only in small populations (N μ_{non} ≤ 0.01), can mean time to nonfunctionaliztion be simply estimated to be on the

Table 3: Simulation results for probabilities of neofunctionalization of duplicate genes with different population sizes *

N	DNR_LINK	DNR_FREE	HI_LINK	HI_FREE	Eq_1	Eq_2
100	0.0164	0.0158	0.0182	0.018	0.0392	0.0392
200	0.0236	0.0242	0.0302	0.0394	0.0741	0.077
500	0.0684	0.063	0.0734	0.092	0.1667	0.1832
1000	0.1152	0.1018	0.166	0.2502	0.2857	0.3406
2000	0.1552	0.1646	0.3378	0.6118	0.4444	0.5966
5000	0.5396	0.8696	0.7022	0.9962	0.6667	0.9549
10000	0.9596	0.9998	0.9824	1	0.8	0.9999
20000	0.9996	1	0.9998	1	0.8889	1
50000	1	1	1	1	0.9524	1
100000	1	1	1	1	0.9756	1

* Other genetic parameters are $\mu_{neo} = 10^{-6}$, $\mu_{non} = 10^{-4}$, and s = 0.01. Simulation repeats 5000 times. DNR_LINK and DNR_FREE are simulation results on linked and unlinked duplicated loci under the DNR selective model, respectively; HI_LINK and HI_FREE are simulation results on linked and unlinked duplicated loci under the HI selective model, respectively; Eq_1 and Eq_2 are predictions from Equation 1 and 2 of Walsh (1995), respectively.

Table 4: Simulation results for probabilities of neofunctionalization of duplicate genes with different positive selection coefficients *

s	DNR_LINK	DNR_FREE	HI_LINK	HI_FREE	Eq_1	Eq_2
10^{-6}	0.0036	0.0036	0.004	0.0052	0.01	0.0004
10^{-5}	0.004	0.0044	0.0054	0.0052	0.0109	0.004
2×10^{-5}	0.003	0.0046	0.0048	0.0074	0.012	0.008
5×10^{-5}	0.0054	0.0062	0.0066	0.0092	0.0157	0.0198
10^{-4}	0.0104	0.008	0.0098	0.0186	0.0226	0.039
2×10^{-4}	0.0156	0.0184	0.0206	0.037	0.0392	0.0762
5×10^{-4}	0.0444	0.067	0.0566	0.1174	0.0909	0.1774
10^{-3}	0.0778	0.1602	0.1338	0.6564	0.16667	0.3177
2×10^{-3}	0.1604	0.339	0.244	0.9198	0.2857	0.5212
5×10^{-3}	0.3502	0.6584	0.486	0.9872	0.5	0.8161
0.01	0.5412	0.8628	0.6986	0.9958	0.6667	0.9549
0.02	0.752	0.976	0.8848	0.9982	0.8	0.9963
0.05	0.8356	0.9988	0.989	1	0.9091	1
0.1	0.926	0.9998	0.9986	1	0.9524	1

* Other genetic parameters are $\mu_{neo} = 10^{-6}$, $\mu_{non} = 10^{-4}$, and N = 5000. Simulation repeats 5000 times. DNR_LINK and DNR_FREE are simulation results on linked and unlinked duplicated loci under the DNR selective model, respectively; HI_LINK and HI_FREE are simulation results on linked and unlinked duplicated loci under the HI selective model, respectively; Eq_1 and Eq_2 are predictions from Equation 1 and 2 of Walsh (1995), respectively.

order of the reciprocal of degenerative mutation rate for gene duplication - $\sim 1/(2\ \mu_{non})$ [20,22,23]. However, it increases when population size is larger (roughly $N\ \mu_{non} > 0.1$), especially for unlinked gene duplication [15,16,20-23]. For unlinked haploinsufficient gene duplication, T_{non} is prolonged dramatically even in a modest population ($0.1 < N\ \mu_{non} \leq 1$) [15,16]. The underlying mechanism is that under recombination the frequency of original (or wild-type) allele is kept high at both duplicated loci, which is a mathematical process and was named originalization [15,16]. High frequency of original allele (x_0) in the population retards nonfunctionalization apparently, because at nonfunctionalization x_0 must be 0. In nature populations, population sizes are usually not small (i.e. N_e from bacteria is about 10^8~10^9, N_e from yeasts is 10^7~10^8, and N_e from mammals is about 10^4~10^5) [24], so T_{non} is usually larger than expected in previous studies (~10^6 generations).

Homogenization results from not only gene conversion, but also originalization

Homogenization is often argued to originate from gene conversion. However, in this study, it is observed that under recombination originalization can also result in homogenization. This result is obtained from the principles of traditional population genetics, under a theoretical framework completely different from gene conversion. In our previous studies on originalization, the effect of gene conversion was neglected. Moreover, in originalization, the wild-type allele is buffered with high frequency on both duplicated loci, which retards the divergence of duplicate genes, while in gene conversion, it is not certain that the wild-type allele is converted on duplicated loci. And during gene conversion, d_n (the rate of non-synonymous nucleotide substitution) and d_s (the rate of synonymous nucleotide substitution) of duplicate genes are both small. However, in originalization, under purifying selection, d_n of duplicate genes are small while d_s are large. This prediction might be applicable to distinguish the effect of originalization from that of gene conversion on genomic evolution.

P_{neo} cannot be expected to be small in natural populations although the rate of advantageous mutation is much small compared with that of degenerative mutation

The rate of degenerative mutation is usually much larger than that of advantageous mutation. So under neutrality, the probability of fixation of advantageous mutations at a locus is much smaller than that of degenerative mutations. This prediction is still hold on for gene duplication under weak selection [13,14]. As shown in Equation 1 from Walsh (1995) [13] and our simulation results (Table 3), for slightly positive selection ($Ns < 0.5$), P_{neo} is equal to $\sim\mu_{neo}/\mu_{non,}$ regardless of recombination. However, under strong positive selection, in larger populations ($N\ \mu_{non} >$

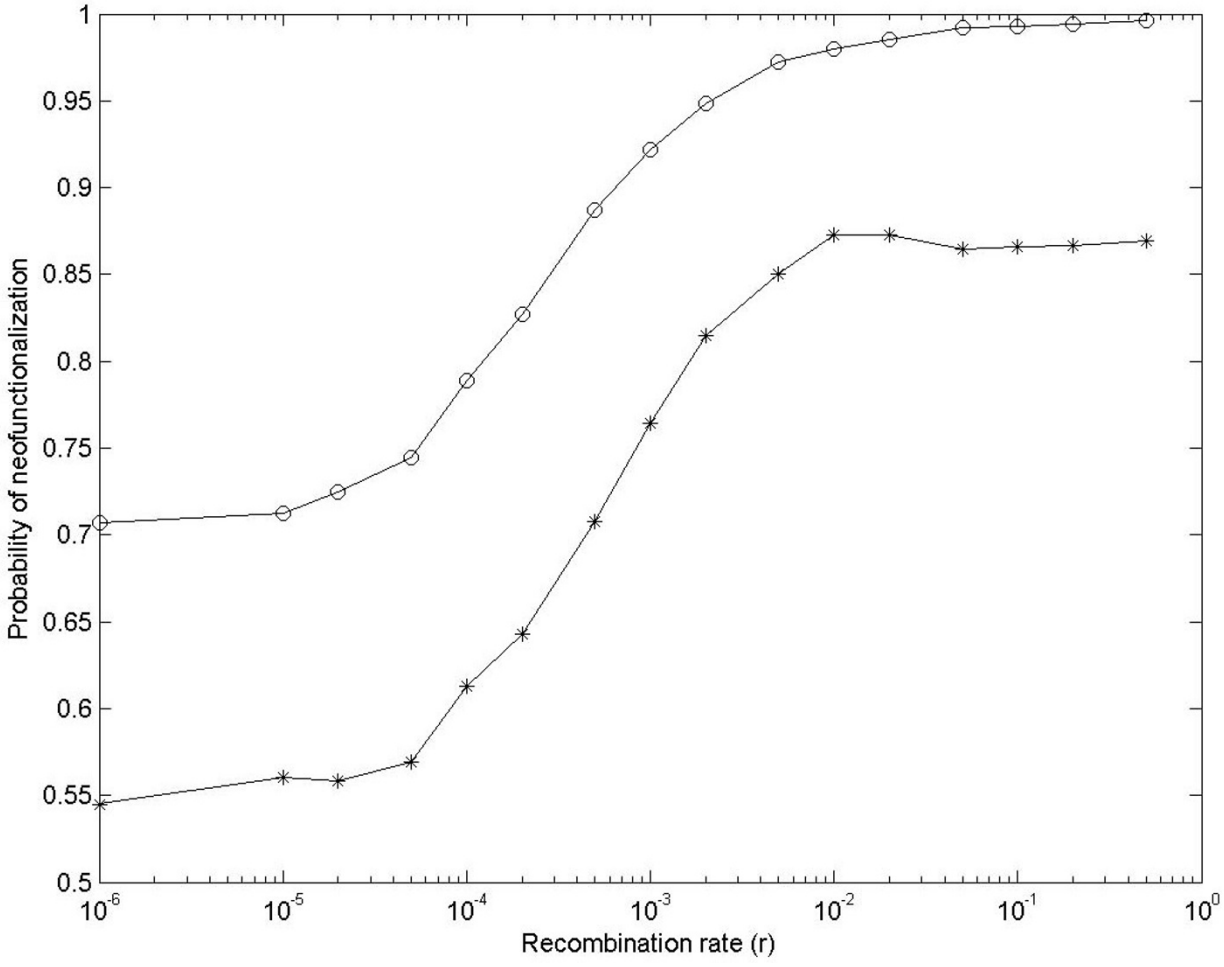

Figure 5 Simulation results for the probability of neofunctionalization for gene duplication with recombination rate. Assume N = 5000, $\mu_{neo} = 10^{-6}$ and $\mu_{non} = 10^{-4}$. Star and Circle spots are simulation results under the DNR and HI selective models, respectively.

0.1) P_{neo} becomes larger under recombination than that under linkage (see Table 3; and Ref. [13]). The underlying mechanism is that recombination provokes the loss of degenerative mutations and the maintenance of wild-type allele at both duplicated loci in the population. The high frequency of wild-type allele facilitates the arising and accumulating of advantageous mutation, so P_{neo} is enlarged. In this way, the power of positive selection is amplified under recombination.

When the evolution of gene duplication is considered in relation to population subdivision (even speciation), the conclusion of P_{neo} enlarged under recombination can be reinforced. When advantageous mutations are slightly selective, each of them is buffered in the population at a low frequency for a prolonged period under recombination by originalization. If environments under which subpopulations live are changing and different, they might provide different strong positive selections, under which advantageous alleles might quickly be fixed at the duplicated loci in subpopulations because of shortened T_{neo}. Therefore, P_{neo} of duplicate genes in nature populations might be larger than expected before.

At the genic level speciation is a differential process accompanied by differential adaptations [25]. It has often been argued that genomic rearrangement resulting from random loss of duplicate genes might cause passive reproductive isolation and then speciation [3,20,26]. Here our results further suggest that via originalization different kinds of neofunctionalizations for duplicate genes among subdivided populations might also contribute to speciation.

Methods

Methods of simulation and numerical analyses have been described in detail in our previous studies [15,16].

Appendix

Consider a pair of duplicated loci on the same chromosome in a random mating, diploid population without considering genetic drift. Let y_0, y_1, y_2, y_3, y_4, y_5, y_6, y_7, y_8 be the frequencies of chromosomal haplotypes, "00", "01", "02", "10", "11", "12", "20", "21", "22", respectively. The fitness of individual genotypes is shown in Table 2. Under the DNR selective model, $s_1 = 0$; Under the HI selective model, $s_1 = 1$. Because $y_0+y_1+y_2+y_3+y_4+y_5+y_6+y_7+y_8 = 1$, only 7 of them are independent. Here we focus on the first 7 frequencies. Therefore, mean population fitness (w) and differential changes of chromosomal haplotype frequencies are given by

$$
\begin{aligned}
w &= y_0^2 + 2y_0y_1 + y_1^2 + 2y_0y_2 + 2sy_0y_2 + 2y_1y_2 + 2sy_1y_2 \\
&+ y_2^2 + 2sy_2^2 + 2y_0y_3 + 2y_1y_3 \\
&+ 2y_2y_3 + 2sy_2y_3 + y_3^2 + 2y_0y_4 + 2y_1y_4 - 2s_1y_1y_4 + 2y_2y_4 \\
&+ 2sy_2y_4 - 2s_1y_2y_4 - 2ss_1y_2 \\
&y_4 + 2y_3y_4 - 2s_1y_3y_4 + 2y_0y_5 + 2sy_0y_5 + 2y_1y_5 + 2sy_1y_5 \\
&- 2s_1y_1y_5 - 2ss_1y_1y_5 + 2y_2y_5 \\
&+ 4sy_2y_5 - 2s_1y_2y_5 - 4ss_1y_2y_5 + 2y_3y_5 + 2sy_3y_5 \\
&- 2s_1y_3y_5 - 2ss_1y_3y_5 + 2y_0y_6 + 2sy_0 \\
&y_6 + 2y_1y_6 + 2sy_1y_6 + 2y_2y_6 + 4sy_2y_6 + 2y_3y_6 \\
&+ 2sy_3y_6 + 2y_4y_6 + 2sy_4y_6 - 2s_1y_4y_6 - \\
&2ss_1y_4y_6 + 2y_5y_6 + 4sy_5y_6 - 2s_1y_5y_6 - 4ss_1y_5y_6 \\
&+ y_6^2 + 2sy_6^2 + 2y_0y_7 + 2sy_0y_7 + 2y_1 \\
&y_7 + 2sy_1y_7 - 2s_1y_1y_7 - 2ss_1y_1y_7 + 2y_2y_7 + 4sy_2y_7 \\
&- 2s1y_2y_7 - 4ss_1y_2y_7 + 2y_3y_7 + 2s \\
&y_3y_7 - 2s_1y_3y_7 - 2ss_1y_3y_7 + 2y_6y_7 + 4sy_6y_7 - 2s_1y_6y_7 \\
&- 4ss_1y_6y_7 + 2y_0y_8 + 4sy_0y_8 + 2 \\
&y_1y_8 + 4sy_1y_8 - 2s_1y_1y_8 - 4ss_1y_1y_8 + 2y_2y_8 + 6sy_2y_8 \\
&- 2s_1y_2y_8 - 6ss_1y_2y_8 + 2y_3y_8 + 4 \\
&sy_3y_8 - 2s_1y_3y_8 - 4ss_1y_3y_8 + 2y_6y_8 + 6sy_6y_8 \\
&- 2s_1y_6y_8 - 6ss_1y_6y_8;
\end{aligned}
$$

$$
\begin{aligned}
y_0' &= [y_0^2 + y_0y_1 + (1+s)y_0y_2 + y_0y_3 + ry_1y_3 \\
&+ r(1+s)y_2y_3 + (1-r)y_0y_4 + (1-r)(1+s)y_0y_5 \\
&+ (1+s)y_0y_6 + r(1+s)y_1y_6 + r(1+2s)y_2y_6 \\
&+ (1-r)(1+s)y_0y_7 + (1-r)(1+2s)y_0y_8] / w - \\
&y_0 - 2(\mu_{non} + \mu_{neo})y_0
\end{aligned}
$$

$$
\begin{aligned}
y_1' &= [y_0y_1 + y_1^2 + (1+s)y_1y_2 + (1-r)y_1y_3 + ry_0y_4 \\
&+ (1-s_1)y_1y_4 + r(1+s)(1-s_1)y_2y_4 + (1- \\
&r)(1+s)(1-s_1)y_1y_5 + (1-r)(1+s)y_1y_6 + r(1+s)y_0y_7 \\
&+ (1+s)(1-s_1)y_1y_7 + r(1+2s)(1- \\
&s_1)y_2y_7 + (1-r)(1+2s)(1-s_1)y_1y_8] / w - y_1 \\
&+ \mu_{non}y_0 - (\mu_{non} + \mu_{neo})y_1
\end{aligned}
$$

$$
\begin{aligned}
y_2' &= [(1+s)y_0y_2 + (1+s)y_1y_2 + (1+2s)y_2^2 \\
&+ (1-r)(1+s)y_2y_3 + (1-r)(1+s)(1-s_1)y_2y_4 \\
&+ r(1+s)y_0y_5 + r(1+s)(1-s_1)y_1y_5 \\
&+ (1+2s)(1-s_1)y_2y_5 + (1-r)(1+2s)y_2y_6 + (1-r)(1 \\
&+ 2s)(1-s_1)y_2y_7 + r(1+2s)y_0y_8 + r(1+2s)(1-s_1)y_1y_8 \\
&+ (1+3s)(1-s_1)y_2y_8] / w - y_2 + y_0 \\
&\mu_{neo} - (\mu_{neo} + \mu_{non})y_2
\end{aligned}
$$

$$
\begin{aligned}
y_3' &= [y_0y_3 + (1-r)y_1y_3 + (1-r)(1+s)y_2y_3 \\
&+ y_3^2 + ry_0y_4 + (1-s_1)y_3y_4 + r(1+s)y_0y_5 + (1+ \\
&s)(1-s_1)y_3y_5 + (1+s)y_3y_6 + r(1+s)(1-s_1)y_4y_6 \\
&+ r(1+2s)(1-s_1)y_5y_6 + (1-r)(1+s)(1- \\
&s_1)y_3y_7 + (1-r)(1+2s)(1-s_1)y_3y_8] / w - y_3 \\
&+ y_0\mu_{non} - (\mu_{non} + \mu_{neo})y_3
\end{aligned}
$$

$$
\begin{aligned}
y_4' &= [ry_1y_3 + (1-r)y_0y_4 + (1-s_1)y_1y_4 \\
&+ (1-r)(1+s)(1-s_1)y_2y_4 + (1-s_1)y_3y_4 + r(1+s)(1 \\
&- s_1)y_1y_5 + (1-r)(1+s)(1-s_1)y_4y_6 \\
&+ r(1+s)(1-s_1)y_3y_7] / w - y_4 + \mu_{non}(y_1 + y_3)
\end{aligned}
$$

$$
\begin{aligned}
y_5' &= [r(1+s)y_2y_3 + r(1+s)(1-s_1)y_2y_4 + (1-r)(1+s)y_0y_5 \\
&+ (1-r)(1+s)(1-s_1)y_1y_5 + (1 \\
&+ 2s)(1-s_1)y_2y_5 + (1+s)(1-s_1)y_3y_5 \\
&+ (1-r)(1+2s)(1-s1)y_5y_6 + r(1+2s)(1-s_1)y_3 \\
&y_8] / w - y_5 + y_2\mu_{non} + y_3\mu_{neo}
\end{aligned}
$$

$$
\begin{aligned}
y_6' &= [(1+s)y_0y_6 + (1-r)(1+s)y_1y_6 + (1-r)(1+2s)y_2y_6 \\
&+ (1+s)y_3y_6 + (1-r)(1+s)(1 \\
&s_1)y_4y_6 + (1-r)(1+2s)(1-s_1)y_5y_6 + (1+2s)y_6^2
\end{aligned}
$$

$$
\begin{aligned}
y_7' &= [r(1+s)y_1y_6 + r(1+s)(1-s_1)y_4y_6 \\
&+ (1-r)(1+s)y_0y_7 + (1+s)(1-s_1)y_1y_7 + (1-r)(1 \\
&+ 2s)(1-s1)y_2y_7 + (1-r)(1+s)(1-s_1)y_3y_7 \\
&+ (1+2s)(1-s_1)y_6y_7 + r(1+2s)(1-s_1)y_1 \\
&y_8] / w - y_7 + y_6\mu_{non} + y_1\mu_{neo}
\end{aligned}
\qquad \text{(A-1)}
$$

where r is recombination rate between duplicated loci, s is positive selective coefficient, μ_{neo} is advantageous mutation rate and μ_{non} is degenerative mutation rate.

Authors' contributions

CX conceived of the study, carried out the most works, and drafted the manuscript. YXF participated in the design of the study. RH and SQL performed some simulation works. All authors read and approved the final manuscript.

Authors' information

Cheng Xue, GuangDong Institute for Monitoring Laboratory Animals, and Key Laboratory of Laboratory Animals in GuangDong, 105 Road Xingang West, Guangzhou, 510260, China. E-mail: lflf27@yahoo.com.cn

Ren Huang, GuangDong Institute for Monitoring Laboratory Animals, and Key Laboratory of Laboratory Animals in GuangDong, 105 Road Xingang West, Guangzhou, 510260, China. E-mail:
labking@sohu.com
Shu-Qun Liu, Laboratory for Conservation and Utilization of Bio-resources, Yunnan University,
Yunnan, China. E-mail:
shuqunliu@gmail.com
Yun-Xin Fu, Laboratory for Conservation and Utilization of Bio-resources, Yunnan University, Yunnan, China, and Human Genetics Center, School of Public Health, University of Texas at Houston, Houston, Texas USA. E-mail:
Yunxin.fu@uth.tmc.edu

Acknowledgements
We thank anonymous reviewers for their valuable comments. The publication of this paper is financially supported by Guangdong Natural Science Foundation 9151026005000002 and funds from Yunnan University

Author Details
[1]GuangDong Institute for Monitoring Laboratory Animals, Guangzhou, China, [2]Key Laboratory of Laboratory Animals in GuangDong, Guangzhou, China, [3]Laboratory for Conservation and Utilization of Bio-resources, Yunnan University, Yunnan, China and [4]Human Genetics Center, School of Public Health, University of Texas at Houston, Houston, Texas, USA

Received: 18 August 2009 Accepted: 9 June 2010
Published: 9 June 2010

References

1. Long M-Y, Betran E, Thornton K, Wang W: **The origin of new genes: glimpses from the young and old.** *Nature Reviews Genetics* 2003, **4:**865-875.
2. Semon M, Wolfe K: **Consequences of genome duplication.** *Curr Opin Genet Dev* 2007, **17:**505-512.
3. Conant G, Wolfe K: **Turning a hobby into a job: how duplicated genes find few functions.** *Nature Reviews Genetics* 2008, **9:**938-950.
4. Studer R, Robinson-Rechavi M: **How confident can we be that ortholog are similar, but paralogs differ?** *Trends Genet* 2009, **25:**210-216.
5. Li W-H, Yang J, Gu X: **Expression divergence between duplicate genes.** *Trends Genet* 2005, **21:**602-607.
6. He X, Zhang J: **Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution.** *Genetics* 2005, **169:**1157-1164.
7. Zhang J: **Evolution by gene duplication: an update.** *Trends Eco Evo* 2003, **18:**292-298.
8. Ohno S: **Evolution by Gene Duplication.** Springer-Verlag, New York; 1970.
9. Jaillon O, Aury J, Brunet F, Petit J, Stange-Thomann N, Mauceli E, Bouneau L, Fischer C, Ozouf-Costaz C, Bernot A, Nicaud S, Jaffe D, Fisher S, Lutfalla G, Dossat C, Segurens B, Dasilva C, Salanoubat M, Levy M, Boudet N, Castellano S, Anthouard V, Jubin C, Castelli V, Katinka M, Vacherie B, Biémont C, Skalli Z, Cattolico L, Poulain J, *et al.*: **Genome duplication in the teleost fish Tetraodon nigroviridis reveals the early vertebrate proto-karyotype.** *Nature* 2004, **431:**946-57.
10. Hughes M, Hughes A: **Evolution of duplicate genes in a tetraploid animal, *Xenopus laevis*.** *Mol Biol Evol* 1993, **10:**1360-1369.
11. Wolfe K, Shields D: **Molecular evidence for an ancient duplication of the entire yeast genome.** *Nature* 1997, **387:**708-713.
12. Kellis M, Birren B, Lander E: **Proof and evolutionary analysis of ancient genome duplication in the yeast Saccharomyces cerevisiae.** *Nature* 2004, **428:**617-624.
13. Walsh J: **How often do duplicated genes evolve new functions?** *Genetics* 1995, **139:**421-428.
14. Walsh J: **Population-genetic models of the fates of duplicate genes.** *Genetica* 2003, **118:**279-294.
15. Xue C, Fu Y: **Preservation of duplicate genes by originalization.** *Genetica* 2009, **136:**69-78. DOI: 10.1007/s10709-008-9311-5
16. Xue C, Fu Y: **Mean time to resolution of gene duplication.** *Genetica* 2009, **136:**119-126. Doi: 10.1007/s10709-008-9319-x
17. Chapman B, Bower J, Feltus F, Paterson A: **Buffering of crucial functions by paleologous duplicated genes may contribute cyclicality to angiosperm genome duplication.** *Proc Natl Acad Sci USA* 2006, **103:**2730-2735.
18. Kincaid D, Cheney W: **Numerical Analysis: Mathematics of Scientific Computing.** Third edition. Brooks/Cole Pub. Co., Pacific Grove; 2002.
19. Sawyer S, Parsch J, Zhang Z, Hartl D: **Prevalence of positive selection among nearly neutral amino acid replacements in Drosophila.** *Proc Nat Acad Sci* 2007, **104:**6504-6510.
20. Li W-H: **Rate of gene silencing at duplicate loci: a theoretical study and interpretation of data from tetraploid fishes.** *Genetics* 1980, **95:**237-258.
21. Takahata N, Maruyama T: **Polymorphism and loss of duplicate gene expression: A theoretical study with application to the tetraploid fish.** *Proc Natl Acad Sci USA* 1979, **76:**4521-4525.
22. Watterson G: **On the time for gene silencing at supplicate loci.** *Genetics* 1983, **105:**745-766.
23. Lynch M, Force A: **The probability of duplicate gene preservation by subfunctionalization.** *Genetics* 2000, **154:**459-473.
24. Lynch M, Conery J: **The origins of genome complexity.** *Science* 2003, **302:**1401-1404.
25. Wu C-I: **The genic view of the process of speciation.** *J Evol Biol* 2001, **14:**851-865.
26. Lynch M: **Gene duplication and evolution.** *Science* 2002, **297:**945-947.

doi: 10.1186/1471-2156-11-46
Cite this article as: Xue *et al.*, Recombination facilitates neofunctionalization of duplicate genes via originalization *BMC Genetics* 2010, **11**:46