High Fitness Costs and Instability of Gene Duplications Reduce Rates of Evolution of New Genes by Duplication-Divergence Mechanisms

Marlen Adler,[1] Mehreen Anjum,[‡,1] Otto G. Berg,[2] Dan I. Andersson,[1] and Linus Sandegren*,[1]

[1]Department of Medical Biochemistry and Microbiology, Uppsala University, Uppsala, Sweden

[2]Department of Cell and Molecular Biology, Uppsala University, Uppsala, Sweden

[‡]Present address: Institute of Biochemistry and Biotechnology, University of the Punjab, Quaid-e-Azam Campus, Lahore, Pakistan

***Corresponding author:** E-mail: linus.sandegren@imbim.uu.se.

Associate editor: Miriam Barlow

Abstract

An important mechanism for generation of new genes is by duplication-divergence of existing genes. Duplication-divergence includes several different submodels, such as subfunctionalization where after accumulation of neutral mutations the original function is distributed between two partially functional and complementary genes, and neofunctionalization where a new function evolves in one of the duplicated copies while the old function is maintained in another copy. The likelihood of these mechanisms depends on the longevity of the duplicated state, which in turn depends on the fitness cost and genetic stability of the duplications. Here, we determined the fitness cost and stability of defined gene duplications/amplifications on a low copy number plasmid. Our experimental results show that the costs of carrying extra gene copies are substantial and that each additional kilo base pairs of DNA reduces fitness by approximately 0.15%. Furthermore, gene amplifications are highly unstable and rapidly segregate to lower copy numbers in absence of selection. Mathematical modeling shows that the fitness costs and instability strongly reduces the likelihood of both sub- and neofunctionalization, but that these effects can be offset by positive selection for novel beneficial functions.

Key words: **gene amplification, fitness cost, evolution of new genes, *Escherichia coli*.**

Introduction

Organisms can acquire new genes by at least three different processes, including horizontal gene transfer, de novo origination from noncoding sequences, and duplication-divergence of existing genes. The latter process has emerged as a major contributor to generating functional diversification of genes in all types of organisms (Ohno 1970; Hughes 1994; Bergthorsson et al. 2007; Dittmar and Liberles 2011). Duplication-divergence includes several different types of submodels that differ with regard to the mechanism of duplication and the relative roles played by selection and drift in maintaining the duplicated state and driving divergence of the duplicated gene copies. The fate of any duplication in a population will depend on both its potential benefits as well as the cost associated with carrying the duplication. The most likely fate of a duplicated gene is nonfunctionalization where one gene copy accumulates mutations that eventually lead to gene inactivation. A second model, subfunctionalization, posits a mechanism that involves a combination of accumulation of initially neutral mutations that are ultimately maintained by negative selection where the original function is distributed between two partially functional genes that complement each other (Hughes 1994; Force et al. 1999; Lynch 2000; Lynch and Force 2000a; Dittmar and Liberles 2011). Finally, the neofunctionalization model involves the development of a new function in one of the duplicated copies while maintaining the old function in another copy (Ohno 1970; Bergthorsson et al. 2007; Dittmar and Liberles 2011; Näsvall et al. 2012). For some of the proposed models (e.g., the classical neofunctionalization and subfunctionalization models), it is assumed that duplications are noncostly and stable whereas other models (e.g., the IAD model, discussed later) state that the amplified state is costly and unstable and therefore has to be maintained by continuous selection for the process to proceed. We recently tested experimentally the IAD model (Innovation-Amplification-Divergence) in which a new function occurs before the duplication and functionally different gene copies evolve under continuous selection (Hendrickson et al. 2002; Francino 2005; Bergthorsson et al. 2007). This study showed that enzymatic specialization could occur in the laboratory in less than 3,000 generations under conditions of strong and continuous selection (Näsvall et al. 2012).

Despite its importance in assessing the likelihood of different new gene evolution models, few precise experimental measurements of duplication costs are available (Pettersson et al. 2009; Reams et al. 2010). The costs of duplications have many components and a priori one could imagine costs manifesting at four different levels: (i) costs associated with carrying and replicating the duplicated DNA, (ii) costs due to gene expression and production of RNA and protein, (iii) costs due to the expressed protein being involved in an energy-requiring metabolic reaction (e.g., flagellar proteins running the ATP-demanding flagellar rotation (Koskiniemi et al. 2012), and (iv) costs due to increased levels of an RNA or protein that lead to imbalances in gene dosage and, for example, improper gene regulation or unwanted molecular interactions. At present we do not know the relative contributions of these costs for duplications but it is likely that (i) is of small importance because DNA and RNA synthesis constitute very minor costs as compared with protein synthesis (Neidhardt et al. 1990).

Another related question regards the intrinsic genetic stability of a duplication/amplification. Often, duplications/amplifications are formed by unequal crossing-over mechanisms or rolling-circle amplification resulting in tandem arrays of the duplicated region. Generally these arrays are intrinsically unstable because of the presence of long, directly repeated regions with perfect homology (i.e., the duplicated region) that will allow homologous recombination and segregational loss of the amplified region down to one copy (Anderson and Roth 1977; Anderson and Roth 1979; Anderson and Roth 1981; Andersson and Hughes 2009; Hastings et al. 2009). However, if the duplication mechanism generates duplicated copies that are inversely oriented or where the copies are located at widely separated sites (e.g., on different chromosomes), the intrinsic instability of the amplified state might be negligible.

Here, we determined both the cost of carriage of defined tandem amplifications and the intrinsic instability of plasmid-borne antibiotic resistance genes. Our results show that these arrays are highly unstable and rapidly segregate to lower copy numbers in absence of selection. In addition, the cost of carrying the copies is high and for this particular region corresponds to an approximately 0.15% reduction in fitness for each additional kilo base pairs of DNA, showing that there is substantial counter-selection against the amplified state. This finding suggests that the common assumption of cost-free duplication, as in the subfunctionalization model and the classical model of neofunctionalization, is an oversimplification and that the likelihood of evolution via either of these pathways is greatly increased if positive selection is present.

Results

Selection for Mutant Strains with Amplification of Plasmid-Borne β-Lactamase Genes

To perform measurements of the costs of gene amplifications and to determine their intrinsic instability, we generated a set of strains with variable copy numbers of a specific region. To this end, we took advantage of plasmid-borne beta-lactamase genes (*bla*) that are located in a plasmid region with high amplification rates and for which the copy number of the *bla* genes is directly correlated with the level of resistance to β-lactam antibiotics. The used plasmid, pUUH239.2, is a low copy number plasmid (1–2 copies per chromosome) with a size of approximately 220 kbp (Sandegren et al. 2011; Adler et al. 2012). We previously showed that an increased copy number of the plasmid-borne *bla* genes can be obtained by selecting for mutant strains that can grow on increasing concentrations of the antibiotic meropenem (Adler et al. 2012).

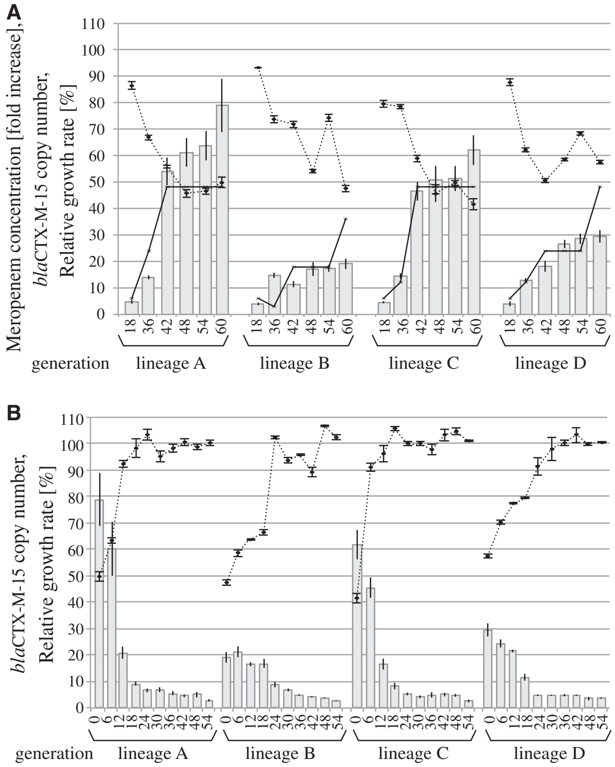

Fig. 1. Copy number variation of *bla*$_{\text{CTX-M-15}}$. (*A*) Increased carbapenem tolerance following meropenem passage as fold increases over the MIC of DA24338 (solid lines), the relative growth rate in percent in the absence of meropenem compared with DA24338 (dashed lines), and the gene copy numbers of *bla*$_{\text{CTX-M-15}}$ (bars). (*B*) Loss of gene arrays in absence of selection. Each lineage corresponds to the average of 4 parallel experiments (16 lineages in total). The gene copy numbers of *bla*$_{\text{CTX-M-15}}$ (bars) and the relative growth rate in the absence of meropenem in percent compared with DA24338 are illustrated. Error bars are SEM.

Four lineages of strain DA24338 were continuously serially passaged in Mueller-Hinton (MH) broth supplemented with increasing concentrations of meropenem (1×, 2×, 4×, and 6× the minimal inhibitory concentration [MIC]). After overnight incubation, cells were further passaged from the cultures with the highest concentration of antibiotic that allowed growth. All lineages were serially passaged 10 times (corresponding to 60 generations). After 60 generations, the tolerated meropenem concentration increased to 18 mg/l (lineage B) and 24 mg/l (lineages A, C, D). This procedure generated a set of strains with different resistance and copy numbers of the plasmid-borne *bla* genes that were then used for further studies.

Copy Number of Amplified Arrays

A DNA extract was used as template to determine the average gene copy number of *bla*$_{\text{CTX-M-15}}$ in the cycled populations, with an unamplified gene, *pemK*, as internal control on the resistance plasmid. In all lineages, the number of *bla*$_{\text{CTX-M-15}}$ gene copies increased greatly (fig. 1*A*; gray bars). After 60 generations, lineage A acquired 80 copies, C acquired 60 copies, and D acquired 30 copies, and these

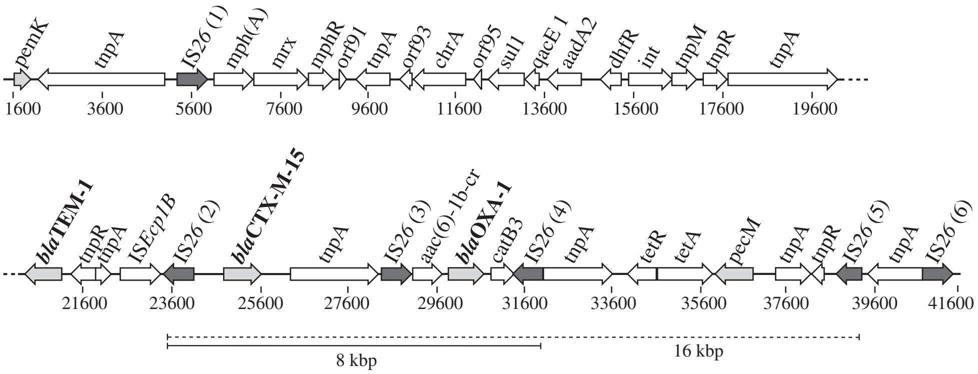

FIG. 2. Size of the amplified unit within the pUUH239.2 resistance cassette. Coding regions are shown as arrows. Genes targeted by real-time PCR are shaded light gray, IS26 elements dark gray, and other genes white. The solid line indicates the 8-kbp amplified region in lineages A–C and the dashed line the 16-kbp amplified region in lineage D.

lineages grew at concentrations 48-fold above the starting concentration. Lineage B acquired approximately 20 copies of $bla_{\text{CTX-M-15}}$, which allowed growth at 36-fold above the MIC of the DA24338 starting strain (fig. 1*A*; solid lines).

Fitness Costs of Extra Gene Copies

To determine the fitness cost of extra gene copies, we measured the growth rates of meropenem-cycled populations immediately after removing the antibiotic to reduce the risk of segregation of the amplified array (fig. 1*A*; dashed line). A clear reduction of relative growth rates can be seen for all populations, and the mutant growth rates ranged from 42% to 58% of the growth rate of the starting strain. The relative growth rates correlated well with the change in $bla_{\text{CTX-M-15}}$ copy numbers where increasing copy numbers result in proportionally slower growth in absence of antibiotic (fig. 1*A*).

Loss Rates of Amplified Arrays in the Absence of Selection Pressure

To measure the loss rate of the amplified arrays, each lineage with varying starting copy numbers (for lineages A–D 80, 20, 60, and 30 copies, respectively) was divided into four new lineages and passaged without selection (i.e., no antibiotic present) for the amplified region. As gene amplifications are intrinsically unstable due to the presence of directly repeated perfect homology, lower copy number segregants are rapidly formed, and if the amplified state poses a fitness cost on the bacterium these more fit segregants will rapidly take over the population. Thus, the loss rate from the population will be determined by both the intrinsic loss rate and the growth rate differences between clones with different copy numbers. A rapid loss of amplification could be seen within 18 generations for all amplified lineages (fig. 1*B*; bars). At the same time, the relative growth rates increase correspondingly in response to the decreased copy numbers (fig. 1*B*; dashed lines). The fitness of all lineages was restored to the level of the starting strain after 18–30 generations of growth.

Amplification Endpoints Are Located in Directly Oriented IS26 Elements

To allow an estimation of the fitness costs per kilo base pair of amplified DNA, we determined the endpoints and size of the amplified unit. Directly repeated IS-elements provide very efficient recombination points for gene amplification because of their size (1 kbp) and perfect homology. The resistance cassette in pUUH239.2 contains six IS26 elements in different orientations (fig. 2). Amplification endpoints were determined by measuring the relative copy number at different positions in the resistance cassette compared with the copy number of the *pemK* gene situated outside of the amplified resistance cassette. In the evolved lineages A–C, the upstream amplification endpoint was situated in IS26(2) and the downstream endpoint at IS26(4), resulting in amplification of an 8-kbp region (fig. 2). Lineage D had a more complex arrangement, with 30 copies of the region between IS26(2) and IS26(4) and 15 copies of the region between IS26(4) and IS26(5) (fig. 2).

Determination of Recombination Rates of Amplified Arrays and Fitness Cost of Each Copy

To determine the recombination rate and cost per additional array copy, we used the model for homologous recombination as explained in detail previously (Pettersson et al. 2005; Pettersson et al. 2009). Briefly, the probability that an array of *m* copies is transformed to one of *n* copies in each recombination event is

$$p(n\,|\,m) = \begin{cases} \frac{2m-n}{m^2}; & m \leq n \leq 2m \\ \frac{n}{m^2}; & 1 \leq n \leq m \end{cases} \qquad (1)$$

Recombination is assumed to occur with rate

$$v_m^{\text{rec}} = \frac{k_{\text{rec}}(m-1)}{1 + k_{\text{rec}}(m-1)} \qquad (2)$$

per replication. This rate has been chosen such that it increases linearly with array size for small *m* and is limited to at most one per replication for large arrays. It can be noted

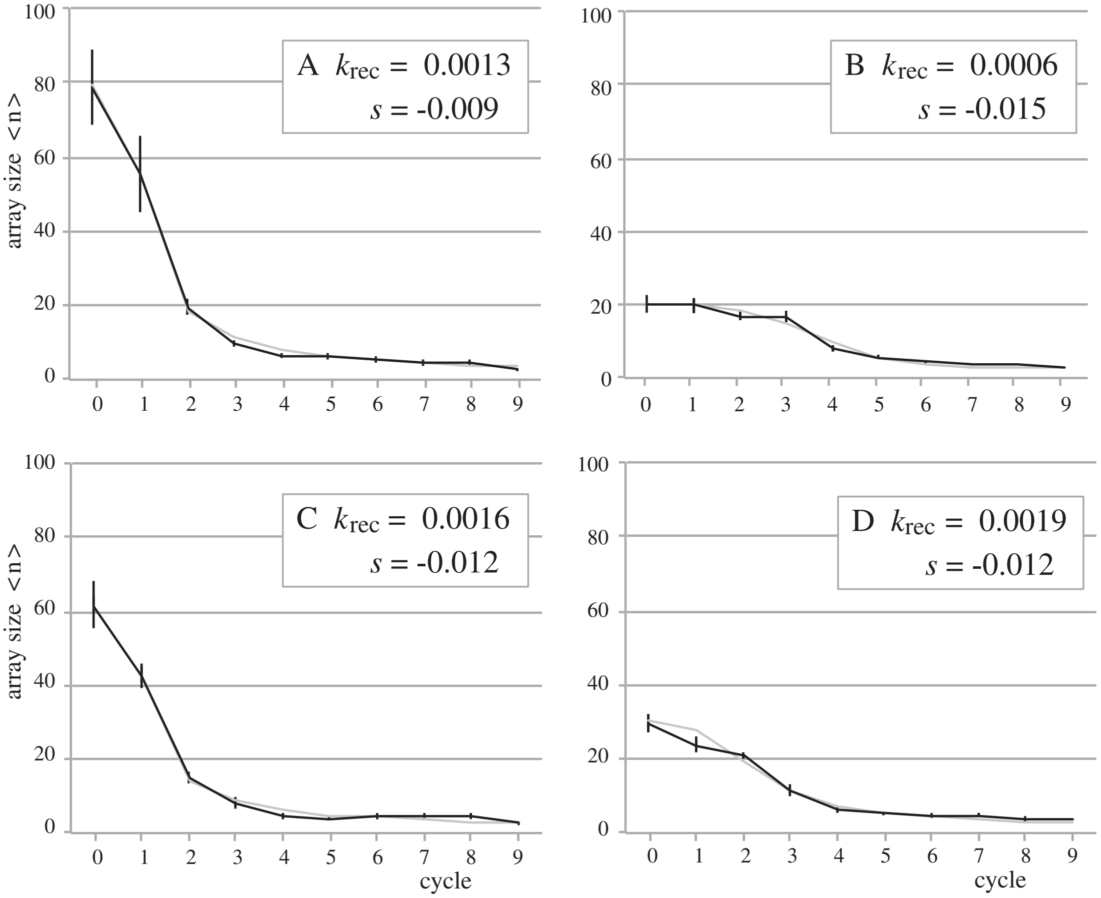

Fig. 3. Determination of fitness cost and recombination rates for segregation of gene amplifications. Average array size, $n = \sum_n nf_n$, after a given number of growth cycles. Black lines show the experimental results as averages for the four replicates in each lineage A, B, C, and D and gray lines are the simulated results using the best-fit values for k_{rec} and s as shown. Error bars are SEM.

that k_{rec} is a parameter of the model with the consequence that the rate of loss of a duplication is

$$k_{loss} = v_2^{rec} p(1\,|\,2) = \frac{k_{rec}}{1+k_{rec}}\frac{1}{4} \approx \frac{1}{4}k_{rec} \qquad (3)$$

Furthermore, it is assumed that the extra gene copies incur a physiological cost, such that the growth rate is reduced by a factor (1 + s), where $s < 0$, for each extra copy of the genes. Thus the fitness of a variant with m copies relative to wild type with only one copy is

$$w_m = (1+s)^{m-1} \qquad (4)$$

The recombination model is embedded in a population-dynamic model where variants with different array size compete through their different growth rates. Assuming that arrays of size n (=1, 2, 3 . . .) are present in a fraction f_n of the population at some time, the average fitness of the population at that time will be $W = \sum_n w_n f_n$. In each generation of the population these fractions will change by

$$\Delta f_n = \left(\frac{w_n}{W} - 1\right) f_n + \sum_{m=2}^{\infty} w_m^{rec} p(n\,|\,m) \frac{w_m}{W} f_m - v_n^{rec} \frac{w_n}{W} f_n \qquad (5)$$

This is a fully deterministic model which is justified by the fact that changes in copy number are fast and the population is very large; it is also justified by the very small differences observed between the different replicates for each lineage. Recombination is reciprocal and leads with equal probability to an increase or a decrease in array size. The fast reduction in array size observed is instead driven by the fitness cost. Due to this fitness cost, the array size will remain limited and the summation to infinity in equation (5) can be cut off at a suitably chosen upper value.

We have simulated the process for meropenem passaged lineages, starting from a given array size and applying equation (5). The parameters s and k_{rec} were chosen to best mimic the observed decline in array size. Each growth cycle involves a 100-fold increase in population size, corresponding to approximately 6.6 generations. Very good agreement with the measured decay in array sizes in the different lineages could be achieved within a relatively narrow range of parameter values, where $s \approx -0.012 \pm 0.003$ and $k_{rec} \approx 0.0013 \pm 0.0006$ (fig. 3). This value for s agrees with an independent estimate based on the measured fitness values as a function of copy number (60 data points from fig. 1). These give a good agreement with equation (4) using a load per extra copy corresponding to $s = -0.014$. Similarly, considering only the 17 data points with

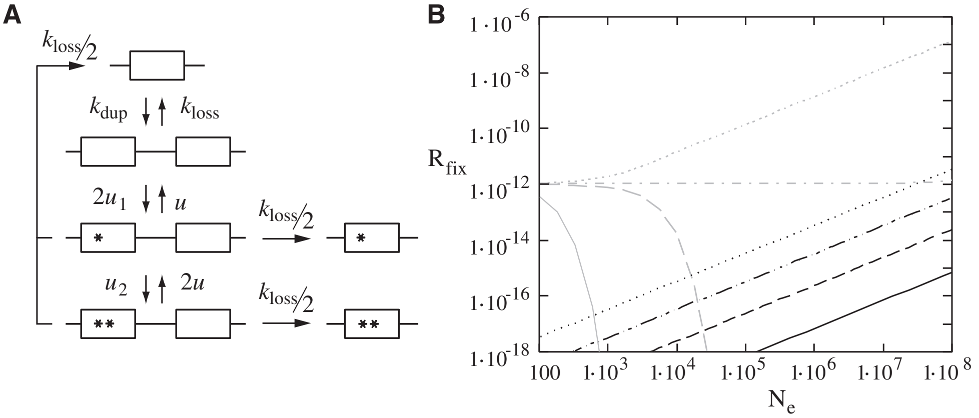

Fig. 4. (*A*) Two-step neofunctionalization model. Back mutations (rate u) have been neglected relative to the loss rate. The state with a duplication carries the cost $s < 0$. As long as an unmutated copy remains, the mutations are assumed positively selected, the first with s_1 and the second with s_2 relative the unmutated single-copy state. The mutated single-copy states are assumed lethal. (*B*) Rate of neofunctionalization as a function of effective population size. Parameter values used: $f_2 = 10^{-2}$, $u_1 = u_2 = 5 \times 10^{-10}$, $k_{loss} = 3 \times 10^{-4}$, and s_1 and s_2 as indicated for the different curves. Gray curves are for single-mutation fixation from equation (8): solid $s_1 = -10^{-2}$, dashed $s_1 = 0$, dash-dot $s_1 = k_{loss}$, and dotted $s_1 = 10^{-3}$. Black curves are for double-mutation fixation from equation (10): solid $s_1 = -10^{-2}$ and $s_2 = 10^{-3}$, dashed $s_1 = 0$ and $s_2 = 10^{-3}$, dash-dot $s_1 = 0$ and $s_2 = 10^{-2}$, and dotted $s_1 = 0$ and $s_2 = 10^{-1}$.

copy number $m < 5$, the estimated load per extra copy would be $s = -0.010$. However, the data resolution is not good enough to estimate the load for a single duplication and it is not impossible that this could correspond to $s \approx 0$. Thus, with the exception of a possible threshold effect for very low copy numbers, equation (4) describes the measured fitness quite well. We have also tested the assumption that recombination is linear in copy number by replacing $(m-1)$ by $(m-1)^{\xi}$ at both positions in equation (2). Then for $\xi = 0.5$, 1, 1.5, and 2, good fits with the measured decay in array sizes can be achieved with $k_{rec} = 10^{-2}$, 10^{-3}, 10^{-4}, and 10^{-5}, respectively. The estimated s remains essentially the same, but the estimated k_{rec} leads to a duplication loss rate (eq. 3) in the range, $k_{loss} = k_{rec}/4 = 2.5 \times 10^{-3} - 2.5 \times 10^{-6}$, which is smaller than the measured values for this kind of tandem array, where $k_{loss} \approx 4 \times 10^{-2} - 4 \times 10^{-3}$ (Pettersson et al. 2009; Reams et al. 2010). Thus, a recombination model with weaker than linear ($\xi \leq 0.4$ if $k_{loss} \geq 0.004$) dependence on array size appears likely. In the calculations that follow below, we will stay with the estimate from the linear model (eq. 2) giving $k_{loss} = 3 \times 10^{-4}$, bearing in mind that this may be an overestimate of the stability of a tandem duplication.

Consequences of Fitness Costs on New Gene Models

When de novo duplications are included in the model, an extra term, $-k_{dup}f_1/W$ and $+k_{dup}f_1/W$, respectively, should be added to Δf_1 and Δf_2 in equation (5). With these additions, equation (5) can be used to calculate the fraction of the population that carries a duplication in a steady state of duplication and loss. This gives

$$f_2 = \frac{k_{dup}}{k_{dup} + k_{loss} - s} \tag{6}$$

Even if $s = 0$, longer arrays are less common than duplications and they will be neglected in the calculations below. For the particular duplications studied here, $k_{loss} = k_{rec}/4 = 3 \times 10^{-4}$. To this should be added the rate of gene destruction due to other kinds of mutation (nonfunctionalization); this is expected to be negligible relative to loss due to recombination. Furthermore, $s = -0.01$ and $k_{dup} \sim 10^{-3} - 10^{-5}$ per generation (Reams et al. 2012). This would give $f_2 \approx 10^{-1} - 10^{-3}$. Experimentally, it is found that $f_2 \approx 10^{-2} - 10^{-6}$ for tandem duplications of this kind; below we will use the estimate $f_2 < 10^{-2}$.

Neofunctionalization

A duplicated gene is thought to be a substrate for the evolution of new gene function by picking up mutations in one of the two redundant copies (Ohno 1970; Bergthorsson et al. 2007; Dittmar and Liberles 2011; Näsvall et al. 2012). Thus, if we assume that a new gene function can be created by a single mutation (with rate u_1) in either of the two gene copies (fig. 4*A*), the rate of appearance in the population would be $2u_1Nf_2$. The new gene would be subject to the same loss rate, k_{loss}, as the original duplicate. To have any significant chance of spreading in the population, the new function must have sufficient positive selection, s_1, relative to the single-copy variant. Then the effective selection would be $s_1 - k_{loss}$ and the fixation probability in the haploid is (Kimura 1962).

$$P_{fix} = \frac{N_e}{N} \frac{2(s_1 - k_{loss})}{1 - \exp[-2N_e(s_1 - k_{loss})]} \tag{7}$$

Here, N_e is the effective population size and N is the actual size. Thus, the total rate of fixation of the mutated duplicate (new gene) is

$$R_{fix} = 2u_1Nf_2P_{fix} = \frac{4u_1f_2N_e(s_1 - k_{loss})}{1 - \exp[-2N_e(s_1 - k_{loss})]} \tag{8}$$

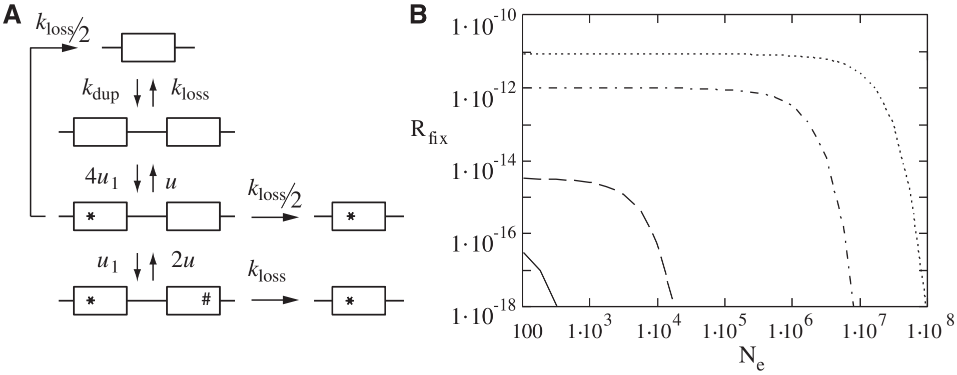

Fig. 5. (*A*) Subfunctionalization model. The complementary mutations (* and #) are assumed to occur with the same rate u_1 and with equal probability in either gene copy; in the diagram above they are considered interchangeable. In the final state, one gene with * and one with #, the copies have lost one function each, but together they perform like the original. The final state is absorbing if back mutations (u) are neglected. The single-copy states with a mutation (* or #) are assumed lethal. (*B*) Rate of subfunctionalization from equation (11) as a function of effective population size. Parameter values used: $f_2 = 10^{-2}$, $u_1 = u_2 = 5 \times 10^{-9}$; solid curve $k_{loss} = 3 \times 10^{-4}$ and $s = -10^{-2}$; dashed $k_{loss} = 3 \times 10^{-4}$ and $s = 0$; dash-dot $k_{loss} = 10^{-6}$ and $s = 0$; dotted $k_{loss} = 10^{-7}$ and $s = 0$.

In the calculations below we will use the estimates $k_{loss} = 3 \times 10^{-4}$ and $f_2 < 0.01$. Mutations that will introduce a new selected gene function are probably rare and $2u_1 \sim 10^{-9}$ is not unreasonable (Gordo et al. 2011). Thus, if $N_e \,|\, s_1 - k_{loss} \,| < 1$, the fixation rate of a new gene through tandem duplication followed by adaptation would be $R_{fix} < 2 \times 10^{-11}$ per generation. With strong positive selection for the new function, $N_e(s_1 - k_{loss}) >> 1$, the fixation would be faster by this factor; for example, $R_{fix} < 2 \times 10^{-6}$ if $N_e = 10^8$ and $s_1 - k_{loss} = 0.001$ (fig. 4*B*).

If the first mutation is not sufficiently strongly selected (i.e., if $s_1 << k_{loss} - 2u_1$), fixation could occur if a second mutation has sufficiently strong positive selection s_2. In this case, the expected stationary presence of individuals in the population with a singly mutated gene copy would be

$$N_2^* \approx Nf_2 \frac{2u_1}{2u_1 + k_{loss} - s_1} \qquad (9)$$

Each of these individuals can pick up a second mutation with rate u_2, which can be fixed with the probability as given in equation (7) but with s_1 replaced by s_2.

$$\begin{aligned} R_{fix}^{(2)} &= u_2 N_2^* P_{fix} \\ &= \frac{4u_1 u_2 f_2 N_e (s_2 - k_{loss})}{(2u_1 + k_{loss} - s_1)(1 - \exp[-2N_e(s_2 - k_{loss})])} \end{aligned} \qquad (10)$$

Thus, as can be seen in figure 4*B* (dotted gray line), a strong selection already for the first mutation, $N_e(s_1 - k_{loss}) >> 1$, is necessary for efficient gene creation through duplication and adaptation. This requires both a large effective population size and $s_1 > k_{loss}$. Also, s_1 is the net positive selection after the cost, s, of carrying the duplication has been accounted for.

Subfunctionalization

A gene can get its function distributed over different genes that together provides the full function (Hughes 1994; Force et al. 1999; Lynch and Force 2000a, 2000b). Let us consider the case of two functions, each of which can be deactivated by a single mutation. Both functions are required for viability of the cell. After a duplication of the gene, the copies can lose by mutation one function each, but the two together can still carry out the required functions. Once the functions have been divided on different genes, each can be further adapted such that the two work better than the original one. Here, we will consider only a minimal model for the first step where the functions have been divided but no further fine-tuning has occurred. The model (fig. 5*A*) is based on duplication followed by a two-step mutation mechanism and the result resembles equation (10). The main difference is that here the mutations are assumed neutral and the duplication will carry the cost, $s < 0$, throughout. For simplicity, the two mutations are assumed to have the same rate, u_1. In the first step, either mutation can occur in either copy giving the total rate $4u_1$. In the second step, the other mutation can occur in the unmutated copy only. Thus,

$$R_{fix}^{subf} = \frac{8f_2 u_1^2 N_e (s - k_{loss})}{(4u_1 + k_{loss} - s)(1 - \exp[-2N_e(s - k_{loss})])} \qquad (11)$$

As $s - k_{loss} < 0$, the best scenario is for $N_e \,|\, s - k_{loss} \,| << 1$, that is, a very small effective population size. Furthermore, if $s = -0.01$, $k_{loss} = 3 \times 10^{-4}$, $2u_1 = 2u_2 = 10^{-8}$, and $f_2 < 0.01$, this gives $R_{fix} < 10^{-16}$ per generation (solid line in fig. 5*B*). Here we have chosen the mutation rates an order of magnitude larger than in Eqs. (8–10) to account for the fact that a mutation that destroys a function is more likely than one that creates one. If instead $s = 0$, one finds $R_{fix} < 10^{-14}$ per generation (dashed line in fig. 5*B*). For a relatively stable duplication with $s = 0$ and $k_{loss} = 10^{-7}$, one finds $R_{fix} < 10^{-11}$ (dotted line in fig. 5*B*); thus the high loss rate alone could be sufficient to block significant subfunctionalization via tandem duplication.

As calculated, the inverse of the fixation rates above (Eqs. 8, 10, and 11) correspond to the expected waiting times before a variant appears that is destined to become fixed. Most new variants—selected or not—leave no long-term descendants in the population. Given fixation, the time required to

Table 1. Experimental Evidence for Fitness Costs of Duplications from Different Organisms.

Organism	Size (kbp)[a]	Cost of Duplication (s-value)	Average Cost per 1 kbp (*s*-value)	Reference
Escherichia coli	8	1.2×10^{-2}	1.5×10^{-3}	This work
Salmonella enterica	20	1×10^{-2}	0.5×10^{-3}	Reams et al. (2010)
S. enterica	72	$<3 \times 10^{-2}$	$<0.4 \times 10^{-3}$	Pettersson et al. (2009)
S. enterica	131	3×10^{-2}	0.2×10^{-3}	Reams et al. (2010)
S. enterica	137	9×10^{-2}	0.7×10^{-3}	Pettersson et al. (2009)
S. enterica	330	12×10^{-2}	0.4×10^{-3}	Pettersson et al. (2009)
S. enterica	518	20×10^{-2}	0.4×10^{-3}	Pettersson et al. (2009)
S. enterica	1,246	6×10^{-2}	0.05×10^{-3}	Pettersson et al. (2009)
Caenorhabditis elegans	2.4–13.9	Not quantified[b]	NA[c]	Lipinski et al. (2011)
Saccharomyces cerevisiae	Variable	Not quantified	NA	Katju et al. (2009)
Drosophila melanogaster	Variable	Not quantified	NA	Langley et al. (2012)

[a]Size of the duplicated region studied.
[b]Cost inferred but not quantified.
[c]Not applicable.

actually penetrate the whole population is quite short ~$(1/|s - k_{\text{loss}}|)\ln(N_e)$ generations in the case of the counter-selected ($s < 0$) variant in equation (11). Similarly, for the positively selected ($s_1 > k_{\text{loss}}$) variant described in equation (8), the conditional time required to penetrate the whole population is also short, ~$[1/(s_1 - k_{\text{loss}})]\ln(N_e)$. Furthermore, the calculations start from a state of duplication/loss equilibrium; the time to reach this equilibrium is also short, ~$1/(k_{\text{dup}} + k_{\text{loss}} - s)$ generations. Thus, $1/R_{\text{fix}}$ in these models corresponds to the expectation time for neofunctionalization or subfunctionalization, respectively.

Discussion

In this study, we measure the fitness costs of defined amplified arrays and their rate of segregation (instability). The main unit of amplification examined here covers 8 kbp of DNA and includes three full-length antibiotic resistance genes (*bla*$_{\text{CTX-M-15}}$, *aac*(6)-1b-cr, and *bla*$_{\text{OXA-1}}$), one truncated resistance gene (*catB3*), one transposase gene (*tnpA*), and one IS*26* element. Two flanking IS*26* elements provided homology for recombination and as a result an IS*26* element forms the join point between each amplified unit. During selection for increased antibiotic resistance, this unit could amplify up to 80 copies and when selection was relieved it rapidly segregated back to a few copies. To determine the recombination rate and cost per additional array copy, we used a previously developed model for homologous recombination (Pettersson et al. 2005; Pettersson et al. 2009). In this model it is assumed that the recombination rate (both loss and gain of copies) increases linearly with increasing array size and that each extra gene copy incur a physiological cost that reduces the growth rate. The recombination model is combined with a population-dynamic model where variants with different copy numbers compete through their different growth rates. This deterministic model has two free parameters, the cost for each additional copy (s) and the recombination rate (k_{rec}). We experimentally measured the copy loss rate from several independent clones (in the absence of antibiotic selection) with different starting copy numbers and simulated which parameter values of s and k_{rec} would best fit the observed decline in copy number. Very good agreement with the measured array sizes in the different lineages was seen within a rather narrow range of parameter values, where $s \approx -0.012$ (±0.003) and $k_{\text{rec}} \approx 0.0013$ (±0.0006). Thus, each extra copy of 8-kbp DNA incurs a fitness cost of −1.2% and the rate of recombination is 0.13% per cell and generation (fig. 3).

How do these costs compare to previous studies and what are the implications of these fitness costs with regard to evolution of new genes? Using the observed costs, addition of 1 kbp of duplicated DNA would result in a 0.15% (*s*-value = 1.5×10^{-3}) reduction in fitness. Data from other studies (Pettersson et al. 2009; Reams et al. 2010), with less precise measurements of fitness costs and larger, less defined duplications, suggest that for duplications in the size range 20–1,246 kbp the fitness cost (*s*-value) for each added kilo base pair ranges between 0.05×10^{-3} and 1.5×10^{-3} with a median value of 0.4×10^{-3} (table 1). Sizes of duplications can vary between a few kbp up to Mbp in eubacteria (e.g., *Salmonella enterica* and *Escherichia coli*) (Anderson and Roth 1981; Andersson and Hughes 2009; Hastings et al. 2009) with similar sizes in eukaryotes such as *Caenorhabditis elegans*, *Drosophila melanogaster*, and *Saccharomyces cerevisiae* (Katju and Bergthorsson 2013). It is therefore likely, even for very small duplications (1 kbp), that the cost is visible to selection in many organisms (table 1). Apart from these direct experimental determinations of costs, other lines of evidence support the notion that duplications are generally deleterious (Katju and Bergthorsson 2013). For example, the large discrepancy in mutation accumulation experiments and bioinformatics-based estimates of gene duplication rates is most easily explained by selection against duplicated genes. Even if the cost of duplication were to be negligible in multicellular eukaryotes with large genomes, tandem duplications may be as unstable as those studied here. If so, they would form very unlikely substrates for both neo- and subfunctionalization. Stable duplications would still be subject to loss due to mutational inactivation (nonfunctionalization), but on a time scale much longer than that for recombination.

The mathematical modeling demonstrates the importance of fitness costs of duplications and their instability on the probability of the neo- and subfunctionalization models. Applying the experimentally determined values of loss rates and fitness costs of duplication carriage make neofunctionalization seem very improbable. Thus, for efficient gene creation through duplication and divergence, it is either required that the benefit of the first adaptive mutation is greater than the combined negative effect of instability (k_{loss}) and the fitness costs of carrying the whole duplication or that the function of the initial mutation could be strengthened by gene dosage which in turn would increase the mutational target for further adaptive mutations. Alternatively, the cost and instability of the duplication can be balanced by selection for increased gene dosage allowing long-term maintenance of the duplicated region. This is the case for the Innovation-Amplification-Divergence (IAD) model which posits that a gene has a weak trace secondary function in addition to its primary function (Hendrickson et al. 2002; Hooper and Berg 2003; Francino 2005; Bergthorsson et al. 2007; Näsvall et al. 2012). If a change in the environment makes the fortuitous side activity beneficial, positive selection favors common duplication mutants that express more of the enzyme. The fitness cost and inherent instability of tandem duplications are thus off-set by positive selection for the secondary activity. As the copy number increases, so does the target size for beneficial mutations, facilitating their accumulation. Improved variants may be further amplified while less improved variants may be lost. As beneficial mutations accumulate, positive selection to keep the amplified state is relaxed, until a variant that provide enough secondary activity on its own arises, and the amplification segregates. The result is a new pair of paralogous genes, where one copy has the new improved function whereas the other copy retains the original function. With regard to the subfunctionalization, the modeling and experimental data suggest that this process is extremely unlikely for the estimated fitness costs and segregation rates of the 8-kbp duplication studied here. However, if the duplication is cost-free and fully stable (i.e., not a tandem duplication), the subfunctionalization model is feasible (fig. 5*B*). This is the case in most previous modeling of subfunctionalization (Force et al. 1999; Lynch 2000; Lynch and Force 2000a; Dittmar and Liberles 2011).

The costs of duplications could be manifested at several different levels, and for this study and other studies the exact physiological reason(s) are poorly defined. One obvious cost would be carriage and gene expression of the extra DNA. The additional costs for DNA and RNA synthesis of duplicated DNA are minor in comparison to the costs for protein synthesis and can therefore be ignored. Thus, of the total energy and mass expenditure for polymerization of all types of macromolecules, protein synthesis accounts for >95% if one includes the cost of amino acid biosynthesis, tRNA charging, and polypeptide formation on the ribosome (Neidhardt et al. 1990). If one assumes that gene expression scales approximately with size of the extra DNA, the predicted fractional cost for protein synthesis of the duplication studied here would be 8 kbp/4.8 Mbp $\approx$ 0.17%. This is 7-fold lower than the cost observed which would imply that either is this region relatively more highly expressed than expected from its size or there are other costs involved. However, for several of the other duplications listed in table 1, the cost that is experimentally determined fits well with that obtained from calculating the predicted fractional cost of gene expression (i.e., protein synthesis). Obviously this is a very rough calculation, but it suggests that gene expression costs in terms of energy and mass are substantial and can explain the costs of many duplications. However, other costs that are generally much harder to define and quantify might also be important for certain duplicated regions. For example, the duplicated region might encode functions that are involved in energy-consuming metabolic reactions (e.g., futile ATP expenditure [Koskiniemi et al. 2012]) or imbalances in gene dosage could confer reduced fitness by generating improper gene regulation or unwanted molecular interactions. Experimental determination of the relative contributions of these processes to the total cost is not straightforward, and a proper determination of gene expression costs would require that the proteins included in the duplication are functionally inactivated to prevent "downstream" costs due to the normal activity of the protein while maintaining gene expression costs. In conclusion, our modeling and experimental data suggest that the fitness cost of gene amplifications will substantially reduce the likelihood of new gene functions originating through either neofunctionalization or subfunctionalization without a selective pressure to retain the amplified state.

Materials and Methods

Bacteria and Growth Conditions

Porin deficient strain DA23325 (F-, λ-, *ilvG*-, *rbf-50*, *rph-1*, Str^{R}, Rif^{R}, Nal^{R}, *ompC*::FRT scar, *ompF*::FRT scar) was constructed by deleting *ompC* and *ompF* using lambda red-mediated recombination (Datsenko and Wanner 2000). Conjugation of ESBL-plasmid pUUH239.2 (Sandegren et al. 2011) into DA23325 created DA24338 (F-, λ-, *ilvG*-, *rbf-50*, *rph-1*, Str^{R}, Rif^{R}, Nal^{R}, *ompC*::FRT scar, *ompF*::FRT scar/pUUH239.2). Throughout the study, MH broth (Difco) was used and supplemented with 1.5% agar (Oxoid) for plates. The antibiotic concentrations used were cefotaxime 25 mg/l, erythromycin 25 mg/l, and various concentrations of meropenem as required for the specific experiments. Etest (bioMérieux) was used to determine minimal inhibitory concentrations (MICs) at 37 °C, according to the manufacturer's description.

Enrichment of Population of Bacteria with β-Lactamase Gene Amplifications in Broth

DA24338 was cycled in MH broth supplemented with the starting concentration of 0.5 mg/l meropenem and 2-fold, 4-fold, and 6-fold the starting concentration. The following day 10 µl from a densely grown culture with the highest antibiotic concentration were inoculated into new 1 ml cultures with fresh MH medium supplemented with antibiotic concentrations 1-fold, 2-fold, 4-fold, and 6-fold of the concentration from the previous culture. The selection was carried out for 10 days (60 generations). Samples were prepared for real-time polymerase chain reaction (PCR) and frozen in 10% DMSO at −80 °C after day 3, 6, 7, 8, 9, and 10.

Serial Passage for Determining Gene Array Segregation

Each meropenem-resistant population was split up into four new lineages and cycled in MH broth supplemented with erythromycin to keep selection for pUUH239.2 but relax selection for amplified β-lactamase gene arrays. This cycling was continued for 54 generations. Daily one sample was saved in 10% DMSO at −80 °C and another sample was prepared for real-time PCR.

Determination of Gene Copy Number and Amplification Endpoints

Real-time PCR technique was used to determine gene copy numbers on pUUH239.2. The MiniOpticon real-time PCR system (Bio-Rad, Hercules, CA, USA) was used to detect the fluorescent signal of iQ SYBR green supermix (Bio-Rad) binding to double-stranded DNA. Samples of passaged cultures were boiled for 10 min, pelleted at 17,000 × g, and the supernatant was used as DNA template. The β-lactamase genes *bla*$_{\text{CTX-M-15,}}$ *bla*$_{\text{OXA-1}}$, and *bla*$_{\text{TEM-1}}$ were targeted, and *pemK* outside of the resistance region was used as an internal control. Primers binding to *pecM* were used to determine the amplification endpoints. The primers used for amplification were *E. coli* RT ctx-m-15: forward 5′-TCGGTTCGCTTTCACTTTTCTT-3′, reverse 5′-AGTCTGGGTGTGGCATTGA-3′; *E. coli* RT pemK: forward 5′-GCCTGTTGTTGTGCCCGTA-3′, reverse 5′-TTTTCCGCCCCGTGCTTT-3′; *E. coli* RT oxa-1: forward 5′-TGGTGATCGCATTTTTCTTGGCT-3′, reverse 5′-ACGGATGGTTTGAAGGGTTTAT-3′; *E. coli* RT tem-1: forward 5′-TGAATGAAGCCATACCAAACG-3′, reverse 5′-ATCCGCCTCCATCCAGTC-3′; *E. coli* RT pecM: forward 5′-CGTGGCGCTGTTGGTGTTGA-3′, reverse 5′-AGGCGGTAAAGGTGAGCAGA-3′. The gene copy number was calculated as follows: gene copy number = $2^{[\mathrm{CT(control)-CT(target)}]}$.

Growth Rate Measurements

Growth rates were measured at 37 °C in MH broth using a Bioscreen C reader (Labsystems) measuring the OD$_{600}$ every 4 min. OD$_{600}$ values between 0.02 and 0.08, where growth was observed to be exponential, were used for the calculation. The relative growth rate was calculated as the derived growth rate divided by the growth rate of DA24338 from the same experiment.

Acknowledgments

This work was supported by grants from the Swedish Research Council (to D.I.A. and L.S.), Carl Tryggers Stiftelse (to L.S.), and European Commission project EvoTAR (to D.I.A.).

References

Adler M, Anjum M, Andersson DI, Sandegren L. 2012. Influence of acquired β-lactamases on the evolution of spontaneous carbapenem resistance in Escherichia coli. *J Antimicrob Chemother.* 68:51–59.

Andersson DI, Hughes D. 2009. Gene amplification and adaptive evolution in bacteria. *Annu Rev Genet.* 43:167–195.

Anderson P, Roth J. 1981. Spontaneous tandem genetic duplications in *Salmonella typhimurium* arise by unequal recombination between rRNA (rrn) cistrons. *Proc Natl Acad Sci U S A.* 78:3113–3117.

Anderson RP, Roth JR. 1977. Tandem genetic duplications in phage and bacteria. *Annu Rev Microbiol.* 31:473–505.

Anderson RP, Roth JR. 1979. Gene duplication in bacteria: alteration of gene dosage by sister-chromosome exchanges. *Cold Spring Harb Symp Quant Biol.* 43:1083–1087.

Bergthorsson U, Andersson DI, Roth JR. 2007. Ohno's dilemma: evolution of new genes under continuous selection. *Proc Natl Acad Sci U S A.* 104:17004–17009.

Datsenko KA, Wanner BL. 2000. One-step inactivation of chromosomal genes in Escherichia coli K-12 using PCR products. *Proc Natl Acad Sci U S A.* 97:6640–6645.

Dittmar K, Liberles D. 2011. Evolution after gene duplication. Hoboken (NJ): John Wiley & Sons.

Force A, Lynch M, Pickett FB, Amores A, Yan Y-L, Postlethwait J. 1999. Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* 151:1531–1545.

Francino MP. 2005. An adaptive radiation model for the origin of new gene functions. *Nat Genet.* 37:573–577.

Gordo I, Perfeito L, Sousa A. 2011. Fitness effects of mutations in bacteria. *J Mol Microbiol Biotechnol.* 21:20–35.

Hastings PJ, Lupski JR, Rosenberg SM, Ira G. 2009. Mechanisms of change in gene copy number. *Nat Rev Genet.* 10:551–564.

Hendrickson H, Slechta ES, Bergthorsson U, Andersson DI, Roth JR. 2002. Amplification-mutagenesis: evidence that "directed" adaptive mutation and general hypermutability result from growth with a selected gene amplification. *Proc Natl Acad Sci U S A.* 99:2164–2169.

Hooper SD, Berg OG. 2003. On the nature of gene innovation: duplication patterns in microbial genomes. *Mol Biol Evol.* 20:945–954.

Hughes AL. 1994. The evolution of functionally novel proteins after gene duplication. *Proc R Soc Lond B Biol Sci.* 256:119–124.

Katju V, Bergthorsson U. 2013. Copy-number changes in evolution: rates, fitness effects and adaptive significance. *Front Genet.* 4:273.

Katju V, Farslow JC, Bergthorsson U. 2009. Variation in gene duplicates with low synonymous divergence in *Saccharomyces cerevisiae* relative to *Caenorhabditis elegans*. *Genome Biol.* 10:R75.

Kimura M. 1962. On the probability of fixation of mutant genes in a population. *Genetics* 47:713–719.

Koskiniemi S, Sun S, Berg OG, Andersson DI. 2012. Selection-driven gene loss in bacteria. *PLoS Genet.* 8:e1002787.

Langley CH, Stevens K, Cardeno C, Lee YC, Schrider DR, Pool JE, Langley SA, Suarez C, Corbett-Detig RB, Kolaczkowski B, et al. 2012. Genomic variation in natural populations of *Drosophila melanogaster*. *Genetics* 192:533–598.

Lipinski KJ, Farslow JC, Fitzpatrick KA, Lynch M, Katju V, Bergthorsson U. 2011. High spontaneous rate of gene duplication in *Caenorhabditis elegans*. *Curr Biol.* 21:306–310.

Lynch M. 2000. The evolutionary fate and consequences of duplicate genes. *Science* 290:1151–1155.

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

Lynch M, Force AG. 2000b. The origin of interspecific genomic incompatibility via gene duplication. *Am Nat.* 156:590–605.

Näsvall J, Sun L, Roth JR, Andersson DI. 2012. Real-time evolution of new genes by innovation, amplification, and divergence. *Science* 338:384–387.

Neidhardt FC, Ingraham JL, Schaechter M. 1990. Physiology of the bacterial cell: a molecular approach. Sunderland (MA): Sinauer Associates Inc.

Ohno S. 1970. Evolution by gene duplication. Springer Verlag.

Pettersson ME, Andersson DI, Roth JR, Berg OG. 2005. The amplification model for adaptive mutation: simulations and analysis. *Genetics* 169: 1105–1115.

Pettersson ME, Sun S, Andersson DI, Berg OG. 2009. Evolution of new gene functions: simulation and analysis of the amplification model. *Genetica* 135:309–324.

Reams AB, Kofoid E, Kugelberg E, Roth JR. 2012. Multiple pathways of duplication formation with and without recombination (RecA) in *Salmonella enterica*. *Genetics* 192:397–415.

Reams AB, Kofoid E, Savageau M, Roth JR. 2010. Duplication frequency in a population of *Salmonella enterica* rapidly approaches steady state with or without recombination. *Genetics* 184: 1077–1094.

Sandegren L, Linkevicius M, Lytsy B, Melhus A, Andersson DI. 2011. Transfer of an *Escherichia coli* ST131 multiresistance cassette has created a Klebsiella pneumoniae-specific plasmid associated with a major nosocomial outbreak. *J Antimicrob Chemother*. 67: 74–83.