eLIFE elifesciences.org

RESEARCH ARTICLE

Intrinsic adaptive value and early fate of gene duplication revealed by a bottom-up approach

Guillermo Rodrigo[1,2]*, Mario A Fares[1,2,3†]

[1]Instituto de Biología Molecular y Celular de Plantas, CSIC – UPV, Valencia, Spain; [2]Instituto de Biología Integrativa y de Sistemas, CSIC – UV, Paterna, Spain; [3]Trinity College Dublin, University of Dublin, Dublin, Ireland

Abstract The population genetic mechanisms governing the preservation of gene duplicates, especially in the critical very initial phase, have remained largely unknown. Here, we demonstrate that gene duplication confers per se a weak selective advantage in scenarios of fitness trade-offs. Through a precise quantitative description of a model system, we show that a second gene copy serves to reduce gene expression inaccuracies derived from pervasive molecular noise and suboptimal gene regulation. We then reveal that such an accuracy in the phenotype yields a selective advantage in the order of 0.1% on average, which would allow the positive selection of gene duplication in populations with moderate/large sizes. This advantage is greater at higher noise levels and intermediate concentrations of the environmental molecule, when fitness trade-offs become more evident. Moreover, we discuss how the genome rearrangement rates greatly condition the eventual fixation of duplicates. Overall, our theoretical results highlight an original adaptive value for cells carrying new-born duplicates, broadly analyze the selective conditions that determine their early fates in different organisms, and reconcile population genetics with evolution by gene duplication.

DOI: https://doi.org/10.7554/eLife.29739.001

***For correspondence:** guillermo.rodrigo@csic.es

†Deceased

Competing interests: The authors declare that no competing interests exist.

Funding: See page 17

Received: 20 June 2017
Accepted: 04 January 2018
Published: 05 January 2018

Reviewing editor: Diethard Tautz, Max-Planck Institute for Evolutionary Biology, Germany

Introduction

Gene duplication has enthralled researchers for decades due to its link to the emergence of major evolutionary innovations in organisms of ranging complexity (***Ohno, 1970***). The key aspect to deeply understand this process concerns the early stage, when the fate of the new-born gene is decided (***Innan and Kondrashov, 2010***). A classical theory predicts the fixation of duplicated genes in the population under neutral selective conditions (i.e. by random genetic drift; ***Kimura, 1983***; ***Lynch and Conery, 2003***). Hence, the loss of the new-born gene is the most common evolutionary fate. Once a duplicate is fixed, it is generally accepted that genetic redundancy leads to relaxed selection constraints over one or both gene copies, which increases the load in mutations (***Lynch and Conery, 2000***; ***Keane et al., 2014***). In rare occasions, this evolutionary process leads to the origin of a novel, previously unexplored function by one of the gene copies (***Conant and Wolfe, 2008***).

However, because gene duplication can impose a cost to the cell by requiring additional resources for expression (***Wagner, 2005***; ***Lynch and Marinov, 2015***; ***Price and Arkin, 2016***), especially in simple organisms, purifying selection could preclude that fixation. Gene duplication can also unbalance tightly regulated pathways that are instrumental for the cell (***Papp et al., 2003***; ***Birchler et al., 2005***), leading to diseases in complex organisms (***Tang and Amon, 2013***). A possible rationale that has been long recognized is that those duplicated genes that were fixed in the population immediately contributed with an adaptive value to the organism (***Innan and Kondrashov, 2010***). Even

though, it is still stunningly unclear to what extent natural selection could also take part in the process that drives the fixation, and also initial maintenance, of duplicated genes according to population genetics (**Lynch, 2007**).

Two basic hypotheses have been proposed to explain the selective advantage of duplicated genes. First, a higher gene expression level resulting from duplication could be favorable (**Riehle et al., 2001**). This hypothesis requires that the ancestral system (pre-duplication) is far from the optimal operation point; as far as to assert that nearby 100% expression increase is beneficial. This seems plausible in extreme circumstances, but not in routine environments for which the organism should be adapted (**King and Masel, 2007**). It is then not surprising that many of the reported examples in which a greater gene copy number is favorable relate to sporadic, mainly stressing environments (**Riehle et al., 2001**; **Gonzalez et al., 2005**). Arguably, if a duplicate were fixed in one of these environments, it would be rapidly removed by purifying selection once the extreme circumstance ceased. Moreover, beneficial single-point mutations occurring in the *cis*-regulatory region of the gene of interest would be mostly sufficient to face several environmental changes (**Wray, 2007**). Thus, this model is insufficient to clarify the origin of most duplications, although it could explain some particular cases.

Second, the functional backup provided by the second gene copy upon duplication may allow the rapid accumulation of beneficial mutations, either to develop a novel function (**Zhang et al., 1998**; **Bergthorsson et al., 2007**), or to escape from the conflict of optimizing alternative functions (**Hittinger and Carroll, 2007**; **Des Marais and Rausher, 2008**). The positive selection of these mutations may of course occur, as suggested by the dN/dS values (>1) reported for different genomic sequences (**Han et al., 2009**; **Fischer et al., 2014**). This requires, nevertheless, that the frequency of cells carrying a second gene copy in the population increases to a point at which a mutation in the duplicate is likely to be found; a condition that is not met during the critical very initial phase following duplication (**Lynch et al., 2001**). Therefore, such adaptive processes, although important for the long-term maintenance of duplicates, do not contribute much to increase their fixation probabilities.

In addition to these two hypotheses, it has been proposed that gene duplication could allow compensating for errors in the phenotypic response due to a loss of expression caused by genotypic or phenotypic mutations (**Clark, 1994**; **Nowak et al., 1997**; **Wagner, 1999**). This model needs to invoke high error rates to have an impact at the population level from the beginning, and then to reach prevalence of genotypes with duplication by overcoming genetic drift. Errors in phenotype could also be caused by stochastic fluctuations in gene expression (**Elowitz et al., 2002**; **Balázsi et al., 2011**), with gene duplication eventually reducing the amplitude of such fluctuations (**Kafri et al., 2006**; **Lehner, 2010**; **Rodrigo and Poyatos, 2016**). But this strategy works on average, that is, duplication may warrant more accuracy when multiple decisions in gene expression are considered. Thus, it is not obvious whether an individual (or some) with duplication is able to invade a population, especially in a fluctuating environmental context. This is a key, largely unexplored question that may preclude the support of this idea. Other mechanistic models have been proposed beyond the demand for increased expression or the accumulation of beneficial mutations (**Innan and Kondrashov, 2010**), yet do not convincingly resolve the main population genetic dynamical issue.

In this work, we tested the idea of error buffering to reveal the adaptive value that gene duplication has per se. Subsequently, we developed a comprehensive model to explain the early fate of duplicates compatible with population genetics (**Lynch et al., 2001**; **Lynch, 2007**), global gene expression patterns (**Qian et al., 2010**; **Gout and Lynch, 2015**; **Cardoso-Moreira et al., 2016**; **Lan and Pritchard, 2016**), and unexpected gene copy number variation rates (**Reams et al., 2010**; **Schrider et al., 2013**). To this end, instead of performing a conventional sequence analysis (top-down approach), we followed a very precise quantitative framework, based on biochemistry, to study the goodness of having a second gene copy for the cell without functional divergence (bottom-up approach). Using a gene of *Escherichia coli* (*lacZ*) as a model system from which to apply our theory, we showed, without loss of generality, that the sum of two different, partially correlated responses allows reducing gene expression inaccuracies (**Rodrigo and Poyatos, 2016**); inaccuracies that are a consequence of the inherently stochastic nature of all molecular reactions underlying gene expression (**Raser et al., 2004**; **Carey et al., 2013**) and suboptimal gene regulation (**Dekel and Alon, 2005**; **Price et al., 2013**). Here, we considered intrinsic and extrinsic noise sources (**Elowitz et al., 2002**), that is, stochastic fluctuations that are specific of a gene and fluctuations that are unspecific, so gene duplication is expected to only buffer intrinsic fluctuations. In turn, cell fitness

can weakly increase on average, if such errors in gene expression are costly (**Wang and Zhang, 2011**); that is, a stochastic fluctuation may take the system far from the optimal operation point if the system is deterministically centered in this point), and then genotypes with duplication can be fixed in the population. We further studied the genetic and environmental conditions that are more favorable for the selection of gene duplication.

Results

Quantitative biochemical view of a fitness trade-off

In cellular systems, fitness trade-offs arise because beneficial actions involve costs. Fitness is a complex figure integrating multiple components, so the enhancement of one component (vital attribute) usually comports the reduction of another component (e.g. stress resistance vs. reproductive success; **Casanueva et al., 2012**). This is critically revealed when the environment changes, as the relevance of each component mostly depends on the external conditions. Such components can be described in different ways according to the problem. A paradigmatic and simple fitness trade-off emerges when a given enzyme needs to be expressed to metabolize a given nutrient present in the environment (**Figure 1a,b,c**). On the one hand, the cell growth rate (here taken as a metric of fitness; **Elena and Lenski, 2003**) increases as long as the enzyme metabolizes the nutrient. On the other hand, the enzyme expression produces a cost to the cell (i.e. reduces its growth rate). Therefore, the enzyme expression needs to be very precise to warrant an optimal or near-optimal behavior (cost-benefit analysis). To solve this issue, regulations (mainly transcriptional) evolved to link enzyme expression inside the cell with nutrient amount available in the environment. An example of this paradigmatic system is the well-known lactose utilization network of *E. coli* (**Jacob and Monod, 1961**), where lactose (nutrient, environmental molecule) activates, through inhibition of LacI (transcription factor), the production of LacZ (enzyme). We used this model system to apply a theoretical framework (see Materials and methods) in order to reveal the intrinsic adaptive value of gene duplication under a fitness trade-off, as this system has been quantitatively characterized (**Dekel and Alon, 2005**; **Kuhlman et al., 2007**; **Eames and Kortemme, 2012**).

Cell fitness increases monotonically with lactose dose (following a Michaelis-Menten kinetics), but presents an optimum with LacZ expression (**Figure 1d**). This is because lactose does not introduce a cost into the system, but LacZ does. Here, we simply considered a cost function based on LacZ expression (i.e. more expression, more cost), with a marginal cost of 0.036 in the units of the model (**Dekel and Alon, 2005**). However, it would be more precise to have a cost function based on lactose permease (LacY) activity (**Eames and Kortemme, 2012**), another gene in the *lac* operon in charge of the uptake, rather than on LacZ expression. The regulation of the system appears to be quite accurate, as the actual and optimal dose-response curves roughly match (**Figure 1e**). By generating different dose-response curves with values of x_0 (lactose EC_{50} on LacZ) between 0.01 and 1 mM, we found that most of them deviate from the optimal one ($p = 0.02$; Euclidean distance as a metric). This entails great phenotypic plasticity of the cell to cope with lactose variations. However, plasticity is not equal for all environmental changes. Whilst the system (in terms of LacZ expression or cell fitness) reaches optimal sensitivity at intermediate doses, it is quite insensitive at very low or very high doses, where lactose-LacZ information transfer falls down (**Figure 1f**).

Gene duplication helps to better resolve the fitness trade-off

The LacZ expression in *E. coli* involves a variety of noisy actions, such as the LacI expression, the LacI-DNA binding, the RNA polymerase-DNA binding, and the transcriptional elongation process (**Elowitz et al., 2002**; **Raser et al., 2004**; **Carey et al., 2013**). The resulting stochastic fluctuations in expression can have an impact on fitness (**Figure 2**). Using a simple mathematical model, we simulated the stochastic LacZ expression of the wild-type system for a varying lactose dose (**Figure 3a, b**). The magnitudes of the stochastic fluctuations were chosen as to end in typical variations of lactose EC_{50} of 10–100%, up or down, resulting in values of gene expression noise, around 0.5, compatible with experimental results (**Elowitz et al., 2002**). At a given dose, these simulations would correspond to different single-cell responses. We also considered a system with two copies of the *lacZ* gene, with total expression equal to the previous one-copy system, and simulated its stochastic response (**Figure 3c**). For the moment, we ensured gene dosage sharing to evaluate in a

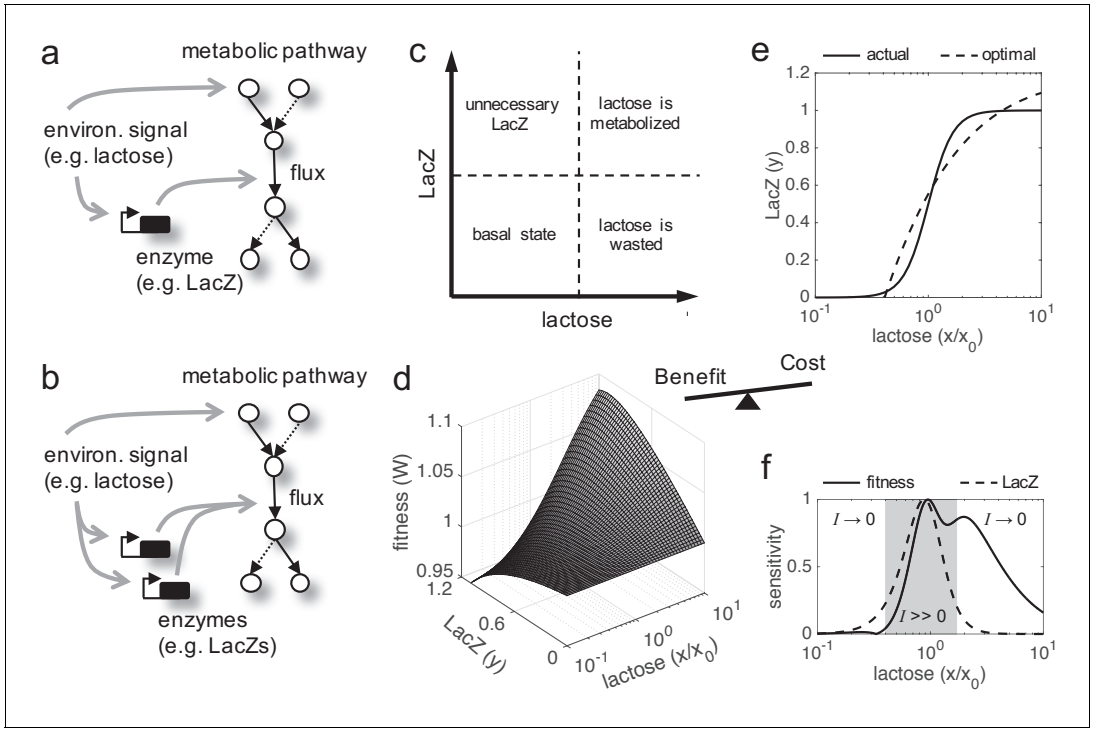

Figure 1. Fitness trade-off related to metabolic benefit and expression cost. (**a**) Scheme of a paradigmatic genetic system, coupling regulation and metabolism, where a given environmental signal determines the physiology of the cell. The environmental molecule can be metabolized by the cell, and it can also activate transcriptionally the expression of enzymes. A particular case is the lactose utilization system of *E. coli*. (**b**) Scheme of the same system with gene duplication. (**c**) Illustrative chart of the fitness trade-off showing four different cellular regimes. When the signal molecule (lactose) is not present in the medium, the expression of the enzyme (LacZ) is not required. However, when the signal molecule is present, the enzyme is required for its metabolic processing. (**d**) Fitness (W) landscape as a function of lactose (contributing to the benefit, x denotes its concentration) and LacZ (contributing to both the benefit and the cost, y denotes its concentration). This was experimentally determined. x_0 denotes the lactose EC_{50} on LacZ expression, so x/x_0 is a normalized lactose concentration. (**e**) Dose-response curve between lactose and LacZ. The solid line corresponds to the actual regulation (experimentally determined), whilst the dashed line corresponds to a hypothetical optimal regulation (obtained by imposing $dW/dy = 0$). (**f**) Sensitivity to changes in lactose dose, either in fitness (dW/dx, solid line) or in LacZ (dy/dx, dashed line), characterizing the nonlinear phenotypic plasticity of the cell. Each curve is normalized by its maximum. This also measures sensitivity to molecular noise. The region where information transfer is high is shaded.

DOI: https://doi.org/10.7554/eLife.29739.002

quantitative way the goodness of having a second gene copy for the cell without invoking the need for more expression. We observed that the system with gene duplication produces a more accurate response (i.e. a response closer to the deterministic one), highlighting the role of gene copy number in noise buffering (***Rodrigo and Poyatos, 2016***).

In addition, we calculated the proposed fitness function for each single-cell response. Small gene expression inaccuracies (e.g. an excess of enzyme for the available substrate) can be perceived as a consequence of a hill-like fitness landscape in terms of the genotype-environment interaction (***Figure 1d***). To properly compare how each system of study resolves the fitness trade-off, we then calculated the selection coefficient for each response. We found a skewed distribution, peaked at 0 and with a positive mean of 0.08% (***Figure 3d***). This entails that phenotypic responses generated by duplicated genes give, on average, higher fitness values than responses generated by singleton genes. To better illustrate this fact, we represented cell fitness as a function of LacZ expression (***Figure 3e***), uncovering two reasons by which gene duplication is adaptive. In first place, the variance of the stochastic fluctuations (noise) in gene expression is reduced by 50% upon duplication (***Wang and Zhang, 2011***); when only intrinsic fluctuations are considered). However, when both intrinsic and extrinsic fluctuations are considered, the variance is reduced by 15–25%. In any case, this increases fitness on average, because the system displays a near-optimal behavior in the deterministic regime, thus fluctuations are costly. In second place, the population response upon

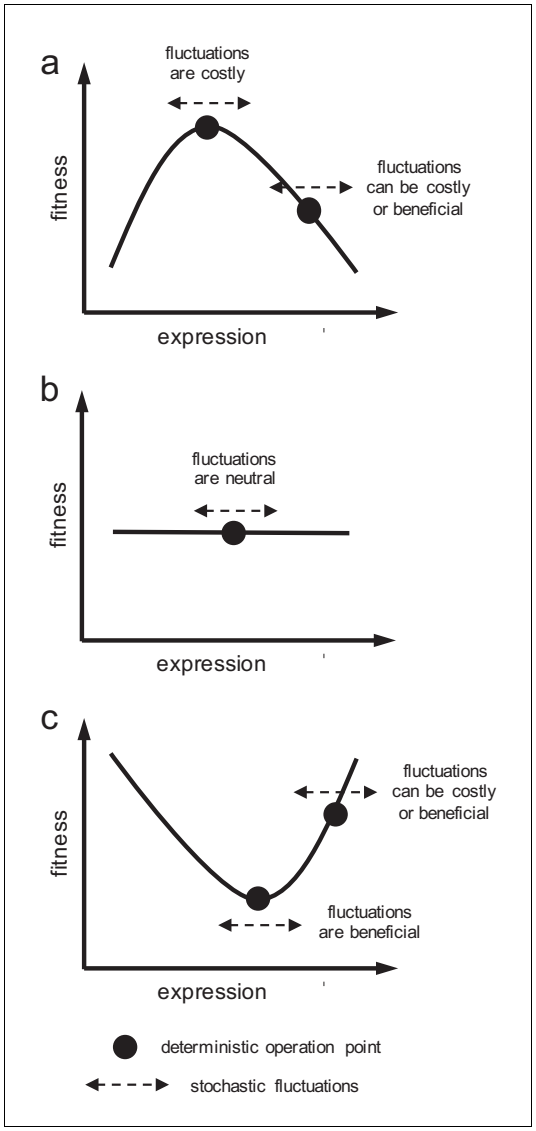

Figure 2. Schematics of cell fitness as a function of gene expression. Fitness function can (**a**) present a maximum, (**b**) be flat, or (**c**) present a minimum. Depending on the local shape, stochastic fluctuations in expression can be costly, beneficial, or neutral.
DOI: https://doi.org/10.7554/eLife.29739.003

duplication is slightly closer to the optimal operation point (***Figure 3e,f***). The model-based median dose-response curve (corresponding to the experimental response at the population level) is sigmoidal and has a Hill coefficient of 4 (***Dekel and Alon, 2005***). This results in a slope (LacZ vs. normalized lactose) of 1, calculated as $n/4$ at x_0 (n is the Hill coefficient). This slope is higher than the slope coming from the optimal dose-response curve, which is 0.47 at x_0. However, when duplication is considered (maintaining the same expression levels), the median dose-response curve shows a slope of 0.75 (corresponding to an effective Hill coefficient of 3) also at x_0 (***Figure 3f***). This is because, in this case, the actual dose-response curve is more nonlinear than the optimal one, a feature that can indeed be amended by genetic redundancy (***Gammaitoni, 1995***; ***Rodrigo and Poyatos, 2016***).

Finally, we calculated how much selection exists, on average, as a two-dimensional function of the magnitude of intrinsic noise and the concentration of lactose in the medium (***Figure 3g***). This highlights the fundamental link between noise reduction in gene expression and selective advantage (cell fitness). More in detail, we found that the higher the intrinsic noise, the higher the adaptive value of gene duplication. This is because intrinsic noise generates the required heterogeneity between the responses of the two gene copies to limit large stochastic fluctuations in the total gene expression. We also found that there is a maximal adaptive value of gene duplication at intermediate lactose doses, where the sensitivity of the system is the highest (***Figure 1f***). Out of this regime, the stochastic fluctuations, according to our simple mathematical model, have less impact on the phenotype (***Blake et al., 2006***).

Gene duplication can be positively selected in a population thanks to more accurate responses

If gene duplication enhances cell fitness on average, viz., by reducing gene expression inaccuracies, it would be expected a positive selection of this trait in a population (***Kimura, 1983***). To verify this assumption, we performed experiments of in silico evolution (see Materials and methods), where a mixed population of cells carrying singletons and duplicates was monitored, considering equal LacZ expression in both types of cells (***Figure 4a***). The population was left to evolve without introducing any bias, with time-dependent stochastic fluctuations in gene expression uncorrelated from cell to cell. For simplicity, we simulated a scenario of experimental evolution (***Elena and Lenski, 2003***; ***Dekel and Alon, 2005***), although the dynamics in nature might be more complex. We found that the frequency of cells carrying duplicates in the population increases with time, and that such an increase is well predicted by population genetic dynamics with the mean selection coefficient (***Figure 4b***). Notably, this points out that this parameter, which can be mathematically calculated

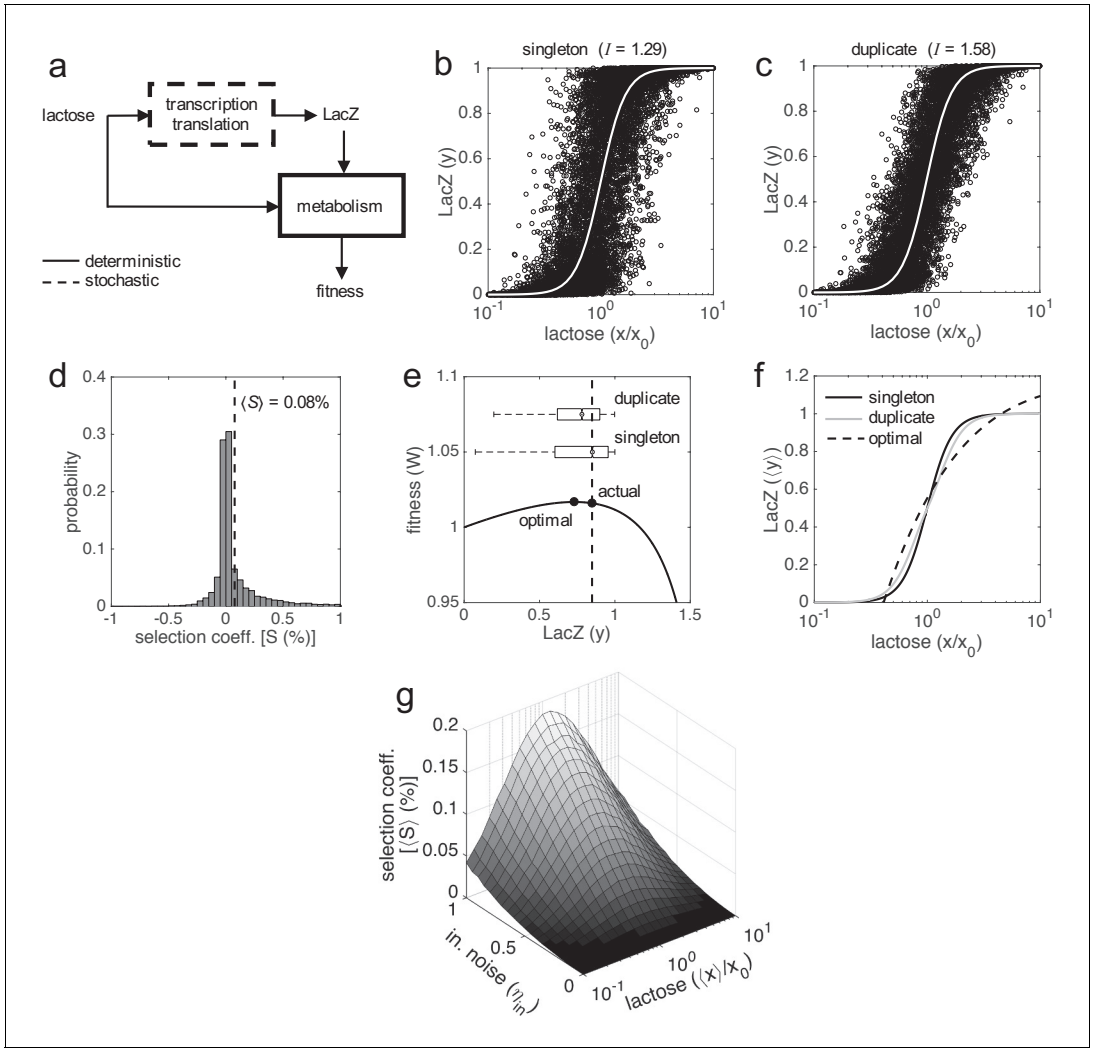

Figure 3. Selective advantage of gene duplication. (**a**) Block diagram of the system. Gene expression is calculated by means of a stochastic function, whilst fitness by means of a deterministic one. (**b, c**) Single-cell responses at different lactose doses (stochastic simulations, noise amplitudes of η_{in} = 0.5 and η_{ex} = 0). Lactose and LacZ concentrations are denoted by *x* and *y*, respectively. The solid white line corresponds to the deterministic simulation. In b) the genotype contains a single copy of *lacZ* gene, whilst in c) it contains two copies. The value of mutual information (*I*) is shown in both cases: 1.29 bits of information in case of a singleton and 1.58 bits in case of a duplicate (about 25% increase in fidelity, significance assessed by a *z*-test, p ≈ 0 with 10^4 points). (**d**) Selection coefficient (*S*) of a genotype with two copies of *lacZ* gene over another with just one copy. The mean selection coefficient is shown (dashed line). Skewness coefficient of 2.63. *W* values calculated from *x*, *y* values shown in b, c). (**e**) Fitness (*W*) as a function of LacZ (constant *x* = 0.2 mM), showing the distributions of expression (boxplots) in case of one or two gene copies. The actual LacZ expression is shown (dashed line). (**f**) Dose-response curve between lactose concentration and the median LacZ expression (⟨*y*⟩). The solid lines correspond to the actual responses in case of one (black) or two (gray) gene copies (η_{in} = 0.5 and η_{ex} = 0), whilst the dashed line corresponds to the optimal response. (**g**) Mean selection coefficient (⟨*S*⟩) landscape of gene duplication as a function of the median lactose dose (⟨*x*⟩, fluctuating dose) and the amplitude of intrinsic noise (η_{in}, with fixed η_{ex} = 0.3). In all these plots, the expression levels of the duplicates with respect to the singletons are equal ($y_{max,1} = y_{max,2} = 0.5$).

DOI: https://doi.org/10.7554/eLife.29739.004

a priori, is sufficient to capture all the complexity underlying the stochastic evolutionary dynamics of the system (***Hegreness et al., 2006***).

In addition, we studied the effect of the magnitude of molecular noise. We distinguished between intrinsic and extrinsic noise (***Elowitz et al., 2002***). As predicted from our previous results, we found that the higher the intrinsic noise of the system, the higher the frequency of gene duplication in the population (***Figure 4c***). By contrast, the higher the extrinsic noise, the lower the frequency (***Figure 4c***), as this type of noise affects in the same way the responses of the two copies. Note that there is no gain following duplication when only extrinsic noise is considered. Furthermore, we

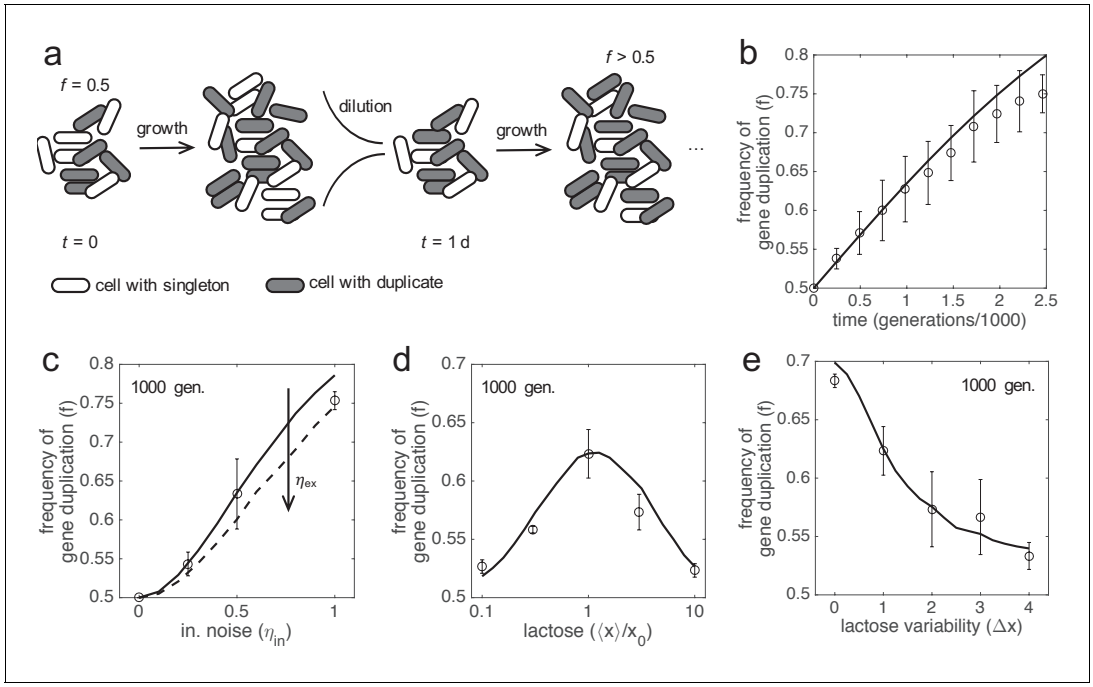

Figure 4. In silico evolution experiments. (**a**) Scheme of an evolutionary procedure, where serial dilution passages are applied, to assess the performance in a cell population of a genotype with two copies of *lacZ* gene over another with just one copy. (**b**) Time-dependent frequency of cells with gene duplication (*f*). Open circles and error bars correspond to experiments of in silico evolution (mean and standard deviation of three replicates) with an initial frequency of $f_0 = 0.5$, fluctuating lactose dose, and noise levels of $\eta_{in} = 0.5$ and $\eta_{ex} = 0$. The solid line corresponds to the theoretical prediction. (**c**) *f* at 1000 generations (f_{1000}) as a function of the amplitude of intrinsic noise (η_{in}). Experiments and prediction with $f_0 = 0.5$ and $\eta_{ex} = 0$. The dashed line corresponds to the theoretical prediction with $\eta_{ex} = 1$. (**d**) f_{1000} as a function of the median lactose dose ($\langle x \rangle$). Experiments and prediction with $f_0 = 0.5$, $\eta_{in} = 0.5$ and $\eta_{ex} = 0.5$. (**e**) f_{1000} as a function of the lactose fluctuation amplitude (Δx). $\Delta x = 0$ corresponds to constant lactose dose. Experiments and prediction with the same values of f_0, η_{in} and η_{ex} as in d). Three replicates were also considered in c, d, e). In all these plots, the expression levels of the duplicates with respect to the singletons are equal ($y_{max,1} = y_{max,2} = 0.5$).

DOI: https://doi.org/10.7554/eLife.29739.005

studied the effect of the environment (lactose dose). As predicted, we found an intermediate median dose at which the frequency of gene duplication in the population is the highest (***Figure 4d***). We also found that the higher the variance, the lower the frequency (***Figure 4e***). This is because, when lactose fluctuates from very low to very high doses, the signal-to-noise ratio is large enough to warrant a relatively accurate response with just one gene copy (***Hansen et al., 2015***). Of relevance, the population genetic dynamics in all these cases, with the corresponding mean selection coefficients, correctly explained the reported frequencies.

Fixation is conditioned by the unexpected recurrence of formation and deletion of duplicates in a population

Gene duplicates can be spontaneously produced, through different mechanisms (***Hastings et al., 2009***), at very high rates in the cell. These rates, measured from experiments of mutation accumulation, go from 10^{-4} dup./gene/gen. in prokaryotes (***Reams et al., 2010***) to 10^{-7} dup./gene/gen. in higher eukaryotes (***Schrider et al., 2013***). Once produced, most of these duplicates are deleted as they are unstable, with a rate that appears to be higher than the formation rate (***Reams et al., 2010***; ***Schrider et al., 2013***). In the particular case of the *lacZ* gene, we have a formation rate of $3{\cdot}10^{-4}$ dup./gene/gen. and a deletion rate of $4.4{\cdot}10^{-2}$ -/gene/gen. (in a single bacterial cell; data for *Salmonella enterica*). Therefore, gene duplication can be understood as a recurrent process that reaches an equilibrium point given by the ratio between the formation and deletion rates (***Figure 5a***), neglecting fitness effects. This equilibrium point would be lower if fitness effects (mostly detrimental) were considered. This entails about $2{\cdot}10^5$ cells carrying *lacZ* duplicates in a typical *E. coli* population of $2{\cdot}10^8$ cells in nature (***Lynch et al., 2016***; that is, frequency of about 0.1%). This surprising scenario

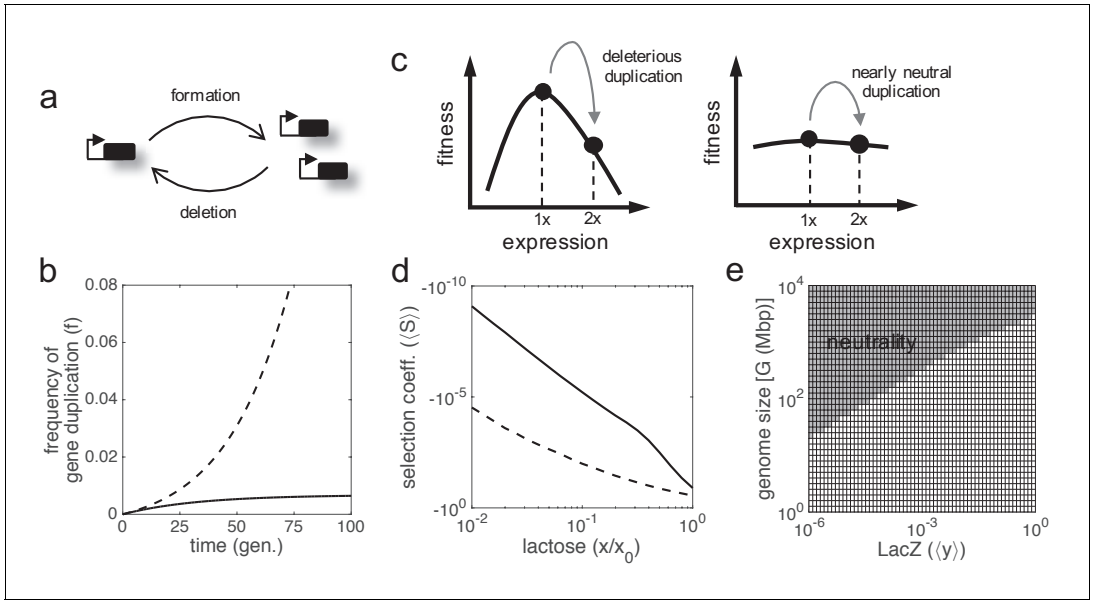

Figure 5. Gene duplication leading to double expression. (**a**) Scheme of the formation-deletion balance in gene duplication. (**b**) Time-dependent frequency of cells with gene duplication (*f*) when the formation and deletion rates of a second *lacZ* copy are considered. Sequence remodeling was not taken into account. The solid line corresponds to a scenario of neutrality, whilst the dashed line corresponds to a scenario of positive selection (with $S = 10\%$). (**c**) Schematics of fitness as a function of expression showing the effect of gene duplication. Two scenarios are considered: deleterious duplication (left; hill-like fitness landscape) and nearly neutral duplication (right; quasi-flat fitness landscape). (**d**) Mean selection coefficient ($\langle S \rangle$) as a function of lactose dose upon *lacZ* duplication doubling gene expression ($y_{max,1} = y_{max,2} = 1$). The solid line corresponds to noise levels of $\eta_{in} = \eta_{ex} = 0.3$ (moderate), whilst the dashed line corresponds to $\eta_{in} = \eta_{ex} = 1$ (high). (**e**) Identification of effectively neutral selective conditions (when $|\langle N \rangle \cdot \langle S \rangle| < 1$, region shaded) in terms of gene expression (y) and genome size (G), which determines the effective population size ($\langle N \rangle$). In this context, no benefit was considered ($a = 0$), with moderate noise levels.

DOI: https://doi.org/10.7554/eLife.29739.006

has an immediate consequence, viz., duplicated genes cannot be fixed in the population by drift under neutral selective conditions (***Figure 5b***); a result already anticipated (***Clark, 1994***) in clear discrepancy with the conventional wisdom (***Lynch, 2007***). Indeed, the formation-deletion balance would always take the system to the same equilibrium point.

However, the preceding argument only focuses on a static picture, ignoring the dynamics of the genetic process. In bacteria (*lacZ* gene), the time to reach the equilibrium point is about 68 generations (three times the inverse of the deletion rate), which is a relatively short transient period. By contrast, in flies (*Drosophila melanogaster*), the formation rate is of 10^{-7} dup./gene/gen. and the deletion rate of 10^{-6} -/gene/gen. (***Schrider et al., 2013***). Although this would yield equilibrium frequencies up to 10%, the transient periods would be longer than 10^6 generations (0.2 Ma in natural conditions; ***Pool, 2015***). Fixation could then happen by drift, as their effective population sizes are of 10^6 flies (***Lynch et al., 2016***), although not persistently. Note that the inverse of this number indeed specifies an upper limit for the deletion rate. In addition, the formation-deletion balance could be shifted if further genome rearrangements affecting duplicated genes were considered, such as gene relocation (about 10^{-11} fixed rearr./gene/gen. for *D. melanogaster*; ***Ranz et al., 2001***). In effective terms, gene relocation would reduce the deletion rate, and, consequently, fixation would be more likely (***Wong and Wolfe, 2005***). Such a relocation would also shift the intrinsic-extrinsic noise balance toward more uncoupled responses (***Becskei et al., 2005***), which could enhance the benefit by intrinsic noise reduction.

Most of the new-born duplicates lead to increased expression and are costly for the cell

So far, we have demonstrated that a duplicated gene offers a selective advantage provided the total gene expression level is maintained, with one or two copies (gene dosage sharing). However, this condition is not usually met during the critical very initial phase, when the duplicate has just born. In

general, we can assume that the expression level is doubled upon duplication, although this may vary due to the particular position in the chromosome of the duplicated gene and the type of cell (**Stranger et al., 2007**). Certainly, an increase of expression due to gene duplication is detrimental in most environments (**Figure 5c,d**; **Price and Arkin, 2016**), thus positive or neutral selective conditions are difficult to invoke to explain the fixation of these type of genotypic changes, mainly in prokaryotes and lower eukaryotes (**Lynch and Marinov, 2015**). For instance, at constant 0.13 mM lactose, we obtained mean selection coefficients between −28% (at very high noise levels) and −1% (at no noise) upon duplication of the *lacZ* gene (assuming double expression), which yield negligible fixation probabilities (almost 0) for a sufficiently large bacterial population. It can be argued, nevertheless, that the cost of over-expression decreases as long as the genome size increases (**Lynch and Marinov, 2015**). This assumption, together with the negative correlation between complexity and population size (**Lynch and Conery, 2003**), makes effectively neutral selective conditions plausible to rationalize the fixation of duplicates that are expressed (e.g., essential genes) in higher eukaryotes (**Figure 5e**; **Makino et al., 2009**).

Only in absence of lactose, when the enzyme is not needed, the duplication is strictly neutral (no benefit, no cost due to regulation). But neutral selective conditions can be reached *de facto* if the absolute value of the selection coefficient is lower than the inverse of the effective population size (**Kimura, 1983**). This condition is challenging for prokaryotes, as their population sizes are very large (**Lynch and Conery, 2003**). In our particular case, we obtained mean selection coefficients in the order of -10^{-10} (at moderate noise levels) when the nutrient amount is scarce (1 µM lactose), which could favor the fixation of a *lacZ* duplicate by genetic drift.

Gene dosage sharing upon duplication, fitness increase on average, and estimation of the fixation probability

Can a cell carrying a new-born duplicate that is expressed (in principle, in an operation point close to a local optimum) overcome the cost of an additional copy and then invade the population without invoking the need for more expression (to face an extreme environment)? We here predicted that the genetic variability existing in a population would allow reaching adaptive gene duplications (**Figure 6a**). Mutations in the *cis*-regulatory region of the *lacZ* gene may change its wild-type expression level. According to previous results (**Otwinowski and Nemenman, 2013**), the distribution of mutations in terms of maximal promoter activity is peaked at 1, but skewed to the left (**Figure 6b**). This indicates that about 10% of them yield cells with nearby 50% lower expression. Thus, if a gene duplication event occurred in one of these cells, the genotypic change would be selectively advantageous (**Figure 6c**). The frequency of such cells in the population depends, of course, on the mutation rate; the greater the ability to generate genetic diversity, the higher the chances to reach adaptive duplications. For *E. coli*, where the per base mutation rate is of 10^{-10} mut./bp/gen. (**Lee et al., 2012**), this frequency can be estimated in 10^{-9} (i.e. 0.2 mutants with nearby 50% lower expression per generation in a natural population of $2{\cdot}10^{8}$ cells). Hence, the probability that a duplication and such a mutation concur in the same cell in a generation (duplication after promoter mutation) is of 10^{-4} (=$0.2{\cdot}10^{-3}/2$; i.e., 1 suitable concurrence each 10^{4} generations).

In particular, at constant 0.13 mM lactose, we obtained a relatively high mean selection coefficient of 0.19% when the wild-type expression is recovered upon duplication (in a highly noisy scenario). However, the selection coefficient has to be greater than the duplication deletion rate to ensure fixation (**Figure 5b**); a condition that is not met here. Certainly, the high deletion rates observed in bacteria (**Reams et al., 2010**) protect them from acquiring genetic redundancy (perhaps, this is why *lacZ* is not duplicated in *E. coli* despite this may be beneficial). In other local genetic contexts, also in bacteria, the deletion rate of a *lacZ* duplicate can be as low as $4.1{\cdot}10^{-4}$ -/gene/gen. (**Reams et al., 2012**). In this scenario, a selection coefficient of 0.19% would lead to fixation. We then estimated a global fixation probability of $3{\cdot}10^{-7}$ (= $2{\cdot}15{\cdot}10^{-4}{\cdot}10^{-4}$; **Figure 6d**; see Materials and methods). Remarkably, our estimation is much higher than $5{\cdot}10^{-9}$, the fixation probability under hypothetical neutrality (**Kimura, 1983**).

A fitness increase on average due to expression noise reduction could also lead to the fixation of duplicates in eukaryotes, as nothing prevents assuming the same positive selective conditions (**Raser et al., 2004**; **Hansen et al., 2015**), which now largely outperform the duplication deletion processes. For *D. melanogaster*, for instance, where the per base mutation rate is of $5{\cdot}10^{-9}$ mut./bp/gen. (**Schrider et al., 2013**), and complete gene duplications have little impact on fitness

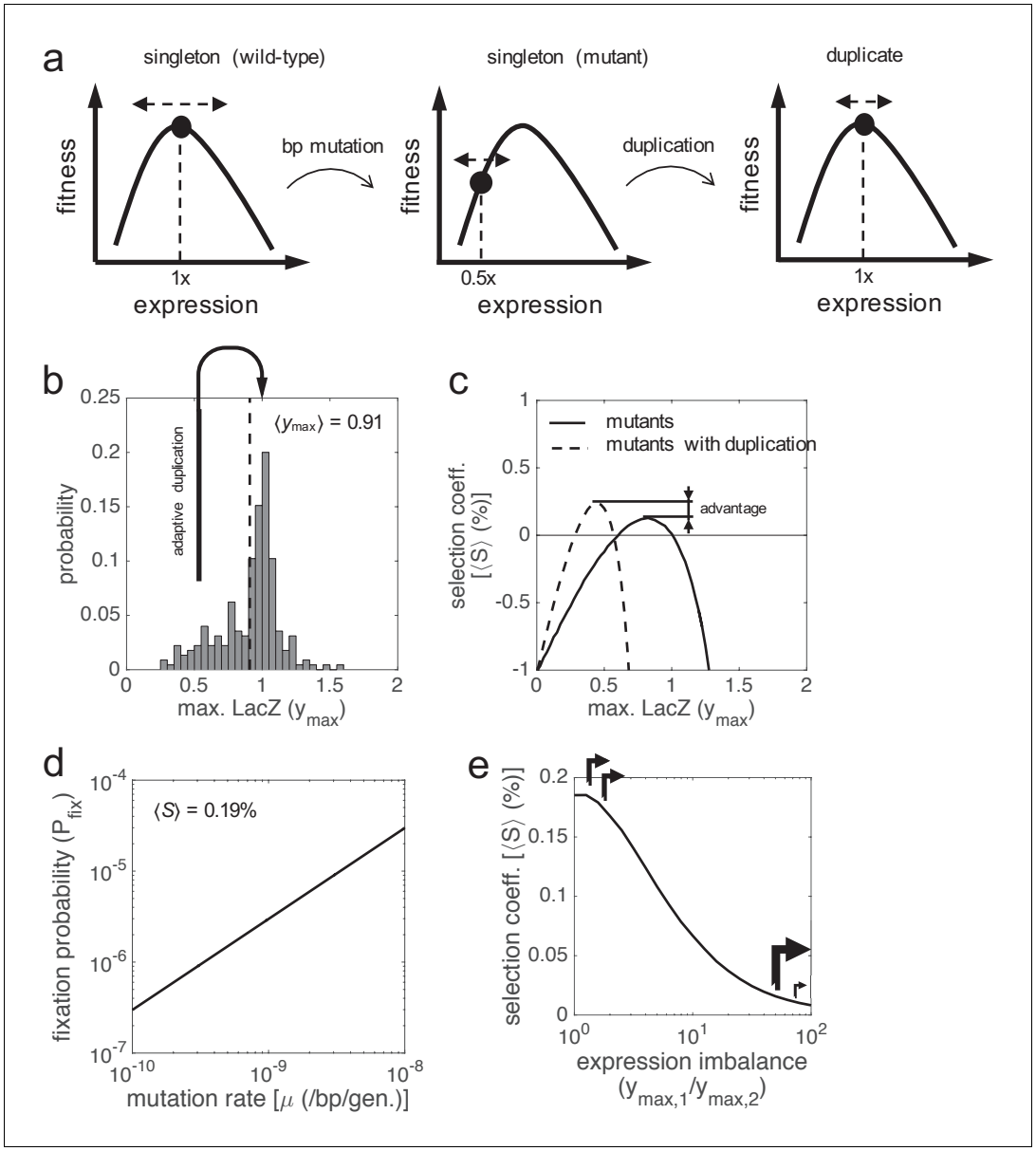

Figure 6. Gene duplication leading to maintained expression. (**a**) Schematics of fitness as a function of expression showing a path to reach adaptive gene duplications without the need for more expression. Two steps are considered: first a base-pair mutation that reduces in half the expression level, and then a duplication that recovers the ancestral level. (**b**) Distribution of the activity of *lac* promoter mutants based on experimental data, as the maximal LacZ expression (y_{max}, irrespective of lactose dose). The mean activity is shown (dashed line). Skewness coefficient of −0.68. (**c**) $\langle S \rangle$ of the promoter mutants versus the wild-type system (solid line), with fluctuating lactose dose and high noise levels. The dashed line corresponds to the comparative between promoter mutants that duplicated the *lacZ* gene and the wild-type system. (**d**) Fixation probability (P_{fix}) of gene duplication as a function of the mutation rate of the cell (μ), with $\langle S \rangle = 0.19\%$ and $\langle N \rangle = 2{\cdot}10^8$. (**e**) $\langle S \rangle$ as a function of the expression imbalance between the two *lacZ* copies ($y_{max,1}$ / $y_{max,2}$), when the system recovers its ancestral expression levels ($y_{max,1} = y_{max,2} = 0.5$), with constant $x = 0.13$ mM and high noise levels. Arrows illustrate the corresponding promoter strengths.

DOI: https://doi.org/10.7554/eLife.29739.007

(*Emerson et al., 2008*; note that other genome rearrangements not affecting entire genes are significantly deleterious), we estimated that 0.05 mutants with nearby 50% lower expression and up to 10^5 duplicants of the gene of interest would be found in the natural population. Hence, the probability of concurrence in the same organism (duplication after promoter mutation) would be of $2.5 \cdot 10^{-3}$. Consequently, the global fixation probability would be of 10^{-5}; again, higher than the one under hypothetical neutrality (*Kimura, 1983*).

Maintenance of a duplicate upon fixation in the population

A forthcoming change in lactose dose would be highly detrimental if a second *lacZ* copy were fixed in the population either under neutrality due to insignificant expression or under strong selection due to expression demand. In the former case, an increase of lactose would be detrimental; in the latter, a decrease would. Consequently, either the elimination of the duplicate by purifying selection (*Lynch and Conery, 2000*) or the accumulation of mutations that lower the LacZ expression to recover the ancestral phenotype (*Force et al., 1999*; *Qian et al., 2010*) would be promoted; with clonal interference in the case of asexual populations (*Rozen et al., 2002*; *Desai et al., 2007*). In the latter case, the two gene copies could be maintained in the genome for long time by buffering of costly stochastic fluctuations of intrinsic nature if they held similar expression levels (*Figure 6e*; *Gout and Lynch, 2015*); otherwise the gain in accuracy decreased. Conversely, if a second *lacZ* copy were fixed according to the path shown in *Figure 6a* under weak selection, it would be safe from changes in lactose dose.

The genomic inspection of organisms in which genetic drift is not, in principle, a suitable force to drive the fixation of duplicates (e.g. bacteria or yeast; *Lynch and Marinov, 2015*) gave us some empirical insight, despite the masking produced by subsequent evolutionary trajectories. In many cases of duplication, there is no a significant increase in total expression (e.g. duplicates in *Saccharomyces cerevisiae vs.* singletons in *Schizosaccharomyces pombe*; *Qian et al., 2010*). Thus, either duplicates were fixed by dosage in a definite environment to then return to ancestral expression levels, or duplicates were fixed by other means. In any case, the preservation of the ancestral function in the second copy is expected (*DeLuna et al., 2008*). Whether noise reduction was actually relevant for some fixations or not is hard to say without conducting an experimental approach to measure variability and selection (revealing the fitness landscape; *Figure 2*); notwithstanding, it seems a plausible mechanism according to our results, already put forward with the computational analysis of gene expression patterns (*Lehner, 2010*) and metabolic flux balances (*Wang and Zhang, 2011*) in yeast.

If dosage mattered at some point, the function encoded by the duplicated gene would be more important at the time of duplication than today. In *E. coli*, for example, genes *fsaA* and *fsaB* are paralogs, with high sequence (69%) and functional similarity, coding for a genuine fructose-6-phosphate aldolase (*Sánchez-Moreno et al., 2012*). The relevance of this enzyme for today *E. coli* is unclear, suggesting that *fsaB* might have been fixed by dosage in past habitats in which rare sugars were frequent. However, if noise were the critical aspect, the system would present some regulation to link environment with phenotype and the function would be of routine for the cell. In particular, *E. coli* expresses two redundant gluconokinases, encoded in genes *gntK* and *idnK* (51% of sequence identity), to face environments in which gluconate is the carbon source due to glucose oxidation (*Vivas et al., 1994*). Similar to the regulation of *lacZ* by lactose (*Jacob and Monod, 1961*), gluconate activates the expression of *gntK* and *idnK* by inhibiting the transcriptional repressor GntR (*Afroz et al., 2014*). Again, there would be a trade-off between metabolic benefit and expression cost (*Figure 1c*; read gluconate instead of lactose and GntK/IdnK instead of LacZ). Arguably, duplication might have been fixed in this case to cope with gene expression inaccuracies, especially when GntR produces bimodal responses (captured in single-cell experiments; *Afroz et al., 2014*).

A comprehensive model compatible with population genetics to explain the early fate of gene duplications

Taking all our results together, we formulated a comprehensive model to explain the early fate (viz., fixation or elimination) of gene duplications (*Figure 7*). Notably, this model is compatible with population genetics, involving positive and neutral selective conditions (*Lynch, 2007*). On the one hand, a significant number of duplicates could be fixed by genetic drift only in complex organisms (i.e.

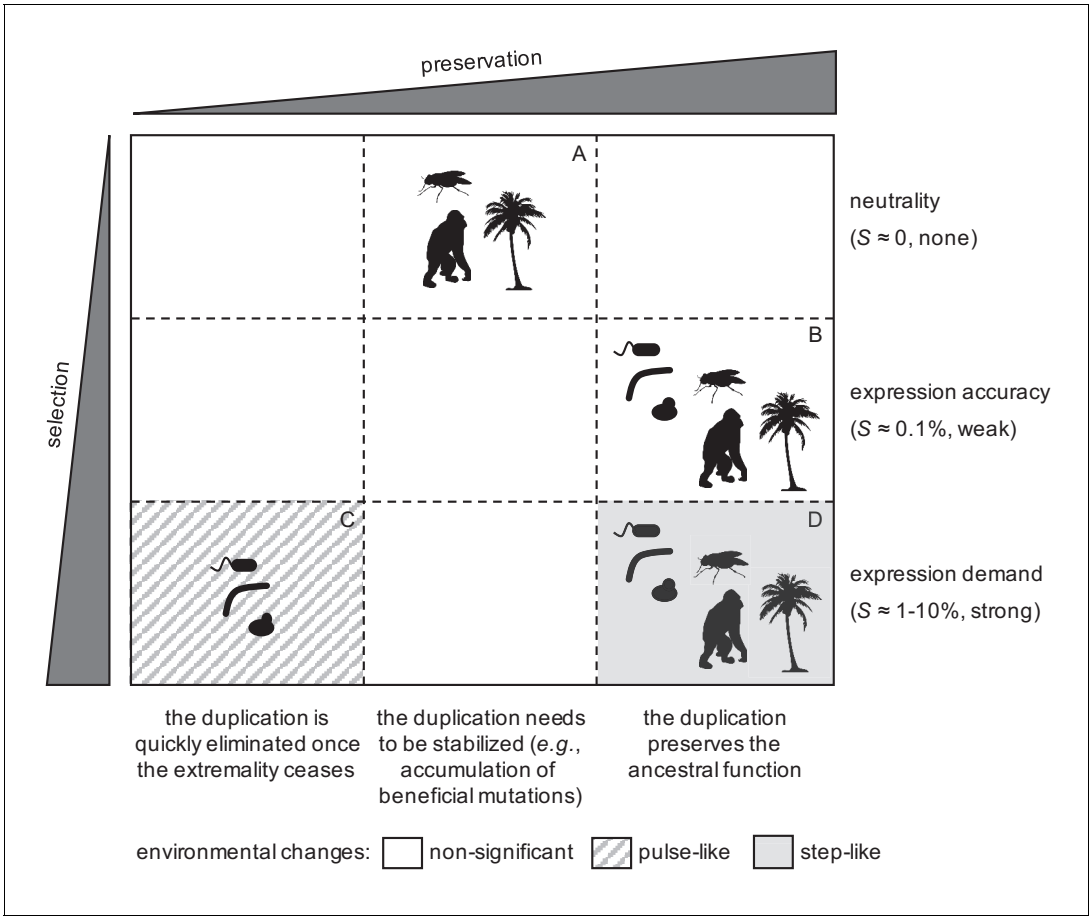

Figure 7. General model to explain the fixation of duplicated genes as a function of the degree of selection in the population and preservation in the genome for long time. Representative silhouettes correspond to bacteria (prokaryotes), yeasts (lower eukaryotes), insects, plants, and mammals (higher eukaryotes).

DOI: https://doi.org/10.7554/eLife.29739.008

higher eukaryotes; sector A in ***Figure 7***). This would be due to their increased ability to allocate additional resources for expression (***Lynch and Marinov, 2015***), and their apparently reduced duplication deletion rate with respect to the inverse of the population size (***Schrider et al., 2013***). However, these fixed duplications would not be stable, due to the formation-deletion balance (***Reams et al., 2010***), and then, for a long-term preservation, they would require the accumulation of beneficial mutations (***Han et al., 2009***), or the relocation of the second copy in the genome to prevent its deletion (***Ranz et al., 2001***). This would lead to late fates of sub- or neo-functionalization (***Force et al., 1999***; ***Conant and Wolfe, 2008***).

On the other hand, positive selection could drive the fixation of duplicates in both complex and simple organisms. When the environmental changes were relatively rapid, only organisms with short generation times (i.e., prokaryotes and lower eukaryotes) could fix duplications (sector C in ***Figure 7***; ***Riehle et al., 2001***). However, such duplications would be quickly eliminated from the population afterwards (once the environment changed again), as the genome rearrangement rates are orders of magnitude higher than the per base mutation rates (***Reams et al., 2010***). By contrast, when a given environmental change were prolonged, any organism, irrespective of its generation time, could fix duplications (sector D in ***Figure 7***; ***Emerson et al., 2008***). In this case, they would be under strong positive selection, and, consequently, they would be preserved for long time. Furthermore, all organisms could fix duplications by producing more accurate responses (sector B in ***Figure 7***), without the need of significant environmental changes; provided the gene of interest were noisily expressed (***Elowitz et al., 2002***; ***Raser et al., 2004***), and the duplication deletion rate were lower

than the weak selective advantage. In the very long term, these weak selective conditions could also allow the exploration of novel functions, as they ensure the preservation of duplicates, without invoking fortuitous exploration in the ancestral state (**Bergthorsson et al., 2007**), and with amplification when the advantage provided by the narrowed novel function were higher than the advantage by noise reduction.

Discussion

The inherently stochastic nature of gene expression is certainly an evolutionary driver when it is linked to cell fitness to dictate the selection of particular genetic architectures (**Batada and Hurst, 2007**; **Maamar et al., 2007**). Our results demonstrate that gene duplication can be positively selected as an architecture that allows enhancing information transfer in genetic networks (i.e. mitigation of expression errors; **Rodrigo and Poyatos, 2016**). Accordingly, the genetic robustness indeed observed upon the accumulation of genetic redundancy (**Keane et al., 2014**) would be more a consequence than a selective trait (**Kafri et al., 2006**). Certainly, by aggregating the responses of two genes, intrinsic fluctuations can be mitigated, but not fluctuations of extrinsic nature. This way, duplication would be more favorable in scenarios in which intrinsic noise is preponderant. The balance between intrinsic and extrinsic noise depends on the particular environmental conditions and the regulatory structures in which the gene in embedded. Intrinsic noise can be significant when the medium is rich in nutrients, the expression levels are low, and no further regulations affect the gene (**Swain et al., 2002**). For example, competence in *Bacillus* is mainly governed by intrinsic noise (**Maamar et al., 2007**). To follow our model, noise has to mainly impinge the regulation of the system, that is, disturb the link between the signal molecule and gene expression (**Blake et al., 2006**). Moreover, our results highlight that a population genetic model with the mean selection coefficient is enough to explain the complex, stochastic evolutionary dynamics of duplication fixation. Of note, the reported intrinsic adaptive value, which cannot be captured by sequence analyses, was derived from basic mathematical models of gene regulation and cell fitness (**Dekel and Alon, 2005**).

Notably, we anticipated a series of testable results by following our theory of error buffering upon duplication. First, the gene expression level is indicative of the fixation path. The theory requires that gene expression is roughly maintained (i.e. gene dosage sharing, duplicates vs.. singletons), with the aim of minimizing deleterious fitness effects. This would hold for several fixed duplicates in different organisms (**Qian et al., 2010**; **Gout and Lynch, 2015**; **Cardoso-Moreira et al., 2016**; **Lan and Pritchard, 2016**), although most of the formed duplicates would be under strong purifying selection due to the cost of over-expression, as already proposed (**Lynch and Conery, 2000**). By contrast, those fixed duplicates showing increased gene expression levels would reflect the effect of genetic drift (**Lynch and Conery, 2003**) or positive selection for dosage after prolonged environmental changes (e.g. the case of flies; **Emerson et al., 2008**; **Cardoso-Moreira et al., 2016**).

Second, noisy genes are expected to be more duplicable (e.g. as it seems to happen in yeast; **Lehner (2010)**; **Dong et al., 2011**) when noise has deleterious fitness effects. Indeed, the gain experimented by the system upon duplication is greater when gene expression inaccuracies are significant (**Rodrigo and Poyatos, 2016**). This would explain the TATA box enrichment in the *cis*-regulatory regions of duplicated genes, as these genetic motifs are associated to high plasticity (i.e. high sensitivity to multiple environmental changes) and high gene expression noise by inducing transcriptional bursts (**Blake et al., 2006**; **Lehner, 2010**). Note that if noise were beneficial (e.g. as a survival strategy in fluctuating environments; **Acar et al., 2008**), duplication would not be favored. Moreover, we might argue that essential genes would be less duplicable (**He and Zhang, 2006**) as a consequence of their reduced gene expression noise (**Batada and Hurst, 2007**). Genes under the control of regulatory structures that buffer noise (e.g. negative feedbacks) would not be duplicable either (**Warnecke et al., 2009**). However, this consideration should be taken with caution, as genes not essential a priori could be duplicated and then, upon fixation, accumulate beneficial mutations (**Han et al., 2009**) to ensure preservation for long time, resulting a posteriori in essential genes due to functional diversification (as it seems in the case of mammals; **Makino et al., 2009**).

Third, the local genetic context would be highly determinant of the fixation of a duplicate (**Reams et al., 2012**), explaining why some genes are more duplicable than others in scenarios of apparent neutrality (hot spots; **Perry et al., 2006**). Moreover, duplicates would be much shorter lived in prokaryotes than in eukaryotes (**Lynch and Conery, 2003**), due to the differences of orders

of magnitude in the duplication deletion rates. After all, the precise experimental determination of the molecular rates of gene copy number variation would unveil to what extent natural selection has actually rivaled random genetic drift to shape complexity along the course of life history (**Rodrigo, 2017**).

These predictions involve, nevertheless, some limitations. On the one hand, due to a simplified mathematical model not considering the many molecular/genetic attributes that impinge implicitly on gene expression, such as promoter sequence-dependent noise levels (**Metzger et al., 2015**), response coupling due to genetic proximity (**Becskei et al., 2005**), or recursive fitness-expression dependence (**Klumpp et al., 2009**). On the other hand, due to the difficulty to provide direct empirical evidence supporting the fixation of duplicates by reducing intrinsic noise. In this regard, we expect to carry out in the future an experimental approach (**Dekel and Alon, 2005**; **Keane et al., 2014**) complementary to this theoretical study. Despite these edges, our results complete a mechanistic model in which duplicates are fixed either by genetic drift (no selection) or by gene dosage (strong selection) with the addition of a new principle, viz., reduction of gene expression inaccuracies upon duplication can result in a weak selective advantage.

Materials and methods

Fitness function

The *lac* operon of *E. coli* (**Jacob and Monod, 1961**) was considered as a biological model system from which to apply a mathematical framework, and cell growth rate was taken as a metric of fitness (W; **Elena and Lenski, 2003**). In this particular case, the benefit function reads $B = a{\cdot}y{\cdot}x \,/\, (k + x)$, where a accounts for the increase in growth rate due to lactose utilization (x denotes its concentration; y denotes the normalized LacZ expression), and k is the Michaelis-Menten constant. In addition, the cost function reads $C = b{\cdot}y \,/\, (h - y)$, where b accounts for the decrease in growth rate due to LacZ expression, and h for the maximal resources available in the cell (**Dekel and Alon, 2005**). Thus, the fitness function reads $W = W_0{\cdot}(1 + B - C)$, where W_0 is the cell growth rate in absence of lactose ($x = 0$). Note that this model underestimates the adaptive ability of the bacterium by not considering the effect of LacY. Moreover, the normalized LacZ expression, in the deterministic regime, is given by $y = x^n/(x_0{}^n + x^n)$, where x_0 is the lactose regulatory constant, and n the Hill coefficient (accounting for the regulatory sensitivity). In this model, LacZ is not expressed in absence of lactose. If $y > h$, we assumed $W = 0$. All parameter values were experimentally fitted, resulting in $W_0 = 1\ \mathrm{h}^{-1}$, $a = 0.17$, $k = 0.40$ mM, $b = 0.036$, $h = 1.80$, $x_0 = 0.13$ mM, and $n = 4$ (**Dekel and Alon, 2005**). The optimal LacZ expression (y_{opt}) was obtained by imposing $dW/dy = 0$, resulting in $y_{opt} = h - [b{\cdot}h{\cdot}(k + x) \,/\, (a{\cdot}x)]^{1/2}$.

Stochastic gene expression

The normalized LacZ expression in presence of molecular noise was modeled, in steady state, as $y = y_{max}{\cdot}(x{\cdot}z_1{\cdot}z_0)^n \,/\, [x_0{}^n + (x{\cdot}z_1{\cdot}z_0)^n]$, where y_{max} is the maximal expression level (in general, $y_{max} = 1$), and z_1 and z_0 random variables accounting for intrinsic and extrinsic noise sources, respectively. Here, they were log-normally distributed [with mean 0 for both $\log(z_1)$ and $\log(z_0)$, and standard deviation η_{in} for $\log(z_1)$ and η_{ex} for $\log(z_0)$]. This accounts for the noisy de-repression of the promoter and subsequent expression due to lactose. Note that whilst LacZ can show a bistable expression pattern with non-metabolizable synthetic compounds (**Ozbudak et al., 2004**), its expression is monostable with lactose (**van Hoek and Hogeweg, 2006**). For simplicity, the transient LacZ expression was overlooked, and the noise levels were considered constant during a cell cycle. The median response of a population is denoted by $\langle y \rangle$.

Typical values characterizing the magnitude of the stochastic fluctuations (η_{in} and η_{ex}) range between 0.1 and 0.5. They lead to values of gene expression noise (understood as the coefficient of variation) between 0.26 and 0.72 (in the case of $\eta_{in} = \eta_{ex}$ and $x = x_0$), in agreement with experimental reports (**Elowitz et al., 2002**).

Gene duplication

The combined expression of two genes coding for LacZ in presence of molecular noise was modeled as $y = y_{max,1}{\cdot}(x{\cdot}z_1{\cdot}z_0)^n \,/\, [x_0{}^n + (x{\cdot}z_1{\cdot}z_0)^n] + y_{max,2}{\cdot}(x{\cdot}z_2{\cdot}z_0)^n \,/\, [x_0{}^n + (x{\cdot}z_2{\cdot}z_0)^n]$, where z_2 is a random

variable accounting for intrinsic noise on the second copy, with the same distribution as for z_1 (z_1 and z_0 as before). Note that whilst extrinsic fluctuations (z_0) are common, intrinsic fluctuations (z_1 and z_2) are independent for each gene copy (**Elowitz et al., 2002**). Moreover, the expression levels of the duplicates with respect to the singletons can be adjusted with the values of $y_{max,1}$ and $y_{max,2}$, with $y_{max,1} = y_{max,2} = 0.5$ for equal total expression, and $y_{max,1} = y_{max,2} = 1$ for double expression.

In addition, the bacterial model was modified to simulate the effect of gene duplication in organisms of different complexity. For that, the parameter h in the cost function was set in terms of the genome size (G, in Mbp of haploid genome), simply as $h \approx 0.36 \cdot G$ (e.g. $G \approx 5$ for *E. coli*, or $G \approx 3000$ for *H. sapiens*), assuming that complex organisms have more resources to accommodate new gene expressions (**Lynch and Marinov, 2015**). The effective population size (here denoted by $\langle N \rangle$), determinant of the fixation of new genotypes, was also set in terms of G, resulting in $\langle N \rangle \approx 3 \cdot 10^9 / G^{1.44}$; an equation roughly inferred from previously reported estimates (**Lynch and Conery, 2003**).

Information transfer

Mutual information (I) was used as a metric to characterize information transfer by considering the system as a communication channel between the environmental molecule (lactose) and the functional protein (enzyme, LacZ) resulting from gene expression. I was calculated as previously done (**Rodrigo and Poyatos, 2016**), between $\log(x)$ and y. To model the variation of lactose, a random variable log-normally distributed was considered [with mean 0 and standard deviation 1, otherwise specified, for $\log(x/x_0)$]. The median lactose dose is denoted by $\langle x \rangle$, and the fluctuation amplitude, denoted by Δx, corresponds to the standard deviation of $\log(x)$. To compare statistically two I values, we followed the approximation proposed by **Cellucci et al. (2005)** to obtain an equivalent correlation coefficient, and then the Fisher's r-to-z transformation.

Genotype-phenotype map

Here, the LacZ expression defines the phenotype of the cell (i.e. its metabolic capacity), and for the wild-type genotype it is lactose dependent through the LacI regulation (**Jacob and Monod, 1961**). Because differences in fitness are very small, the normalized expression (y) was assumed independent of it (**Klumpp et al., 2009**). Potential beneficial mutations are those that change the *lac* promoter activity (the *cis*-regulatory regulatory region of LacZ, of about 10^2 bp). According to an analysis of a large library of mutants (**Kinney et al., 2010**) resulting in a linear model of categorical variables (**Otwinowski and Nemenman, 2013**), the distribution of maximal LacZ expression upon single-point mutations was inferred. For simplicity, no epistatic interactions were taken into account, although they could matter. Mutations were also assumed to affect only the mean expression level and not the noise, even though this latter might happen (**Metzger et al., 2015**).

In silico evolution

A medium with maximal capacity for $N = 10^5$ cells was considered, and serial dilution passages were simulated (**Elena and Lenski, 2003**), with a dilution factor of $D = 100$ (in terms of volume, with deterministic dominance). The dilution period was set to 1 d. Lactose also varied with the same period. The doubling time of a given cell was $1/W$, with W calculated from the stochastic LacZ expression. In case of no saturation, the cell volume increased as $2^{W \cdot t}$, where t is the time in h. Because doublings occur in about 1 h, the number of generations per passage is bounded to $\log_2(D) = 6.64$. Two genotypes were put in competition: one with a single copy of LacZ, the other with two copies. No mutations were allowed to occur.

Population genetics

In scenarios of competition between two subpopulations (i.e. two different genotypes), the ratio between them (r) reads $r = r_0 \cdot 2^{S \cdot t}$, where r_0 is the initial ratio, S the selection coefficient, and t the time measured in generations (**Hegreness et al., 2006**). By setting W and W' the fitness values of each genotype (with $W' > W$), the selection coefficient is calculated as $S = W'/W - 1$. When fitness changes over time, the mean selection coefficient ($\langle S \rangle$) is used. The frequency of the genotype with advantage in the population is $f = 1/(1 + 1/r)$. The dynamics of a punctual beneficial mutant appeared in an evolutionary experiment of serial dilution passages, with maximal population size N and dilution factor D, is given by $r = 2^{S \cdot t} / \langle N \rangle$, where $\langle N \rangle = N / D^{1/2}$ is the geometric mean

population size (also considered the effective population size; **Lewontin and Cohen, 1969**). The fixation probability is $P_{fix} = 2S$, and the characteristic fixation time $t_{fix} = \log_2(\langle N \rangle^2)/S$. Note that the time for 50% invasion of the population is $t_{half\text{-}fix} = \log_2(\langle N \rangle)/S = t_{fix}/2$. However, we have $P_{fix} = 1/\langle N \rangle$ and $t_{fix} = 2\langle N \rangle$ for a neutral mutant (**Kimura, 1983**).

By contrast, if multiple beneficial mutants are recurrently produced at rate μ_b, the dynamics is given by $r = \mu_b \cdot N \cdot 2^{S \cdot t} / [S \cdot \log(D) \cdot \langle N \rangle] \approx \mu_b \cdot 2^{S \cdot t} / S$, as in each passage $\mu_b \cdot N$ different mutants are generated (valid for $\mu_b \cdot N > 1$; **Desai et al., 2007**). Because mutants are now recurrent, $P_{fix} = 1$, and the characteristic fixation time reads $t_{fix} = \log_2[\langle N \rangle \cdot S / \mu_b]/S$. When m different mutations accumulate successively, $t_{fix} \approx t_{fix}(m) + t_{half\text{-}fix}(m\text{-}1) + \ldots + t_{half\text{-}fix}(1)$, that is, a subsequent mutation can start its fixation when the preceding mutation has invaded the 50% of the population (**Lang et al., 2013**). If $\mu_b \cdot N << 1$, the system can be treated as in the case of a punctual beneficial mutation, and the dynamics can be written as $r = 2^{S \cdot (t - T)} / \langle N \rangle$, with a delay of $T = \log_2(D) / (\mu_b \cdot N)$, the mean number of generations required to produce a mutant, and $P_{fix} = 2S$.

Moreover, in case of gene duplication, if multiple beneficial mutants are recurrently produced at rate μ_c, and deleted at rate μ_d, the dynamics is given by $r \approx \mu_c \cdot 2^{S' \cdot t} / S'$, with $S' = S - \mu_d$ as an effective selection coefficient (valid for $\mu_c \cdot N > 1$, and $S > \mu_d$). Again, if $\mu_c \cdot N << 1$, the system can be treated as in the case of a punctual beneficial mutation, with $P_{fix} = 2S'$. If $S << \mu_d$, the stationary solution can be approached by $r \approx \mu_c / \mu_d$ for effectively neutral mutations, or by $r \approx \mu_c / (\mu_d - S)$ for deleterious mutations.

Genetic diversity

The per base mutation rate of *E. coli* is $\mu = 10^{-10}$ mut./bp/gen. (**Lee et al., 2012**). Cultures of this bacterium may reach population sizes up to $N = 10^9$ cells ($\langle N \rangle = 2 \cdot 10^8$). This means, on average, 0.02 ($= \mu \cdot \langle N \rangle$) mutants of a given base pair in the population. The number of base pairs, mainly in the *cis*-regulatory regulatory region, whose mutation reduces in half the expression of a gene of interest can be estimated in 10 (based on data for *lacZ*). Thus, $\mu_b = 10 \cdot \mu$, which means 0.2 ($= \mu_b \cdot \langle N \rangle$) mutant of this type in the population on average. This frequency may even be higher if we not only consider the mutations in the *lac* promoter, but also the mutations in the coding region, or affecting the activity of its regulators (e.g. CRP; **Kinney et al., 2010**).

In addition, for the *lacZ* gene, its duplication formation rate is of $\mu_c = 3 \cdot 10^{-4}$ dup./gene/gen., and its duplication deletion rate of $\mu_d = 4.1 \cdot 10^{-4}$–$4.4 \cdot 10^{-2}$ -/gene/gen. (**Reams et al., 2010**; **Reams et al., 2012**). In absence of lactose, duplications are neutral ($S = 0$), which means, on average, a duplication frequency in the population of 0.68–42% [$= \mu_c / (\mu_c + \mu_d)$]. By contrast, in presence of lactose, duplications are deleterious ($S \approx -28\%$), and then the average duplication frequency is of 0.09–0.11% [$= \mu_c / (\mu_c + \mu_d - S)$]. Note that the deletion rates are difficult to estimate experimentally, as this requires starting from a genotype with new-born (mostly unstable) duplications, albeit they are essential to properly understand the fixation process.

Availability of resources

A Matlab code to model gene expression (y) and cell fitness (W) and a C++ code to perform the in silico evolution (as described above) are freely available for download at https://sourceforge.net/projects/rodrigo-duplications/files (**Rodrigo, 2017b**). A copy is archived at https://github.com/elifesciences-publications/rodrigo-duplications.

Acknowledgements

This work was supported by grants BFU2015-66894-P (to GR) and BFU2015-66073-P (to MAF) from the Spanish Ministry of Economy (MINECO/FEDER), and also by grant GVA/2016/079 from the Generalitat Valenciana (to GR).

Additional information

Funding

Funder	Grant reference number	Author
Ministerio de Economía y Competitividad	BFU2015-66894-P	Guillermo Rodrigo
Ministerio de Economía y Competitividad	BFU2015-66073-P	Mario A Fares
Generalitat Valenciana	GVA/2016/079	Guillermo Rodrigo

The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

Author contributions

Guillermo Rodrigo, Conceptualization, Formal analysis, Validation, Methodology, Writing—original draft; Mario A Fares, Validation, Writing—original draft

Author ORCIDs

Guillermo Rodrigo https://orcid.org/0000-0002-1871-9617

Decision letter and Author response

Decision letter https://doi.org/10.7554/eLife.29739.011

Author response https://doi.org/10.7554/eLife.29739.012

Additional files

Supplementary files

- Transparent reporting form

DOI: https://doi.org/10.7554/eLife.29739.009

References

Acar M, Mettetal JT, van Oudenaarden A. 2008. Stochastic switching as a survival strategy in fluctuating environments. *Nature Genetics* **40**:471–475. DOI: https://doi.org/10.1038/ng.110, PMID: 18362885

Afroz T, Biliouris K, Kaznessis Y, Beisel CL. 2014. Bacterial sugar utilization gives rise to distinct single-cell behaviours. *Molecular Microbiology* **93**:n/a–1103. DOI: https://doi.org/10.1111/mmi.12695, PMID: 24976172

Balázsi G, van Oudenaarden A, Collins JJ. 2011. Cellular decision making and biological noise: from microbes to mammals. *Cell* **144**:910–925. DOI: https://doi.org/10.1016/j.cell.2011.01.030, PMID: 21414483

Batada NN, Hurst LD. 2007. Evolution of chromosome organization driven by selection for reduced gene expression noise. *Nature Genetics* **39**:945–949. DOI: https://doi.org/10.1038/ng2071, PMID: 17660811

Becskei A, Kaufmann BB, van Oudenaarden A. 2005. Contributions of low molecule number and chromosomal positioning to stochastic gene expression. *Nature Genetics* **37**:937–944. DOI: https://doi.org/10.1038/ng1616, PMID: 16086016

Bergthorsson U, Andersson DI, Roth JR. 2007. Ohno's dilemma: evolution of new genes under continuous selection. *PNAS* **104**:17004–17009. DOI: https://doi.org/10.1073/pnas.0707158104, PMID: 17942681

Birchler JA, Riddle NC, Auger DL, Veitia RA. 2005. Dosage balance in gene regulation: biological implications. *Trends in Genetics* **21**:219–226. DOI: https://doi.org/10.1016/j.tig.2005.02.010, PMID: 15797617

Blake WJ, Balázsi G, Kohanski MA, Isaacs FJ, Murphy KF, Kuang Y, Cantor CR, Walt DR, Collins JJ. 2006. Phenotypic consequences of promoter-mediated transcriptional noise. *Molecular Cell* **24**:853–865. DOI: https://doi.org/10.1016/j.molcel.2006.11.003, PMID: 17189188

Cardoso-Moreira M, Arguello JR, Gottipati S, Harshman LG, Grenier JK, Clark AG. 2016. Evidence for the fixation of gene duplications by positive selection in Drosophila. *Genome Research* **26**:787–798. DOI: https://doi.org/10.1101/gr.199323.115, PMID: 27197209

Carey LB, van Dijk D, Sloot PM, Kaandorp JA, Segal E. 2013. Promoter sequence determines the relationship between expression level and noise. *PLoS Biology* **11**:e1001528. DOI: https://doi.org/10.1371/journal.pbio.1001528, PMID: 23565060

Casanueva MO, Burga A, Lehner B. 2012. Fitness trade-offs and environmentally induced mutation buffering in isogenic C. elegans. *Science* **335**:82–85. DOI: https://doi.org/10.1126/science.1213491, PMID: 22174126

Cellucci CJ, Albano AM, Rapp PE. 2005. Statistical validation of mutual information calculations: comparison of alternative numerical algorithms. *Physical Review E* **71**:066208. DOI: https://doi.org/10.1103/PhysRevE.71.066208, PMID: 16089850

Clark AG. 1994. Invasion and maintenance of a gene duplication. *PNAS* **91**:2950–2954. DOI: https://doi.org/10.1073/pnas.91.8.2950, PMID: 8159686

Conant GC, Wolfe KH. 2008. Turning a hobby into a job: how duplicated genes find new functions. *Nature Reviews Genetics* **9**:938–950. DOI: https://doi.org/10.1038/nrg2482, PMID: 19015656

Dekel E, Alon U. 2005. Optimality and evolutionary tuning of the expression level of a protein. *Nature* **436**:588–592. DOI: https://doi.org/10.1038/nature03842, PMID: 16049495

DeLuna A, Vetsigian K, Shoresh N, Hegreness M, Colón-González M, Chao S, Kishony R. 2008. Exposing the fitness contribution of duplicated genes. *Nature Genetics* **40**:676–681. DOI: https://doi.org/10.1038/ng.123, PMID: 18408719

Des Marais DL, Rausher MD. 2008. Escape from adaptive conflict after duplication in an anthocyanin pathway gene. *Nature* **454**:762–765. DOI: https://doi.org/10.1038/nature07092, PMID: 18594508

Desai MM, Fisher DS, Murray AW. 2007. The speed of evolution and maintenance of variation in asexual populations. *Current Biology* **17**:385–394. DOI: https://doi.org/10.1016/j.cub.2007.01.072, PMID: 17331728

Dong D, Yuan Z, Zhang Z. 2011. Evidences for increased expression variation of duplicate genes in budding yeast: from cis- to trans-regulation effects. *Nucleic Acids Research* **39**:837–847. DOI: https://doi.org/10.1093/nar/gkq874, PMID: 20935054

Eames M, Kortemme T. 2012. Cost-benefit tradeoffs in engineered lac operons. *Science* **336**:911–915. DOI: https://doi.org/10.1126/science.1219083, PMID: 22605776

Elena SF, Lenski RE. 2003. Evolution experiments with microorganisms: the dynamics and genetic bases of adaptation. *Nature Reviews Genetics* **4**:457–469. DOI: https://doi.org/10.1038/nrg1088, PMID: 12776215

Elowitz MB, Levine AJ, Siggia ED, Swain PS. 2002. Stochastic gene expression in a single cell. *Science* **297**:1183–1186. DOI: https://doi.org/10.1126/science.1070919, PMID: 12183631

Emerson JJ, Cardoso-Moreira M, Borevitz JO, Long M. 2008. Natural selection shapes genome-wide patterns of copy-number polymorphism in Drosophila melanogaster. *Science* **320**:1629–1631. DOI: https://doi.org/10.1126/science.1158078, PMID: 18535209

Fischer I, Dainat J, Ranwez V, Glémin S, Dufayard JF, Chantret N. 2014. Impact of recurrent gene duplication on adaptation of plant genomes. *BMC Plant Biology* **14**:151. DOI: https://doi.org/10.1186/1471-2229-14-151, PMID: 24884640

Force A, Lynch M, Pickett FB, Amores A, Yan YL, Postlethwait J. 1999. Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* **151**:1531–1545. PMID: 10101175

Gammaitoni L. 1995. Stochastic resonance and the dithering effect in threshold physical systems. *Physical Review E* **52**:4691–4698. DOI: https://doi.org/10.1103/PhysRevE.52.4691, PMID: 9963964

Gonzalez E, Kulkarni H, Bolivar H, Mangano A, Sanchez R, Catano G, Nibbs RJ, Freedman BI, Quinones MP, Bamshad MJ, Murthy KK, Rovin BH, Bradley W, Clark RA, Anderson SA, O'connell RJ, Agan BK, Ahuja SS, Bologna R, Sen L, et al. 2005. The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility. *Science* **307**:1434–1440. DOI: https://doi.org/10.1126/science.1101160, PMID: 15637236

Gout JF, Lynch M. 2015. Maintenance and loss of duplicated genes by dosage subfunctionalization. *Molecular Biology and Evolution* **32**:2141–2148. DOI: https://doi.org/10.1093/molbev/msv095, PMID: 25908670

Han MV, Demuth JP, McGrath CL, Casola C, Hahn MW. 2009. Adaptive evolution of young gene duplicates in mammals. *Genome Research* **19**:859–867. DOI: https://doi.org/10.1101/gr.085951.108, PMID: 19411603

Hansen AS, O'Shea EK, O'Shea EK. 2015. Limits on information transduction through amplitude and frequency regulation of transcription factor activity. *eLife* **4**:e06559. DOI: https://doi.org/10.7554/eLife.06559, PMID: 25985085

Hastings PJ, Lupski JR, Rosenberg SM, Ira G. 2009. Mechanisms of change in gene copy number. *Nature Reviews Genetics* **10**:551–564. DOI: https://doi.org/10.1038/nrg2593, PMID: 19597530

He X, Zhang J. 2006. Higher duplicability of less important genes in yeast genomes. *Molecular Biology and Evolution* **23**:144–151. DOI: https://doi.org/10.1093/molbev/msj015, PMID: 16151181

Hegreness M, Shoresh N, Hartl D, Kishony R. 2006. An equivalence principle for the incorporation of favorable mutations in asexual populations. *Science* **311**:1615–1617. DOI: https://doi.org/10.1126/science.1122469, PMID: 16543462

Hittinger CT, Carroll SB. 2007. Gene duplication and the adaptive evolution of a classic genetic switch. *Nature* **449**:677–681. DOI: https://doi.org/10.1038/nature06151, PMID: 17928853

Innan H, Kondrashov F. 2010. The evolution of gene duplications: classifying and distinguishing between models. *Nature Reviews Genetics* **11**:4–108. DOI: https://doi.org/10.1038/nrg2689, PMID: 20051986

Jacob F, Monod J. 1961. Genetic regulatory mechanisms in the synthesis of proteins. *Journal of Molecular Biology* **3**:318–356. DOI: https://doi.org/10.1016/S0022-2836(61)80072-7, PMID: 13718526

Kafri R, Levy M, Pilpel Y. 2006. The regulatory utilization of genetic redundancy through responsive backup circuits. *PNAS* **103**:11653–11658. DOI: https://doi.org/10.1073/pnas.0604883103, PMID: 16861297

Keane OM, Toft C, Carretero-Paulet L, Jones GW, Fares MA. 2014. Preservation of genetic and regulatory robustness in ancient gene duplicates of Saccharomyces cerevisiae. *Genome Research* **24**:1830–1841. DOI: https://doi.org/10.1101/gr.176792.114, PMID: 25149527

Kimura M. 1983. *The Neutral Theory of Molecular Evolution*. New York: Cambridge University Press . DOI: https://doi.org/10.1017/CBO9780511623486

King OD, Masel J. 2007. The evolution of bet-hedging adaptations to rare scenarios. *Theoretical Population Biology* **72**:560–575. DOI: https://doi.org/10.1016/j.tpb.2007.08.006, PMID: 17915273

Kinney JB, Murugan A, Callan CG, Cox EC. 2010. Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence. *PNAS* **107**:9158–9163. DOI: https://doi.org/10.1073/pnas.1004290107, PMID: 20439748

Klumpp S, Zhang Z, Hwa T. 2009. Growth rate-dependent global effects on gene expression in bacteria. *Cell* **139**:1366–1375. DOI: https://doi.org/10.1016/j.cell.2009.12.001, PMID: 20064380

Kuhlman T, Zhang Z, Saier MH, Hwa T. 2007. Combinatorial transcriptional control of the lactose operon of Escherichia coli. *PNAS* **104**:6043–6048. DOI: https://doi.org/10.1073/pnas.0606717104, PMID: 17376875

Lan X, Pritchard JK. 2016. Coregulation of tandem duplicate genes slows evolution of subfunctionalization in mammals. *Science* **352**:1009–1013. DOI: https://doi.org/10.1126/science.aad8411, PMID: 27199432

Lang GI, Rice DP, Hickman MJ, Sodergren E, Weinstock GM, Botstein D, Desai MM. 2013. Pervasive genetic hitchhiking and clonal interference in forty evolving yeast populations. *Nature* **500**:571–574. DOI: https://doi.org/10.1038/nature12344, PMID: 23873039

Lee H, Popodi E, Tang H, Foster PL. 2012. Rate and molecular spectrum of spontaneous mutations in the bacterium Escherichia coli as determined by whole-genome sequencing. *PNAS* **109**:E2774–E2783. DOI: https://doi.org/10.1073/pnas.1210309109, PMID: 22991466

Lehner B. 2010. Conflict between noise and plasticity in yeast. *PLoS Genetics* **6**:e1001185. DOI: https://doi.org/10.1371/journal.pgen.1001185, PMID: 21079670

Lewontin RC, Cohen D. 1969. On population growth in a randomly varying environment. *PNAS* **62**:1056–1060. DOI: https://doi.org/10.1073/pnas.62.4.1056, PMID: 5256406

Lynch M, Ackerman MS, Gout JF, Long H, Sung W, Thomas WK, Foster PL. 2016. Genetic drift, selection and the evolution of the mutation rate. *Nature Reviews Genetics* **17**:704–714. DOI: https://doi.org/10.1038/nrg.2016.104, PMID: 27739533

Lynch M, Conery JS. 2000. The evolutionary fate and consequences of duplicate genes. *Science* **290**:1151–1155. DOI: https://doi.org/10.1126/science.290.5494.1151, PMID: 11073452

Lynch M, Conery JS. 2003. The origins of genome complexity. *Science* **302**:1401–1404. DOI: https://doi.org/10.1126/science.1089370, PMID: 14631042

Lynch M, Marinov GK. 2015. The bioenergetic costs of a gene. *PNAS* **112**:15690–15695. DOI: https://doi.org/10.1073/pnas.1514974112

Lynch M, O'Hely M, Walsh B, Force A. 2001. The probability of preservation of a newly arisen gene duplicate. *Genetics* **159**:1789–1804. PMID: 11779815

Lynch M. 2007. The frailty of adaptive hypotheses for the origins of organismal complexity. *PNAS* **104**:8597–8604. DOI: https://doi.org/10.1073/pnas.0702207104, PMID: 17494740

Maamar H, Raj A, Dubnau D. 2007. Noise in gene expression determines cell fate in Bacillus subtilis. *Science* **317**:526–529. DOI: https://doi.org/10.1126/science.1140818, PMID: 17569828

Makino T, Hokamp K, McLysaght A. 2009. The complex relationship of gene duplication and essentiality. *Trends in Genetics* **25**:152–155. DOI: https://doi.org/10.1016/j.tig.2009.03.001, PMID: 19285746

Metzger BP, Yuan DC, Gruber JD, Duveau F, Wittkopp PJ. 2015. Selection on noise constrains variation in a eukaryotic promoter. *Nature* **521**:344–347. DOI: https://doi.org/10.1038/nature14244, PMID: 25778704

Nowak MA, Boerlijst MC, Cooke J, Smith JM. 1997. Evolution of genetic redundancy. *Nature* **388**:167–171. DOI: https://doi.org/10.1038/40618, PMID: 9217155

Ohno S. 1970. *Evolution by Gene Duplication*. New York: Springer Verlag . DOI: https://doi.org/10.1007/978-3-642-86659-3

Otwinowski J, Nemenman I. 2013. Genotype to phenotype mapping and the fitness landscape of the E. coli lac promoter. *PLoS One* **8**:e61570. DOI: https://doi.org/10.1371/journal.pone.0061570, PMID: 23650500

Ozbudak EM, Thattai M, Lim HN, Shraiman BI, Van Oudenaarden A. 2004. Multistability in the lactose utilization network of Escherichia coli. *Nature* **427**:737–740. DOI: https://doi.org/10.1038/nature02298, PMID: 14973486

Papp B, Pál C, Hurst LD. 2003. Dosage sensitivity and the evolution of gene families in yeast. *Nature* **424**:194–197. DOI: https://doi.org/10.1038/nature01771, PMID: 12853957

Perry GH, Tchinda J, McGrath SD, Zhang J, Picker SR, Cáceres AM, Iafrate AJ, Tyler-Smith C, Scherer SW, Eichler EE, Stone AC, Lee C. 2006. Hotspots for copy number variation in chimpanzees and humans. *PNAS* **103**:8006–8011. DOI: https://doi.org/10.1073/pnas.0602318103, PMID: 16702545

Pool JE. 2015. The mosaic ancestry of the drosophila genetic reference panel and the D. melanogaster reference genome reveals a network of epistatic fitness interactions. *Molecular Biology and Evolution* **32**:3236–3251. DOI: https://doi.org/10.1093/molbev/msv194, PMID: 26354524

Price MN, Arkin AP. 2016. A theoretical lower bound for selection on the expression levels of proteins. *Genome Biology and Evolution* **8**:1917–1928. DOI: https://doi.org/10.1093/gbe/evw126, PMID: 27289091

Price MN, Deutschbauer AM, Skerker JM, Wetmore KM, Ruths T, Mar JS, Kuehl JV, Shao W, Arkin AP. 2013. Indirect and suboptimal control of gene expression is widespread in bacteria. *Molecular Systems Biology* **9**:660. DOI: https://doi.org/10.1038/msb.2013.16, PMID: 23591776

Qian W, Liao BY, Chang AY, Zhang J. 2010. Maintenance of duplicate genes and their functional redundancy by reduced expression. *Trends in Genetics* **26**:425–430. DOI: https://doi.org/10.1016/j.tig.2010.07.002, PMID: 20708291

Ranz JM, Casals F, Ruiz A. 2001. How malleable is the eukaryotic genome? Extreme rate of chromosomal rearrangement in the genus Drosophila. *Genome Research* **11**:230–239. DOI: https://doi.org/10.1101/gr.162901, PMID: 11157786

Raser JM, O'Shea EK, O'Shea EK. 2004. Control of stochasticity in eukaryotic gene expression. *Science* **304**: 1811–1814. DOI: https://doi.org/10.1126/science.1098641, PMID: 15166317

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. DOI: https://doi.org/10.1534/genetics.112.142570, PMID: 22865732

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. DOI: https://doi.org/10.1534/genetics.109.111963, PMID: 20083614

Riehle MM, Bennett AF, Long AD. 2001. Genetic architecture of thermal adaptation in Escherichia coli. *PNAS* **98**: 525–530. DOI: https://doi.org/10.1073/pnas.98.2.525, PMID: 11149947

Rodrigo G, Poyatos JF. 2016. Genetic redundancies enhance information transfer in noisy regulatory circuits. *PLoS Computational Biology* **12**:e1005156. DOI: https://doi.org/10.1371/journal.pcbi.1005156, PMID: 27741249

Rodrigo G. 2017. Evolutionary impact of copy number variation rates. *BMC Research Notes* **10**:393. DOI: https://doi.org/10.1186/s13104-017-2741-3, PMID: 28797272

Rodrigo G. 2017b. rodrigo-duplications. *SourceForge*.https://sourceforge.net/projects/rodrigo-duplications/files

Rozen DE, de Visser JA, Gerrish PJ. 2002. Fitness effects of fixed beneficial mutations in microbial populations. *Current Biology* **12**:1040–1045. DOI: https://doi.org/10.1016/S0960-9822(02)00896-5, PMID: 12123580

Sánchez-Moreno I, Nauton L, Théry V, Pinet A, Petit J-L, de Berardinis V, Samland AK, Guérard-Hélaine C, Lemaire M. 2012. FSAB: A new fructose-6-phosphate aldolase from escherichia coli. cloning, over-expression and comparative kinetic characterization with FSAA. *Journal of Molecular Catalysis B: Enzymatic* **84**:9–14. DOI: https://doi.org/10.1016/j.molcatb.2012.02.010

Schrider DR, Houle D, Lynch M, Hahn MW. 2013. Rates and genomic consequences of spontaneous mutational events in Drosophila melanogaster. *Genetics* **194**:937–954. DOI: https://doi.org/10.1534/genetics.113.151670, PMID: 23733788

Stranger BE, Forrest MS, Dunning M, Ingle CE, Beazley C, Thorne N, Redon R, Bird CP, de Grassi A, Lee C, Tyler-Smith C, Carter N, Scherer SW, Tavaré S, Deloukas P, Hurles ME, Dermitzakis ET. 2007. Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science* **315**:848–853. DOI: https://doi.org/10.1126/science.1136678, PMID: 17289997

Swain PS, Elowitz MB, Siggia ED. 2002. Intrinsic and extrinsic contributions to stochasticity in gene expression. *PNAS* **99**:12795–12800. DOI: https://doi.org/10.1073/pnas.162041399, PMID: 12237400

Tang YC, Amon A. 2013. Gene copy-number alterations: a cost-benefit analysis. *Cell* **152**:394–405. DOI: https://doi.org/10.1016/j.cell.2012.11.043, PMID: 23374337

van Hoek MJ, Hogeweg P. 2006. In silico evolved lac operons exhibit bistability for artificial inducers, but not for lactose. *Biophysical Journal* **91**:2833–2843. DOI: https://doi.org/10.1529/biophysj.105.077420, PMID: 16877514

Vivas EI, Liendo A, Dawidowicz K, Istúriz T. 1994. Isolation and characterization of the thermoresistant gluconokinase from Escherichia coli. *Journal of Basic Microbiology* **34**:117–122. DOI: https://doi.org/10.1002/jobm.3620340207, PMID: 8014844

Wagner A. 1999. Redundant gene functions and natural selection. *Journal of Evolutionary Biology* **12**:1–16. DOI: https://doi.org/10.1046/j.1420-9101.1999.00008.x

Wagner A. 2005. Energy constraints on the evolution of gene expression. *Molecular Biology and Evolution* **22**: 1365–1374. DOI: https://doi.org/10.1093/molbev/msi126, PMID: 15758206

Wang Z, Zhang J. 2011. Impact of gene expression noise on organismal fitness and the efficacy of natural selection. *PNAS* **108**:E67–E76. DOI: https://doi.org/10.1073/pnas.1100059108, PMID: 21464323

Warnecke T, Wang GZ, Lercher MJ, Hurst LD. 2009. Does negative auto-regulation increase gene duplicability? *BMC Evolutionary Biology* **9**:193. DOI: https://doi.org/10.1186/1471-2148-9-193, PMID: 19664220

Wong S, Wolfe KH. 2005. Birth of a metabolic gene cluster in yeast by adaptive gene relocation. *Nature Genetics* **37**:777–782. DOI: https://doi.org/10.1038/ng1584, PMID: 15951822

Wray GA. 2007. The evolutionary significance of cis-regulatory mutations. *Nature Reviews Genetics* **8**:206–216. DOI: https://doi.org/10.1038/nrg2063, PMID: 17304246

Zhang J, Rosenberg HF, Nei M. 1998. Positive Darwinian selection after gene duplication in primate ribonuclease genes. *PNAS* **95**:3708–3713. DOI: https://doi.org/10.1073/pnas.95.7.3708, PMID: 9520431