Compensatory Drift
and the Evolutionary Dynamics of Dosage-sensitive Duplicate Genes

Ammon Thompson*, †‡, Harold H. Zakon*, †, Mark Kirkpatrick*

*Department of Integrative Biology, University of Texas, Austin, TX 78712, USA
†Department of Neuroscience, University of Texas, Austin, TX 78712, USA
‡Current address: Department of Evolution and Ecology, University of California, Davis, CA 95616, USA
Short running title: Duplicate Gene Expression Evolution
Key Words or Phrases: Duplication, Expression evolution, Dosage balance, Neofunctionalization, Whole genome duplication

Corresponding author: Ammon Thompson

Phone: 801-318-6628

E-mail: ammthompson@ucdavis.edu

Department of Integrative Biology
University of Texas
Austin, TX 78712 USA

Present address:
Department of Ecology and Evolution
University of California
Davis, CA 95616
Abstract

Dosage balance selection preserves functionally redundant duplicates (paralogs) at the optimum for their combined expression. Here we present a model of the dynamics of duplicate genes co-evolving under dosage balance selection. We call this the “compensatory drift” model. Results show that even when strong dosage balance selection constrains total expression to the optimum, expression of each duplicate can diverge by drift from its original level. The rate of divergence slows as the strength of stabilizing selection, the mutation effect size, and/or the population size increases. We show that dosage balance selection impedes neofunctionalization early after duplication but can later facilitate it. We fit this model to data from sodium channel duplicates in ten families of teleost fish; these include two convergent lineages of electric fish in which one of the duplicates neofunctionalized. Using the model, we estimated the strength of dosage balance selection for these genes. The results indicate that functionally redundant paralogs may still undergo radical functional changes after a prolonged period of compensatory drift.
INTRODUCTION

The fate of duplicate genes is characterized by two extremes: degeneration, and the origin of biological novelty. Early models for the evolutionary dynamics of duplicates suggested that typically one member of a duplicate pair would quickly degenerate into a nonfunctional pseudogene (Haldane 1933; Ohno 1970). More rarely, a duplicate may instead evolve a novel function in a process called neofunctionalization (Muller 1936; Ohno 1970; Ohta 1987). The time scale for either pseudogenization or neofunctionalization is expected to be on the order of a few million years (Lynch and Conery 2000).

Recent research indicates, however, that the evolutionary dynamics for many duplicates are not so simple (Walsh 1995; Force et al. 1999; Papp et al. 2003; Walsh 2003; He and Zhang 2005; Rastogi and Liberles 2005; Scannell and Wolfe 2008; Qian et al. 2010; Kondrashov 2012). Some genes are dosage sensitive, meaning that a change in their copy number alters expression and disrupts the stoichiometric balance of their gene products with those of other genes. Duplicates of dosage sensitive genes will typically only fix in a population if they originate in a whole genome duplication (WGD), where all interacting partners duplicate together. Selection to maintain the stoichiometric relations between the products of duplicate genes, termed “dosage balance selection,” can preserve duplicates as functionally redundant copies for prolonged periods of time (Birchler et al. 2001; Veitia et al. 2002; Papp et al. 2003; Birchler et al. 2005; Aury et al. 2006; Blomme et al. 2006; Stranger et al. 2007; Qian et al. 2008; Edger et al. 2009; Freeling and Thomas 2006; Makino and McLysaght 2010; Konrad et al. 2011; Birchler and Veitia 2012; McGrath et al. 2014a).

Recent data on a pair of sodium channel duplicates in teleost fishes are consistent with the expectations of the dosage balance hypothesis (Thompson et al. 2014). The two duplicates, also called paralogs, have been conserved in muscle cells for over 300 million years since the teleost-specific WGD. In two independent lineages of electric fish, however, only one of the sodium channels is expressed in muscle cells. The other duplicate neofunctionalized and now plays a key role in the electric organ (Novak et al. 2006; Zakon et al. 2006; Arnegard et al. 2010). These convergent neofunctionalization events happened on a very slow time scale, more than 100 million years after duplication (Arnegard et al. 2010; Lavoué et al. 2012; Betancur-R et al. 2013). The phylogenetic context for the evolution of the duplicates is shown in Figure 1.
Thompson et al. (2014) proposed that in the teleost ancestor, the duplicates were preserved after WGD by dosage balance selection. They hypothesized that under this selective constraint, one paralog gradually drifted to lower expression levels while the other compensated by evolving higher expression. Eventually, one paralog contributed so little to its original function that it could be neofunctionalized in the electric organ without major compromise to muscles. This mode of evolution may also explain comparative expression patterns observed in some ciliates (Gout and Lynch 2015) as well as some mammals (Lan and Pritchard 2015). This hypothesis raises theoretical and quantitative issues not previously explored. Can dosage balance selection in fact maintain duplicates for hundreds of millions of years? Will this mode of evolution produce comparative patterns in a phylogeny that are distinct from other models? And how does this evolutionary process impact the likelihood of neofunctionalization?

Here, we develop a model for evolution of paralog expression under dosage balance selection. It envisions a process, which we call “compensatory drift”, in which paralogs diverge by weakly selected mutations that fix largely by drift. The model shows how key genetic parameters determine the time scale over which duplicates are maintained before one is lost or neofunctionalizes. The evolutionary dynamics are determined by just two compound parameters. The first is a speed parameter that relates mutation, selection, and random genetic drift to the rate at which the duplicates’ expression diverges. The second is a threshold parameter that determines the point at which expression of one duplicate is sufficiently low that it is largely relieved from dosage-balance constraints and free to evolve a novel function. The model predicts two phases of evolution. In the initial phase, the difference in expression between functionally identical paralogs drifts randomly while their combined expression remains nearly constant. In the second phase, the expression threshold is reached and one of the duplicates quickly accrues function-altering substitutions.

We fit the compensatory drift model to data from Thompson et al. (2014) on the expression of sodium channel duplicates in ten families of teleost fishes. Our estimate for the speed parameter is consistent with what is known about the biological parameters that feed into it, suggesting compensatory drift is a plausible model for sodium channel evolution. Our estimate for the threshold parameter is to our knowledge the first available. Finally, we demonstrate that dosage balance selection can greatly enhance the probability of neofunctionalization, compared to the classic neutral scenario. These results suggest that whole
genome duplication, and other contexts in which dosage balance selection acts, may be a particularly rich source of genetic novelty for geologically long periods of time.

THE MODEL

After duplication, stabilizing selection favors an optimal total expression of two paralogs. A mutation that affects expression of either one will either increase or decrease fitness depending on whether it brings total expression closer to or further from the optimum. Mutations also experience random genetic drift, and so there is a nonzero probability that both mildly deleterious and beneficial mutations will be established.

We visualize compensatory drift as a series of fixation events that change the expression of the duplicates. A schematic of the process is shown in Figure 2. The two paralogs evolve in an anti-correlated pattern. Mutations in one duplicate can move the total expression away from its optimum. Compensatory mutations in the other duplicate tend to move total expression closer to the optimum. The result is that total expression remains close to the optimum, while the difference in their expression fluctuates randomly. The state of the population at any time is described by the total expression of the two duplicates, and the difference in expression between them. If expression evolves to a point at which one of the duplicates produces the bulk of the gene product, selection is no longer strong enough to prevent function-altering substitutions from accruing in the paralog with lower expression. This threshold can be interpreted as either the point where pseudogenization quickly occurs or when the benefit of neofunctionalization outweighs the fitness trade-off from loss of the ancestral function.

Assumptions

The expression levels of two duplicates are denoted as $p_1$ and $p_2$. We assume that stabilizing selection acts on the sum of expression, $A = p_1 + p_2$. The fitness function acting on $A$ is proportional to a normal distribution with mean equal to the optimum for total expression, $\theta$, and with variance $\omega^2$ (which are assumed constant in time). The variance parameter determines the strength of selection, where larger values of $\omega^2$ indicate weaker stabilizing selection. No selection acts on the difference in expression, $D = p_1 - p_2$.

Mutations occur in the regulatory regions of each of the four gene copies at a rate $\mu$ per generation. They evolve according to a Fisher-Wright model of drift and selection. Mutations
enter the population at a rate of $4N\mu$, where $N$ is the population size. Their effects on expression are additive. The effect of a given mutation on $p_1$ or $p_2$, which we denote $\delta$, is normally distributed with mean 0 and variance $\sigma^2_m$. We therefore assume that the distribution of mutational effects is constant in time and independent of a gene’s current level of expression. Biologically, this means that the regulation of expression is free of complicated forms of epistasis.

New mutations are either lost or fix under the combined forces of selection and drift. We assume that mutation is weak ($4N\mu \ll 1$) so that there is a negligible chance that more than one mutation will be segregating. (We will return to this point in the Results, which suggest the model may also be a good approximation when that assumption is violated.) Evolution thus proceeds by a series of fixation events at the two loci. This is a Poisson process, and the waiting time between mutations is exponentially distributed with mean of $1/(4N\mu)$ generations.

We calculate the fixation probability for each mutation using Kimura’s (1962) diffusion approximation:

$$P_{\text{fix}} = \frac{1 - e^{-2s}}{1 - e^{-4Ns}}.$$

(1)

Here, $s$ is the selection coefficient of the new mutation:

$$s \approx -\frac{\delta^2}{2\omega^2}.$$

(2)

Equation (2) is an approximation that neglects the deviation of the population from the optimum $\theta$. The approximation is valid when the standard deviation of mutational effects is large relative to the typical deviation from the optimum. We verified the accuracy of the approximation using parameter values consistent with the data on teleost sodium channels from Thompson et al. (2014) (see File S1).

We assume that when the duplication occurs, the two paralogs have equal expression and their total expression is optimal ($D = 0$ and $A = \theta$). As evolution proceeds, expression of the duplicates will eventually fall to threshold level, denoted $p^*$, while its paralog rises to $\theta - p^*$. At this point, the paralog with lower expression rapidly either becomes a pseudogene or neofunctionalizes. That threshold is represented in our model by a critical difference in the
expression of the duplicates, \( D^* = \theta - 2p^* \). If \( D \) reaches either \( D^* \) or \(-D^*\), then one or the other duplicate loses its original function.

**Evolutionary dynamics**

Our goal is to determine the probability distribution for expression levels at times following the duplication event. Simulations reveal that under plausible parameter values, evolutionary trajectories are confined to values of \( A \) very close to \( \theta \) (see File S1). This suggests the dynamics can be approximated by a one-dimensional diffusion in the expression difference, \( D = p_1 - p_2 \).

We write the probability density of \( D \) at time \( t \) following the duplication as \( F(D, t) \). The evolution of the density function is described by

\[
\frac{\partial}{\partial t} F(D, t) = \frac{1}{2} \sigma_D^2 \frac{\partial^2}{\partial D^2} F(D, t). \tag{3}
\]

This is the heat equation (Cox and Miller 1965), where \( \sigma_D^2 \) is the diffusion parameter. This parameter determines the speed at which \( D \) evolves, and it equals the rate of increase in variance of the probability distribution \( D \) per generation. File S2 shows that

\[
\sigma_D^2 = k \mu \omega^3 \sigma_m N^{3/2}. \tag{4}
\]

Here, \( k \) is a constant that is independent of the model’s parameters. It is difficult to calculate analytically, so we determined its value (\( k \approx 1.543 \)) numerically (see File S2).

From equation 4, we gain insight on the impact of biological parameters on the speed at which \( D \) evolves. Imagine that we follow a set of evolutionary lineages that began to diverge independently after the duplication event. The variance in the distribution of \( D \) initially increases at a constant rate and is equal to \( \sigma_D^2 t \) at \( t \) generations after duplication. Thus the diffusion rate \( \sigma_D^2 \) sets the speed of divergence, as illustrated in Figure 3. Equation 4 shows how the biological parameters affect this speed. The speed is reduced by larger population sizes. Larger values of \( N \) cause a greater number of mutations to enter the population in each generation but also increase the efficiency of purifying selection; the net result is that a smaller number of mutations fix (see Equation 1). Equation 4 also shows that the speed of divergence increases with higher mutation
rates (larger $\mu$) and decreased strengths of selection (larger $\omega^2$). A somewhat counterintuitive result is that the speed of divergence declines as the average effect size of mutations ($\sigma_m$) increases. This is because larger mutations are more likely to be strongly deleterious and therefore very unlikely to fix.

To summarize the model: the probability density of $D$ evolves according to Equation 3, with initial condition $D = 0$ at $t = 0$ and with absorbing barriers at $\pm D^*$. Before doing any further analysis, Equation 3 tells us a simple but important fact — although the model is based on six underlying biological parameters ($\mu, \omega, \sigma_m, N, \theta,$ and $p^*$), the evolutionary dynamics are governed by only two: the speed parameter $\sigma_D^2$, and the threshold $D^*$.

The solution for the density function of $D$ is:

$$F(D, t) = \frac{1}{\sqrt{2\pi \sigma_D^2 t}} \sum_{n=-\infty}^{\infty} \left\{ \exp \left( -\frac{(D+4nD^*)^2}{2\sigma_D^2 t} \right) - \exp \left( -\frac{(D+(4n-2)D^*)^2}{2\sigma_D^2 t} \right) \right\}$$

for $-D^* < D < D^*$ (Cox and Miller 1965). The probability that one of the duplicates has either been lost as a pseudogene or neofunctionalized after $t$ generations is

$$P_{\text{loss}}(t) = 1 - \int_{-D^*}^{D^*} F(D, t) dD$$

$$= 4 \sum_{n=0}^{\infty} \left\{ \Phi \left( \frac{(4n+3)D^*}{\sqrt{\sigma_D^2 t}} \right) - \Phi \left( \frac{(4n+1)D^*}{\sqrt{\sigma_D^2 t}} \right) \right\},$$

where $\Phi$ denotes the standard normal cumulative distribution function (Cox and Miller 1965).

With Equation 6 we can infer how varying the biological parameters in Equation 4 impact the expected life span of duplicate genes. Figure 4 shows the result of varying each biological parameter of the model on the time scale of duplicate loss. These results imply that the time between duplication and loss can be very long, especially for large populations, genes under strong dosage constraints (small $\omega$), and genes with high expression (large $D^*$). To get some idea of the time scale, we can calculate the number of generations needed to reach a probability of 1/2 that one of the duplicates is lost, using parameter values that are plausible for the electric fish.
clades: a population size $N = 10^4$ and a mutation rate $\mu = 10^{-5}$. The strength of stabilizing selection ($\omega^2 = 81 \sigma_m^2$) is such that $90\%$ of mutations are strongly deleterious ($|Ns| > 1$) and so have negligible chance of fixation. The threshold is $D^* = 5 \sigma_m$, which means that following duplication the threshold $p^*$ could in principle be reached with the fixation of just five mutations of typical size. (As we will see, however, that does not happen because the vast majority of mutations are eliminated by dosage balance selection.) Under these assumptions, we find from Equation 6 that after 1.7 billion generations there is still a $50\%$ probability that neither gene will have been lost. Thus dosage balance selection can maintain functional paralogs for very long evolutionary periods. If we decrease dosage balance selection strength such that half of mutations are nearly neutral ($\omega^2 = 2.3 \times 10^3 \sigma_m^2$), the amount of time drastically decreases to just 11 million generations.

**FITTING THE MODEL TO DATA**

In order to assess the plausibility of this model and to estimate parameters of biological interest, we fit this model to data on the expression of sodium channel duplicates from Thompson et al. (2014). The data are the relative expression levels of the two teleost-specific paralogs in ten families of fishes sampled broadly across the entire teleost clade and the phylogenetic relations between those families (Figure 1). The parameters being fit are the diffusion rate $\sigma_D^2$ and the threshold for gene loss $D^*$. We used Approximate Bayesian Computation (ABC) because it allows inferences about models that are too complicated for statistical frameworks such as likelihood (Tavare et al. 1997; Beaumont et al. 2002; Beaumont 2010). The basic approach is to compare summary statistics measured from simulated data to the same statistics measured from the real data. Estimates for the parameters are given by the values that produce simulated datasets most similar to the real data. In practice, this is accomplished by choosing values for the model parameters from prior distributions, simulating data using the model with those values, and comparing the summary statistics that result with those from the real data. The parameter values used in the simulation are rejected from the posterior distribution if the summary statistics from the real and simulated datasets are not sufficiently similar.

We simulated the evolution of expression on the phylogeny under the model described above. The output of the simulation gave the identities of the lineages (if any) that lost one of the
paralogs to muscle function, and the relative expression of the two paralogs for those lineages that have not. These results were compared to the actual data using two types of summary statistics. The first, which is binary, is determined by whether neofunctionalization occurred in the same locations on the tree as is observed in the data. We rejected all simulations in which that pattern was not observed. The second kind of summary statistic were the independent contrasts (Felsenstein 1985) at the nodes of the phylogeny for the relative expression of the duplicates in the nonelectric fishes. We rejected simulations if the Euclidean distance of the independent contrasts between the real and the simulated data exceeded a threshold. Further details are given in Supporting Information, File S3.

Including the electric fish data in the analysis upwardly biases our estimate of the probability of neofunctionalization. (The families of fishes in the dataset are not randomly chosen: it intentionally includes the only two families in which neofunctionalization is known.) To address this issue, we performed ABC analysis both with the electric fish and without them. Excluding the electric fish biases the estimate in the opposite direction, and therefore the two analyses give boundaries for our estimates of model parameters.

The joint posterior distributions for the diffusion rate $\sigma_D^2$ and the threshold for gene loss $D^*$ from the two analyses are shown in Figure 5. The distributions are quite similar. On a log-log plot, the values of log $\sigma_D^2$ and log $D^*$ are strongly correlated. The data are consistent with either small values of the speed parameter $\sigma_D^2$ and the threshold $D^*$, or with large values of both those parameters.

We can use published information about absolute gene expression levels to refine the likely range of values for these parameters. Promoter and enhancer mutation studies suggest that gene expression levels may be on the order of 10 $\sigma_m$ to 100 $\sigma_m$ (Melnikov et al. 2012; Patwardhan et al. 2012; Metzger et al. 2015). The data from Thompson et al. (2014), in conjunction with estimates of the distribution of transcript levels in eukaryotic cells (Mortazavi et al. 2008; Islam et al. 2010; Schwannhauser et al. 2011; Marguerat et al. 2012), suggest that a conservative lower limit for $D^*$ is 3 $\sigma_m$ (see File S3 for details). Letting $D^*$ vary between 3 $\sigma_m$ and $10^2 \sigma_m$, we used the linear regressions shown in Fig. 5 to determine a range of plausible values for $\sigma_D^2$. We estimate that if only 3 substitutions of typical size are needed to reduce a paralog’s expression to the threshold ($D^* = 3 \sigma_m$), then the expected value of $\sigma_D^2$ is $5.4 \times 10^{-9} \sigma_m^2$/yr. If expression is
much larger, such that 100 substitutions of typical size are required to reach threshold, then the expected value of $\sigma^2_D$ is $1.5 \times 10^{-5} \sigma_m^2/yr$.

We explored what these results imply about the biological parameters on which the model is based. We began by estimating the strength of dosage balance selection on the sodium channels. We assumed the range of values for $\sigma^2_D$ just described, that $\mu$ lies between $10^{-6}$ and $10^{-4}$/allele/generation, that $N$ lies between $10^4$ and $10^6$, and that there is one generation per year. Equation 4 and those parameter values then imply that the variance of the fitness function, $\omega^2$, is between $11 \sigma_m^2$ and $4.6 \times 10^6 \sigma_m^2$. We can use plausible expression levels derived from the studies cited above to estimate how efficient dosage balance selection is at removing expression mutations in terms of transcripts per cell. If the sodium channels are expressed at 50 transcripts per cell and $\sigma_m$ is 5% of that expression level, for example, then the estimated values of $\omega^2$ imply that mutations that change sodium channel expression by more than 5.3 transcripts per cell are efficiently eliminated by dosage balance selection.

Next, we asked about the properties of mutations that fix. We simulated the compensatory drift process using the parameter values cited in the previous paragraph (see File S1 for details). These results show that for small values of $D^*$ ($= 3 \sigma_m$) and strong dosage balance selection ($\omega^2 = 11 \sigma_m^2$), 97% of mutations are removed by selection that would otherwise fix. On average, mutations that fix change expression by only 0.02 $\sigma_m$, and some 9,000 substitutions occur before one of the duplicates becomes a pseudogene or neofunctionalizes. For a larger value of $D^*$ ($100 \sigma_m$) and very weak selection ($\omega^2 = 4.6 \times 10^6 \sigma_m^2$), only 17% of mutations are prevented from fixing by dosage balance selection. The effect of the average mutation that fixes is 0.7 $\sigma_m$, and 8,000 substitutions occur before the threshold is reached. We emphasize that these estimates are very rough, but they are to our knowledge the first for these important evolutionary parameters.

We find that if dosage balance selection is strong ($\omega$ not very much bigger than $\sigma_m$), then the parameter estimates for the sodium channels are consistent with the assumptions of one-dimensional diffusion approximation. With weak selection, however, the approximation breaks down. This is because total expression can deviate substantially from the optimum and so the dynamics are not well-approximated by a one-dimensional diffusion. Our model therefore describes the evolutionary dynamics of these sodium channel duplicates if $D^*$ and $\omega$ are not very much larger than $\sigma_m$, but would be more accurately modeled by a two-dimensional diffusion.
model if they are not. It may be difficult to develop analytic results for that model, but it could be studied numerically.

Stochastic simulations suggest that our results are surprisingly robust to the assumption that no more than one mutation segregates at any given time (that is, \(4N\mu << 1\)). Simulations of a Wright-Fisher model show that mutations that fix do so largely as a neutral process. The distribution of fitness effects for fixed mutations is shown in Supporting Information, Figure S1. For the parameter values simulated, the mean value of \(|N_s|\) is between 0.15 and 0.24, and it is very rare for mutations to fix with \(|N_s| > 1\). We ran simulations in which the mutation rate varied over more than four orders of magnitude. When \(N\mu\) equals 1, the most common allele is typically at a frequency of only about 50\% (Figure S2). Nevertheless, the substitution rate is very close to what our model predicts (Figure S3). That behavior is also consistent with a model in which mutations that segregate at appreciable frequencies are entirely neutral. The results of the simulations only begin to significantly depart from the expectations of our model when \(N\mu > 1\). In sum, our analytic results may apply when mutation rates are higher than the approximations assume.

**NEOFUNCTIONALIZATION AND COMPENSATORY DRIFT**

Because dosage balance selection can maintain duplicates for long evolutionary periods, it may be more likely that neofunctionalization will occur than it does when dosage balance is weak or absent (Force et al. 1999; Papp et al. 2003; Aury et al. 2006; Hughes et al. 2007; Scannell and Wolfe 2008; Thompson et al. 2014; Gout and Lynch 2015). To explore that idea further, we extended our model by adding two new kinds of mutations. The first is a loss-of-function mutation that renders one of the duplicates a pseudogene. The probability that it fixes is again given by the fitness function used in the main model. The second kind of mutation neofunctionalizes one of the duplicates. It suffers the same fitness cost as a loss-of-function mutation, but also benefits by a 0.1\% fitness gain from its new function.

We compared the frequency of neofunctionalization in three simulated populations (File S1) evolving under dosage balance selection that ranged from strong to very weak: \(\omega^2 = 10^2 \sigma_m^2, 10^4 \sigma_m^2,\) and \(10^6 \sigma_m^2\). For all three simulations, mutations that alter expression were ten times more frequent than pseudogenizing mutations, and pseudogenizing mutations were \(10^3\) times more frequent than neofunctionalizing mutations. The population size was \(N = 10^4\), the
mutation rate was $\mu = 10^{-5}$ mutations / allele / generation, and the optimal expression was $\theta = 5 \sigma_m$.

We found that neofunctionalization is greatly facilitated by dosage balance selection. Figure 6 shows that when dosage balance selection is stronger, duplicate genes are preserved for longer and more mutations occur before a duplicate is lost. In consequence, neofunctionalization happens nearly 10 times more often than when dosage balance selection is very weak. Neofunctionalization is most likely when expression falls inside a window of values in which the cost of losing the original function is smaller than the benefit of gaining the new function. In this window, a mutation that causes pseudogenization is still too deleterious to fix. Equation 4 shows that stronger dosage balance selection slows the rate of compensatory drift and thus increases the amount of time the population spends in this evolutionary window. These results suggest that dosage balance selection greatly diminishes the evolutionary potential of paralogs early after duplication, but after a long period of compensatory drift greatly facilitates the acquisition of a new adaptive function.

**DISCUSSION**

We formalized a model of the evolutionary process that we call compensatory drift. This model shows how dosage balance selection on duplicate genes (paralogs) can lead to neofunctionalization some tens to hundreds of millions of years after duplication. Dosage balance selection constrains the combined expression of both paralogs to an optimum, but not the expression of the individual genes. That allows the relative expression of the paralogs to drift apart by the fixation of mutations with small effects. The speed at which this divergence occurs is determined by the diffusion rate $\sigma_D^2$, which in turn is a function of several biological parameters. Our results show that stronger dosage balance selection and larger mutational effects on expression slows divergence because a greater fraction of mutations are strongly deleterious and so have virtually no chance of fixation. Larger populations also decrease divergence because they enhance the efficiency of selection and so eliminate a larger fraction of mutations.

Simulations of compensatory drift reveal that dosage balance selection can improve the odds that neofunctionalization occurs rather than pseudogenization. If a novel function yields a slight advantage while having a large trade-off with the ancestral function, dosage balance selection can still improve the chances of neofunctionalization, but only after a long period of
compensatory drift. As expression of one paralog declines, the strength of selection to maintain its original function diminishes. It reaches a level at which mutations that pseudogenize the gene are still strongly deleterious but mutations that neofunctionalize are beneficial. Our results show the probability of neofunctionalization is increased when the added expression of duplicates is high, dosage balance constraints are strong, and population sizes are large.

We fit the model to data on sodium channel duplicates in teleost fishes. We estimate that the diffusion rate $\sigma^2_D$ lies between $5.4 \times 10^{-9} \sigma^2_m / \text{year}$ and $1.5 \times 10^{-5} \sigma^2_m / \text{year}$, where $\sigma^2_m$ is the variance of mutation effect sizes. The square root of $\sigma^2_D$ is roughly equal to the amount of divergence that accumulates in a lineage per year. This implies duplicates diverge between $7 \times 10^{-5} \tilde{\delta} / \text{year}$ and $4 \times 10^{-3} \tilde{\delta} / \text{year}$, where $\tilde{\delta}$ is the average effect that a mutation has on the amount of gene product produced by a duplicate. About $8,000 – 9,000$ substitutions occur before the threshold is reached. This number seems large, but it is not inconceivable. Summing up all the genetic elements that can affect expression (promoters, enhancers, microRNAs, post-translational regulators, etc.), there are many mutational targets for expression changes. Indeed, high rates of enhancer gain and loss (enhancer turnover) have been seen in several taxa (Schmidt et al. 2010; Domene et al. 2013; Paris et al. 2013; Arnold et al. 2014). Dosage balanced duplicates may undergo more rapid enhancer turnover than singleton genes since compensation is possible at two different loci. A last consideration is that the time span involved is long, on the order of $10^8$ generations. In any event, our inferences about numbers of substitutions are very imprecise and the actual number may be much smaller. In the future we expect larger datasets of comparative paralog expression will emerge and will allow greater precision in parameter estimates using methods of analysis such as ABC.

This work builds on earlier hypotheses about the evolution of dosage-sensitive duplicates. Aury et al. (2006) proposed that expression of duplicates evolves by compensatory changes, which can greatly delay the pseudogenization or neofunctionalization of one of the pair. Later work suggested this process leads to a random walk along a line of equal combined expression, a process that could explain comparative gene expression patterns observed in disparate lineages of organisms (Thompson et al. 2014; Gout and Lynch 2015; Lan and Pritchard 2015). Other researchers suggested that gene loss in a duplicated network would cause imbalances and thus put positive selective pressure for loss of other duplicates in the same network, leading to concerted duplicate inactivation (Papp et al. 2003; Hughes et al. 2007; Konrad et al. 2011).
Under compensatory drift, the eventual loss of a duplicate may not have much impact on other genes in its network since its paralog will already be producing (almost) all the gene product needed.

Several lines of evidence are consistent with dosage balance selection after whole genome duplication (WGD). In contrast to classical models in which redundant duplicates evolve neutrally (Ohno 1970; Walsh 1995; Force et al. 1999; Lynch and Conery 2000), dosage balance selection will cause both genes to be essential immediately after duplication. WGD does not disrupt dosage balance and therefore many preserved duplicates originating in a WGD may evolve under dosage balance selection. The paramecium *P. tetraurelia* has undergone three WGDs in its evolutionary history. In a large proportion of the duplicates from the most recent WGD, both members of the pair are evolving under strong purifying selection, and this proportion declines in time (Aury et al. 2006). This pattern indicates that many genes are dosage-sensitive and evolve under dosage balance selection, but that eventually selection to conserve function is lost for one of the duplicates. Other examples come from vertebrates. Some 100 my after a WGD in the ancestor of salmonid fishes, about half of the duplicates are retained, and one quarter of those are still similar in expression and sequence (Berthelot et al. 2014). In a WGD that happened in the ancestor of teleost fishes about 300 my ago, many duplicate pairs persisted for over 200 my before a member of the pair was lost (Blomme et al. 2006; Brunet et al. 2006; Sato et al. 2009). Delayed loss of duplicates long after WGD is also seen in paramecium species (McGrath et al. 2014b). Together these patterns indicate that many duplicates after WGDs are dosage-sensitive and evolve in two phases: an initial prolonged phase where both duplicates evolve under selection that conserves function, and a later phase in which a duplicate is lost. This later phase could be due to a paralog drifting to low expression and may be the stage at which a redundant gene is most likely to evolve a new function.

Additional predictions flow from the compensatory drift model. Duplicate pairs should persist longer if their total expression is high because more mutations must fix to reach the expression threshold $p^*$ (that is, $D^*$ is larger). (See Fig. 4 for the impact of increasing $D^*$ on the time until duplicate loss.) Both yeast and *Paramecium* show just that pattern: there is a positive correlation between expression levels and the longevity of duplicated genes following WGD (Seoighe et al. 1999; Aury et al. 2006; Gout et al. 2010; McGrath et al. 2014b). To explain this pattern, Gout et al. (2010) argued that stabilizing selection on total expression is stronger on
dosage-sensitive duplicates that have high levels of expression. That idea is consistent with our model: the speed at which expression of paralogs diverges becomes slower as the strength of selection increases. The model also makes predictions about patterns of subfunctionalization of dosage-balanced duplicates. When duplicates are expressed in different cell types under different regulation, compensatory drift can occur in parallel in the two cell types, occasionally leading to subfunctionalized expression. Finally, our model makes predictions about phylogenetic patterns. We expect the member of a duplicate pair that has neofunctionalized in a lineage to have lower expression than its paralog in closely related lineages where neofunctionalization has not occurred (Anderson et al. 2009; Thompson et al. 2014). Recent data support this prediction (Gout and Lynch 2015).

Compensatory drift may play an important role in two other evolutionary contexts. Dosage balance selection can act on gene duplicates that do not arise by WGD. Selection for increased expression can fix a duplicated gene in a population (Kondrashov 2012). Subsequently, there is stabilizing selection favoring the new, higher expression optimum. Once that level is reached, the expression can diverge by compensatory drift as described by our model. Second, compensatory drift can act on the transcription rate and translation rate for a gene evolving under stabilizing selection for expression. An important difference with duplicate genes is that transcription and translation rates cannot completely compensate for each other. Qualitatively, however, we expect to see similar evolutionary dynamics.

In our model, neofunctionalization happens after a long period of compensatory drift. Alternatively, a novel gene function could pre-date the duplication event as a minor pleiotropic effect that is not optimized because of tradeoffs. Under the escape-from-adaptive-conflict model, duplicates are freed from these tradeoffs, allowing one of them to become rapidly optimized for the alternative function (Conant and Wolfe 2008; Des Marais and Rausher 2008). However, if one of the gene’s functions requires both duplicates to contribute expression, then compensatory drift would have to occur before one duplicate can escape from the adaptive conflict.

Compensatory drift is related to but distinct from quantitative subfunctionalization (QS). That process describes how, following duplication, degenerative mutations accumulate by drift in each paralog until their total expression declines to a minimum total level necessary for viability (Force et al. 1999, Stoltzfus 1999, Force and Lynch 2000, Hahn 2009; Qian et al. 2010). Compensatory drift, in contrast, is the divergence of expression in paralogs that have already
reached optimal expression under dosage-balance selection. A second difference between the processes is that under compensatory drift, half of the mutations that fix increase expression, while under QS none of them do. Despite these differences, there are also important similarities. Both processes can dramatically increase the probability that a gene neofunctionalizes. The two processes could operate in succession. Following the tandem duplication of a gene, expression of each duplicate can decline until both paralogs are necessary to produce the minimal expression that is needed. The duplicates can then diverge through compensatory drift.

Dosage balance selection may provide opportunities for adaptation long after whole genome duplication occurs. When one of a duplicate pair of genes drifts to a low level of expression, a period of incubation occurs during which it can evolve a new function. As illustrated by duplicates of sodium channel genes in teleost fishes, down-regulation of dosage-sensitive duplicates may be a common pre-adaptation in many diversifying gene families. Compensatory drift thus may still be facilitating adaptation very long after the two WGDs that occurred near the root of the vertebrate tree.

Acknowledgements

The authors thank Laura Crothers for comments on the manuscript. This research was funded by NSF DEB-1311521 to AT and NSF grant DEB-0819901 to MK. The authors acknowledge the Texas Advanced Computing Center (TACC) at The University of Texas at Austin for providing HPC resources that have contributed to the research results reported within this paper. URL: http://www.tacc.utexas.edu
References


Supporting Information

File S1: Description of the stochastic simulations.

File S2: Details of the derivation of the analytical model of compensatory drift.

File S3: Description of the ABC algorithm for Bayesian estimation of model parameter values.
Fig. 1. Sodium channel expression and phylogenetic relationships of ten teleost fish species. The families represented here span almost the entire teleost clade. The relative expression in skeletal muscle of the voltage-gated sodium channel genes, *Scn4aa* and *Scn4ab*, is represented with pie charts for each of the ten species. Thick red segments on the branches leading to the two electric fishes, *E. veriscens* and *G. petersii*, indicate the likely times where electric organs evolved and *Scn4aa* neofunctionalized (Arnegard et al. 2010; Betancur-R et al. 2013).
Figure 2. Schematic of the coevolution of paralog expression under compensatory drift. Axes show the expression of the duplicate genes, and fitness is represented with a grey-scale. The dashed diagonal shows the maximal fitness where $p_1 + p_2 = \theta$. The dotted vertical and horizontal lines show the expression thresholds where a duplicate loses its original function. A sequence of several consecutive expression changes are shown with numbered circles.
Figure 3. The speed of divergence in the expression of duplicate genes. Examples of the evolution of expression difference, $D = p_1 - p_2$, when the diffusion parameter $\sigma_D^2$ is small (top) and large (bottom). Each panel shows sample trajectories and the final probability distribution for $D$ (at right). In the top panel, the final distribution of $D$ is approximately normal; neofunctionalization and pseudogenization are rare. In the bottom panel, neofunctionalization is frequent (large rectangles outside the thresholds at $\pm D^*$).
Figure 4. The probability of gene loss after duplication. The thick curve shows the probability of loss, either through pseudogenization or neofunctionalization, through time with standard parameter values: the selection strength is $\omega^2 = 10^4$, the standard deviation of the mutation effect size is $\sigma_m = 1$, the population size is $N = 10^4$, the mutation rate is $\mu = 10^{-5}$, and the expression threshold is $D^* = 100$. Other curves show results when individual parameter values are doubled.
Figure 5. Joint posterior distribution of the diffusion rate parameter, $\sigma_D^2$, and the expression threshold, $D^*$ estimated from sodium channel expression in teleost fish. The joint distributions from two ABC analyses using the expression data from Thompson et al. (2014). In one analysis, the two lineages of electric fish are included, and in the other they are not. The linear relationship between the parameters from the two analyses is very similar. The regression lines are $\log \sigma_D^2 = 2.04 \log D^* - 20.5$ with the electric fish, and $\log \sigma_D^2 = 2.06 \log D^* - 21.3$ without them.
Fig. 6. Dosage balance selection and the probability of neofunctionalization. The top panel shows the number of mutations (a proxy for time) that occur before one of the duplicates neofunctionalizes or pseudogenizes. Dots show the mean and the whiskers show one standard deviation. The bottom panel compares the frequency of neofunctionalization for three strengths of dosage balance selection, from strong to very weak. Results are based on $10^3$ simulations.