Supporting Methods

Methods for Grouping Genes into Families. We first verified genes that were free of repetitive elements. Next, we linked genes into families by following Rost's criterion, as adapted and described in detail by Gu *et al.* (1). This approach provides a stringent criterion for family membership, and only closely related genes were considered members of the same family. Note that therefore only recent gene duplicates were studied, probably indicating gene families with high rates of duplication. In our neutral duplication-birth-death model, high duplication rates necessitate higher rates of expression loss to maintain equilibrium.

Methods for Estimating Gene Family Phylogenies. We first aligned translated sequences by using ClustalW 1.81 (2) before back-translating to nucleotides for phylogenetic analysis by using transAlign (3). Using Modeltest (4), we determined the best-fit maximum likelihood model of the majority of gene families (GTR + I + Γ) and assumed that model for phylogenetic analyses of all families by using PAUP* 4.0b10 (5). To root gene family trees, we first added several potential outgroup genes (lacking expression data) to each ingroup family. These multiple potential outgroup genes were found from the *Drosophila melanogaster* and *D. pseudobscura* genomes by using Blast. We then constructed a phylogeny of the ingroup plus multiple candidate outgroup genes and rooted that composite phylogeny at the midpoint. With this strategy, some candidate outgroups fell within the ingroup, and we did not use those as outgroups. The remaining candidate outgroups thus would be more distant from the ingroup genes than the ingroup genes are from each other. We then used the single closest outgroup gene as the outgroup in the final phylogenetic analysis for each gene family. We included in our final analysis only gene families of size three or greater, for which expression data were present for all gene family members.

Rates of Gain and Loss for Individual Gene Families. The main text considers the rate of loss and gain of gene expression estimated across all gene families simultaneously (Fig. 2). This approach assumes a single rate of gain and a single loss rate for all gene families. However, this estimate could be incorrect if individual gene families have different rates of gain and loss of gene expression regions. To discount this possibility, we examined rates of gain and loss for each gene family separately. We plotted the likelihood functions for each family individually and determined whether faster rates of loss showed higher likelihood values than higher rates of gain. In 24 gene families, likelihood values were higher for higher rates of loss. Only two families showed the opposite result, and a single family showed equal likelihood values for a given rate of gain or loss (Fig. 5). These results indicate that the conclusion of faster rate of loss is robust to the existence of different rates of evolution in different gene families.

Duplication-Birth-Death Model. An important question raised by our study is whether gene expression can persist even when loss of individual expression domains is more common than gain. In the main text, we argued that duplication of genes and their regulatory elements introduces a situation where expression regions of duplicate genes share a single evolutionary origin. After duplication, the expression regions may be lost separately. This process may lead to a situation where loss of expression regions balances origin by gain of new regions and duplication of existing regions. To formalize this logic and examine this hypothesis, we constructed a duplication-birth-death model.

The model has three rate parameters: activation rate of new expression regions for a gene, repression rate of expression regions for a gene, and gene duplication rate. The model assumes that all of these rates are constant during evolution. When a gene duplicates, all of its expression regions are initially conserved in the duplicates, but subsequent activation or repression of expression regions may occur. If a gene loses expression in all regions, it becomes a pseudogene that we assume cannot regain function. Because we are testing for the possibility of long-term maintenance of gene expression, we allow both duplicates or all members of any gene family to be lost.

We investigated equilibrium conditions of a simple model describing birth and death of gene expression regions, coupled with gene duplication. More formally, the model is described by:

$dN_r / dt = (\delta - \alpha - \beta)\, N_r + \alpha\, N_{r-1} + \beta\, N_{r+1}$, **[1]**

where α = expression region activation rate, β = expression region repression rate, δ = gene duplication rate, and N_r = the number of genes expressed in r regions. As such, equilibrium is given by $dN_r / dt = 0$, because there is no change in the total number of genes expressed in r regions. This equation can be expressed as a transition matrix, M, where M_{ij} is the contribution of N_i to N_j with i and j taking values from 1 to $r = r_\mathrm{max}$. As an example, the domain transition matrix when r_{max} is 4 is:

$$M = \begin{pmatrix} \delta - \beta(1+\gamma) & \beta & 0 & 0 \\ \gamma\beta & \delta - \beta(1+\gamma) & \beta & 0 \\ 0 & \gamma\beta & \delta - \beta(1+\gamma) & \beta \\ 0 & 0 & \gamma\beta & \delta - \beta \end{pmatrix}$$

where $\gamma = \alpha / \beta$, the ratio of gain to loss. Note that M_{44} is equal to δ-β meaning that there is no decrease of N_5 as a result of gain of a region to N_6.

To investigate this model, we set $\alpha / \beta = 0.5$ based on likelihood estimates from fly data described in the main text. Solving the equation requires assumptions about the maximum number of possible expression regions in which a single gene could be expressed (r_max). As such, we assumed several different values of r_max. For each value of r_max examined, we found a single biologically meaningful equilibrium by solving Determinant[M] = 0 (Fig. 6). Other solutions were not biologically realistic because they required negative values of N_r (negative numbers of genes is not biologically possible). Each equilibrium solution was a fixed ratio of the rate of gene duplication (δ) to the rate of expression region loss (β). In other words, maintaining equilibrium with higher rates of domain loss requires higher rates of gene duplication. This ratio changed with different values of $r_{\max}$, but the ratio (β / δ) converged to ≈11.5 when $r_{\max}$ had a value of ≈50 or higher (Fig. 7).

Using this simple evolutionary model, we found that conditions indeed exist whereby genes may evolve at a dynamic equilibrium, even when expression regions are lost more commonly than gained. One way to understand how equilibrium is possible is to examine two types of parameter values that do not result in equilibrium. First, if expression loss is too rapid, all genes become extinct. Second, if expression gain is too rapid, all genes become expressed everywhere. Because living organisms possess genes that are differentially expressed, there may be equilibrium between gain and loss of expression regions, or at least a long-term steady state. Our model confirms that a dynamic equilibrium can be maintained indefinitely, even with high rates of loss of expression regions. Put simply, the same number of expression regions introduced by gene duplication and expression region gain must be lost by repression to maintain equilibrium.

However, this equilibrium is unstable, and deviations from the equilibrium ratio (β / δ) could result in extinction of all expression or infinite expression, the two stable equilibria. In living, evolving lineages, nonneutral processes like natural selection probably contribute to maintaining expression of genes in only some regions, but these processes are not included in our model. For example, loss of critical expression regions is prevented probably by natural selection. With the important caveat of natural selection in mind, the neutral model indicates a general and predictive feature that organisms or gene families with higher rates of gene duplication also should show increased rates of expression region loss. Increased rates of repression also may lead to higher specialization of genes, which also may translate to morphological specialization and morphological and ecological diversity (6, 7).

1. Gu, Z., Nicolae, D., Lu, H. H. & Li, W. H. (2002) *Trends Genet* **18,** 609-613.

2. Higgins, D. G., Bleasby, A. J. & Fuchs, R. (1992) *Comp. Appli. Biosci.* **8,** 189-191.

3. Bininda-Emonds, O. R. P. (2005) *BMC Bioinformatics* **6**, 156.

4. Posada, D. & Crandall, K. A. (1998) *Bioinformatics* **14,** 817-818.

5. Swofford, D. L., Waddell, P. J., Huelsenbeck, J. P., Foster, P. G., Lewis, P. O. & Rogers, J. S. (2001) *Syst. Biol.* **50,** 525-359.

6. Ohno, S. (1970) *Evolution by Gene Duplication* (Springer, New York).

7. Carroll, S. B., Grenier, J. K. & Weatherbee, S. D. (2005) *From DNA to Diversity: Molecular Genetics and the Evolution of Animal Design* (Blackwell, Oxford).