File S6: Determining the correlation c between gene expression in old and new states

The variable c from the main text quantifies the statistical association between the random variables g_i^N and g_i^O. The purpose of the following calculation is to establish the relationship between c and the Pearson product-moment correlation coefficient R between these random variables. Briefly, c is identical to R except for a scaling constant, and the only reason to use c instead of R is that it makes central mathematical expressions such as equation (21) from File S5 simpler. The following calculations rely on the assumption that $R \geq 0$, which is justified by the empirical observation that gene expression states in different environments or cell types are generally positively correlated (File S7).

The variances of the two random variables g_i^N and g_i^O compute as $\mathbb{V}(g_i^N) = f^N(1-f^N)$ and $\mathbb{V}(g_i^O) = f_1^O(1-f_1^O)$. Their covariance equals $Cov(g_i^N, g_i^O) = \mathbb{E}(g_i^N g_i^O) - \mathbb{E}(g_i^N)\mathbb{E}(g_i^O) = P(g_i^N = 1|g_i^O = 1)f_1^O - f^N f_1^O$, where $\mathbb{E}(x)$ denotes the expectation of the random variable x. To compute $P(g_i^N = 1|g_i^O = 1)$, I use the *Ansatz* that this quantity can be described by a linear function $f(c)$ $(0 \leq c \leq 1)$, such that (i) $f(0) = f^N$ if the two random variables are uncorrelated (in which case $P(g_i^N = 1|g_i^O = 1) = P(g_i^N = 1) = f^N$), and (ii) $f(1) = 1$ if the two random variables are perfectly correlated (in which case $P(g_i^N = 1|g_i^O = 1) = 1$). The linear function that fulfills these constraints is $f(c) = f^N + c(1-f^N)$. With this function, we get $Cov(g_i^N, g_i^O) = (f^N + c(1-f^N))f_1^O - f^N f_1^O = c(1-f^N)f_1^O$. The Pearson correlation coefficient R between g_i^N and g_i^O then calculates as

$$R = \frac{Cov(g_i^N, g_i^O)}{\sqrt{\mathbb{V}(g_i^N)\mathbb{V}(g_i^O))}} \tag{26a}$$

$$= \frac{c(1-f^N)f_1^O}{\sqrt{f^N(1-f^N)f_1^O(1-f_1^O)}} \tag{26b}$$

$$= c\sqrt{\frac{(1-f^N)f_1^O}{f^N(1-f_1^O)}} \tag{26c}$$

In sum, this calculation shows that c and R are linearly related, with a proportionality constant that equals the right-most expression of this equation.

I next discuss how I estimated c for experimental genome-wide gene expression data. One challenge here is that the quantity f_1^O can be experimentally measured, but the quantity f^N cannot. However, because $f^N = f_1^N - f_{01}^N + f_{10}^N = f_1^N - \Delta_m^N$, the fraction f_N of optimally expressed genes in a new state can be approximated by the fraction f_1^N of actually expressed genes, as long as the excess of wrongly active over wrongly inactive genes in the new state is not very large. This is the case for the population genetic parameters explored here. (In addition, if Δ_m is stochastically independent of f_1, then one can substitute f_1^N for f^N in the above calculation.) Under this assumption, I have calculated c for various experimental data sets. To estimate c for any one data set that records the number of expressed genes or transcripts in two environments or cell states a and b, denote as f^a, f^b, and f^{ab} the fractions of genes expressed in state a, in state b, and in both states, respectively. With this notation $Cov(g_i^N, g_i^O) = f^{ab} - f^a f^b$, $R = (f^{ab} - f^a f^b)/\sqrt{f^a(1 - f^a) f^b(1 - f^b)}$, and $c = R\sqrt{[f^b(1 - f^a)]/[(1 - f^b) f^a]}$. Notice that R is invariant to the exchange of a and b, but c is not, which means that one needs to make an arbitrary assumption as to whether a is the old or the new state (as I did in my data analysis). However, for the data sets I examined, values of c obtained under either assumption are generally close to one another (not shown).