

File S1: At least half of all genes are expressed in most environments and tissues.

An important quantity affecting the consequences of misregulation is the fraction f^N of genes that should be expressed under optimal adaptation, and in particular the sign of $2f^N - 1$. It is thus important to find out whether more than half of all genes are typically expressed in a state to which an organism is well adapted, such as an evolutionarily old environment or tissue. Useful data is provided by genome-wide mRNA gene expression data on well-established tissues. Figure S1A shows the fraction of genes expressed in seven human and mouse tissues, which ranges from 0.64 to 0.75 (human), and from 0.49 to 0.56 (mouse), depending on the tissue.

In single-celled organisms the fraction of expressed genes can be even higher. For example, RNA-seq data shows that more than 90 percent of genes are expressed in *E. coli* (rich medium), *Mycobacterium tuberculosis*, and *Vibrio cholerae* (minimal medium) (Haas et al., 2012). In the yeast *Scheffersomyces stipitis* RNA-seq data shows that 90.1 percent (5243 of 5816) genes are expressed in glucose-containing rich medium. The same yeast is also well-adapted to ferment xylose, on which it expresses 91.5 percent (5323=5176+147) of its 5816 genes (Yuan et al., 2011). In the soil bacterium *Bacillus subtilis*, tiling microarray data show that 75% of all genes (3189 of 4244) are expressed in rich medium, and 72.4% (3074 of 4244) in minimal medium (Rasmussen et al., 2009).

Unfortunately, data like these only show the fraction f_1 of actually expressed genes, not the fraction f of genes that are expressed under optimal adaptation, i.e., if misregulation is absent (in either an old or new state). However, we can take advantage of the fact that f and f_1 are linked by Δ_m as $f = f_1 - f_{01} + f_{10} = f_1 - \Delta_m$. My population simulations show that under realistic selection strengths Δ_m does not usually exceed 0.2 and can be much smaller or negative when p_B is small. In addition, because unicellular eukaryotes and prokaryotes have smaller genomes, shorter non-protein coding regions, less complex promoters, fewer transcription factors, and longer transcription factor binding sites (van Nimwegen, 2006; Stewart and Plotkin, 2012; Lynch, 2007; Sengupta et al., 2002), misregulation may be less prevalent, and the fraction f_1 of expressed genes in any one cell state may approximate more closely the fraction of genes that is expressed under optimal adaptation.

Together, these considerations suggest that f_1 usually exceeds 0.5 and often by a substantial margin. Allowing for Δ_m as high as 0.2, $f \approx 0.3$ would be a reasonable lower bound for f (i.e., for both f^N and f^O).

References

- Haas, B. J., M. Chin, C. Nusbaum, B. W. Birren, and J. Livny (2012). How deep is deep enough for RNA-Seq profiling of bacterial transcriptomes? *BMC Genomics* 13, 734.
- Lynch, M. (2007). *The origins of genome architecture*. Sunderland, MA: Sinauer.
- Rasmussen, S., H. B. Nielsen, and H. Jarmer (2009). The transcriptionally active regions in the genome of *Bacillus subtilis*. *Molecular Microbiology* 73(6), 1043–1057.
- Sengupta, A. M., M. Djordjevic, and B. I. Shraiman (2002). Specificity and robustness in transcription control networks. *Proceedings of the National Academy of Sciences of the United States of America* 99(4), 2072–2077.
- Stewart, A. J. and J. B. Plotkin (2012). Why transcription factor binding sites are ten nucleotides long. *Genetics* 192(3), 973–985.
- van Nimwegen, E. (2006). Scaling laws in the functional content of genomes. In *Power Laws, Scale-Free Networks and Genome Biology*, pp. 236–253. Springer.
- Yuan, T., Y. Ren, K. Meng, Y. Feng, P. Yang, S. Wang, P. Shi, L. Wang, D. Xie, and B. Yao (2011). RNA-Seq of the xylose-fermenting yeast *Scheffersomyces stipitis* cultivated in glucose or xylose. *Applied Microbiology and Biotechnology* 92(6), 1237–1249.