www.sciencemag.org/cgi/content/full/1140247/DC1

Supporting Online Material for

Genetic Properties Influencing the Evolvability of Gene Expression

Christian R. Landry,* Bernardo Lemos,* Scott A. Rifkin, W. J. Dickinson, Daniel L. Hartl

*To whom correspondence should be addressed. E-mail: clandry@post.harvard.edu (C.R.L.); blemos@oeb.harvard.edu (B.L.)

Published 24 May 2006 on *Science* Express
DOI: 10.1126/science.1140247

This PDF file includes:

SUPPORTING ONLINE MATERIAL

MATERIALS AND METHODS

Mutation accumulation

Eight independent lines were derived from a single haploid ancestor derived from FY10 (Mat-*a*, *leu2Δ1*, *ura3-52) (1)*, a strain isogenic with the reference strain s288c. These lines were maintained on standard YPD (yeast extract-peptone-dextrose) solid plates. Approximately every 20 generations, a single colony was used to produce new colonies in the next generation by streaking and plating cells. This propagation scheme represents a single-cell bottleneck every 20 generations, resulting in an effective population size of ~10 as calculated using the harmonic mean of the population sizes. At this population size, the fixation or loss of most mutations will be determined by random genetic drift and not natural selection. Only selection coefficients larger than the reciprocal of the population size (10%) have a significant effect on the fate of new mutations, and these are expected to be rare (*2, 3*). There is no direct estimate of the rate of nucleotide substitution in yeast, but the per-genome mutation rate has been estimated to be around 0.003 in *S. cerevisiae (4)*. Pairs of strains that have diverged for 8,000 generations are therefore expected to differ by ~30 nucleotide substitutions. Other perturbations are also likely to contribute to gene expression evolution in those lines. For instance, the rate of duplication of a gene is of the same order of magnitude as the rate of mutation per nucleotide site (*5*). High rates of expansion and contraction of low complexity regions, which are abundant in the yeast genome (*6*) and are known to contribute to phenotypic diversity (*7*), may also contribute to evolution of gene expression. Finally, at least two of these lines became diploid during the course of the experiment (W. K. Thomas, unpublished).

Gene expression analysis

Four mutation accumulation lines (MA-lines) and the ancestral strain were expression profiled. Two fresh colonies per line were grown in parallel on each of two separate days for a total of four replicates. A colony was diluted in 5 ml of YPD and the equivalent of 25,000 cells was used to inoculate 40 ml of liquid YPD. Cells were then grown overnight at 30°C and 180 rpm in 500 ml flasks and harvested by centrifugation in the morning at an optical density between 0.8 and 1. Cultures were then centrifuged at 3,500 rpm for 20 minutes at room temperature and the pellets were flash frozen in liquid nitrogen. Two samples of each strain from two separate days were pooled prior to RNA extraction. There were a total of 2 extractions per strain. Frozen cell pellets were suspended and extracted with hot acid phenol/chloroform extraction. Total RNA was ethanol precipitated, washed and resuspended in TE buffer. RNA quality was confirmed by spectrophotometric analyses with A_{260}/A_{280} ratios ~ 2. cDNA synthesis, labeling and hybridization were performed using 3DNA Array 50 Kit Version 2 (Genisphere Inc, Hatfield, PA) according to the manufacturer's protocol using 25 ug of total RNA. The samples were hybridized on arrays containing 6,388 unique probes (Qiagen Operon, Valencia, CA) printed on poly-L-lysine coated slides (Erie, Portsmouth, NH) according to standard protocols (www.microarray.org) and blocked according to Diehl et al. (*8*). A loop design was used, as it maximizes the number of replicates for the number of arrays used (*9, 10*). All the samples (four derived and the ancestral strain) were compared to each other, for a total of four replicates per strain (fig. S5). The arrays were scanned on an Axon GenePix 4000B Scanner (Axons Instrument, Molecular Devices, Sunnyvale, CA) and the images were analyzed using GenePix Pro 5 (Axons Instrument, Molecular Devices, Sunnyvale, CA). Spots of poor quality were flagged manually and eliminated from the downstream analyses. Only spots with foreground intensity higher than the background intensity plus two standard deviations were conserved. A total of 5,688 genes met these criteria. Normalization of the raw intensity was performed using methods implemented in the library Limma of the statistical software R (*11-13*). In order to make ratios of intensities independent of the absolute signal, background-subtracted intensities were normalized for each quadrant independently using the method loess and this in order to make data consistent across the array. Finally, the distributions of intensities across the experiment were normalized to have the same median-absolute-deviation by scaling the log-ratios. Raw data were deposited to the NCBI GEO database, series reference number GSE7537.

Significant changes in gene expression were assessed using Bayesian statistics as implemented in BAGEL (*10, 14*) using default parameters, which assume that all nodes have the same error variance. We estimated changes among the four MA-lines using the Bayesian posterior probability (BPP) of differential expression. We estimated the proportion of false positive tests at different BPP thresholds by randomizing the data matrix and running BAGEL on the randomized data (*15*). The proportion of false positives estimated this way is presented in Fig. 1B for different BPP thresholds. A threshold BPP of 0.99 was chosen because it best minimized type I and type II errors. BAGEL returns relative expression among the strains. In order to measure the rate of evolution of gene expression among those mutation accumulation lines, we estimated mutational variance (V_m) (*16*) using the variance of the log-transformed gene expression estimates from BAGEL of the four MA-lines, in an approach similar to (*17*). The relative fold changes (Fig. 1C) were calculated as the maximum expression level over the minimum expression level among the four MA lines.

SUPPORTING TEXT

In order to test which classes of genes, if any, were associated with high or low V_m , we separated genes with significant variation among the lines into high and low V_m classes (top 50 percentile (n = 1016), lowest 50 percentile (n = 1015)). We then assigned those genes to Gene Ontology Biologcial process and Molecular Function classes using Super GO-Slim (Sachharomyces Genome Database, SGD: http://db.yeastgenome.org/cgi-bin/GO/goTermMapper). Within each of those classes, genes were assigned to categories described in Fig. S2. Since these categories can be overlapping, we compared the distribution of number of genes of the low and high V_m classes in each GO category separately and corrected for multiple testing as follows. Each gene assignment to one GO category can be seen as a success or a failure. We therefore used a 2-sample test for equality of proportions (χ^2) to test for differences in success rate for the assignment of genes to each category for the low and high V_m groups. The results are presented in Fig. S2.

In order to investigate which features of the yeast regulatory network may affect the neutral rate of gene expression evolution, we obtained data from several public sources. First, we mapped the transcription factor binding sites using a map of regulatory elements that combines Chip-chip experiments (*18*) and computational predictions available from ftp.stanford.edu/pub/yeast/chromosomal_feature/scerevisiae_regulatory.gff (accessed on August 8, 2006). In this database, the identity of DNA-binding proteins and the position of the binding sites are provided. We assigned a binding site to a gene when it was within 1kb upstream of the translation start site using the position of ORFs available through SGD. mRNA abundance was obtained from (*19*) and corrected protein abundance (Ave. YEPD) and noise in protein abundance in rich medium (DM_YEPD) from (*20*). DM is a corrected measure of noise in tagged-protein abundance that allows for direct comparisons of levels of noise among proteins without confounding factors such as protein abundance. Data on TATA box and gene expression responsiveness (“plasticity”) across environments were obtained from (*21*). Note that expression responsiveness is directly linked to the abundance of mRNA of the gene. However, protein noise integrates both transcription noise and translation noise. The presence of a TATA-box is known to influence transcription efficiency (*22*), which is directly related to noise in transcription (*23*). This may explain the stronger association between V_m in gene expression and protein expression noise for genes with a TATA-box (Supplementary Table 4).

Genes involved in stress response were identified as follows. Gene Ontologies were obtained from the Saccharomyces Genome Database (ftp://ftp.yeastgenome.org/yeast/data_download/literature_curation/orf_geneontology.tab). Stress related genes were identified as genes with GO Biological Process containing the term stress: response to osmotic stress, response to oxidative stress, response to salt stress and response to stress.

In order to estimate the *trans*-mutational target size (fraction of other genes in the genome that affect the expression of the focal gene) we obtained gene expression levels from 300 perturbations from (*24*). In this experiment, expression profiling was performed on more than 300 genetic perturbations in *S. cerevisiae*. We eliminated 23 experiments because they were either chemical perturbations or perturbations performed with titrable promoters and not gene deletions. For each gene in the database, we computed the probability of differential expression or *trans*-mutational target as the fraction of genetic perturbations that changed the expression level of the focal gene using *P*-values < 0.01 (*24*).

In an attempt to partition the variation in V_m across genes, we constructed a Generalized Linear Model (log-transformed GLM with gaussian link) in R (*11*) including *trans*-mutational target size, presence of a TATA-box and *cis*-mutational target size as predictive variables the. This analysis shows that both the size of the *trans*-mutational target and the presence of a TATA-box independently explain a significant fraction of the variance, while the *cis*-target size has no significant effect when the other two properties are considered. However, the relative importance of the TATA box and the *trans*-mutational target size cannot be ascertained using these models as collinearity makes the GLM (Type I) results dependent on the input order of these two variables. Nonetheless, GLM results are unequivocal in showing that part of the effect due to *trans*-mutational target size is completely independent from the effect due to TATA. The results of the analysis of deviance are presented in supplementary Table 2 and supplementary Table 3.

Finally, we investigated whether genes having specific transcription factors binding (TFB) sites were enriched among genes with significant expression changes across our mutation-accumulation lines. Genes bound by the same transcription factor are often co-regulated and this analysis may point to functionally related or coregulated genes varying similarly in expression in each line presumably due to their sharing of a *trans*-factor that has been disrupted. In order to do this, we computed the number of genes with significant variation across mutation-accumulation lines that contain at least one binding site in its promoter (85 transcription factors tested; only genes with known binding sites are included, n = 1202). The expected value is computed as the fraction of genes in the genome with a given TFB multiplied by the total number of genes with evidence for differential expression at P < 0.99. No enrichment is evident after a bonferroni correction for multiple testing, thus suggesting that the effects we observed cannot be linked to an overwhelmingly large effect due to the disruption of any transcription factor exclusively.

Supplementary Table 1: Genes sensitive to spontaneous mutations are enriched for TATA-containing genes. Distribution of the number of genes with significant differential expression among the four MA lines separated according to the presence or absence of a TATA box in the *cis* regulatory region. The genes that change in expression level among the four lines are significantly enriched for TATA-box containing genes (Fisher's exact test, $P = 2.5 \times 10^{-16}$). Only genes for which there is information on the presence or absence of a TATA-box are included.

	TATA-containing	TATA-less	Total
Significant changes	351	881	1232
No significant changes	363	1839	2202
Total	714	2720	3434

Supplementary Table 2:

Analysis of deviance (Gaussian, link: identity) of V_m

Model: $\log V_m$ = *trans-target size* + *TATA* + *cis-mutational target size*

	Deviance	Residual Df	Residual Deviance	F	*P*-value
Null model		408	820.42		
Trans-mutational target size	116.62	407	703.8	73.77	< 2.2E-16
TATA	59.39	406	644.4	37.57	2.10E-09
Cis-mutational target size	4.15	405	640.25	2.63	0.11

Supplementary Table 3:

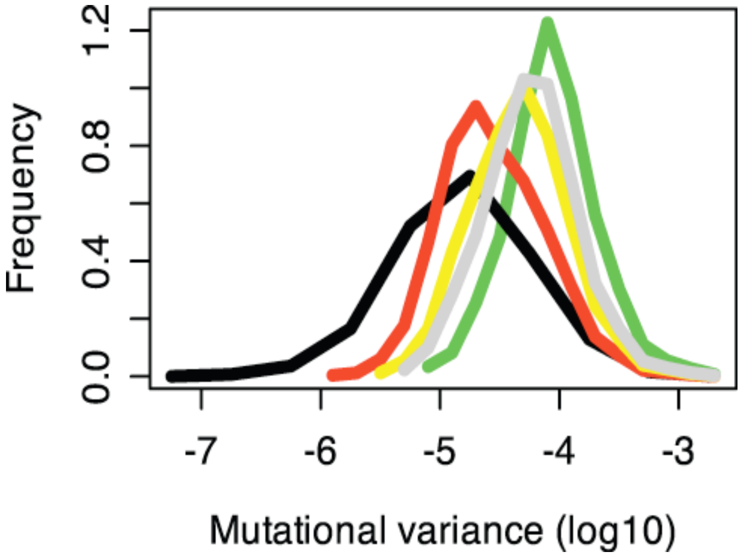
Analysis of deviance (Gaussian, link: identity) of V_m

Model: $\log V_m$ = *cis-mutational target size + TATA + trans-target size*

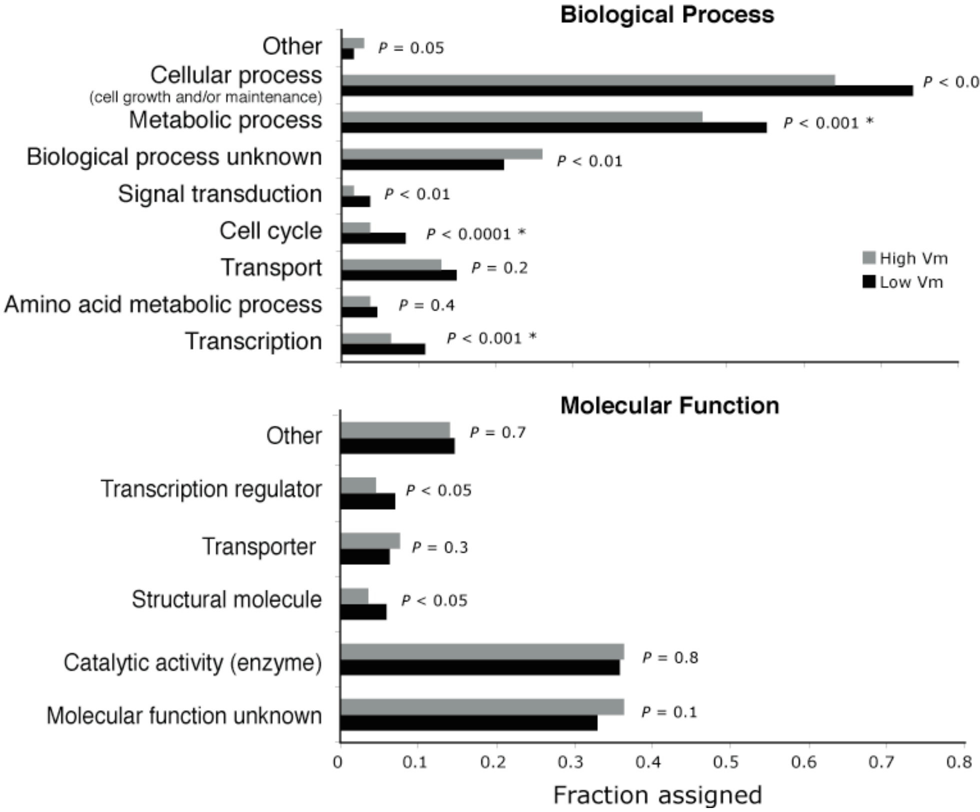
	Deviance	Residual Df	Residual Deviance	F	*P*-value
Null model		408	820.42		
cis-mutational target size	4.03	407	816.39	2.5477	0.1112
TATA	111.65	406	704.74	70.6245	7.41E-16
Trans-mutational target size	64.49	405	640.25	40.7942	4.65E-10

Supplementary Table 4. Associations between V_m, expression noise and plasticity for genes with and without a TATA box. Spearman rank correlations (ρ), P-values (P), and sample size (N) are shown.

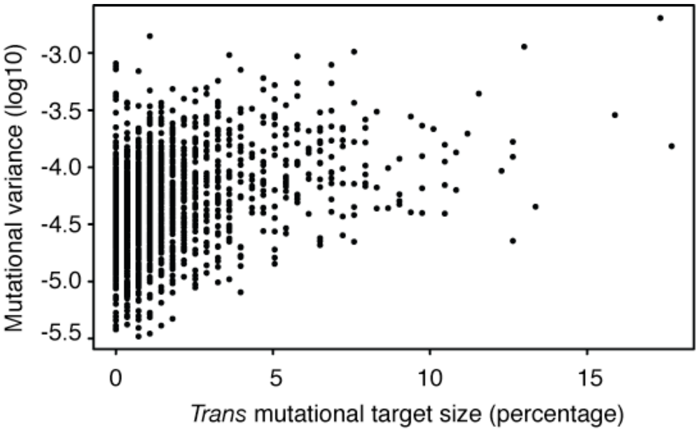
	V_m x plasticity	V_m x expression noise
TATA-containing	$\rho = 0.55$, $P = 2 \times 10^{-16}$ $N = 351$	$\rho = 0.44$, $P = 3 \times 10^{-8}$ $N = 146$
TATA-less	$\rho = 0.21$, $P = 7 \times 10^{-10}$ $N = 881$	$\rho = 0.06$, $P = 0.23$ $N = 368$



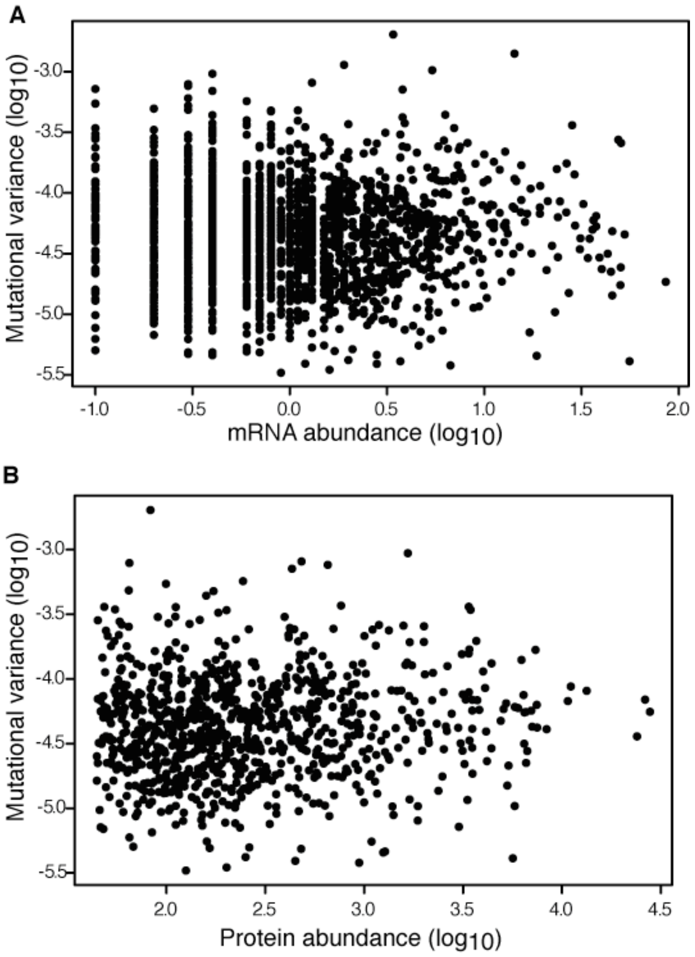
Supplementary Figure 1: Distribution of the V_m of gene expression at different threshold of significance (Bayesian Posterior Probability, BPP). Black: all genes, median: 1.5 × 10^{-5}; Red: BPP > 0.95, median : 2.6 × 10^{-5}; Yellow: BPP > 0.99, median : 4.7 × 10^{-5}; Grey: BPP > 0.995, median: 5.7 × 10^{-5}; Green: BPP > 0.999, median: 8.4 × 10^{-5}.



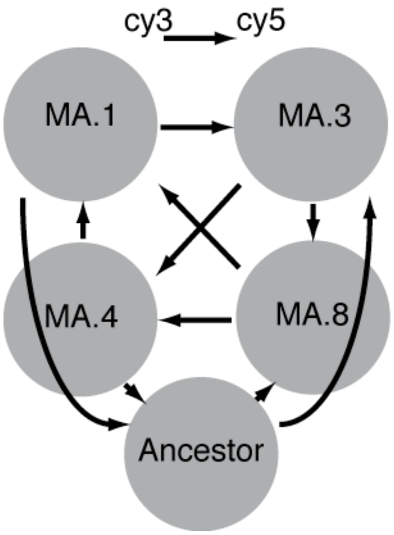
Supplementary Figure 2: Classes of genes associated with high and low V_m in gene expression. Genes with significant changes in gene expression among the four lines were separated in two groups representing the first half (low V_m) and second half (high V_m) of the ranked V_{ms}. P-values of the 2-sample test for equality of proportions are presented. * indicates significance after Bonferroni correction (at $\alpha = 0.05/15$).



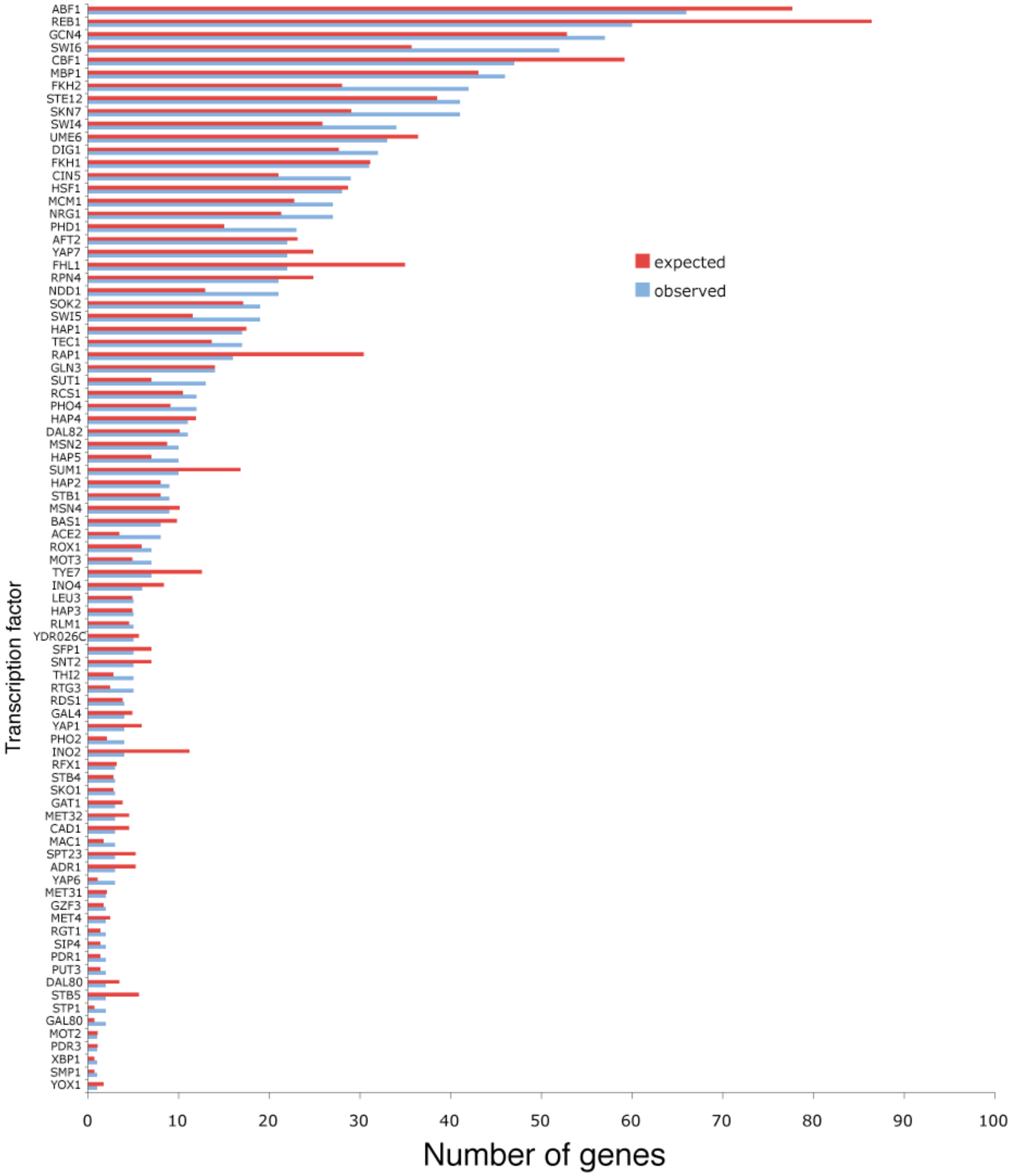
Supplementary Figure 3: Mutational variance of gene expression correlates with the size of the *trans*-mutational target. Raw data used in Fig. 2B are presented here.



Supplementary Figure 4: (**A**) Mutational variance is independent of gene mRNA abundance. Gene expression levels were obtained from (*19*) and estimated using SAGE. Spearman correlation, *r* = -0.02, *P* = 0.47. (**B**) Mutational variance is independent of protein abundance as measured through the fluorescence of labeled proteins (*20*). Spearman correlation, *r* = 0.06, *P* = 0.055.



Supplementary Figure 5: Experimental design of two-color competitive microarray comparisons. Arrows indicate direct comparisons of samples on an array. Every line was compared to every other line in a complete loop design.



Supplementary Figure 6: Enrichment of genes that change in expression level among the four lines for transcription factor binding sites (TFB). The observed values are computed as the number of genes containing at least one TFB in its promoter (only genes with known TFB are included, n = 1202) and the expected value is derived from the fraction of genes in the genome with a given TFB multiplied by the total number of genes observed. Significant at α = 0.05: INO2, P = 0.0315; ACE2, P = 0.0161; SUT1, P = 0.0233;RAP1, P = 0.0089;SWI5, P = 0.0283;NDD1, P = 0.0252; FHL1, P = 0.0281; PHD1, P = 0.0403;SKN7, P = 0.0265; FKH2, P = 0.0081;SWI6, P = 0.0063; REB1, P = 0.0045; No enrichment is significant after correction for multiple testing α = 0.05/85.

Supporting references

1. J. W. Thatcher, J. M. Shaw, W. J. Dickinson, *Proc. Natl. Acad. Sci. U S A* **95**, 253 (1998).
2. T. Ohta, *Nature* **246**, 96 (1973).
3. D. L. Hartl, A. G. Clark, *Principles of population genetics* (Sinauer Associates, Sunderland, MA, ed. 3rd, 1997), pp. xiii, 542.
4. J. Drake, *Proc. Nat. Acad. Sci. USA* **88**, 7160 (1997).
5. M. Lynch, J. S. Conery, *Science* **290**, 1151 (2000).
6. M. A. DePristo, M. M. Zilversmit, D. L. Hartl, *Gene* **378**, 19 (2006).
7. K. J. Verstrepen, A. Jansen, F. Lewitter, G. R. Fink, *Nat. Genet.* **37**, 986 (2005).
8. F. Diehl, S. Grahlmann, M. Beier, J. D. Hoheisel, *Nucleic Acids Res.* **29**, E38 (2001).
9. M. K. Kerr, G. A. Churchill, *Biostatistics* **2**, 183 (2001).
10. J. P. Townsend, J. W. Taylor, *Methods Enzymol.* **395**, 597 (2005).
11. R. Ihaka, R. Gentleman, *Journal of Computational and Graphical Statistics* **5**, 299 (1996).
12. G. Smyth, in *Bioinformatics and Computational Biology Solutions using R and Bioconductor* R. Gentleman, V. Carey, S. Dudoit, R. Irizarry, W. Huber, Eds. (Springer, New York, 2005).
13. G. Smyth, T. Speed, *Methods* **31**, 265 (2003).
14. J. P. Townsend, D. L. Hartl, *Genome Bio.l* **3**, RESEARCH0071 (2002).
15. C. D. Meiklejohn, J. P. Townsend, *Brief. Bioinform.* **6**, 318 (2005).
16. M. Lynch, B. Walsh, *Genetics and analysis of quantitative traits* (Sinauer, Sunderland, Ma., 1998), pp. xvi, 980.
17. D. R. Denver *et al.*, *Nat. Genet.* **37**, 544 (2005).
18. C. T. Harbison *et al.*, *Nature* **431**, 99 (2004).
19. F. C. Holstege *et al.*, *Cell* **95**, 717 (1998).
20. J. R. Newman *et al.*, *Nature* **441**, 840 (2006).
21. I. Tirosh, A. Weinberger, M. Carmi, N. Barkai, *Nat. Genet.* **38**, 830 (2006).
22. J. M. Raser, E. K. O'Shea, *Science* **304**, 1811 (2004).
23. J. M. Raser, E. K. O'Shea, *Science* **309**, 2010 (2005).
24. T. R. Hughes *et al.*, *Cell* **102**, 109 (2000).