Supplementary Material

Evolution of gene sequence and gene expression are not correlated in yeast

Itay Tirosh[1] and Naama Barkai[2]

[1]Department of Molecular Genetics and [2]Department of Physics of Complex Systems, Weizmann Institute of Science, Rehovot 76100, Israel
Corresponding author: Barkai, N. (naama.barkai@weizmann.ac.il).

Supplementary Methods

Expression Divergence

Datasets of yeast expression divergence were taken from seven independent studies. Six of these studies were previously published [1–6]. Another study examined the response of three yeast species (*S. cerevisiae*, *S. paradoxus* and *S. mikatae*) to mating pheromone and is now under review [7]. Three of these studies defined a continuous measure for expression divergence [1,3,5]; we normalized these datasets by subtracting their means and dividing by their standard deviations. Another study defined a set of conserved and a set of divergent genes [7]; divergent and conserved genes were given expression divergence values of 1 and -1, respectively. Finally, the three remaining studies defined only a set of divergent genes [2,4,6]; these genes were given expression divergence values of 1 and the remaining genes were given negative expression divergence values which were chosen such that the average of each dataset is zero. The combined measure of expression divergence was generated by averaging these seven datasets over all genes that had a value in at least 5 datasets.

These seven datasets employed *S. cerevisiae* microarrays to measure expression of different strains *of S. cerevisiae* or different Saccharomyces species. However, we note that lower hybridization due to sequence mismatches could not account for our observations: in the four datasets that examined expression levels, lower hybridization signals would lead to higher expression divergence of genes with more coding-sequence changes, and thus artificially inflate the correlation between coding-sequence and expression divergence. The other three studies examined expression ratios, and therefore lower hybridization should not influence their results of hybridization ratios.

Expression divergence of other organisms included bacteria [8], flies [9], human versus chimpanzee [10] and human versus mouse [11]; these datasets were processed as previously described [3,8].

Sequence Divergence

To compare expression divergence with coding-sequence divergence, we used the rate of non-synonymous substitutions (*Ka*) as a measure of sequence divergence. *Ka* values are frequently normalized by the rate of synonymous substitutions (*Ks*) to get an estimation of the strength of purifying selection. However, the expression divergence data reflects total divergence and cannot be normalized in such a way, and thus we reasoned that the unnormalized *Ka* values are more appropriate in this analysis. Yet, replacing the *Ka* values with normalized *Ka/Ks* ratios gives similar results (see below).

Ka values were taken from Wall *et al.* [12] which compared 4 yeast species; To obtain approximately normal distribution the data was $\log_2$ transformed [12], mean subtracted and std normalized. To verify that our results do not depend on the exact method for determining coding-sequence divergence we considered both Ka and Ka/Ks as defined by:

(i) Wall *et al.* [12] which used multiple alignments of *S. cerevisiae*, *S. paradoxus*, *S. mikatae* and *S. bayanus*,

(ii) Kellis *et al.* [13] which used pairwise alignments of *S. cerevisiae* with either one of the three other species (three different datasets).

(iii) We performed a similar analysis of the four yeast species using PAML codonml with the global clock assumption.

None of these measures had a significant positive correlation with any of the expression divergence datasets.

We also examined the sequence divergence between two strains of *S. cerevisiae* (S288c and YJM789) [14]. Coding-sequence divergence (1-%identity in amino acid sequence), was not significantly positively correlated with any of the datasets of expression divergence.

Coding-sequence divergence (Ka) of human-chimp was taken from ref. 10 and human-mouse from the Ensembl database (http://www.ensembl.org/). *Ka* for fly species and bacteria were calculated by the maximum-likelihood PAML codonml with the global clock assumption. Flies: we compared the three *Drosophila* species *D. melanogaster*, *D. simulans* and *D. yakuba*. Sequences were taken from FlyBase (http://flybase.bio.indiana.edu/) and from the geneID predictions [15] (http://genome.imim.es/genepredictions/index.html). Bacteria: we compared *Escherichia coli K12*, *Shigella sonnei* and *Shigella flexneril*, taken from (ftp://ftp.ncbi.nlm.nih.gov/genomes/Bacteria/) using the tree topology (*E. coli*, (*S. sonnei*, *S. flexneri*)). In each case, orthologs were found by reciprocal blastp [16] best matches and aligned with clustalw. To avoid alignment mistakes we excluded all alignments with high percentage of gaps (>10%).

Additional genetic properties

We examined the Pearson correlations of coding-sequence and expression divergence with eight genetic properties. Only genes with values of both sequence and expression divergence (combined dataset) were used. To approximate normal distribution, skewed properties were log-transformed ($x' = \log_2(x+k)$), where k was chosen to maximize the correlation with a linear fit at a normal probability plot. To control for the dependency among these properties we also used partial correlations, in which the correlation between sequence or expression divergence and a single property is examined while all seven other properties are controlled for, and multiple regression analysis that simultaneously estimates the influence of all factors (Fig. S1). Using the partial correlations and multiple regression analysis, we also examined the correlation between coding-sequence and expression divergence after controlling for specific factors (TATA or essentiality) or all factors combined. In all cases, coding-sequence and expression divergence were either not correlated or significantly negatively correlated.

Ks

We used the adjusted Ks from Wall *et al.*, which is normalized to eliminate the effect of codon usage on synonymous substitution rate [12].

Essentiality

Essentiality was defined as one minus the minimal deletion mutant's growth rate among five growth media [17]; essential genes [12] were given an essentiality value of one. Growth rates of heterozygote deletion strains were taken from Deutschbauer *et al.* [18]. The 500 non-essential genes with lowest growth rates were compared to all other non-essential genes.

Protein–protein interactions

Comprehensive datasets of interactions from multiple sources were taken from Yu *et al.* [19]. and von Mering *et al.* [20]. Both datasets give highly significant correlations with coding-sequence divergence (r<-0.2 and $p<10^{-30}$ in both cases) but lower and more variable correlations with expression divergence (r=-0.15,-0.05 for Yu *et al.* and von Mering *et al.*, respectively). The correlations with expression divergence were further reduced in the multiple regression analysis and were no longer significant (Fig. S1). We combine the two datasets by averaging them and present the results for the combined dataset in Figure 1 and Figure S1.

Protein abundance

Protein abundance was taken from Ghaemmaghami *et al.* [21]. Similar results were obtained with either mRNA abundance [22] or Codon Adaptation Index (CAI) [23].

Promoter elements

TATA-boxes were defined as in ref. 3; TATA-containing and TATA-less genes were given values of 1 and -1, respectively. Cis-target size was defined by the number of transcription factor binding sites from Harbison *et al.* [24] (dataset of p<0.001 and no conservation criteria).

Trans-target size

Following Landry *et al.* [5] we defined Trans-target size as the number of deletion mutants in which a gene is significantly (p<0.05) regulated [25].

Functional sites

The number of PROSITE [26] annotations of each yeast gene.

Unicellular versus multicellular organisms

We found several differences between divergence in unicellular and multicellular organisms. First, coding-sequence and expression divergence were found to be positively correlated in several studies of multicellular organisms [10, 27–32], although others found no significant correlations [27, 33, 34]. Second, while TATA box is positively correlated only with expression divergence in yeast, it is also positively correlated with coding-sequence divergence in multicellular organisms: we compared the coding-sequence divergence datasets of human-mouse and *Drosophila* species, as described above, with predictions of TATA-containing genes [3]; TATA-containing genes had significantly higher sequence divergence than TATA-less genes in both cases ($p<10^{-7}$). Third, while essentiality is correlated both with coding-sequence divergence and expression divergence in yeast, it is only significantly correlated with coding-sequence divergence in mammals: we estimated essentiality of mouse genes as described in Liao *et al.* [30] and compared the human-mouse divergence of essential and non-essential genes. Only coding-sequence divergence was significantly associated with essentiality ($p<0.05$).

These discrepancies may indicate that it is difficult to estimate expression divergence in multicellular organisms due to the diversity of cell types, each having unique expression patterns. Alternatively, this diversity of cell types may lead to genuine differences between the evolution of unicellular and multicellular organisms that are reflected by these differential associations. The evolutionary forces acting on each gene depend on the tissues in which that gene is active [10, 35]. Each tissue may impose different evolutionary constraints and these may similarly influence the rates of coding-sequence and gene expression divergence, thus leading to a positive correlation. This effect could also possibly account for the differential associations with TATA and essentiality: (i) TATA-containing genes tend to be tissue-specific and could thus be subjected to weaker negative selection and diverge faster. (ii) Essentiality could constrain the expression of a gene only in the tissue in which it is important and this could be overlooked when examining few cell-types.

References

1 Townsend, J.P. *et al.* (2003) Population genetic variation in genome-wide gene expression. *Mol. Biol. Evol.* 20, 955–963

2 Brem, R.B. *et al.* (2002) Genetic dissection of transcriptional regulation in budding yeast. *Science* 296, 752–755

3 Tirosh, I. *et al.* (2006) A genetic signature of interspecies variations in gene expression. *Nat. Genet.* 38, 830–834

4 Landry, C.R. *et al.* (2006) Genome-wide scan reveals that genetic variation for transcriptional plasticity in yeast is biased towards multi-copy and dispensable genes. *Gene* 366, 343–351

5 Landry, C.R. *et al.* (2007) Genetic Properties Influencing the Evolvability of Gene Expression. *Science* 317, 118–121

6 Fay, J.C. *et al.* (2004) Population genetic variation in gene expression is associated with phenotypic variation in Saccharomyces cerevisiae. *Genome Biol.* 5, R26

7 Tirosh, I. *et al.* On the relation between promoter divergence and gene expression evolution. *Mol. Syst. Biol.* (in press)

8 Le Gall, T. *et al.* (2005) Selection-driven transcriptome polymorphism in *Escherichia coli*/Shigella species. *Genome Res.* 15, 260–268

9 Ranz, J.M. *et al.* (2003) Sex-dependent gene expression and evolution of the Drosophila transcriptome. *Science* 300, 1742–1745

10 Khaitovich, P. *et al.* (2005) Parallel patterns of evolution in the genomes and transcriptomes of humans and chimpanzees. *Science* 309, 1850–1854

11 Su, A.I. *et al.* (2004) A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl. Acad. Sci. U. S. A.* 101, 6062–6067

12 Wall, D.P. *et al.* (2005) Functional genomic analysis of the rates of protein evolution. *Proc. Natl. Acad. Sci. U. S. A.* 102, 5483–5488

13 Kellis, M. *et al.* (2003) Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* 423, 241–254

14 Wei, W. *et al.* (2007) Genome sequencing and comparative analysis of Saccharomyces cerevisiae strain YJM789. *Proc. Natl. Acad. Sci. U. S. A.* 104, 12825–12830

15 Parra, G. *et al.* (2000) GeneID in Drosophila. *Genome Res.* 10, 511–515

16 Yuan, G.C. *et al.* (2005) Genome-scale identification of nucleosome positions in *S. cerevisiae*. *Science* 309, 626–630

17 Steinmetz, L.M. *et al.* (2002) Systematic screen for human disease genes in yeast. *Nat. Genet.* 31, 400–404

18 Deutschbauer, A.M. *et al.* (2005) Mechanisms of haploinsufficiency revealed by genome-wide profiling in yeast. *Genetics* 169, 1915–1925

19 Borneman, A.R. *et al.* (2007) Divergence of transcription factor binding sites across related yeast species. *Science* 317, 815–819

20 von Mering, C. *et al.* (2002) Comparative assessment of large-scale data sets of protein-protein interactions. *Nature* 417, 399–403

21 Ghaemmaghami, S. *et al.* (2003) Global analysis of protein expression in yeast. *Nature* 425, 737–741

22 Beyer, A. *et al.* (2004) Post-transcriptional expression regulation in the yeast Saccharomyces cerevisiae on a genomic scale. *Mol. Cell. Proteomics* 3, 1083–1092

23 Sharp, P.M. and Li, W.H. (1987) The codon Adaptation Index–a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res.* 15, 1281–1295

24 Harbison, C.T. *et al.* (2004) Transcriptional regulatory code of a eukaryotic genome. *Nature* 431, 99–104
25 Roberts, C.J. *et al.* (2000) Signaling and circuitry of multiple MAPK pathways revealed by a matrix of global gene expression profiles. *Science* 287, 873–880
26 Hulo, N. *et al.* (2006) The PROSITE database. *Nucleic Acids Res.* 34, D227–D230
27 Jordan, I.K. *et al.* (2005) Evolutionary significance of gene expression divergence. *Gene* 345, 119–126
28 Sartor, M.A. *et al.* (2006) A new method to remove hybridization bias for interspecies comparison of global gene expression profiles uncovers an association between mRNA sequence divergence and differential gene expression in *Xenopus*. *Nucleic Acids Res.* 34, 185–200
29 Gibson, G. *et al.* (2004) Extensive sex-specific nonadditivity of gene expression in *Drosophila melanogaster*. *Genetics* 167, 1791–1799
30 Liao, B.Y. and Zhang, J. (2006) Evolutionary conservation of expression profiles between human and mouse orthologous genes. *Mol. Biol. Evol.* 23, 530–540
31 Lemos, B. *et al.* (2005) Evolution of proteins and gene expression levels are coupled in Drosophila and are independently associated with mRNA abundance, protein length, and number of protein-protein interactions. *Mol. Biol. Evol.* 22, 1345–1354
32 Miller, W. *et al.* (2004) Comparative genomics. *Annu. Rev. Genomics Hum. Genet.* 5, 15–56
33 Matise, T.C. *et al.* (2003) A 3.9-centimorgan-resolution human single-nucleotide polymorphism linkage map and screening set. *Am. J. Hum. Genet.* 73, 271–284
34 Yanai, I. *et al.* (2004) Incongruent expression profiles between human and mouse orthologous genes suggest widespread neutral evolution of transcription control. *OMICS* 8, 15–24
35 Gu, X. and Su, Z. (2007) Tissue-driven hypothesis of genomic evolution and sequence-expression correlations. *Proc. Natl. Acad. Sci. U. S. A.* 104, 2779–2784

Supplementary Tables

Table S1. Correlation between essentiality and different datasets of expression divergence

Dataset	Correlation[a]	P-value
ED1	-0.1419	0
ED2	-0.087	0.03
ED3	-0.10162	$9*10^{-12}$
ED4	-0.04566	0.0003
ED5	-0.06051	$2*10^{-6}$
ED6	-0.06245	$8*10^{-7}$
ED7 (MA[b])	-0.02184	0.11
Combined ED	-0.1467	$2*10^{-25}$

a – the correlation of essentiality with sequence divergence (Ka) is -0.25

b - Mutation Accumulation

Table S2. Correlation between sequence and expression divergence in additional organisms (see Supplementary Methods for details)

Organisms	Correlation
Bacteria (E.coli – Shigella)	-0.054
Flies (*D. melanogaster* – D. simulans)	0.14
Mammals (human – mouse)	0.23
Mammals (human – chimp)	0.087

Supplementary Figures

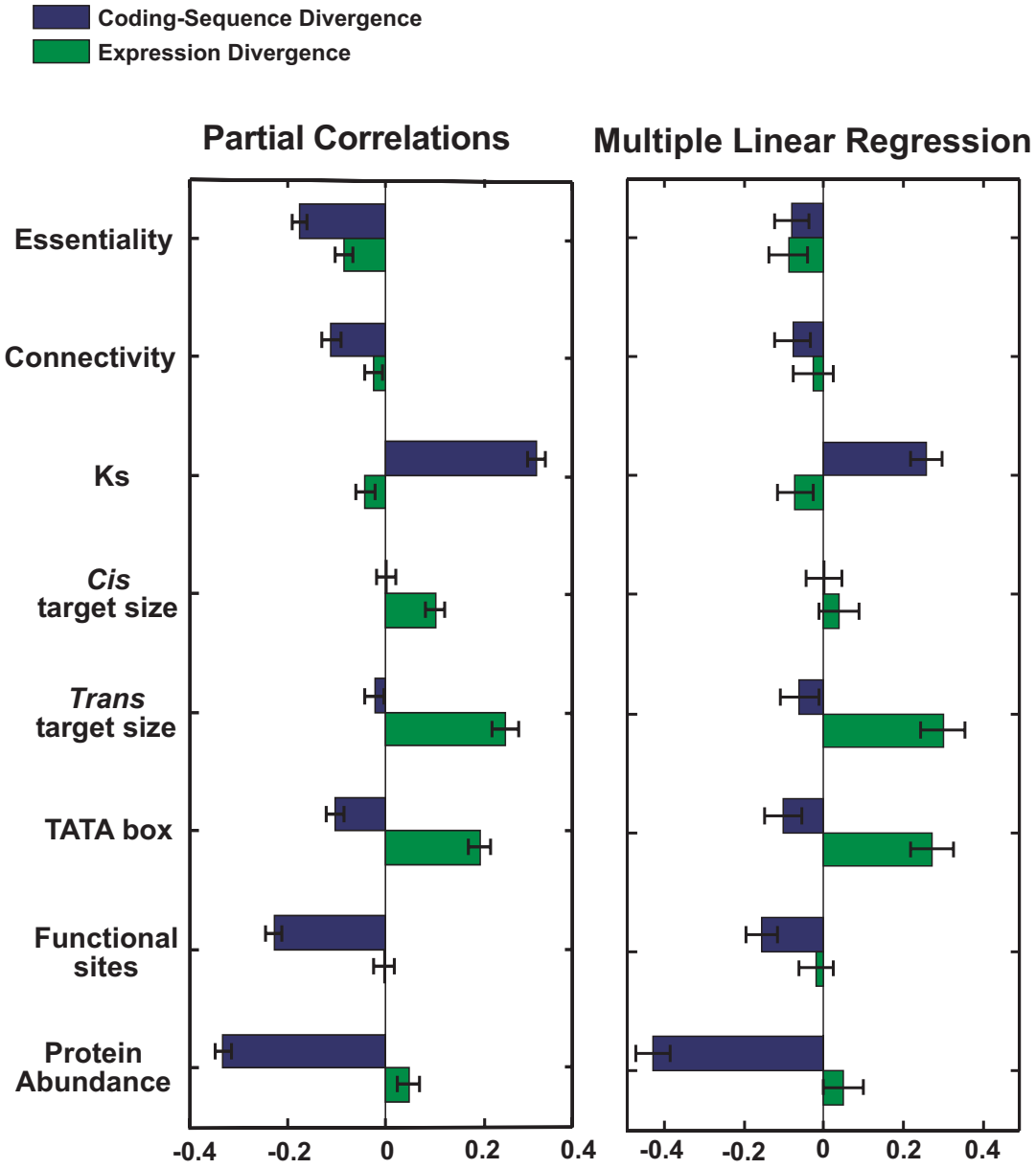

Figure S1. Partial correlation and multiple regression analysis of genetic properties influencing coding-sequence and gene expression divergence. The effects of 8 genetic properties on coding-sequence divergence (blue) or expression divergence (green; combined dataset) were estimated by partial correlations (each property was controlled for the effects of all other properties) or multiple linear regression of all 8 properties.