nature genetics

Differential translation efficiency of orthologous genes is involved in phenotypic divergence of yeast species

Orna Man[1,2] & Yitzhak Pilpel[1]

A major challenge in comparative genomics is to understand how phenotypic differences between species are encoded in their genomes. Phenotypic divergence may result from differential transcription of orthologous genes, yet less is known about the involvement of differential translation regulation in species phenotypic divergence. In order to assess translation effects on divergence, we analyzed ~2,800 orthologous genes in nine yeast genomes. For each gene in each species, we predicted translation efficiency, using a measure of the adaptation of its codons to the organism's tRNA pool. Mining this data set, we found hundreds of genes and gene modules with correlated patterns of translational efficiency across the species. One signal encompassed entire modules that are either needed for oxidative respiration or fermentation and are efficiently translated in aerobic or anaerobic species, respectively. In addition, the efficiency of translation of the mRNA splicing machinery strongly correlates with the number of introns in the various genomes. Altogether, we found extensive selection on synonymous codon usage that modulates translation according to gene function and organism phenotype. We conclude that, like factors such as transcription regulation, translation efficiency affects and is affected by the process of species divergence.

Differences in gene content among diverging organisms often have a role in their phenotypic divergence[1]. Yet it is possible that even shared genes are involved in phenotypic diversity, provided that they are regulated differently across species (see refs. 2–4), as in the role of differential transcription regulation of ribosomal proteins across yeast species, for example[3]. To complement this picture, we examined the role of differential translational efficiency in phenotypic divergence among ten fully sequenced yeast species.

Translational efficiency of genes is commonly gauged by the extent of their adaptation to the tRNA pool[5,6]. It has been observed in several species that *in vivo* concentration of a tRNA bearing a certain anticodon correlates with the number of gene copies coding for this tRNA (for example, in *S. cerevisiae*, Pearson's $r = 0.91$; ref. 7). This facilitates the investigation of the tRNA pools of any fully sequenced species. Using a hidden Markov model (HMM) approach[8], we obtained reliable tRNA gene copy numbers for nine of the ten fully sequenced yeast species (**Supplementary Table 1** online).

We first investigated the evolution of tRNA repertoires among the species. The size of the tRNA repertoire ranges from 133 genes

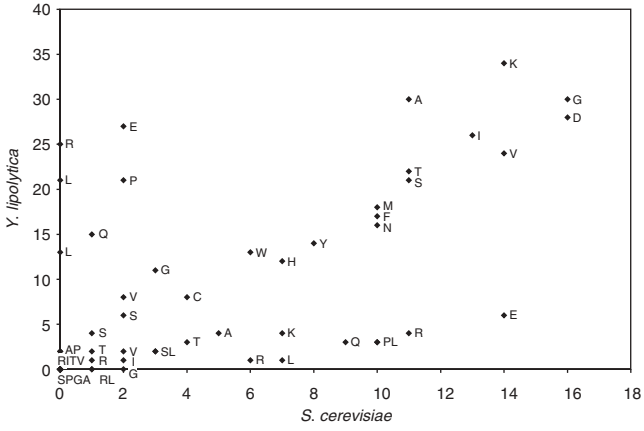

Figure 1 Comparison of the tRNA gene repertoires of *S. cerevisiae* and *Y. lipolytica*. The gene copy numbers for each tRNA and each species were determined from the whole-genome sequence using an HMM-based approach[8] (see Methods). For each anticodon, the gene copy number in *S. cerevisiae* (*x* axis) and *Y. lipolytica* (*y* axis) is shown. The points are annotated with the one-letter symbol of the amino acid the anticodon translates. The Pearson correlation between the two tRNA gene repertoires is 0.58 ($P = 3.48 \times 10^{-6}$). The balance-swaps among anticodons translating glutamic acid (E), proline (P), glutamine (Q), arginine (R) and leucine (L) can be seen clearly. These switches in dominance of tRNA species are accompanied by corresponding changes in the usage of the codons they translate. For example, in *S. cerevisiae*, there are nine genes encoding a tRNA bearing the anticodon for CAA (encoding glutamine) and only one tRNA gene for the anticodon of CAG, the second codon for glutamine. Accordingly, *S. cerevisiae* uses CAA to encode glutamine 79,139 times (69% of the time) and uses CAG only 36,234 times. In *Y. lipolytica*, on the other hand, there are only three genes for tRNAs bearing the anticodon for CAA and 15 genes for the tRNAs bearing the anticodon for CAG. *Y. lipolytica* uses CAA to encode glutamine only 30,507 times (23% of the time), whereas CAG is used 100,228 times.

[1]Department of Molecular Genetics and [2]Department of Structural Biology, The Weizmann Institute of Science, Rehovot 76100, Israel. Correspondence should be addressed to Y.P. (pilpel@weizmann.ac.il).

Received 20 September 2006; accepted 3 January 2007; published online 4 February 2007; corrected after print 28 March 2007; doi:10.1038/ng1967

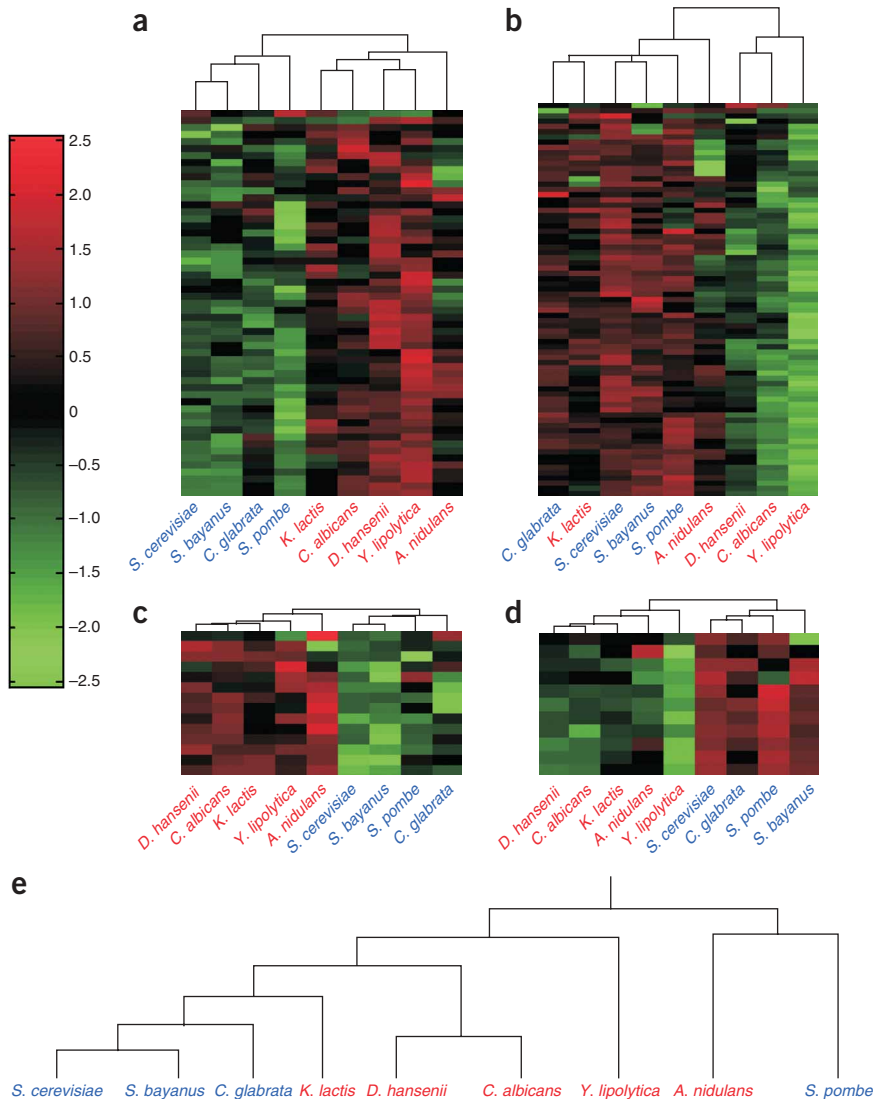

Figure 2 The translation efficiency profiles of mitochondrial and cytosolic ribosomal proteins, glycolysis and the tricarboxylic acid cycle show coherent patterns. (**a**–**d**) The relative translational efficiencies across species (normalized tAI; see Methods) of the mitochondrial ribosomal proteins (MRPs, **a**), cytosolic ribosomal proteins (CRPs, **b**), genes of the tricarboxylic acid (TCA) cycle (**c**) and glycolysis genes (**d**) are shown on heat maps. A color bar indicates the relationship between colors and values. Note that for **b**–**d**, the range of values is only from –2 to 2. Both rows (genes) and columns (species) have been sorted by average linkage hierarchical clustering, with euclidean distance as a distance measure. The clustering of the species is indicated by dendrograms. (**e**) Topology of the phylogenetic tree of the species analyzed based on 18S rRNA[29]. The branch lengths are not proportional to time. In all panels, aerobic species are colored in red, and anaerobic species are colored in blue. The translation efficiency profiles of MRP, TCA cycle and glycolysis genes segregate between the aerobic and anaerobic species.

(*C. albicans*) to 510 genes (*Y. lipolytica*). Despite this wide range, the distribution of gene copy numbers among the different anticodons is highly correlated among the species (**Supplementary Table 1**). This extremely slow evolution of the tRNA repertoire is expected, as dominance shifts among synonymous anticodons may affect translation efficiency of many genes in the genome. We conclude that codon usage in these species evolved against an almost invariant tRNA pool. However, the correlations between the tRNA repertoires of *A. nidulans* and *Y. lipolytica* and those of the other species is much lower than among the rest of the species pairs. Closer examination (**Fig. 1**) uncovers a minority of outlying anticodons that represent cases of dominance shifts between synonymous anticodons. For example, whereas in *Y. lipolytica* (and in *A. nidulans*), the codon CAG for glutamine corresponds to the most abundant tRNA among the tRNAs bearing this amino acid, in *S. cerevisiae* (and in the rest of the species, except *A. gossypii*), the synonymous codon CAA corresponds to the tRNA with the highest copy number. Owing to the pleiotropic effects of changes in the tRNA pool on cellular protein concentrations, we expect that the unusual decoding seen in *Y. lipolytica* and *A. nidulans* arose very gradually, perhaps through an intermediate stage of redundancy, where the old and new major tRNAs coexisted in similar proportions during the time when the coding genes changed their codon preference. The identification of such intermediate redundant stages may require the sequencing of additional genomes. This process may be similar to the dominance switch observed in the promoters of the ribosomal protein genes in yeast that switched between alternative regulatory regimes through a redundant intermediate stage[9].

Although the tRNA pool largely evolves very slowly, it is possible that individual genes and gene modules evolved different extents of adaptation to that pool in the different species. To enable a comparison of translational efficiency of genes among the species, we constructed a matrix of ~2,800 orthologous groups (**Supplementary Table 2** online). In this matrix, the i,jth element contains the inferred translational efficiency of gene i in species j, as calculated by the tRNA adaptation index (tAI[5]), an index based on the availability of each of the tRNAs (see Methods). tRNA availability was approximated by the gene copy number of each tRNA in a procedure that also incorporates codon-anticodon wobble interactions (see the **Supplementary Note** online for discussion of the use of the tAI versus the codon adaptation index[6]). Several lines of evidence indicate that the tAI-based translation efficiency values are biologically significant. First, in *S. cerevisiae*, we found good correlation ($r = 0.63$, $P < 1 \times 10^{-363}$) between the tAI of a gene and its experimentally measured[10] protein expression level (**Supplementary Fig. 1** online). This correlation remains highly significant even when controlling for the mRNA level that could serve as a confounder (**Supplementary Fig. 1**), indicating that among genes with similar transcript levels, higher tAI often corresponds to higher protein abundance. In addition, we found, as expected[11,12], that physically interacting proteins in *S. cerevisiae* show similar translation efficiency profiles not only in this species but also across the entire set of analyzed species (**Supplementary Fig. 2** online). This also implies conservation of many of the protein interactions.

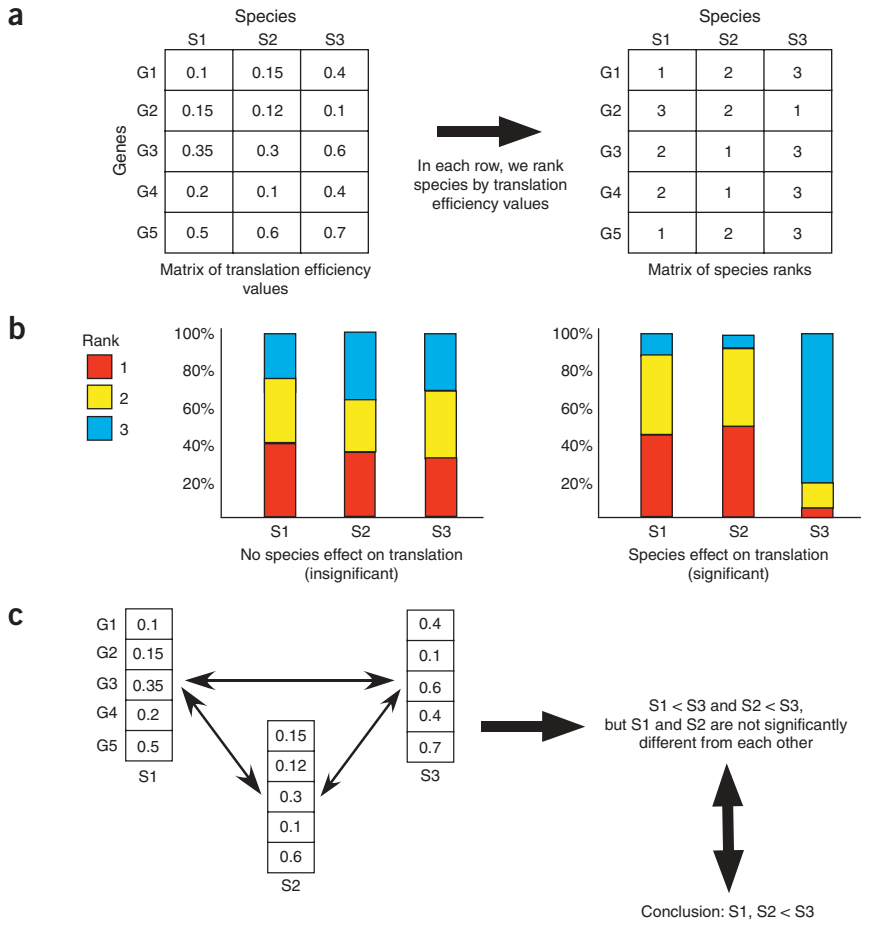

Figure 3 Scheme for testing for a species effect on translation efficiency for a group of genes. The presented scheme can be applied to any group of genes (in our study, we applied it to clusters of genes derived from the hierarchical clustering procedure, as well as groups of genes associated with various GO[14] categories). (**a**) The presence of a species effect on translation efficiency is tested using the Friedman test[13]. This test is based on the ranking of the values in each row; that is, for each gene, the species are ranked according to how efficiently they translate it (approximated by tAI). The matrix of ranks is then summarized and its significance is assessed. (**b**) Stacked histograms of species ranks representing two hypothetical gene sets: one in which there is no species effect on translation efficiency (the ranks are almost equally divided among the species; left) and one (corresponding to the matrix in **a**) where there is such an effect (the third species has an excess of high ranks; right). (**c**) Once a species effect on translation has been established, an attempt is made to discover the source of the signal of differential efficiency using *post hoc* tests. For this purpose, the translation efficiency values for the group of genes in question are compared for all possible pairs of species (see Methods). The conclusion that is drawn from these pairwise comparisons is presented as a species stratification.

Given this comparative translation efficiency matrix, we could examine relationships between gene functions and lifestyle properties of the different yeasts. A recent analysis established a connection among yeast species between aerobic versus anaerobic preference and transcription regulation of the mitochondrial versus the cytosolic ribosomal proteins (MRPs and CRPs, respectively)[3]. It was found that the aerobic yeasts maintain the capacity to coregulate the two types of ribosomes, whereas the anaerobic species lost it. This motivated us to examine the translation efficiency of the MRPs and CRPs in the aerobic and anaerobic species in our data set. We found that each of these two groups of genes showed a markedly coherent, yet markedly different, pattern of relative translational efficiencies across the species (**Fig. 2a**,**b**). Notably, the MRPs showed the highest translational efficiency in the five aerobic species (**Fig. 2a**). This probably reflects the reduced need of facultative anaerobes for MRPs, which translate components of the pathways of oxidative energy metabolism. We also checked whether translation efficiency of enzymes that are either needed for fermentation or respiration segregate according to the anaerobic or aerobic lifestyle of the species. Specifically, we examined the tricarboxylic acid (TCA) cycle and the glycolytic genes (**Fig. 2c**,**d**). As expected, we found the glycolytic enzymes to have maximal translation efficiency in the anaerobes, and the TCA cycle genes show the highest translational efficiency in the aerobes. These results cannot be simply explained by species phylogeny, as in the species tree, both aerobes and anaerobes are not monophyletic (**Fig. 2e**). Specifically, *S. pombe* and the rest of the anaerobic species seem to have converged upon similar translation efficiency profiles of the above genes (**Fig. 2a–d**). In general, clustering the species in the matrix results in only limited resemblance to the underlying species tree (**Supplementary Fig. 3** online), suggesting that evolution of the tAI values reflects more than evolutionary drift.

We next turned to cluster all the genes in the translation efficiency matrix using hierarchical clustering and partitioned the genes to 40 clusters (see Methods and **Supplementary Fig. 3** and **Supplementary Table 2**). We used the Friedman test[13] (**Fig. 3**), a nonparametric analog of the two-way analysis of variance (ANOVA), to test, in each cluster, the null hypothesis that there are no differences in the translational efficiency of orthologous genes (rows) across species (columns). We found that in all 40 clusters, there is a species effect on relative translational efficiencies (the worst P value among the 40 clusters was 7×10^{-5}). We then performed post-hoc tests (**Fig. 3**) using the Wilcoxon signed rank test[13] in order to find all pairs of species for which the genes of the cluster differ significantly in their translational efficiencies. In all clusters, we were able to stratify the species into at least two groups that differed in translational efficiency. The complete results of the Friedman analysis and *post hoc* tests can be found on our project website (see URL in Methods).

In an effort to shed light on phenotypic differences that might be implied by the species stratification in each cluster, we looked for enrichments of terms from the Gene Ontology (GO) database[14] in each of the clusters. We found such enrichments in 22 of the clusters, including the pathways and modules shown in **Figure 2** (**Supplementary Table 2**). **Supplementary Figure 3** shows representative clusters, along with dendrograms depicting similarity between species using the tAI of the genes in each cluster. Here, too, it is apparent that using particular gene sets for species clustering results in significant distortions relative to the phylogenetic species tree. This indicates that translation selection does not evolve merely by genetic drift and

Table 1 Species stratification patterns according to translational efficiencies for various GO categories

Species stratification	GO category	Number of profiles	Friedman test *P*-value
S. cerevisiae, *S. bayanus*, *S. pombe* < *C. glabrata* < *K. lactis*, *D. hansenii*, *C. albicans*, *A. nidulans* < *Y. lipolytica*	Aerobic respiration[a]	50	$<1.11 \times 10^{-16}$
	Mitochondrion[a]	607	$<1.11 \times 10^{-16}$
S. pombe < *S. cerevisiae*, *S. bayanus* < *C. glabrata* < *A. nidulans* < *K. lactis*, *C. albicans* < *D. hansenii* < *Y. lipolytica*	Mitochondrial ribosome[a]	55	$<1.11 \times 10^{-16}$
S. bayanus < *S. cerevisiae*, *S. pombe* < *C. glabrata* < *D. hansenii*, *Y. lipolytica* < *C. albicans* < *K. lactis*, *A. nidulans*	Oxidative phosphorylation[a]	24	$<1.11 \times 10^{-16}$
S. cerevisiae, *S. bayanus*, *S. pombe* < *C. glabrata* < *D. hansenii*, *C. albicans*, *Y. lipolytica*, *A. nidulans* < *K. lactis*	Mitochondrial electron transport chain[a]	17	6.22×10^{-15}
	Respiratory chain complex III (*sensu* Eukaryota)[a]	7	1.41×10^{-6}
S. cerevisiae, *S. bayanus* < *C. glabrata*, *S. pombe* < *K. lactis*, *D. hansenii*, *C. albicans*, *Y. lipolytica*, *A. nidulans*	Tricarboxylic acid cycle[a]	14	5.87×10^{-11}
Y. lipolytica < *C. albicans* < *D. hansenii* < *K. lactis*, *A. nidulans* < *S. bayanus*, *C. glabrata*, *S. pombe* < *S. cerevisiae*	Cytosolic ribosome (*sensu* Eukaryota)	96	$<1.11 \times 10^{-16}$
Y. lipolytica < *K. lactis*, *D. hansenii*, *C. albicans*, *A. nidulans* < *S. cerevisiae*, *S. bayanus*, *C. glabrata*, *S. Pombe*	Glycolysis[a]	11	9.28×10^{-11}
S. cerevisiae, *S. bayanus*, *C. glabrata*, *K. lactis*, *D. hansenii*, *C. albicans*, *Y. lipolytica*, *S. pombe* < *A. nidulans*	Spliceosome complex[a]	41	4.41×10^{-8}
S. cerevisiae, *S. bayanus*, *C. glabrata*, *K. lactis*, *D. hansenii*, *C. albicans*, *Y. lipolytica* < *S. pombe* < *A. nidulans*	Nuclear mRNA splicing, via spliceosome[a]	52	7.83×10^{-8}
	snRNP U1[a]	12	1.76×10^{-5}
S. cerevisiae, *S. bayanus*, *C. glabrata*, *K. lactis*, *D. hansenii*, *C. albicans*, *Y. lipolytica* < *A. nidulans*, *S. pombe*	mRNA processing[a]	79	4.32×10^{-7}
S. cerevisiae, *S. bayanus*, *C. glabrata*, *K. lactis*, *D. hansenii*, *C. albicans*, *A. nidulans*, *S. pombe* < *Y. lipolytica*	Organic acid metabolism[a]	200	0.0024
Y. lipolytica < *K. lactis*, *D. hansenii*, *C. albicans* < *C. glabrata* < *S. cerevisiae*, *S. bayanus*, *A. nidulans*, *S. pombe*	M phase	126	$<1.11 \times 10^{-16}$
	Cell cycle	216	$<1.11 \times 10^{-16}$
A. nidulans, *S. pombe* < *C. albicans* < *C. glabrata*, *K. lactis* < *S. cerevisiae* < *S. bayanus*, *D. hansenii*, *Y. lipolytica*	Ribosome biogenesis	186	$<1.11 \times 10^{-16}$
Y. lipolytica < *S. cerevisiae*, *C. glabrata*, *K. lactis*, *D. hansenii*, *C. albicans*, *A. nidulans*, *S. pombe* < *S. bayanus*	DNA metabolism	285	$<1.11 \times 10^{-16}$
	Transcription	270	$<1.11 \times 10^{-16}$
	Nucleoplasm	224	$<1.11 \times 10^{-16}$
	Nuclear part	608	$<1.11 \times 10^{-16}$
	Nuclear chromosome	104	6.66×10^{-16}
	Transcription from RNA polymerase II promoter	165	4.77×10^{-15}
	DNA replication	78	1.67×10^{-7}
D. hansenii, *Y. lipolytica* < *S. cerevisiae*, *S. bayanus*, *C. glabrata*, *K. lactis*, *C. albicans*, *A. nidulans*, *S. pombe*	Response to stimulus	323	3.00×10^{-14}
	DNA repair	110	8.15×10^{-12}
	Meiosis	62	1.50×10^{-11}
	Response to stress	258	8.55×10^{-10}
	Chromosome, pericentric region	34	7.45×10^{-6}
	Autophagy	21	1.66×10^{-5}
D. hansenii, *C. albicans*, *Y. lipolytica* < *S. cerevisiae*, *S. bayanus*, *C. glabrata*, *K. lactis*, *A. nidulans*, *S. pombe*	Telomere maintenance	168	7.17×10^{-11}
	Meiotic recombination	26	5.07×10^{-7}
A. nidulans < *S. cerevisiae*, *S. bayanus*, *C. glabrata*, *K. lactis*, *D. hansenii*, *C. albicans*, *Y. lipolytica*, *S. pombe*	Vesicle-mediated transport	198	1.26×10^{-9}
	Secretion	138	1.15×10^{-8}
	Golgi vesicle transport	97	5.44×10^{-7}
	ER to Golgi vesicle-mediated transport	55	7.20×10^{-6}
	Actin cytoskeleton	53	4.05×10^{-5}
	Cortical cytoskeleton	38	1.64×10^{-4}
	Threonine metabolism	6	3.97×10^{-4}
S. cerevisiae, *S. bayanus*, *C. glabrata*, *K. lactis*, *Y. lipolytica*, *A. nidulans*, *S. pombe* < *D. hansenii*, *C. albicans*	Proteasome complex (*sensu* Eukaryota)	43	1.94×10^{-9}
C. glabrata, *K. lactis*, *C. albicans*, *Y. lipolytica*, *A. nidulans*, *S. pombe* < *S. cerevisiae*, *S. bayanus*, *D. hansenii*	Helicase activity	56	7.07×10^{-9}
	DNA-directed RNA polymerase activity	31	1.27×10^{-8}
C. albicans, *A. nidulans*, *S. pombe* < *S. cerevisiae*, *S. bayanus*, *C. glabrata*, *K. lactis*, *D. hansenii*, *Y. lipolytica*	RNA modification	49	1.75×10^{-7}
	RNA methyltransferase activity	21	5.40×10^{-6}
C. glabrata < *S. cerevisiae*, *S. bayanus*, *K. lactis*, *D. hansenii*, *C. albicans*, *Y. lipolytica*, *A. nidulans*, *S. pombe*	Pyrophosphatase activity	176	1.34×10^{-5}
	Regulation of pH	17	0.0001

The genes associated with each of the GO terms appearing in the table show a species effect on translation efficiency. Pairwise significant relationships between species have been summarized as a species stratification. For example, for the genes annotated with 'Aerobic respiration', *Y. lipolytica* has the highest translational efficiency, followed by *K. lactis*, *D. hansenii*, *C. albicans* and *A. nidulans* (which do not differ significantly), followed by *C. glabrata*, with *S. cerevisiae*, *S. bayanus* and *S. pombe* having the lowest translational efficiency for these genes. Note that the same stratification may be implied by several GO categories.

[a]Species stratifications implied by GO categories that are related to known phenotypes discussed in the text. For the remaining species stratifications, no phenotypic explanation is currently available.

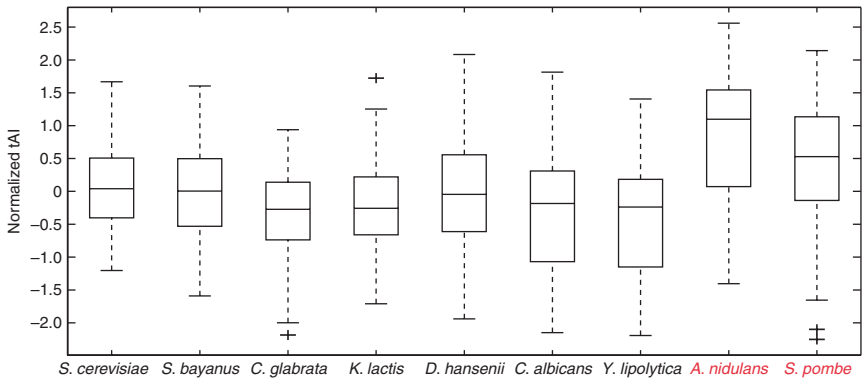

Figure 4 The translation efficiency profiles of genes related to mRNA splicing correlate with the number of introns in the genomes of species. A box plot of the relative translational efficiencies (normalized tAI; see Methods) of the genes annotated with the GO term 'nuclear mRNA splicing, via spliceosome' (52 genes; $P = 7.83 \times 10^{-8}$, Friedman test) is shown. These genes show higher translation efficiency in *A. nidulans* and *S. pombe* (highlighted in red), relative to the other seven species, in concordance with the high proportion of introns in these two species (see text).

implies that different species may have converged upon similar translation efficiencies for gene modules related to physiological traits they have in common.

Despite experimenting with various numbers of clusters, we did not identify a value that gives rise to a simple one-to-one correspondence between GO terms and clusters. We therefore turned to a supervised approach. For each nonredundant GO term (see Methods), we examined the profiles of the group of genes associated with it and, using the Friedman test[13] (**Fig. 3**), checked for a species effect on translational efficiency. We found 571 GO terms that show a significant species effect with 5% false discovery rate (FDR)[15] (see Methods for a comparison with results obtained with randomized data). For 273 out of 571 of these terms, we were able to stratify the species based on the translational efficiency values of the genes into two or more groups (**Fig. 3**) using *post hoc* Wilcoxon signed-rank tests[13] (see **Table 1** and our project website). Consistent with our previous analysis, we observed that for terms related to respiration, there was a tendency for lower translational efficiencies in the anaerobic species, whereas for the term glycolysis, the analysis pointed at higher translational efficiencies for these species (**Table 1**). Notably, according to this analysis, the term 'cytosolic ribosome' showed higher levels of translational efficiency in the anaerobic species than in the aerobic ones (**Table 1**).

The above analysis uncovered additional instances, beyond the aerobic/anaerobic distinction, in which the patterns of translational efficiencies of genes annotated with a certain term could be linked to a known phenotypic difference between the organisms. For example, the orthologous groups annotated with the terms 'nuclear mRNA splicing, via spliceosome' and 'spliceosome complex', as well as other splicing-related terms, showed the highest translational efficiencies in *A. nidulans*, followed by *S. pombe*, whereas the rest of the species showed significantly lower values (**Fig. 4**). Notably, this order corresponds perfectly to the number of intron-containing genes in the various genomes. In *A. nidulans*, 9,227 genes (~85% of the genes) contain a total of 24,824 introns, and in *S. pombe*, 2,256 genes (~50% of the genes) contain a total of 4,736 introns. In contrast, in *S. cerevisiae*, only 266 genes (5% of the genes) contain a total of 275 introns, and in *C. albicans*, the proportion of intron-containing genes is predicted to be only 3% (ref. 16). Thus, we can now predict that the fraction of intron-containing genes in the genomes of the remaining five species, which translate the splicing-related genes relatively inefficiently, is low, perhaps similar to those of *S. cerevisiae* and *C. albicans*.

As an additional example for a relationship between a known species characteristic and the translational efficiency of the relevant gene set, we found that the genes annotated with 'organic acid metabolism' were of higher translational efficiency in *Y. lipolytica* compared with the rest of the species (**Table 1**). This is in line with the use of this yeast for the industrial production of organic acids[17]. Our data also contain some highly significant patterns that we cannot presently explain by known phenotypic differences among the species (see **Table 1**, our project website (listed in Methods), **Supplementary Note** and **Supplementary Fig. 5** online for an example related to the genes of the M phase of the cell cycle).

In conclusion, we have shown here that data derived from the coding sequences of genes and the tRNA repertoire identify genes and pathways underlying specific lifestyle characteristics of organisms. Although transcriptional regulation is obviously important in determining phenotype, the fact that translational efficiency correlates positively with protein levels, even when the mRNA level is constant (**Supplementary Fig. 1**), implies that translation efficiency can enhance the effects of transcription regulation. Our results also show that synonymous codon choices may be under strong selection, adapting the codons to the tRNA pool to different extents depending on the genes' function and the organisms' needs.

The relatively constant tRNA pool found in all species indicates that the co-evolution of the genes and the tRNAs takes place mainly in a distributive fashion at the protein-coding gene level. The same gene in different species adapts itself to different extents to an essentially unmodified tRNA repertoire. Whether the strong correlations observed here are indicative of a direct causal relationship between codon adaptations and species divergence remains unclear. Because other factors (mainly transcription regulation) are also involved, it is possible that changes at the promoter architecture levels[3] may have initiated divergences that could be further intensified by massive adaptations at the translational level. Finally, we anticipate that it should be possible to extend this analysis to additional species, provided that genome analysis of the type performed here (see Methods and **Supplementary Fig. 4** online) indicates that codon usage can be shown to be governed by translational selection.

METHODS

Species analyzed. For this study, we used ascomycotic species whose genomes have been completely assembled according to the National Center for Biotechnology Information (NCBI) and for which we could infer the tRNA gene repertoire reliably: *Saccharomyces cerevisiae, Candida glabrata, Ashbya gossypii, Kluyveromyces lactis, Debaryomyces hansenii, Yarrowia lipolytica* and *Schizosaccharomyces pombe*. We also included three additional species: *Candida albicans*, an important fungal pathogen for which a high-quality gene collection (including tRNA genes) has recently become available[18]; *Saccharomyces bayanus*, a *Saccharomyces* sensu stricto species that diverged from *S. cerevisiae* ~20 million years ago (MYA) and for which the overwhelming majority of ORFs are available[19]; and *Aspergillus nidulans*, a filamentous fungus with a high-quality sequence. Thus, our sample of species spans a wide range of evolutionary

distances, ranging from ~20 MYA between *S. cerevisiae* and *S. bayanus* to 350–1,000 MYA between *S. pombe* and the hemiascomycotic species[20]. As most of the species we used are yeasts, we collectively refer to them in the text as yeast species.

The choice of species for the analyses described in this work is not trivial. On the one hand, remote species may have too few orthologs shared. On the other hand, in very close species, orthologs may present translational efficiencies that are close merely because of phylogenetic relatedness. This means that not all species contribute equally and independently. These two considerations were taken into account when choosing species for analysis. Thus, we excluded from analysis some sensu stricto *Saccharomyces* species (for example, *S. paradoxus*) that are closer to *S. cerevisiae* than *S. bayanus* is, and we included in the analysis only ascomycotic fungi. It is notable that in the present analysis even close species such as *S. bayanus* and *S. cerevisiae* differ in the translational efficiency of genes from some GO categories (**Table 1**), and that in general the true species phylogeny is not trivially reflected in the signals (**Supplementary Fig. 3**).

tRNA gene copy numbers. For all species except *C. albicans* and *S. bayanus*, the tRNA gene copy numbers were obtained by applying tRNAscan-SE software version 1.1 (ref. 8) to the genome sequences. See below for a list of URLs used to obtain sequence information.

For *S. bayanus*, we used the tRNA gene copy numbers of the closely related *S. cerevisiae* (see **Supplementary Methods** online for a discussion of this choice as opposed to deriving the tRNA repertoire of *S. bayanus* from its genome sequence). For *C. albicans*, we extracted the tRNA gene counts from the *Candida* Genome Database (CGD)[21] on August 14, 2005 (see URL below). See **Supplementary Table 1** for the data on gene copy numbers of tRNAs that decode sense codons in each of the species.

Protein and coding sequences. The *C. albicans* protein and coding sequences included in assembly Ca19 were downloaded from the *C. albicans* Research Laboratory website (see URLs section below) on August 14, 2005. This gene set corresponds to the haploid genome of *C. albicans*. *S. cerevisiae* and *S. bayanus* protein and coding sequences were downloaded from the Saccharomyces Genome Database (SGD)[22] on June 16, 2005 (see URL below). For *S. bayanus*, several sequences may correspond to different fragments of the same ORF. We used the annotation of ref. 19 to merge such fragments. Protein, gene sequences and gene structures from release 4 of the *A. nidulans* genome sequence[23] were downloaded from the Broad Institute's *A. nidulans* database (see URLs section below). Using the gene sequences and the corresponding gene structures, we obtained coding sequences for the *A. nidulans* genes.

For the remaining six species, as well as for *C. neoformans*, which was used as an outgroup to establish orthology relationships (see below), protein files were downloaded from Integr8 (ref. 24, and see URL below). The corresponding coding sequences were extracted from the EMBL database[25] (see **Supplementary Methods** for a complete description of the extraction process).

Finally, we removed mitochondrially encoded sequences from all sequence sets.

All coding sequence data sets used are available at our project website (see URL below).

tRNA adaptation index (tAI). The tAI scoring scheme is described in detail in ref. 5. Briefly, the method entails calculating a weight for each of the sense codons, derived from the copy numbers of all the tRNA types that recognize it (including wobble interactions), relying on the observed correlation between tRNA cellular pools and tRNA gene copy numbers[7]. This correlation is in line with a recent observation that the nucleosome has a low predicted affinity for the promoters of many of the tRNA genes, suggesting a constitutive expression with little transcriptional regulation capacity[26]. For a given coding sequence, the tAI value is then the geometric mean of the weights of all its sense codons (stop codons were ignored, when encountered). The tAI of a coding sequence ranges from 0 to 1, with high values corresponding to high levels of translational efficiency. To calculate the tAI for coding sequences, we used the codonR script supplied in ref. 5 (see below for URL), which we modified to include the first codon, as well as other methionines. Files with the tAI values of the coding sequences for the various species are available at our project website (see URL below).

Checking for translational selection in yeast genomes. The application of a translational efficiency index such as tAI[5] to the sequences of a genome is meaningful only if translational selection has had a major part in shaping the codon usage in the genome. We evaluated the extent to which translational selection determines the codon usage of the species in our sample by comparing tAI, which measures conformance to a specific codon bias, to Nc, the effective number of codons, which measures an overall codon bias[27]. Nc reaches its maximal value (theoretically 61, although in practice the equation used for calculating Nc may yield larger values) when codon usage is completely random, and its minimal value (20) when only one codon is used per amino acid. If translational selection were the main force shaping codon usage, sequences with low Nc values would have typically been those that were selected for optimal translation efficiency. However, other factors, such as mutational pressure, may affect codon usage as well. Thus, a strong negative correlation between Nc and tAI would indicate that the main force that shapes codon usage is translational efficiency by means of adaptation to the tRNA pool. Nine of the ten species show a significant negative correlation between Nc and tAI ($P = 0.005$ for *C. albicans*, and $P < 0.001$ for the remaining eight species; see **Supplementary Fig. 4** online). However, in *A. gossypii*, although the correlation was negative, its magnitude was low and insignificant (Pearson's $r = -0.38$; $P = 0.384$; **Supplementary Fig. 4**), suggesting that for this species, tAI would not be a good predictor of expression levels. In this species, changes in protein levels may be achieved in different ways: for example, by raising the levels of transcript. We therefore excluded *A. gossypii* from subsequent analyses. See **Supplementary Methods** for a complete description of how Nc and the significance of the correlation between tAI and Nc were calculated.

Construction of the multispecies array of translational efficiencies. See our project website (URL given below) for a figure describing the construction of the array. *Generation of a table of orthologous groups.* The Multiparanoid program[28] was used to generate a matrix of orthologs in which each row corresponds to a gene and each column to a species. We used *C. neoformans*, a basidiomycotic fungus, as an outgroup. From this matrix, we retained only those orthologous groups (rows in the table) that had representatives from both *S. cerevisiae* and *S. pombe* (2,883 out of 6,226 rows). See **Supplementary Methods** for a complete description of the derivation of the table of orthologs. The table of orthologous groups is available at our project website (URL given below).

Note that if a duplication had occurred after the divergence of *S. pombe* and *A. nidulans* from the remaining species, there would have been more than one gene representing the same species in the same orthologous group (row). As we assume that all genes in a single orthologous group have the same function, we will henceforth refer to them as representing a single gene, even when there is more than one representative per species. *Generation of a matrix of translational efficiencies across species.* We combined the orthologous groups table with the tAI values computed for all ORFs of the nine species to create a matrix of translational efficiencies across species (see **Supplementary Methods** for a complete description of this process). Each of the 2,810 rows in the table will henceforth be referred to as a profile. Each column was standardized so that its mean and s.d. were 0 and 1, respectively. The same standardization was then applied to the rows of the matrix, emphasizing the efficiency of genes relative to their orthologous counterparts, rather than efficiency relative to genes in the same species. Files for the various stages of the data—from raw tAI values without imputed values to the final normalized values used—are available at our project website (see URL below).

Gene Ontology (GO) data. Using the GO database[14] as well as the annotation data at SGD[22], we constructed for each *S. cerevisiae* gene the list of GO terms that describe it (**Supplementary Methods**). Each GO term was then considered to annotate any orthologous group (and corresponding translational efficiency profile) containing a *S. cerevisiae* gene that is associated with this term. If an orthologous group contained more than one *S. cerevisiae* gene, then it was sufficient for one of these genes to be associated with the term in order to associate the whole orthologous group and the corresponding profile with the term.

Intron data. Data regarding introns in the genes of *S. cerevisiae* was obtained from SGD[22] on September 18, 2006. The corresponding data for *S. pombe* were

downloaded from the *S. pombe* Genome Project site at the Sanger Institute and for *A. nidulans* from the *Aspergillus nidulans* database at the Broad Institute (see below for URLs).

Statistical analyses (i) Cluster analysis. Hierarchical clustering of the translational efficiency profiles was performed using the MATLAB/MathWorks package. We used the euclidean distance between normalized profiles as a distance measure and the average linkage algorithm for the construction of the hierarchical tree. The granularity of the clustering was chosen by eye.

(ii) Calculation of functional enrichment for clusters. In each cluster, we checked for the enrichment of each of the nonredundant GO terms, if there was at least one gene within the cluster that was annotated with this term and if that term annotated at least three genes in the whole data set. Enrichment was assessed using the one-sided hypergeometric test. We corrected for multiple testing using the false discovery rate (FDR) method[15] with an FDR of 5%, pooling the results for all clusters. Owing to the hierarchical structure of the gene ontology, the tests in this analysis are nested, making it difficult to account for the multiple testing. We therefore complemented the analysis with an empirical evaluation of the significance of enrichment, based on a permutation test (**Supplementary Methods**).

(iii) Analysis of the species effect on translation efficiency. We used the Friedman test[13], a nonparametric analog of the two-way ANOVA without replicates, to test for a difference of the median translation efficiency between species in a selected group of orthologous groups (**Fig. 3**). This test was applied both to the clusters obtained through hierarchical clusters and to all sets of genes defined by a GO term that annotates all genes in the set, conditional on the set containing at least three genes. We corrected for multiple testing using the FDR method[15] with an FDR of 5%, obtaining a significant species effect on translational efficiency for 571 GO terms. As a control, we repeated the Friedman tests for GO terms after randomizing the genes relative to the lists of GO terms associated with each gene. In contrast to the original assignment, in which there were 571 significant terms with an FDR of 5%, in the randomized case we found only 17 terms with the same FDR.

(iv) *Post hoc* tests. Each set of orthologous groups that was found to be significant using the Friedman test (with an FDR of 5%) was further tested to find the source of difference in medians (**Fig. 3**). For this, we used the Wilcoxon signed-rank test[13], applied to all pairwise comparisons among species (columns), using an FDR[15] of 20% in the correction for multiple tests. Note that for very small sets of genes, the minimal *P* value possible for the Wilcoxon signed-rank test exceeds the threshold set by the multiple testing procedure. Therefore, for such small sets, the comparisons that obtained the minimal *P* value possible for the test were considered significant. For those pairwise comparisons that turned out to be significant, we used the median values for the species in the relevant set of genes to determine the direction of the relationship. These directional pairwise relationships were then used to order the species according to their relative translational efficiency for the set of genes considered.

URLs. The website for our project can be found at http://longitude.weizmann.ac.il/pub/papers/Man2007_tai/suppl/. The *Candida* Genome Database (CGD) can be found at http://www.candidagenome.org/. The *Saccharomyces* Genome Database (SGD) can be found at http://www.yeastgenome.org/. The Integr8 site can be found at http://www.ebi.ac.uk/integr8. The *S. pombe* Genome Project site at the Sanger Institute can be found at http://www.sanger.ac.uk/Projects/S_pombe/. For *A. nidulans*, all protein and genome sequence data and structures were obtained from release 4 of the genome sequence at http://www.broad.mit.edu/annotation/genome/aspergillus_nidulans. The *C. albicans* protein and coding sequences included in assembly Ca19 were downloaded from http://candida.bri.nrc.ca on August 14, 2005. For all species, except *A. nidulans*, *C. albicans*, and *S. bayanus*, chromosome sequences were obtained from GenBank (http://www.ncbi.nlm.nih.gov/Genbank). The codonR script[5] can be downloaded from http://people.cryst.bbk.ac.uk/~fdosr01/tAI/.

Note: Supplementary information is available on the Nature Genetics website.

ACKNOWLEDGMENTS

We thank the Pilpel laboratory for helpful discussions and E. Segal, A. Regev, I. Tirosh, Y. Gilad, N. Barkai, J. Moult and J.L. Sussman for discussions and critical reviews of the manuscript. Y.P. holds the Rothstein Career Development Chair in Genetic Diseases. We thank the Ben May Charitable Trust and EMBRACE, the European Union Network of Excellence in Bioinformatics for grant support.

AUTHOR CONTRIBUTIONS

Y.P. and O.M. conceived the study, and Y.P. supervised the study. O.M. and Y.P. designed the analyses, O.M. performed the analyses and O.M. and Y.P. wrote the paper.

COMPETING INTERESTS STATEMENT

The authors declare that they have no competing financial interests.

Published online at http://www.nature.com/naturegenetics
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions

1. Wolfe, K.H. Comparative genomics and genome evolution in yeasts. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **361**, 403–412 (2006).
2. Gompel, N., Prud'homme, B., Wittkopp, P.J., Kassner, V.A. & Carroll, S.B. Chance caught on the wing: cis–regulatory evolution and the origin of pigment patterns in Drosophila. *Nature* **433**, 481–487 (2005).
3. Ihmels, J. *et al.* Rewiring of the yeast transcriptional network through the evolution of motif usage. *Science* **309**, 938–940 (2005).
4. Powers, D.A. & Schulte, P.M. Evolutionary adaptations of gene structure and expression in natural populations in relation to a changing environment: a multidisciplinary approach to address the million-year saga of a small fish. *J. Exp. Zool.* **282**, 71–94 (1998).
5. dos Reis, M., Savva, R. & Wernisch, L. Solving the riddle of codon usage preferences: a test for translational selection. *Nucleic Acids Res.* **32**, 5036–5044 (2004).
6. Sharp, P.M. & Li, W.H. The codon adaptation index–a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res.* **15**, 1281–1295 (1987).
7. Percudani, R., Pavesi, A. & Ottonello, S. Transfer RNA gene redundancy and translational selection in Saccharomyces cerevisiae. *J. Mol. Biol.* **268**, 322–330 (1997).
8. Lowe, T.M. & Eddy, S.R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* **25**, 955–964 (1997).
9. Tanay, A., Regev, A. & Shamir, R. Conservation and evolvability in regulatory networks: the evolution of ribosomal regulation in yeast. *Proc. Natl. Acad. Sci. USA* **102**, 7203–7208 (2005).
10. Ghaemmaghami, S. *et al.* Global analysis of protein expression in yeast. *Nature* **425**, 737–741 (2003).
11. Fraser, H.B., Hirsh, A.E., Wall, D.P. & Eisen, M.B. Coevolution of gene expression among interacting proteins. *Proc. Natl. Acad. Sci. USA* **101**, 9033–9038 (2004).
12. Lithwick, G. & Margalit, H. Relative predicted protein levels of functionally associated proteins are conserved across organisms. *Nucleic Acids Res.* **33**, 1051–1057 (2005).
13. Rice, J. *Mathematical Statistics and Data Analysis* (Wadsworth Publishing Company, Belmont, Calfiornia, 1995).
14. Harris, M.A. *et al.* The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.* **32**, D258–D261 (2004).
15. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate – a practical and powerful approach to multiple testing. *J. R. S. Soc. Ser. B. Methodol.* **57**, 289–300 (1995).
16. Nantel, A. The long hard road to a completed *Candida albicans* genome. *Fungal Genet. Biol.* **43**, 311–315 (2006).
17. Barth, G. & Gaillardin, C. Physiology and genetics of the dimorphic fungus *Yarrowia lipolytica. FEMS Microbiol Rev* **19**, 219–237 (1997).
18. Braun, B.R. *et al.* A human-curated annotation of the *Candida albicans* genome. *PLoS Genet* **1**, 36–57 (2005).
19. Kellis, M., Birren, B.W. & Lander, E.S. Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae. Nature* **428**, 617–624 (2004).
20. Berbee, M. & Taylor, J. Systematics and evolution. in *The Mycota,* Vol. VIIB (eds. McLaughlin, D., McLaughlin, E. & Lemke, P.) 229–245 (Springer, Berlin, 2001).
21. Arnaud, M.B. *et al.* Sequence resources at the *Candida* Genome Database. *Nucleic Acids Res.* **35**, D452–D456 (2007).
22. Issel-Tarver, L. *et al. Saccharomyces* Genome Database. *Methods Enzymol* **350**, 329–346 (2002).
23. Galagan, J.E. *et al.* Sequencing of *Aspergillus nidulans* and comparative analysis with *A. fumigatus* and *A. oryzae. fumigatus and A. oryzae. Nature* **438**, 1105–1115 (2005).
24. Pruess, M., Kersey, P. & Apweiler, R. The Integr8 project–a resource for genomic and proteomic data. *In Silico Biol.* **5**, 179–185 (2005).
25. Kanz, C. *et al.* The EMBL Nucleotide Sequence Database. *Nucleic Acids Res.* **33**, D29–D33 (2005).
26. Segal, E. *et al.* A genomic code for nucleosome positioning. *Nature* **442**, 772–778 (2006).
27. Wright, F. The 'effective number of codons' used in a gene. *Gene* **87**, 23–29 (1990).
28. Alexeyenko, A., Tamas, I., Liu, G. & Sonnhammer, E.L. Automatic clustering of orthologs and inparalogs shared by multiple proteomes. *Bioinformatics* **22**, e9–e15 (2006).
29. Prillinger, H. *et al.* Phylogeny and systematics of the fungi with special reference to the Ascomycota and Basidiomycota. *Chem. Immunol.* **81**, 207–295 (2002).

Corrigendum: Differential translation efficiency of orthologous genes is involved in phenotypic divergence of yeast species

Orna Man & Yitzhak Pilpel

Nat. Genet. **39**, 415–421 (2007); published online 4 February 2007; corrected after print 28 March 2007

In **Figure 2** of the original version of this paper, panels **c** and **d** were accidentally transposed, resulting in incorrect information in the legend. **Figure 2c** shows genes of the tricarboxylic acid (TCA) cycle, and **Figure 2d** shows glycolysis genes. The error has been corrected in the HTML and PDF versions of the article.

Corrigendum: A common coding variant in *CASP8* is associated with breast cancer risk

Angela Cox, Alison M Dunning, Montserrat Garcia-Closas, Sabapathy Balasubramanian, Malcolm W R Reed, Karen A Pooley, Serena Scollen, Caroline Baynes, Bruce A J Ponder, Stephen Chanock, Jolanta Lissowska, Louise Brinton, Beata Peplonska, Melissa C Southey, John L Hopper, Margaret R E McCredie, Graham G Giles, Olivia Fletcher, Nichola Johnson, Isabel dos Santos Silva, Lorna Gibson, Stig E Bojesen, Børge G Nordestgaard, Christen K Axelsson, Diana Torres, Ute Hamann, Christina Justenhoven, Hiltrud Brauch, Jenny Chang-Claude, Silke Kropp, Angela Risch, Shan Wang-Gohrke, Peter Schürmann, Natalia Bogdanova, Thilo Dörk, Rainer Fagerholm, Kirsimari Aaltonen, Carl Blomqvist, Heli Nevanlinna, Sheila Seal, Anthony Renwick, Michael R Stratton, Nazneen Rahman, Suleeporn Sangrajrang, David Hughes, Fabrice Odefrey, Paul Brennan, Amanda B Spurdle, Georgia Chenevix-Trench, The Kathleen Cunningham Foundation Consortium for Research into Familial Breast Cancer, Jonathan Beesley, Arto Mannermaa, Jaana Hartikainen, Vesa Kataja, Veli-Matti Kosma, Fergus J Couch, Janet E Olson, Ellen L Goode, Annegien Broeks, Marjanka K Schmidt, Frans B L Hogervorst, Laura J Van't Veer, Daehee Kang, Keun-Young Yoo, Dong-Young Noh, Sei-Hyun Ahn, Sara Wedrén, Per Hall, Yen-Ling Low, Jianjun Liu, Roger L Milne, Gloria Ribas, Anna Gonzalez-Neira, Javier Benitez, Alice J Sigurdson, Denise L Stredrick, Bruce H Alexander, Jeffery P Struewing, Paul D P Pharoah & Douglas F Easton, on behalf of the Breast Cancer Association Consortium

Nat. Genet. **39**, 352–358 (2007); published online 11 February 2007; corrected 10 April 2007

In the version of this article initially published, there was an error that affected the calculations of the odds ratios, confidence intervals, between-study heterogeneity, trend test and test for association for SNP *ICAM5* V301I in **Table 1** (*ICAM5* V301I); genotype counts in **Supplementary Table 2** (*ICAM5*; ICR_FBCS and Kuopio studies) and minor allele frequencies, trend test and odds ratios for heterozygotes and rare homozygotes in **Supplementary Table 3** (*ICAM5*; ICR_FBCS and Kuopio studies). The corrected rows from each table are reproduced below.
The errors in **Table 1** have been corrected in the PDF version of the article. The errors in supplementary information have been corrected online.

Table 1 Summary odds ratios and 95% confidence intervals for nine polymorphisms and breast cancer risk

SNP	Between-study heterogeneity	Test for association	Trend test	Analysis model	Heterozygote OR (95% c.i.)	Rare homozygote OR (95% c.i.)
ICAM5 V301I	0.57	0.54	0.98	Fixed effects	1.02 (0.97, 1.07)	0.99 (0.93, 1.05)
rs1056538				Random effects	1.02 (0.97, 1.07)	0.99 (0.93, 1.05)

Supplementary Table 2 Genotype counts among cases and controls by study

			Controls				Cases			
Gene	SNP	Study	AA	Aa	aa	Total	AA	Aa	aa	Total
ICAM5	rs1056538	ICR_FBCS	207	243	71	521	212	239	68	519
		Kuopio	178	209	46	433	193	206	48	447

Supplementary Table 3 Association between nine polymorphisms and breast cancer risk by study

					Heterozygotes			Rare homozygotes		
Gene	SNP	Study	MAF	Trend test	OR	95% c.i.		OR	95% c.i.	
ICAM5	rs1056538	ICR_FBCS	0.369	0.698	0.960	0.739	1.248	0.935	0.637	1.373
		Kuopio	0.348	0.658	0.909	0.687	1.203	0.962	0.612	1.513