Genome Biology

RESEARCH Open Access

Inferring gene function from evolutionary change in signatures of translation efficiency

Anita Krisko[1], Tea Copic[1], Toni Gabaldón[2,3,5], Ben Lehner[3,4,5] and Fran Supek[2,3,4*]

Abstract

Background: The genetic code is redundant, meaning that most amino acids can be encoded by more than one codon. Highly expressed genes tend to use optimal codons to increase the accuracy and speed of translation. Thus, codon usage biases provide a signature of the relative expression levels of genes, which can, uniquely, be quantified across the domains of life.

Results: Here we describe a general statistical framework to exploit this phenomenon and to systematically associate genes with environments and phenotypic traits through changes in codon adaptation. By inferring evolutionary signatures of translation efficiency in 911 bacterial and archaeal genomes while controlling for confounding effects of phylogeny and inter-correlated phenotypes, we linked 187 gene families to 24 diverse phenotypic traits. A series of experiments in *Escherichia coli* revealed that 13/15, 19/23, and 3/6 gene families with changes in codon adaptation in aerotolerant, thermophilic, or halophilic microbes confer specific resistance to, respectively, hydrogen peroxide, heat, and high salinity. Further, we demonstrate experimentally that changes in codon optimality alone are sufficient to enhance stress resistance. Finally, we present evidence that multiple genes with altered codon optimality in aerobes confer oxidative stress resistance by controlling the levels of iron and NAD(P)H.

Conclusions: Taken together, these results provide experimental evidence for a widespread connection between changes in translation efficiency and phenotypic adaptation. As the number of sequenced genomes increases, this novel genomic context method for linking genes to phenotypes based on sequence alone will become increasingly useful.

Background

The genetic code is redundant, meaning that most amino acids can be encoded by more than one codon. Across diverse species, highly expressed genes tend to use optimal codons to increase the accuracy and speed of translation by ensuring better agreement with the cellular tRNA pools [1-3]. Consequently, codon biases are predictive of expression levels in both natural [4,5] and designed [6,7] gene sequences. This 'translational selection' acting to increase the use of optimal codons is stronger in faster growing microbes with large effective population sizes [8], but it has been shown to be widespread in both prokaryotes and eukaryotes [9-11], allowing the signatures of high gene expression to be detected and compared across species [12,13].

Interestingly, several previous studies have suggested a link between increased translation efficiency in specific groups of orthologous genes and phenotypic change during evolution [14-16]. Examples include increased codon optimization of photosynthesis genes in *Synechocystis* and methanogenesis genes in *Methanosarcina acetivorans* [14], reflecting their trophic preferences, and an increased use of optimal codons in glycolytic enzymes in anaerobic microbes or in the Krebs cycle in aerobes [15]. In nine yeast species, the same trend was observed [16] and, in addition, species adapted either to aerobic or anaerobic growth had consistently higher codon adaptation in the mitochondrial or cytoplasmic ribosomal protein (RP) genes, respectively. This correlation could not be explained by the phylogenetic distribution of (an)aerobes,

* Correspondence: fran.supek@crg.eu
[2]Bioinformatics and Genomics Programme, Centre for Genomic Regulation (CRG), Dr. Aiguader 88, 08003 Barcelona, Spain
[3]Universitat Pompeu Fabra (UPF), 08003 Barcelona, Spain
Full list of author information is available at the end of the article

indicating that mere genetic drift is not sufficient to drive the evolution of translation efficiency [16].

These examples of the coupling of codon usage to adaptive phenotypic variation suggest that it might be possible to systematically infer gene function from evolutionary change in the use of optimal codons. The basis for this argument is that diverse species sharing a common phenotypic trait, such as resistance to high temperature, might show increased expression, via a convergent codon adaptation, in a common set of genes involved in that phenotypic trait. However, four important challenges have so far prevented the large-scale inference of novel translation efficiency-phenotype links: 1) insufficient coverage with genomic sequences necessary to detect a weak evolutionary signal; 2) methodological issues with common approaches for predicting expression from codon biases in certain genomes [5,17,18], and with rescaling the predictions to make them comparable across genomes; 3) difficulties in disentangling the influences of the phylogeny and a particular phenotype; and 4) extensive correlations between different phenotypes. For instance, Archaea are typically obligate anaerobes, and within Bacteria, thermophiles tend to be less commonly aerotolerant than mesophiles. Thus, an observed correlation between a genomic feature and aerotolerance might be an artifact of either thermophilicity or phylogenetic relatedness.

Here, we explicitly address these issues using a novel statistical framework to identify meaningful correlations between phenotypes and signatures of selection for translation efficiency. Our approach generalizes over the previous explanatory models for a few select phenotypes to a broadly applicable framework that generates many testable predictions about the genes involved in adaptation to various environments. We experimentally validate a set of predicted gene-phenotype links for genes involved in three environmental adaptations: resistance to oxidative stress, heat, and high salinity. Moreover, we confirm experimentally that changing the codon usage of a gene can be sufficient to confer the expected stress resistance phenotype. Our approach therefore provides a potentially general strategy for annotating gene function in newly sequenced genomes by identifying genes whose translation efficiency is linked to particular phenotypes, important stress responses, or environmental adaptations.

Results

A novel method links translation efficiency of gene families to phenotypes

The codon usage of individual genes is to a large extent determined by mutational processes unrelated to translational selection [19,20], thus necessitating that these influences be factored out before predicting expression levels from codon biases. To this end, we used a machine learning-based method, which tests whether a given gene's codon usage pattern is more similar to a reference set of highly expressed genes than would be expected from the background nucleotide composition in intergenic DNA [9]. Using this approach, we assigned a categorical high/low expression label to genes in 911 bacterial and archaeal genomes (Figure 1A). Changes in the methodology (see Additional file 1) substantially improved the agreement of the predictions with microarray data in 19 diverse bacterial species (Figure 1B; see Additional file 2). The predicted highly expressed genes (at a false discovery rate (FDR) of $\leq 10^{-12}$, sign test) comprise 4 to 20% of the genome, depending on the genome size (see Additional file 3), and have on average 3.9 times higher microarray signal levels than the rest of the genes ($P = 10^{-47}$ by Mann–Whitney test, median of 19 genomes) (Figure 1B; see Additional file 2). For comparison, the very highly expressed ribosomal protein genes are 6.1 times above the genome average in the 19-genome dataset.

Genomes under weak selection for translation efficiency represent a difficult case for detecting signatures of expression levels in codon biases. In three out of four such genomes, the predictions from our machine learning-based method showed better correlation with mRNA levels than did those obtained by a commonly used approach, the CAI [4] (Figure 1C; combined $P = 0.0053$, one-tailed Kolmogorov-Smirnov test). Importantly, although gene expression levels may change substantially across different conditions, the genome-encoded codon biases are static, and are likely to reflect the gene expression in a typical environment encountered by the organism during evolution [24].

In addition to codon usage, other coding sequence determinants can shape protein levels. For instance, strong secondary structures at the mRNA 5′ end were shown to influence translation efficiency in a library of synthetic gene variants [25]. However, we found no correlation between mRNA 5′ folding free energies and gene expression levels in the 19 evaluated bacterial genomes (median $r = 0.02$ to 0.05; see Additional file 4), in contrast to various codon indices (median $r = 0.22$ to 0.43). This is consistent with mRNA folding being more relevant for highly stable 5′ mRNA structures [7], which we found to occur only infrequently in real genomes (median 13 to 17% of genes, depending on their position in the mRNA; see Additional file 4).

To infer whether increased or reduced translational efficiency of a gene is adaptive in a particular environment or is associated with a particular phenotype, we searched for correlations between the high/low expression levels of orthologous gene groups (as identified in clusters of orthologous groups (COGs) [26]) and the phenotypes or environments annotated to each species. We used a statistical framework based on supervised machine learning

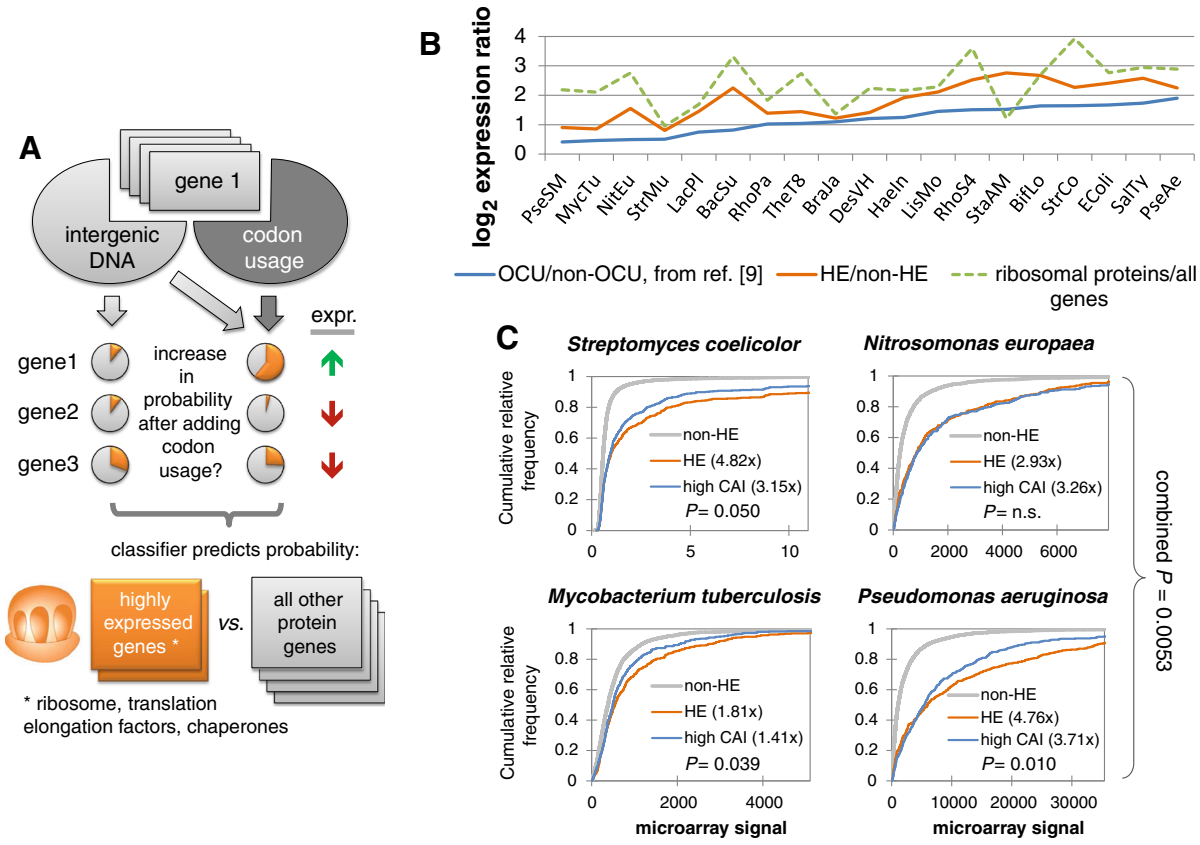


Figure 1 Finding highly expressed (HE) genes in prokaryotic genomes and their enrichment in particular phenotypic groups. (A) HE labels for genes were predicted by comparing the codon usage of each gene with that of a small set of known HE genes, while controlling for local background nucleotide composition determined from the neighboring intergenic DNA. **(B)** Comparison of average microarray signal intensity between the HE genes from this study, and the HE gene group of Supek *et al.* [9] in 19 diverse bacterial genomes, denoted by the UniProt species code on the *x* axis. The machine learning pipeline used here was derived from the methodology of Supek *et al.* [9] (see Additional file 1). **(C)** The predictions of HE genes compared favorably with those of the codon adaptation index (CAI) when evaluated against microarray data in four genomes previously claimed to lack detectable selected codon biases [21-23], representing difficult cases for predicting gene expression from codon usage. The *x* axes range from the minimum to the 99th percentile of microarray signal intensities. The farther away a curve is from the non-HE curve, the better the separation. Numbers in the parenthesis are the ratios of average expression in the HE (or high CAI) genes, compared with the non-HE genes. The 'high CAI' category was defined as containing the same number of genes as the HE category in each genome. *P* values are from a one-tailed Kolmogorov-Smirnov test for HE > high CAI, and were combined using Fisher's method.

that searches within a large set of translation efficiency-phenotype correlations to find the phenotypes that contribute independently to the prediction of translation efficiency, after controlling for all the confounding phenotypes or taxonomic subdivisions (see Additional file 5; summarized in Figure 2A; examples of confounders in Figure 2B). Our method provides predictions for 187 gene families (COGs), which are linked to 24 different phenotypes (see Additional file 5), including the ability to colonize various environments, and plant and mammalian pathogenicity (200 predictions in total).

In 71 of 911 genomes, the detected codon bias did not fully match the optimal codons predicted from genomic tRNA gene composition (see Additional file 6), and it is thus not clear whether translational selection causes the observed signature of high expression in these genomes. tRNA modifications have been hypothesized as a cause for such discrepancies [9,27]. Upon excluding the 71 genome set, we found that our 200 phenotype predictions were highly robust to this factor (see Additional file 7).

Next, for three selected phenotypes, we evaluated these predictions by performing experiments in a series of *Escherichia coli* deletion mutants.

Genes with altered codon adaptation in aerobes protect *E. coli* against oxidative stress

We first focused on genes with differential translation efficiency signatures between 514 aerotolerant microbes and 214 obligate anaerobes. We found that 295 COGs had a significant change in the content of highly expressed (HE) genes (at least twofold enrichment, FDR = 9.6% by

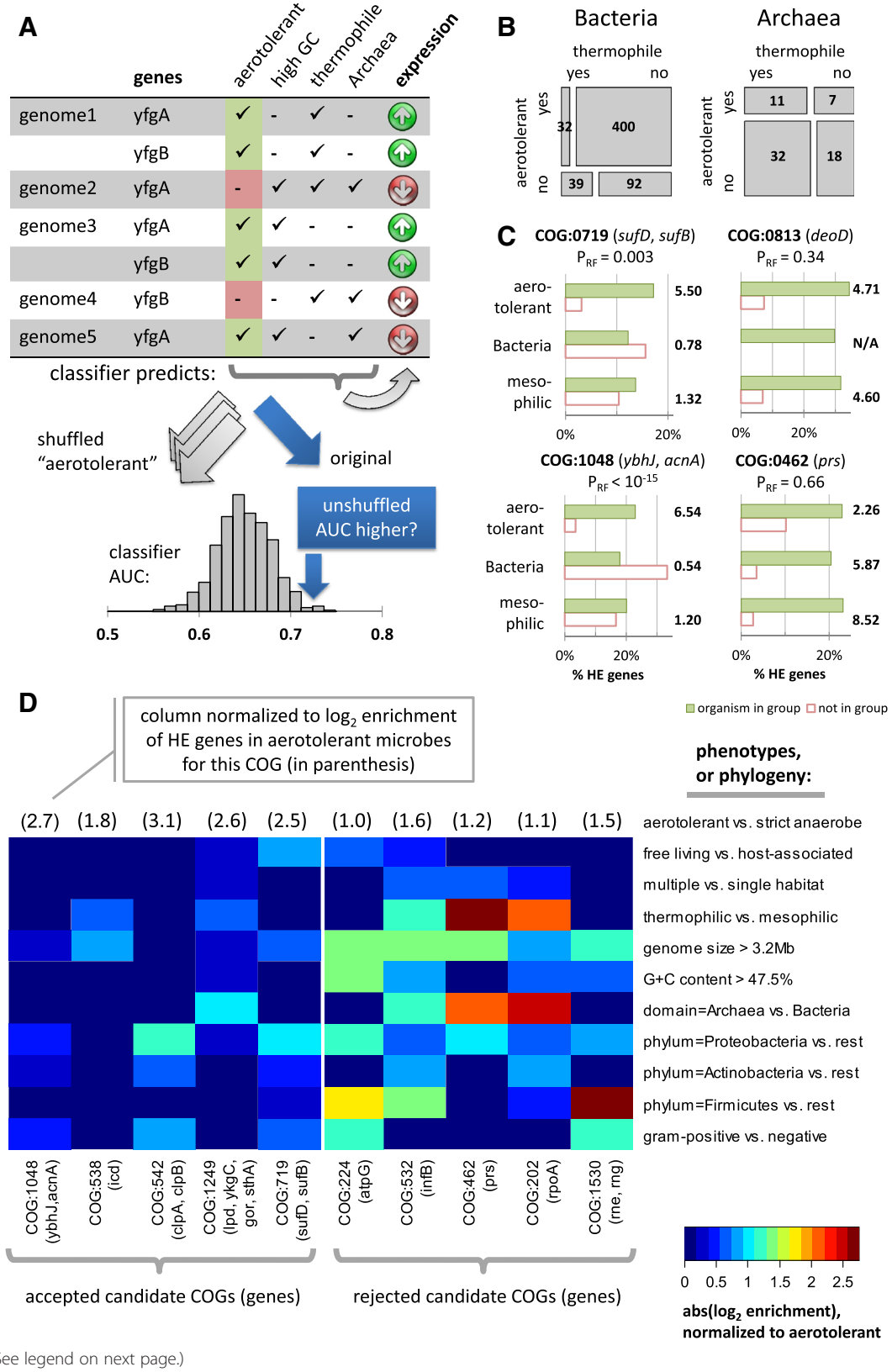


Figure 2 (See legend on next page.)

(See figure on previous page.)
Figure 2 Enrichment for highly expressed (HE) genes in gene families across microbial phenotypic groups. (A) Phenotypes were tested for an independent contribution to predicting expression levels within a gene family, after controlling for 24 other phenotypes, 6 genomic descriptors, and 70 phylogenetic subdivisions using a Random Forest (RF) randomization test (see Additional file 1). **(B)** An example of correlation between two phenotypes (here, thermophilicity and aerotolerance), and their correlation with taxonomy. The area of the rectangles is proportional to the number of genomes in each subgroup (overlaid). **(C)** Enrichment with HE genes in four example clusters of orthologous groups (COG) gene families in aerotolerant microbes versus obligate anaerobes, compared with HE enrichments in two other aerotolerance-correlated traits: genomic G + C content and thermophilicity. The 'accepted' COGs (left) have stronger HE enrichments for aerotolerance than for the other traits, whereas the HE enrichment in the 'rejected' COGs (right) can be more easily explained both by the aerotolerance and by another trait. **(D)** Enrichment of example COGs with HE genes in 10 groups of microbes defined through phenotypic traits, genomic features (GC, size) or taxonomy. The COGs shown all have *Escherichia coli* representative genes, and were found to have at least a twofold enrichment in HE genes in aerotolerant microbes compared with obligate anaerobes ($P < 0.01$, Fisher's exact test). Left block shows the five HE-enriched genes with the most significant *P*-values in the RF randomization test for confounding phenotypes/phylogeny, while the right block shows the genes with the least significant *P*-values in this test. The more significant COGs tended to be less HE-enriched in other phenotypes or phylogenetic groups relative to the HE enrichment in aerotolerant microbes. Thus, the aerotolerant phenotype contains the information about the HE enrichment of genes within these particular COGs that cannot be recovered from the other traits.

Fisher's exact test). Of these, only 23 COGs passed a control to ensure that the enrichment for HE genes could not be explained by the 23 other phenotypes, the 6 genomic features, or the 70 taxonomic subdivisions ($P < 10^{-2}$, Random Forest permutation test). The percentages of HE genes for four example COGs passing or failing this test are shown for select phenotypes in Figure 2C, and the enrichments for a broader set of COGs and phenotypes are shown in Figure 2D. Similarly, a comparison between 296 obligate aerobes and 217 facultative aerobes identified 160 differentially expressed COGs (FDR = 11.8%), with 11 COGs remaining after controlling for confounding factors. In total, 34 differentially expressed COGs were found for the two oxygen-related phenotypes.

Of the 34 COGs, 22 were present in the *E. coli* MG1655 genome, and 15 of these had viable deletion mutants. These *E. coli* strains comprised a biological model system for testing the hypothesis that genes with differential codon adaptation in microbes exposed to varying oxygen levels have a role in resisting the oxidative stress associated with the aerobic lifestyle. All 15 *E. coli* deletion mutants exhibited higher sensitivity to hydrogen peroxide exposure than the wild-type strain (Figure 3A). In particular, nine mutant strains were similarly or more sensitive to 2.5 mM H_2O_2 than the *sodA* strain lacking the Mn-containing superoxide dismutase (≤20% of wild-type survival). Decreased survival of the 15 mutants was observed across a range of H_2O_2 concentrations that spanned almost two orders of magnitude (0.5 mM to 20 mM; see Additional file 8).

To verify that the deletions caused sensitivity to oxidative stress specifically instead of a general frailty of the bacteria, we exposed the mutants to heat and osmotic shock, and found that 13 strains were as resistant as the wild type (≥90% of wild-type survival, Figure 3A). The two remaining non-specifically sensitive strains were *recA*, deficient in the SOS response and in recombination DNA repair, and *lon*, lacking a major protease dealing with clearance of oxidized proteins. Both mutants are known to be sensitive to a variety of different stresses.

Consistent with oxidative stress contributing to growth impairment, all mutants showed increased protein carbonylation levels (Figure 4A), and treatment with the antioxidant N-acetylcysteine rescued the H_2O_2 sensitivity phenotype (Figure 3A). Further, we were able to reverse the phenotype by expressing wild-type copies of the deleted genes (average survival of 15 complemented mutants was 97.2% of wild-type, compared with 18.7% without the plasmid; see Additional file 9) indicating that the observed effect was not due to disrupted regulation of other genes or to background mutations in the deletion strain.

Finally, a literature search yielded additional evidence supporting a role in oxidative stress resistance for 4 of the 13 genes in *E. coli* or other organisms: *sufD* [28-30], *clpA* [31,32], and *gpmM* [33] in bacteria, and the orthologs of *icd* [34] and *gpmM* [35] in mice (see Additional file 10). Two additional genes, *lpd* [36] and *cysD* [37,38], are known to be targets of regulation during aerobiosis or oxidative stress in bacteria. Of the remaining seven genes, three have other known functions (*napF*, *rseC*, and *fre* are all oxidoreductases) while the other four genes (*yaaU*, *yidH*, *ybeQ*, *ybhJ*) are poorly characterized. The *fre* gene has a high-confidence predicted functional interaction with a catalase and a peroxiredoxin (see Additional file 11) in the STRING database [39], based on their correlated expression patterns, and the *cysD* gene to a thioredoxin reductase, based on text mining evidence (see Additional file 11). Interestingly, *ybhJ* is a catalytically inactive paralog of the *E. coli* aconitase enzyme [40], which is also known to act as a superoxide sensor and a regulator of stress response genes [41].

The novel oxidative stress proteins contribute to homeostasis of iron and NAD(P)H

To better describe the specific roles played by these 13 proteins in oxidative stress defense, we measured cytoplasmic

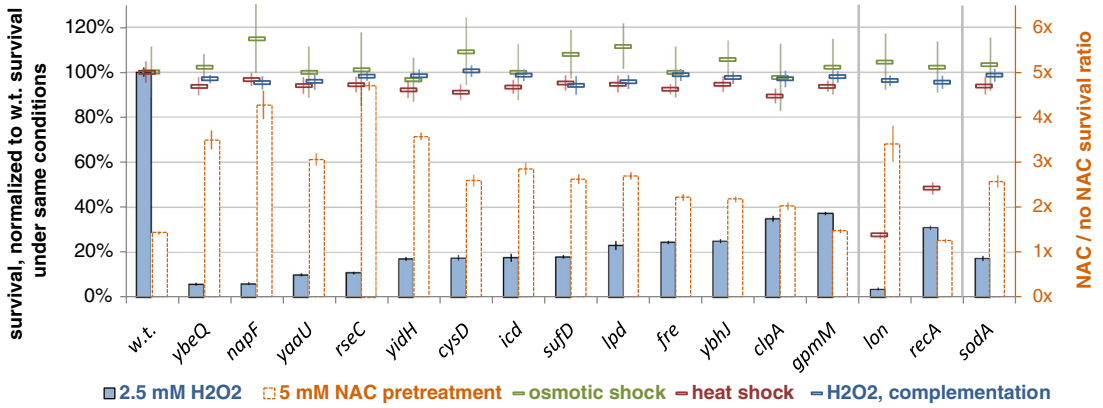


Figure 3 Survival of *Escherichia coli* deletion mutants in genes where codon adaptation was linked to aerobicity. H_2O_2 shock survival of mutants in putative oxidative stress genes (those with changes in codon adaptation in aerotolerant or obligately aerobic microbes, false discovery rate (FDR) ≤9.6% and ≤11.8%, respectively), without or with pre-treatment with N-acetylcysteine (NAC). More H_2O_2 concentrations are shown in Additional file 8. Deleted genes are *E. coli* representatives of clusters of orthologous groups (COGs) with codon adaptation correlated with oxygen in the environment, after controlling for confounding phenotypes or taxonomy. Strains are ordered by H_2O_2 survival, normalized to the wild-type survival under the same stress (13.8% for H_2O_2 after normalization shown as 100% on the plot). The outcome of the NAC rescue experiment is shown as a fold change in H_2O_2 survival over the same strain without NAC (right *x* axis). Additionally, the survival of each strain after heat and osmotic shocks is given for comparison; normalization as above. The strains *lon* and *recA* showed non-specific sensitivity and were thus separated on the plot, alongside *sodA*, which was included as a positive control. Error bars show the 95% C.I. of the mean, determined over at least four replicates.

levels of reactive oxygen species (ROS) and found them to be increased compared with the wild-type in only 3 of the 13 strains: *cysD*, *rseC*, and *yaaU*; Figure 4A), as well as in the *sodA* positive control. Therefore, the majority of these genes appear not to act by detoxifying ROS, but instead by preventing or repairing the damage that ROS cause to cellular components.

Based on their known molecular functions (see Additional file 12) and the lack of increased ROS generation, we hypothesized two possible general roles for these novel genes in oxidative stress resistance: 1) that they function by maintaining the cellular redox state through supporting NAD(P)H production, and 2) that they influence iron homeostasis. These two roles are also suggested by the known functions of the predicted functional interaction partners of these novel genes as presented in the STRING database [39] (see Additional file 13). NAD(P)H levels are known to affect oxidative stress resistance in different ways, including the NADH-driven AhpC enzyme that detoxifies peroxides, or the NADPH-driven regeneration of glutaredoxins and thioredoxins, which reverse oxidative damage to proteins [42]. Iron is well known to aggravate the damaging effects of H_2O_2 through hydroxyl radical-generating reactions [43,44].

Given that oxidative stress is known to upregulate synthesis of NADPH at the expense of NADH in bacteria [45-47], we focused on the former metabolite. We found experimentally that 8 of the 13 deletion mutants did indeed have reduced NADPH levels (Figure 4A, B), including 3 that could be implicated in NADPH production from previous knowledge (*lpd*, *gpmM*, and *icd*; see Additional file 12) and 5 additional genes (*yaaU*, *cysD*, *rseC*, *ybhJ*, and *yidH*). Moreover, pre-treating the bacteria with exogenous NADPH rescued the H_2O_2-sensitive phenotype of all these strains, but none of the strains with normal NADPH levels, (Figure 4A, C), lending support to the hypothesis that the diminished NADPH levels of these eight strains are a contributing factor to the reduced H_2O_2 resistance.

To determine the gene products that might act via regulating iron levels, we measured total cellular iron, and found it to be at higher concentrations relative to the wild type in five of the deletion mutants (*fre*, *sufD*, *rseC*, *lpd*, and *yidH*; Figure 4A, 4B). Three of the deleted genes could be connected to iron-related processes based on previous knowledge (*fre*, *sufD*, and *rseC*; see Additional file 12). A complementary assay using the iron chelator 2,2′-dipyridyl showed that the *rseC*, *fre*, *sufD*, and *yidH* deletion mutants had diminished sensitivity to H_2O_2 after dipyridyl pre-treatment (Figure 4A, B), with some response noted for *lpd*. This outcome corroborated the putative role of these genes in helping maintain iron homeostasis.

While there are many proteins whose translation efficiency could have evolved as adaptation to oxidative stress, our experiments indicate that two important mechanisms that are actually employed are an abundant supply of biological reducing agents and careful management of iron levels.

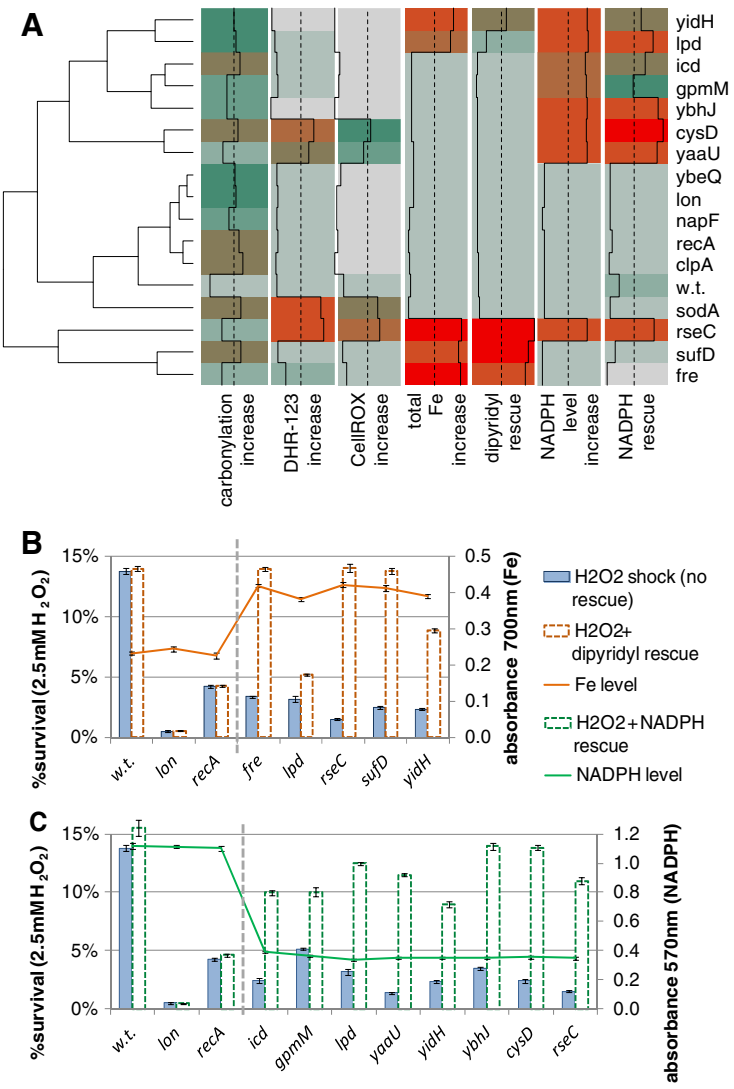


Figure 4 Mechanism of activity of the putative oxidative stress protection genes. (A) Overview of outcomes of seven experimental assays (columns) performed with the wild-type *Escherichia coli* and 15 deletion mutants. A larger value denotes a stronger observed effect; values are adjusted so that 0 signifies no effect and 1 signifies strong effect (values <0 and >1 are possible; for details of normalization for each assay, see Additional file 1). DHR-123 and CellROX are fluorescent dyes that measure reactive oxygen species. 2,2′-dipyridyl is an iron chelator. Genes are clustered based on similarity of the normalized response profiles of the mutants across the assays. Dashed lines denote the median. **(B, C)** A detailed display of the non-normalized measurements of: **(B)** iron levels and survival in the dipyridyl rescue experiment, or **(C)** NADPH levels and survival in the corresponding rescue experiment. Data are shown for wild-type *E. coli*, for *lon* and *recA* mutants (well-investigated genes expected not to act by the examined mechanisms), and those candidate genes in which our experiments support the proposed mechanism of action. Error bars show the 95% CI of the mean, determined over at least four replicates.

Validation of a role for codon optimality in additional phenotypic adaptations

To further investigate the generality of our methodology, we validated the predicted gene-phenotype links for two additional phenotypes: growth at increased temperatures and high salinity. Similarly to the H_2O_2 resistance experiments, we tested whether deletion of the orthologous *E. coli* gene from a COG with altered codon optimality in thermophile genomes proved deleterious after heat shock, while not affecting resistance to H_2O_2 and osmotic stress. Our experiments indicated a heat shock-specific protective role for 19 of 23 candidate genes (>40% decrease in mutant survival; Figure 5A), including the ClpS substrate modulator of the ClpAP chaperone-protease, which is known to direct its activity towards aggregated proteins [48]. Likewise, we also found that *E. coli* strains with deletions in three of six COGs with altered expression in halophiles had greatly decreased osmotic shock survival, but not decreased heat or H_2O_2 stress resistance (Figure 5B). The strongest response was seen in the mutant lacking *yjjB*, a conserved inner

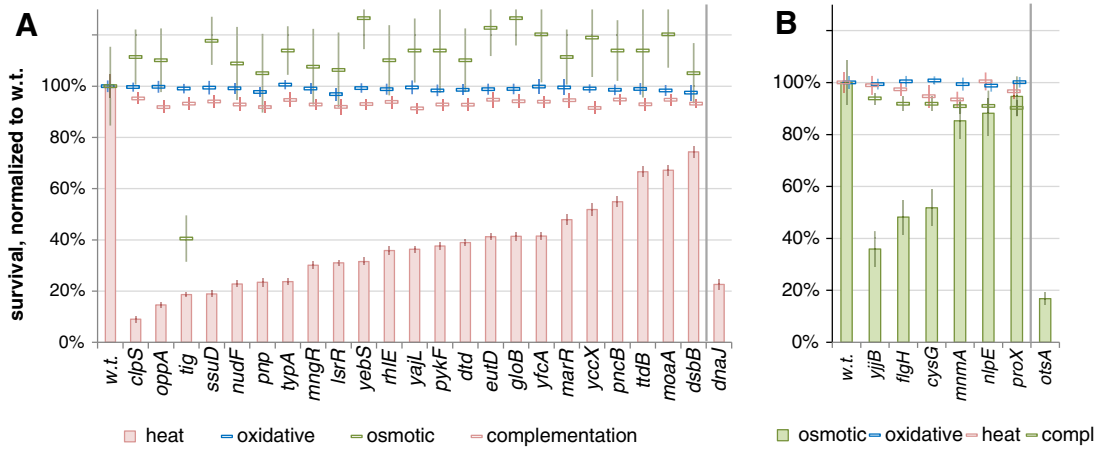


Figure 5 Survival of *Escherichia coli* deletion mutants for genes in which codon adaptation was linked to thermophilicity or halophilicity. Survival of strains with deletions in **(A)** putative heat stress and **(B)** putative osmotic stress genes discovered through differential codon adaptation in thermophilic or halophilic microbes (at false discovery rate (FDR) ≤6.6% and ≤33.9% respectively). Error bars show the 95% CI of the mean, determined over at least four replicates. Strains are ordered by survival, normalized to the wild-type survival under the same stress (23.6% for the heat and 21.3% for the osmotic shock after normalization, both shown as 100% on the plots). The *dnaJ* and *otsA* strains were included as a positive control.

membrane protein of unknown function. For both thermotolerance and osmotic stress resistance, expressing the deleted genes from a plasmid reversed the phenotype of all mutants; average survival was 93.3% and 91.7% of the wild type for the 23 and 6 complemented mutants, respectively (compared with average 37.4% and 67.4% for the deletion mutants) (Figure 5A, B). Thus, just as for oxidative stress resistance, altered translation efficiency across species can be used successfully to identify new genes with other specific functions.

Phenotypic effects of designed gene variants with reduced translation efficiency

While experiments on gene deletion strains can demonstrate the importance of a particular gene for a given phenotype, the natural changes in use of optimal codons are likely to cause less severe effects, such as changes in translation speed and/or accuracy. To show more directly that a change in the translation efficiency of the predicted gene families can bring about a phenotypic change, we selected two *E. coli* genes (*clpS* and *yjjB*) with prominent knockout effects on heat and osmotic shock survival, respectively (Figure 5A, B), and altered the genes' codon usage.

For each gene, we designed three sequence variants with unchanged protein sequence, but with progressively more optimal codons replaced by suboptimal ones (Figure 6A), thus tending toward the lower end of codon adaptation distribution in natural *E. coli* genes (Figure 6B) while still being within the range of observable codon usages. For all tested variants, the survival of *clpS* and *yjjB* deletion mutants complemented with the de-optimized genes was substantially lower than the survival of the wild-type *E. coli*, with a stronger reduction of survival in variants with a larger number of suboptimal codons. Expressing a wild-type gene fully rescued the heat/osmotic shock sensitive phenotype (Figure 6C-E). Of the other coding sequence features known to affect protein levels, secondary structures forming at the 5′ end of the mRNA are known to obstruct translation [25] if they are strong [7]. To rule out this variable, our sequence variants were designed to maximally conserve the original profiles of the mRNA folding free energies along the length of the genes (correlation with the wild-type $r \geq 0.89$ for all variants; see Additional file 14).

Discussion and conclusions

The large number of sequenced prokaryotic and eukaryotic genomes presents an opportunity to better understand how organism physiology adapts to the environment. However, even in model organisms, a significant proportion of genes remains poorly functionally characterized [49]. Computational gene function inference methods can direct experimental work to discover the biological roles of such genes. One approach for predicting gene function is to use the presence/absence patterns of orthologous genes across species [50,51]. However, these 'phylogenetic profiles' capture only a subset of functional relationships [52], even though research efforts continue to gradually improve the methods for mining such data [53]. Similarly to the phylogenetic profiles, the signatures of gene expression levels reflected in codon biases are in principle discoverable in any organism for which the genome sequence is available. Such signatures are known to evolve

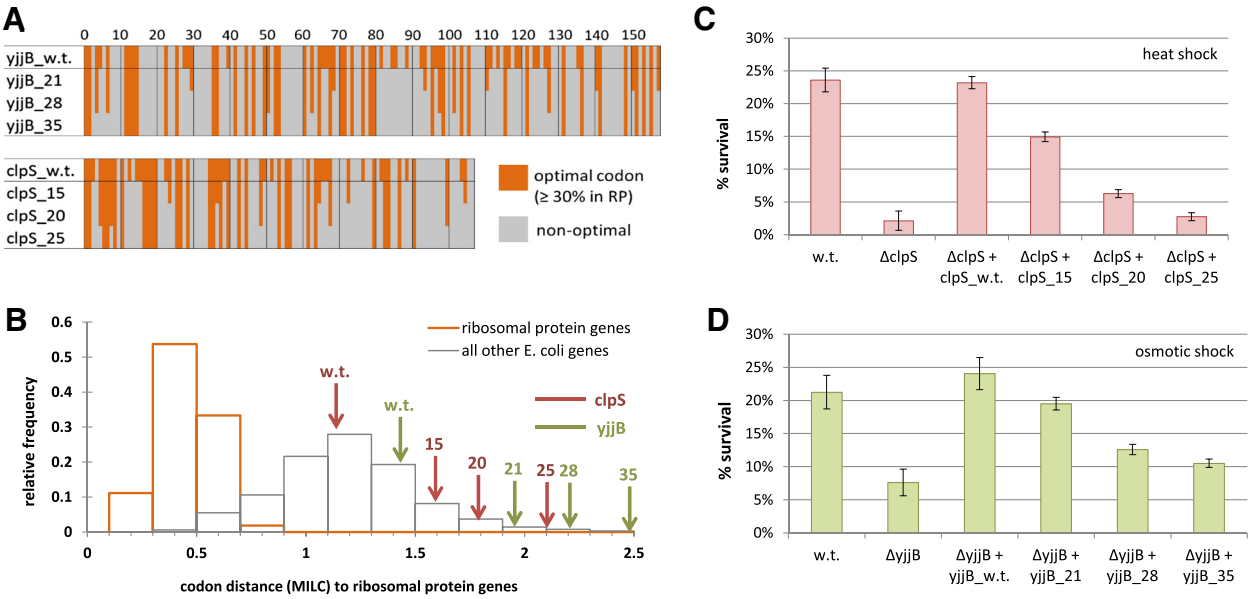


Figure 6 Phenotypes of deletion mutants upon complementation with gene variants with different levels of codon adaptation. (A) Design of *clpS* and *yjjB* gene variants with reduced proportion of optimal codons but with a preserved amino acid sequence. **(B)** Distributions of codon adaptation in *Escherichia coli* ribosomal protein genes (orange histogram) and other genes (grey histogram). The codon adaptation of the designed genes is marked by the arrows. **(C, D)** Heat and osmotic shock survival for the wild-type, the *clpS* and *yjjB* deletion mutants, and the same strains complemented by a wild-type or a gene variant with a reduced codon adaptation. Error bars show the 95% CI of the mean.

in response to the environment within selected gene families and functional modules [9,12,14,16], and also to contribute to speciation in bacteria [54]. In addition, the environment has a global, genome-wide effect on codon biases; organisms adapted to living in multiple habitats exhibit a wider spread of codon usages between genes [55]. Here, we have described a method to systematically exploit the signal present in the synonymous sites of particular genes, while controlling for confounding variables such as phylogenetic proximity or correlated phenotypes. In this way, we were able to discern those correlations likely to correspond to true cause-and-effect relationships between the gene translation efficiency and the phenotype, providing a general scheme for prioritization of gene annotation experiments.

We would expect our predictions to be complementary to existing genome context methods [56]; for instance,, the functional interactions in STRING [39] predicted from gene occurrence in genomes, or the interactions from conserved gene neighborhoods, did not retrieve the same 34 oxidative stress COGs we found (no statistically significant difference from a random gene set; see Additional file 15). This implies that the predictions about oxidative stress genes that we have made and experimentally validated are not easily reachable by phylogenetic profiling or gene neighborhood methods. Interestingly, the set of discovered genes did not include the catalases *katE/G/P*, which are prominent *E. coli* antioxidant proteins. On closer inspection, the catalase COGs were enriched with HE genes in aerotolerant organisms (*katE* 3.4-fold, *katG/P* 3.2-fold), but they had very few gene representatives in aerointolerant microbes (see Additional file 16) and did not reach statistical significance. Therefore, catalases are an example of a gene family whose function is better discoverable by phylogenetic profiling [50], which correlates gene presence/absence with phenotypes (see Additional file 16; Fisher's exact test $P = 10^{-25}$ and $P = 10^{-14}$ for *katE* and *katG/P*, respectively), serving to illustrate the complementary nature of the approaches.

To more systematically evaluate the sensitivity of our method, we performed a cross-validation test that retrieves *E. coli* genes with a known function through similarities of their translation efficiency profiles across different genomes (see Additional file 17). By extrapolation, we estimate that our method could retrieve on average 23% (10 to 29%; see Additional file 17) of the yet undiscovered genes relevant for different environmental responses. For comparison, phyletic profiling retrieves 32% (10 to 48%; see Additional file 17) of such unknown genes in the same setting. Therefore, the translation efficiency profiles possess around three-quarters of the detection power of the established phyletic profiling approach, but draw on an orthogonal and previously unexploited source of evolutionary signal. Moreover, the two methodologies can provide complementary gene function predictions: upon applying the models from the cross-validation test to a broader set of 3,534 *E. coli* genes, 69 genes were assigned a novel role in an environmental response exclusively by the translation

efficiency profiles, while 86 genes were predicted a function exclusively by the phyletic profiles; 101 more genes had predictions by both methods (see Additional file 17).

We have experimentally demonstrated that genes exhibiting differential signatures of translational efficiency in aerobes and anaerobes have a role in defending *E. coli* from H_2O_2-induced oxidative stress. In addition to genes with previously unknown functions, we found genes with known roles in central metabolism to be important for averting oxidative damage, consistent with a known broad metabolic reprogramming of bacteria upon oxidative stress [57,58]. As the aerobicity-related changes in translation efficiency of these gene families across genomes could not be explained by the bacterial and archaeal phylogeny, the aerobes have likely independently evolved convergent codon bias patterns in different lineages under similar selective pressures. Our work, together with the aerobicity-related signatures in codon biases previously found in nine yeasts [16], will help describe the evolution of functional categories for surviving oxidative injury. Such findings may have implications for strategies to modulate the response of pathogens to macrophage assault, or, more broadly, for understanding ROS-induced pathologies and aging in higher organisms.

In addition to aerotolerance, we also predicted and experimentally validated the phenotypic consequences of the thermophilicity-related and halophilicity-related evolutionary codon bias signatures, thus further showing that our approach will be applicable to diverse phenotypes. Furthermore, for representative genes, we complemented the corresponding deletion mutants with synthetic variants that had altered codon optimality. Previous work used designed libraries of gene variants harboring synonymous mutations to demonstrate how they influence the levels of the corresponding protein [6,25]. Similarly, we experimentally changed the translation efficiency of selected genes, but in this instance we demonstrated an organism-level effect of the synonymous changes – a phenotypic change that recapitulates the discovered evolutionary trends. This validation of the pervasive gene codon adaptation-phenotype relationships that we found through computational analysis suggests that changes in translation efficiency may be broadly acting drivers of phenotypic change.

In summary, our work introduces a novel 'genomic context' approach to infer gene function from differences in translation efficiency. We anticipate that the power of this purely sequence-based methodology will grow as the number of fully sequenced genomes, as well as the systematic phenotypic annotations of organisms, increases. To facilitate further experimental work on other putative gene-phenotype connections, we supply the inferred 200 high-confidence links from all COGs to 24 phenotypes (see Additional file 18), as well as a broader set of thousands of predictions obtained at relaxed specificity thresholds.

Materials and methods

Analyzed genomes and predicted gene expression levels

We downloaded 1,275 fully sequenced prokaryotic genomes from the NCBI database [59]. Multiple strains of a single species were removed to counter biases toward commonly re-sequenced species, such as *E. coli* or *Streptococcus pyogenes*, resulting in 912 species-level representatives. Because we later used *E. coli* MG1655 as the experimental model to verify our predictions, its genome was removed from the set of analyzed genomes in order to avoid circularity, leaving a final set of 911 genomes; other *E. coli* strains were also removed.

In brief, the computational pipeline for predicting high/low expression of genes from the genome sequence (Figure 1A; see Additional file 1) involved training an RF classifier [60] to discriminate known HE genes (RPs and chaperones) from the rest of the genome using codon frequencies, and comparing the predictions of the trained RF model against those of a baseline RF model trained on composition of the neighboring intergenic DNA of these genes. This approach is a variant of the methodology presented by Supek *et al.* [9], therefore, we include a detailed description of the computational pipeline (see Additional file 1), as well as a list of the changes from the original version.

Testing for correlation of lifestyles with gene expression

After obtaining the HE or non-HE label for each gene in the 911 genomes, we used Fisher's exact test to call enrichment/depletion for HE genes within COGs in genomes grouped by environment, phenotype, or taxonomy. In particular, for each COG: 1) we tested whether its representatives are more/less frequently HE in 514 aerotolerant microbes compared with the 214 obligate anaerobes (note that some organisms had this phenotype undefined); 2) we checked within the aerotolerant microbes for enrichment/depletion of HE genes in 296 obligate aerobes in comparison to the 217 facultative aerobes; 3) we tested for enrichment/depletion of HE genes in 142 thermophiles versus 643 non-thermophiles; and 4) assessed enrichment/depletion of HE genes in 40 halophiles versus 140 non-halophiles (again, for the majority of organisms, the halophilicity was undefined).

Additionally, the same tests were performed for other descriptions of phenotypes or taxonomy, obtained as follows. Descriptions of the microbes' environments or phenotypes were assembled from the NCBI Entrez Microbial Genome Properties website [61], followed by manual curation, particularly for pathogenicity phenotypes. All properties of interest were encoded as a series of binary ('yes/no' or 'low/high') categorical variables,

possibly with missing values; the two continuous variables – GC content and genome size – were discretized into four classes. In total, this yielded 24 phenotypic features and 6 features describing the GC content and genome size (see Additional file 5). The organisms' taxonomy was handled in a similar manner, where the possible taxonomic subdivisions at the domain, phylum, class and order level were encoded as 'yes'/'no'/'not applicable' categorical variables, yielding 70 features (see Additional file 5), for a total of 100 features per genome.

Thresholds for COG size, enrichment, and statistical significance

We excluded from testing all COG groups with fewer than 20 representative genes in total (counted over all genomes, regardless of the phenotype), or with more than 10,000 genes in total, leaving 4,387 COGs of possible interest.

For each of these COGs, all phenotypes of interest were screened for enrichment with HE genes of two-fold or higher (or ≤0.5-fold depletion) in that specific phenotype. compared with the organisms known not to have the phenotype. These COGs were then tested for statistical significance of the enrichment using Fisher's exact test (two-tailed) at $P < 10^{-2}$. The four phenotypes that we subsequently validated experimentally were: 1) 514 aerotolerant microbes versus 214 obligate anaerobes: 295 of 2,847 tested COGs were significantly enriched/depleted for HE genes; FDR = 9.6%; 2) 296 obligate aerobes versus 217 facultative aerobes: 160 of 1,887 tested COGs were significantly enriched/depleted for HE genes; FDR = 11.8%; 3) 142 thermophiles versus 643 non-thermophiles: 346 of 2,287 tested COGs were significantly enriched/depleted for HE genes; FDR = 6.6%; and 4) 40 halophiles versus 140 non-halophiles: 55 COGs of 1,863 tested COGs were significantly enriched/depleted for HE genes; FDR = 33.9%.

Controlling for confounding effects of other phenotypes and phylogenetic proximity

Even if a strong and highly significant correlation between increased expression in a COG and a phenotype is found, this in itself does not imply a causal relationship between the two variables. A common explanation involves the correlation being due to both variables being causally linked to a third variable (or to more variables). To control for such cases and prioritize the causal relationships within a potentially much larger number of correlations, we introduce a methodology based on supervised machine learning that measures whether a specific phenotype has an independent contribution to predicting gene expression levels, after controlling for all other phenotypes/environments and the phylogenetic relatedness. This computational method relies on the use of a classifier that can infer highly complex relationships involving many different independent variables (here: phenotypes or phylogeny) and one dependent categorical variable (here: the HE/non-HE labels on genes). In other words, the classifier 'learns' to predict gene expression for genes in a certain COG from the phenotypic, environmental, or phylogenetic relatedness of the corresponding organisms. The procedure consists of the following steps:

1) Construct the dataset. For each gene family (here, COG) make a separate dataset that has as many instances (examples) as there are genes in the COG (possibly >1 per genome), and as many independent variables (features) as there are phenotypes and phylogenetic subdivisions (here, 100), plus one dependent variable (class) with the predicted expression levels in the form of HE/non-HE labels.
2) Train the classifier and evaluate the model. Run the classifier and evaluate the accuracy of its predictions (here, using the area under the receiver operating characteristic curve (AUC) score [62]), while employing a cross-validation scheme. This setup penalizes overly complex models that over-fit to noise in the data, while rewarding models that generalize to unknown data better. Here, we used theRF [60] classifier as implemented in the FastRandomForest software [63] that integrates into the Weka Environment for Knowledge Analysis [64].
3) Repeat for randomized datasets. Shuffle a single dependent variable (here, phenotype) while leaving other phenotype/taxonomy-describing variables intact, and repeat the classifier training, and measure the cross-validation AUC score. Repeat this step 30 times while re-shuffling the same variable each time.
4) Test for consistent decrease in AUC score. Calculate a Z-score (number of standard deviations a measurement is away from the mean) for $AUC_{original}$ compared with a distribution of 30 $AUC_{shuffled}$ values. From the Z-score, find a one-tailed P value (using the cumulative distribution function of the normal distribution) that indicates whether the AUC score consistently decreases with randomization of the variable of interest.
5) For all phenotypes/environments of interest, repeat randomization test (steps 3 and 4). Here, these are the two tested aerobicity-related phenotypes; see section 'Testing for correlation of lifestyles to gene expression' above.
6) For all COGs, repeat steps 1 to 5.

The rationale behind the procedure is that shuffling one of the variables will destroy the information that variable might carry and that is relevant for predicting the high/low expression level. If this same information can be

recovered from the other variables (possibly by combining them), the accuracy of classification will not be lowered by the randomization, whereas in cases where the variable in question is informative of the expression level of the genes in a way that cannot be substituted for by the remaining variables, the accuracy of the classification model will be reduced by randomization.

Bacterial strains, growth conditions, and stresses

All the used strains as well as specifics of their construction are listed (see Additional file 19). All strains were derived from wild-type sequenced *E. coli* MG1655 by P1 transduction and/or transformation. Relevant plasmids were purchased from the ASKA library [65]. Bacteria were grown in LB at 37°C, to the mid-exponential phase (OD_{600} = 0.2 to 0.3).

For the H_2O_2 treatment, they were washed in 0.01 M $MgSO_4$ and incubated at 37°C for 20 minutes in the absence and presence of 0.5 mM, 2.5 mM, and 20 mM H_2O_2. Osmotic shock was achieved by exposing exponentially growing *E. coli* to 1 M NaCl (final) for 1 hour, and heat shock was achieved by growing *E. coli* at 56°C for 100 minutes. Viable cell counts were always estimated by plating serial dilutions on LB plates and growing overnight at 37°C.

To test if the mortality of *E. coli* deletion mutants upon exposure to H_2O_2 was caused by the increased ROS production, we performed a rescue experiment using 5 mM N-acetyl cysteine (NAC), a known ROS scavenger. Overnight cultures of *E. coli* were diluted 200 times and grown in the presence ofthe 5 mM NAC until the mid-exponential phase. Cells were then washed and oxidized with 2.5 mM H_2O_2 for 20 minutes, and survival was measured as described above.

Measuring protein carbonylation and reactive oxygen species production

Exponentially growing bacteria were harvested from LB medium. *E. coli* cells were pelleted by centrifugation and resuspended in 10 mM PBS (pH 7.4), supplemented with a mixture of protease inhibitors (Roche, Basel, Switzerland). Resuspended cells were frozen immediately in liquid nitrogen. Cells were broken by a mechanical homogenizer, and centrifuged for 20 minutes at 12,000 × g. Samples were then supplemented with 10 mg/100 μl lipid removal agent (13360-U; Sigma, St. Louis, Missouri, USA), incubated for 1 hour at room temperature with shaking, and centrifuged for 15 minutes at 10,000 × g. The amount of protein in the supernatant was measured by the Lowry method. Protein extracts diluted to 10 μg/ml were loaded into wells (Maxisorb; Nunc, Roskilde, Denmark) and incubated overnight at 4°C to allow proteins to adsorb to the surface, followed by 0.6 mM dinitrophenyl hydrazine (DNPH) derivatization of adsorbed proteins and detection of derivatized dinitrophenol (DNP)-carbonyl by a mouse DNP-specific monoclonal antibody conjugated to horseradish peroxidase. Subsequent incubation with enzyme substrate 3,3′,5,5′-tetramethylbenzidine (TMB; Sigma, St. Louis, Missouri, USA) resulted in a colored product that was quantified using a microplate reader at 450 nm.

ROS levels were determined by labeling *E. coli* strains with 25 μM dihydrorhodamine-123 for 10 minutes in the dark, in the absence or presence of hydrogen peroxide. Cells (approximately 10^6) were then washed in minimal medium, and their fluorescence was measured with excitation at 500 nm and emission at 530 nm. In addition, *E. coli* strains were labeled with CellROX™ Deep Red reagent (Invitrogen, Carlsbad, California, USA) in the absence or presence of hydrogen peroxide. Cells (approximately 10^6) were washed in minimal medium, and their fluorescence was measured with excitation at 630 nm and emission at 665 nm.

Measurement of cellular NADPH and Fe, and rescue experiments

We measured intracellular NADPH level using a commercial kit (Vybrant Cytotoxicity Assay Kit; Molecular Probes, Eugene, Oregon, USA) that is normally used to monitor the release of the enzyme glucose 6-phosphate dehydrogenase (G6PD) from damaged cells. Oxidation of glucose-6-phosphate by G6PD results in the generation of NADPH, which in turn leads to the reduction of resazurin by diaphorase to yield fluorescent resorufin. We took advantage of the second reaction to measure NADPH levels directly, while filtering the cellular extract of each studied strain through a 3 kDa cutoff centrifugal filter (Amicon Ultra; Millipore, Billerica, Massachusets, USA) to prevent the cellular proteins (including G6PD) from creating a background with the reaction mixture. A sample (100 μL) of each cellular filtrate was distributed into wells twice in duplicate, and the level of NADPH was determined as follows. A reaction mixture was prepared by dissolving a lyophilized mixture of diaphorase, glucose-6-phosphate, and $NADP^+$ (Component C of the kit) in 0.5 M Tris buffer pH 7.5 (Component D of the kit). The reaction mixture was then combined with the solution of resazurin so that the final concentration of resazurin was 30 μM (component A). Then, 100 μL of the final mixture was loaded onto the samples distributed in the wells, and incubated at 37°C for 5 hours. The amount of NADPH was measured as the absorbance at 570 nm.

To test which *E. coli* strains were rescued by pretreatment with NADPH, exponentially growing *E. coli* strains were first exposed to 1% v/v toluene in the presence of 10 mM EDTA (known to permeabilize the bacterial membranes to NADPH [66]) and then exposed to 20 μM NADPH dissolved in 10 mM PBS, (pH 7.4). Cells were then treated with H_2O_2, washed in 0.01 M $MgSO_4$,

and incubated at 37°C for 20 minutes in the absence or presence of 2.5 mM H_2O_2. Viable cell counts were estimated by plating serial dilutions on LB plates and growing overnight at 37°C.

We measured the level of cellular iron (both Fe^{2+} and Fe^{3+}) as described by Rad *et al*. [67]. In particular, about 10^7 exponentially growing *E. coli* cells were pelleted and incubated overnight at 110°C without tube caps. After evaporation of liquid, 1 ml of 10 M HCl was added, and samples were incubated for 4 h at 60°C. Next, the content of each tube was diluted twofold with 10 M HCl, and absorbance was measured at 351 nm. To test which *E. coli* strains were rescued by pre-treatment with 2,2′-dipyridyl (iron chelator), exponentially growing *E. coli* strains were exposed to 0.4 mg/ml (final concentration) dipyridyl. Cells were then treated with H_2O_2 as described above, and viable cell counts were estimated by plating serial dilutions on LB plates and growing overnight at 37°C.

Phenotypic effects of introducing synonymous changes in the *clpS* and *yjjB* genes

For each of the two selected *E. coli* genes, we designed three additional variants with synonymous changes: for *clpS*, 15, 20 and 25 optimal codons were replaced with non-optimal ones, and for *yjjB*, 21, 28, and 35 codons were changed. The number of changed codons was chosen to be proportional to the sequence length (*clpS* is 107 codons long and *yjjB* is 158 codons long). The optimality of a codon was defined as its frequency in the *E. coli* RP)genes, normalized to sum to 100% for each amino acid. All introduced changes had to reduce the optimality of the original codon by at least 30% below the original value, while not falling below 3% to avoid the extremely rare codons such as the AGG or AGA arginine codons (0.6% and 0% usage in *E. coli* RP). Therefore, with our gene variants, we aimed to incorporate a large number of moderate changes in codon optimality, rather than a small number of drastic changes, assuring a more even distribution of the codon optimality levels along the length of the gene. In the *yjjB* sequences (including wild-type sequence), we also abolished a HsdR site, AACGTTCCCGTGC, by changing CCC-GT**G**-C to CCC-GT**A**-C (a synonymous change, where one suboptimal valine codon was exchanged for another).

To control for stable secondary structures in the mRNA that may inhibit protein translation independently of the use of optimal codons, we used a script that in each step replaces five (for *clpS*) or seven (for *yjjB*) randomly chosen codons in the sequences with suboptimal ones (while obeying the rules described above), repeating the random selection 100 times, and selecting the variant with the predicted RNA folding energy profile most similar to the original gene. Then, another set of five or seven codons are replaced, again with 100 random samplings, keeping the least changed RNA folding profile, and so on. The RNA folding free energy profiles for the genes were calculated for the 42-nucleotide folding windows using the *hybrid-ss-min* program from the UNAFold 3.6 package [68], with default parameters (NA = RNA, t = 37, [Na+] = 1, [Mg++] = 0, maxloop = 30, prefilter = 2/2). The difference in RNA folding profile between the mutated and the original sequence was computed as the root mean square deviation of folding free energies for all 42-nt windows. All sequences are given in Additional file 20, and the RNA secondary structure folding free energy profiles are given in Additional file 14.

The *clpS* and *yjjB* deletion mutants were complemented with a pJ801 plasmid encoding either the wild-type gene, or the variants with introduced synonymous mutations described above. The plasmids with the appropriate inserts were purchased from DNA2.0 and bore a kanamycine resistance cassette, and the genes were under the control of a rhamnose-inducible promoter. Overnight cultures of *E. coli* strains were diluted 200 times in LB medium, supplemented with 1.5 μM rhamnose and grown for 2 to 3 hours to an OD of 0.2 to 0.3. The *clpS* mutants were exposed to heat shock (100 minutes at 56°C) and *yjjB* mutants to osmotic shock (1 hour at 1 M NaCl and 37°C) and survival measured as for the deletion mutants.

Additional files

Additional file 1: Supplementary Materials and Methods. Contains sections on: creating reference sets of highly expressed genes; the random forest classifier; detecting selection for translational efficiency in genomes; assigning 'highly expressed' labels to individual genes; a figure with a schematic of the computational workflow; comparison with the procedure from Supek *et al*. [9]; gene expression data sources; and normalization of experimental results.

Additional file 2: Agreement with expression data for the predictions about highly expressed (HE) genes, and a comparison with the original 'optimized codon usage' (OCU) method [9]. *P* values are from a Mann-Whitney test for a difference in microarray signal levels between the HE and non-HE genes, or the OCU and non-OCU genes. The 'ratios' were calculated between the average microarray signal of the two groups. The ratio of ribosomal proteins versus whole genome is given for a sense of scale; the ribosomal protein genes are expected to be very highly expressed.

Additional file 3: The relative proportion of highly expressed genes is lower in larger genomes. This correlation was previously explained [9] by different proportions of various gene functional categories in smaller or larger genomes. Many of the functional categories, in turn, tend to have a general preference for higher or lower expression. For instance, larger genomes have a disproportionally increased number of gene regulators, which have a strong tendency to low expression. Smaller genomes, on the other hand, have a higher proportion of ribosomal proteins, whose absolute number is roughly fixed across genomes, regardless of their size.

Additional file 4: Correlations of mRNA 5′ end folding free energies and various codon indices with gene expression levels. The free energies are a measure of the stability of the structures (more negative = more stable) and are calculated in windows of 42 nucleotides in length

on the mRNA sequence using the *hybrid-ss-min* program from UNAFold 3.6 with default parameters, as in [25]. The three 42-nt window positions investigated are: (−4 to 37), found to have a strongest correlation to protein levels [25]; (−20 to 21), a window centered over the start codon; and (−30 to 11), a window centered on the common location of the Shine-Dalgarno sequence at −9 [69]. The −10 kcal/mol figure is the approximate limit for the mRNA folding free energy in 42-nt windows; at negative values below this, the mRNA folding starts to have a considerable effect on translation efficiency [7]. The mRNA coordinates are given relative to the start codon, where 1 is the A in AUG. The codon indices are: CAI [4], B [70], and MILC [5]. RF, probability score obtained from a random forest classifier [9]. All codon indices use the same 'reference set' of known highly expressed genes as used in our analyses (see Supplementary Methods in Additional file 1).

Additional file 5: The 100 features describing each organism which were used in the search for the phenotypes predictive of the changes in translation efficiency within clusters of orthologous groups (COGs). All features are binary variables, and can be undefined for some organisms. We included 70 features describing the phylogeny (left/middle columns) and the 6 features describing genome size and GC content (right column, top) to ensure that correlations detected with the remaining 24 features (phenotypes, right column) could not be explained by the phylogeny or the genomic size/GC. #pos, number of organisms marked as positive for a specific feature; #neg, number of organisms marked as negative for a specific feature.

Additional file 6: Genomes for which the optimal codons inferred from over-representation in highly expressed (HE) genes overall did not match the expected optimal codons inferred from the genomic tRNA repertoire. The nine twofold degenerate amino acids were examined. An optimal codon (HE column) was defined as over-represented at $P < 0.001$ in a Fisher's exact test on codon counts in HE versus the non-HE genes; a non-significant result means no codon is optimal. The codons expected to be optimal from tRNAs tRNA column) are defined in the genomes in which tRNA genes with only one of the two possible anticodons were present (found by tRNAscan-SE), then the codon matching that anticodon by canonical Watson-Crick pairing was considered tRNA-optimal, and the other codon, which uses wobble pairing, was considered tRNA-suboptimal. The table shows 71 (of the 911 total) genomes for which the optimal and the tRNA-optimal codons disagreed in at least 3 of 9of the testable amino acids (# aa column). In 651/911 genomes, there were 0/9 disagreeing amino acids, and 1/9 for a further 135 genomes. Thus, in the 71 genomes, the expression level-related codon bias did not, overall, clearly relate to the tRNA gene repertoire, and may possibly not reflect translational selection, but rather another, unknown factor. We thus excluded the 71 genomes, and re-ran the subsequent analyses to verify if our findings were robust to inclusion of these genomes (see Additional file 7).

Additional file 7: Robustness of the 200 discovered clusters of orthologous groups (COGs)-phenotype links to the exclusion of 71 genomes for which codon biases were not clearly related to the tRNA gene repertoires. Excluded genomes are listed in Additional file 6. **(A)** The $\log_2$ enrichment of the 200 COG-phenotype links with the full set of 911 genomes, and after exclusion of the 71 genomes. **(B)** Same as **(A)**, but limited to the links that we experimentally validated. In the original analysis, a threshold of $\log_2$ enrichment of ≥1 or ≤−1 was a requirement for calling the 200 COG-phenotype links; after excluding the 71 genomes, 195 COG-phenotype links still met this criterion. **(C)** The $\log_{10} P$ value for significance of the enrichment/depletion (two-tailed Fisher's exact test), again compared between the original and the reduced genome sets. **(D)** Same as **(C)**, but only for the COG-phenotype links with $\log_{10}P \geq -6$. In the original analysis, $\log_{10}P \leq -2$ was required for calling the 200 COG-phenotype links; after excluding the 71 genomes, 173/200 links still had $\log_{10}P \leq -2$, and 185/200 still had $\log_{10}P \leq -1.7$ ($P < 0.02$).

Additional file 8: Survival of *Escherichia coli* deletion mutants after oxidative stress induced by different hydrogen peroxide concentrations. Survival after heat and osmotic shock is given for comparison. Deleted genes are on the *x* axis. The *y* axis shows the survival of the mutant, normalized to the survival of the wild type (w.t.) under the same conditions, which was 45.6% for 0.5 mM H_2O_2, 13.8% for 2.5 mM H_2O_2, 4.2% for 20 mM H_2O_2, 23.6% for heat shock, and 21.3% for osmotic shock. The *lon* and *recA* mutants are shown separately as they exhibited a non-specific stress response, being sensitive also to osmotic and heat stress. *sodA* is a known oxidative stress defense gene, serving as a positive control.

Additional file 9: Complementing *Escherichia coli* deletion mutants with wild-type genes. Survival of *E. coli* deletion mutants in the putative oxidative stress response genes with and without the corresponding genes expressed from a plasmid.

Additional file 10: Supporting evidence for putative oxidative stress genes. A survey of the evidence in the literature offering support for the involvement of *sufD*, *clpA*, *icd*, *gpmM*, *lpd*, and *cysD* genes in oxidative stress resistance of various organisms.

Additional file 11: Functional interactions with known oxidative stress genes. Predicted functional interactions between 34 clusters of orthologous groups (COGs) we found to have codon adaptation that correlates with the aerobic lifestyle, and 30 COGs encoding known *Escherichia coli* oxidative stress response proteins. The predicted interactions are from the STRING v9.0 database, using exclusively co-expression (top part of table), or exclusively text mining (bottom part) evidence. Only interactions marked as high confidence by STRING (confidence ≥0.7) are shown.

Additional file 12: Literature data suggesting putative antioxidant mechanism of action assignments. Listed for the *sufD*, *fre*, *rseC*, *gpmM*, *lpd*, and *icd* genes.

Additional file 13: The functional context of the 13 *Escherichia coli* gene representatives of the clusters of orthologous groups (COGs) differentially expressed in aerobic microbes. The genes *recA* and *lon* are not shown because their deletion mutants showed non-specific stress sensitivity (Figure 3). Lines represent the predicted functional interactions from the STRING 9.0 database (medium confidence level, ≥0.4), while dots represent all proteins interacting with at least 1 of the 13 proteins. A large, highly interconnected set of interacting ribosomal proteins is not shown for clarity. The larger, colored dots are proteins annotated with one of the selected functional categories in *E. coli* (right panel). Hollow circles in *fre* or *rseC* or thick border in *napF* denote putative assignments we inferred for these genes from the literature; all other functional annotations were from the Uniprot-GOA (Gene Ontology Annotation) database. All shown functional categories were found to be enriched among the 13 proteins plus interactors at $P < 0.05$ (hypergeometric distribution, corrected for multiple testing) using GeneCodis 2.0. Proximity of the circles in the figure roughly corresponds to their functional similarity, as optimized by the Edge Weighted Spring Embedded layout in Cytoscape 2.8.1, edge weights being derived from interaction confidence levels in STRING.

Additional file 14: Profiles of folding free energies in 42-nucleotide windows along the *clpS* and *yjjB* gene mRNAs. The *x* axes show the starting coordinate (in nucleotides) of the 42-nt window. The folding free energies were calculated using the *hybrid-ss-min* program from UNAFold 3.6 software with default parameters. Alongside each *Escherichia coli* gene (marked 'w.t.'), three variants are given with introduced synonymous changes that reduce codon optimality (Figure 5); the number given after the word 'variant' is the number of codons that have been altered with respect to the wild type. A 14-nt ribosome binding site sequence, AGGAGGUAAAACAU, was prepended before the AUG start codon when determining the folding free energies, as was the case for the actual genes. For each variant, Pearson's correlation coefficient, *r*, and the root mean square deviation (RMSD) are given as measures of similarities of their folding free energy profiles to the wild-type sequence.

Additional file 15: Distributions of predicted functional interactions at different confidence levels. Functional interactions were examined between (1) 30 clusters of orthologous groups (COGs) known to have a role in the oxidative stress response, labeled 'known versus known'; (2) the 'known' group and the 34 COGs found to be differentially expressed between aerotolerant organisms and anaerobes, or between obligate and facultative aerobes, labeled 'diffExpr versus known'; and (3) the 'known' group and a 100 randomly chosen COGs, labeled 'randomSet versus known'. Two of the 34 COGs were also in the 'known' group, and

their functional interactions did not count for the 'diffExpr' group; these were COG0719 (*E. coli sufD* and *sufB* genes) and COG1249 (*E. coli lpd*, *ykgC*, *gor*, and *sthA* genes). The predicted functional interactions are from the STRING v9.0 database [39], the scores vary from 0 to 1; STRING declares interactions between 0.15 and 0.40 to have low confidence, between 0.40 and 0.70 to have medium confidence, and above 0.70 to have high confidence. For details on how the scores are computed for each individual source of data, please refer to references given at the STRING website. The *P* values are from a χ^2 test.

Additional file 16: Relationships of the aerotolerance phenotype with the presence/absence patterns and with the codon adaptation of the catalase genes. Tables show the count of organisms (not genes) with the clusters of orthologous groups (COGs) being absent (first column), present with one or more genes that are all non-highly expressed (HE) (second column), or present with one or more genes of which at least one in the genome is HE (third column). The tables below show the same frequencies, but normalized to the total number of aerotolerant or strictly anaerobic organisms. For both COGs, the presence of the catalases in the genome is strongly and significantly correlated with aerobicity (top right panel for each COG). However, the codon adaptation of the catalases is strongly but not significantly correlated with aerobicity (bottom right panel for each COG), because of the low numbers of strictly anaerobic genomes that have a catalase gene present.

Additional file 17: A cross-validation test of the ability to retrieve functionally related genes, starting from the translational efficiency profiles of clusters of orthologous groups (COGs) across genomes (left panel), or the gene presence/absence profiles (right panel, equivalent to a standard phyletic profiling approach). The test uses *Escherichia coli* K12 genes that are assigned to a COG and that are annotated with one of the five Gene Ontology (GO) categories above, and compares these genes with a sample of other *E. coli* genes that are in COGs but that do not have this GO function assigned. The size of the sample of these 'negative genes' is 19 times the number of 'positive' genes, which thus make up 5% of the combined dataset, mimicking a realistic distribution. Next, a Random Forest model is trained to discriminate the two groups of *E. coli* genes, and tested in a *n*-fold cross-validation scheme (using Weka 3.7.9), where *n* is the number of positive genes for that GO. The plots are precision-recall curves: recall is on the *x* axis, precision on the *y* axis. Importantly, the translation efficiency models (left panel) do not have access to gene presence/absence information, and must discriminate the groups only from the codon biases of the present genes; absent genes are encoded as missing data. The measure of translation efficiency in the profiles is the difference in classifier probabilities of the intergenic DNA versus codon usage data (Figure 1A, left versus right).

Additional file 18: An exhaustive list of the inferred clusters of orthologous groups (COGs)-phenotype links.

Additional file 19: A list of *Escherichia coli* strains used.

Additional file 20: Designed variants of *Escherichia coli clpS* and *yjjB* genes, with progressively more optimal codons replaced by suboptimal ones (Figure 6). The lowercase 'a' in the *yjjB* sequences denotes a replacement of the original G with an A to abolish a HsdR site.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
AK and TC carried out all experimental assays; FS conceived the study, performed the computational analyses, and drafted the manuscript; and AK, TG, and BL participated in the study design and data interpretation, and contributed substantially to the manuscript. All authors read and approved the final manuscript.

Acknowledgements
This work was supported by grants from the Spanish Ministry of Economy and Competitiveness (BFU2009-9618, BFU2008-00365 and 'Centro de Excelencia Severo Ochoa 2013-2017' SEV-2012-0208), an ERC Starting Grant, ERASysBio PLUS, AGAUR, the EMBO Young Investigator Program, and the EMBL-CRG Systems Biology Program. The work of FS was supported in part by Marie Curie Actions and by grant ICT-2013-612944 (MAESTRA). The experimental part of this work was carried out at and financed by the Mediterranean Institute for Life Sciences (MedILS). We are grateful to Ivan Matić for valuable discussions and to Miroslav Radman for feedback on the manuscript.

Author details
[1]Mediterranean Institute for Life Sciences (MedILS), 21000 Split, Croatia. [2]Bioinformatics and Genomics Programme, Centre for Genomic Regulation (CRG), Dr. Aiguader 88, 08003 Barcelona, Spain. [3]Universitat Pompeu Fabra (UPF), 08003 Barcelona, Spain. [4]EMBL/CRG Systems Biology Research Unit, Centre for Genomic Regulation (CRG), Dr. Aiguader 88, 08003 Barcelona, Spain. [5]Institució Catalana de Recerca i Estudis Avançats (ICREA), Pg. Lluis Companys 23, 08010 Barcelona, Spain.

Received: 24 May 2013 Accepted: 3 March 2014
Published: 3 March 2014

References
1. Akashi H: **Synonymous codon usage in Drosophila melanogaster: natural selection and translational accuracy.** *Genetics* 1994, **136:**927–935.
2. Bulmer M: **The selection-mutation-drift theory of synonymous codon usage.** *Genetics* 1991, **129:**897–907.
3. Kanaya S, Yamada Y, Kudo Y, Ikemura T: **Studies of codon usage and tRNA genes of 18 unicellular organisms and quantification of Bacillus subtilis tRNAs: gene expression level and species-specific diversity of codon usage based on multivariate analysis.** *Gene* 1999, **238:**143–155.
4. Sharp PM, Li WH: **The Codon Adaptation Index–a measure of directional synonymous codon usage bias, and its potential applications.** *Nucleic Acids Res* 1987, **15:**1281–1295.
5. Supek F, Vlahovicek K: **Comparison of codon usage measures and their applicability in prediction of microbial gene expressivity.** *BMC Bioinformatics* 2005, **6:**182.
6. Welch M, Govindarajan S, Ness JE, Villalobos A, Gurney A, Minshull J, Gustafsson C: **Design parameters to control synthetic gene expression in Escherichia coli.** *PLoS ONE* 2009, **4:**e7002.
7. Supek F, Smuc T: **On relevance of codon usage to expression of synthetic and natural genes in Escherichia coli.** *Genetics* 2010, **185:**1129–1134.
8. Rocha EPC: **Codon usage bias from tRNA's point of view: redundancy, specialization, and efficient decoding for translation optimization.** *Genome Res* 2004, **14:**2279–2286.
9. Supek F, Škunca N, Repar J, Vlahoviček K, Šmuc T: **Translational selection is ubiquitous in prokaryotes.** *PLoS Genet* 2010, **6:**e1001004.
10. Hershberg R, Petrov DA: **General rules for optimal codon choice.** *PLoS Genet* 2009, **5:**e1000556.
11. Drummond DA, Wilke CO: **Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution.** *Cell* 2008, **134:**341–352.
12. Von Mandach C, Merkl R: **Genes optimized by evolution for accurate and fast translation encode in Archaea and Bacteria a broad and characteristic spectrum of protein functions.** *BMC Genomics* 2010, **11:**617.
13. Karlin S, Brocchieri L, Mrázek J, Kaiser D: **Distinguishing features of δ-proteobacterial genomes.** *Proc Natl Acad Sci* 2006, **103:**11352–11357.
14. Carbone A: **Computational prediction of genomic functional cores specific to different microbes.** *J Mol Evol* 2006, **63:**733–746.
15. Karlin S, Brocchieri L, Campbell A, Cyert M, Mrázek J: **Genomic and proteomic comparisons between bacterial and archaeal genomes and related comparisons with the yeast and fly genomes.** *Proc Natl Acad Sci USA* 2005, **102:**7309–7314.
16. Man O, Pilpel Y: **Differential translation efficiency of orthologous genes is involved in phenotypic divergence of yeast species.** *Nat Genet* 2007, **39:**415–421.
17. Grocock RJ, Sharp PM: **Synonymous codon usage in Pseudomonas aeruginosa PA01.** *Gene* 2002, **289:**131–139.
18. Retchless AC, Lawrence JG: **Quantification of codon selection for comparative bacterial genomics.** *BMC Genomics* 2011, **12:**374.
19. Knight RD, Freeland SJ, Landweber LF: **A simple model based on mutation and selection explains trends in codon and amino-acid usage and GC composition within and across genomes.** *Genome Biol* 2001, **2:**RESEARCH0010.

20. Chen SL, Lee W, Hottes AK, Shapiro L, McAdams HH: **Codon usage between genomes is constrained by genome-wide mutational processes.** *Proc Natl Acad Sci USA* 2004, **101:**3480–3485.
21. Dos Reis M, Savva R, Wernisch L: **Solving the riddle of codon usage preferences: a test for translational selection.** *Nucl Acids Res* 2004, **32:**5036–5044.
22. Sharp PM, Bailes E, Grocock RJ, Peden JF, Sockett RE: **Variation in the strength of selected codon usage bias among bacteria.** *Nucleic Acids Res* 2005, **33:**1141–1153.
23. Carbone A, Képès F, Zinovyev A: **Codon bias signatures, organization of microorganisms in codon space, and lifestyle.** *Mol Biol Evol* 2005, **22:**547–561.
24. Wagner A: **Inferring lifestyle from gene expression patterns.** *Mol Biol Evol* 2000, **17:**1985–1987.
25. Kudla G, Murray AW, Tollervey D, Plotkin JB: **Coding-sequence determinants of gene expression in Escherichia coli.** *Science* 2009, **324:**255–258.
26. Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN, Rao BS, Smirnov S, Sverdlov AV, Vasudevan S, Wolf YI, Yin JJ, Natale DA: **The COG database: an updated version includes eukaryotes.** *BMC Bioinformatics* 2003, **4:**41.
27. Novoa EM, Pavon-Eternod M, Pan T, Ribas de Pouplana L: **A role for tRNA modifications in genome structure and codon usage.** *Cell* 2012, **149:**202–213.
28. Saini A, Mapolelo DT, Chahal HK, Johnson MK, Outten FW: **SufD and SufC ATPase activity are required for iron acquisition during in vivo Fe-S cluster formation on SufB.** *Biochemistry* 2010, **49:**9402–9412.
29. Tokumoto U, Kitamura S, Fukuyama K, Takahashi Y: **Interchangeability and distinct properties of bacterial Fe-S cluster assembly systems: functional replacement of the isc and suf operons in Escherichia coli with the nifSU-like operon from Helicobacter pylori.** *J Biochem* 2004, **136:**199–209.
30. Nachin L, El Hassouni M, Loiseau L, Expert D, Barras F, Nachin L, El Hassouni M, Loiseau L, Expert D, Barras F: **SoxR-dependent response to oxidative stress and virulence of Erwinia chrysanthemi: the key role of SufC, an orphan ABC ATPase, SoxR-dependent response to oxidative stress and virulence of Erwinia chrysanthemi: the key role of SufC, an orphan ABC ATPase.** *Mol Microbiol, Mol Microbiol* 2001, **39:**960–972.
31. Loughlin MF, Arandhara V, Okolie C, Aldsworth TG, Jenks PJ: **Helicobacter pylori mutants defective in the clpP ATP-dependant protease and the chaperone clpA display reduced macrophage and murine survival.** *Microb Pathog* 2009, **46:**53–57.
32. Ekaza E, Teyssier J, Ouahrani-Bettache S, Liautard J-P, Köhler S: **Characterization of Brucella suis clpB and clpAB mutants and participation of the genes in stress responses.** *J Bacteriol* 2001, **183:**2677–2681.
33. Chaturvedi R, Bansal K, Narayana Y, Kapoor N, Sukumar N, Togarsimalemath SK, Chandra N, Mishra S, Ajitkumar P, Joshi B, Katoch VM, Patil SA, Balaji KN: **The multifunctional PE_PGRS11 protein from Mycobacterium tuberculosis plays a role in regulating resistance to oxidative stress.** *J Biol Chem* 2010, **285:**30389–30403.
34. Lee SM, Koh H-J, Park D-C, Song BJ, Huh T-L, Park J-W: **Cytosolic NADP + -dependent isocitrate dehydrogenase status modulates oxidative damage to cells.** *Free Radic Biol Med* 2002, **32:**1185–1196.
35. Kondoh H, Lleonart ME, Gil J, Wang J, Degan P, Peters G, Martinez D, Carnero A, Beach D: **Glycolytic enzymes can modulate cellular life span.** *Cancer Res* 2005, **65:**177–185.
36. Cunningham L, Georgellis D, Green J, Guest JR: **Co-regulation of lipoamide dehydrogenase and 2-oxoglutarate dehydrogenase synthesis in Escherichia coli: characterisation of an ArcA binding site in the lpd promoter.** *FEMS Microbiol Lett* 1998, **169:**403–408.
37. Brown SD, Thompson MR, VerBerkmoes NC, Chourey K, Shah M, Zhou J, Hettich RL, Thompson DK: **Molecular dynamics of the Shewanella oneidensis response to chromate stress.** *Mol Cell Proteomics* 2006, **5:**1054–1071.
38. Pinto R, Tang QX, Britton WJ, Leyh TS, Triccas JA: **The Mycobacterium tuberculosis cysD and cysNC genes form a stress-induced operon that encodes a tri-functional sulfate-activating complex.** *Microbiology* 2004, **150:**1681–1686.
39. Szklarczyk D, Franceschini A, Kuhn M, Simonovic M, Roth A, Minguez P, Doerks T, Stark M, Muller J, Bork P, Jensen LJ, von Mering C: **The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored.** *Nucleic Acids Res* 2011, **39:**D561–D568.
40. Blank L, Green J, Guest JR: **AcnC of Escherichia coli is a 2-methylcitrate dehydratase (PrpD) that can use citrate and isocitrate as substrates.** *Microbiology* 2002, **148:**133–146.
41. Tang Y, Quail MA, Artymiuk PJ, Guest JR, Green J: **Escherichia coli aconitases and oxidative stress: post-transcriptional regulation of sodA expression.** *Microbiology* 2002, **148:**1027–1037.
42. Ritz D, Beckwith J: **Roles of thiol-redox pathways in bacteria.** *Annu Rev Microbiol* 2001, **55:**21–48.
43. Daly MJ, Gaidamakova EK, Matrosova VY, Vasilenko A, Zhai M, Leapman RD, Lai B, Ravel B, Li S-MW, Kemner KM, Fredrickson JK: **Protein oxidation implicated as the primary determinant of bacterial radioresistance.** *PLoS Biol* 2007, **5:**e92.
44. Daly MJ, Gaidamakova EK, Matrosova VY, Vasilenko A, Zhai M, Venkateswaran A, Hess M, Omelchenko MV, Kostandarithes HM, Makarova KS, Wackett LP, Fredrickson JK, Ghosal D: **Accumulation of Mn(II) in deinococcus radiodurans facilitates gamma-radiation resistance.** *Science* 2004, **306:**1025–1028.
45. Singh R, Mailloux RJ, Puiseux-Dao S, Appanna VD: **Oxidative stress evokes a metabolic adaptation that favors increased NADPH synthesis and decreased NADH production in Pseudomonas fluorescens.** *J Bacteriol* 2007, **189:**6665–6675.
46. Sandoval JM, Arenas FA, Vásquez CC: **Glucose-6-phosphate dehydrogenase protects Escherichia coli from tellurite-mediated oxidative stress.** *PLoS ONE* 2011, **6:**e25573.
47. Grose JH, Joss L, Velick SF, Roth JR: **Evidence that feedback inhibition of NAD kinase controls responses to oxidative stress.** *Proc Natl Acad Sci USA* 2006, **103:**7601–7606.
48. Dougan DA, Reid BG, Horwich AL, Bukau B: **ClpS, a substrate modulator of the ClpAP machine.** *Mol Cell* 2002, **9:**673–683.
49. du Plessis L, Škunca N, Dessimoz C: **The what, where, how and why of gene ontology—a primer for bioinformaticians.** *Brief Bioinform* 2011, **12:**723–735.
50. Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO: **Assigning protein functions by comparative genome analysis: protein phylogenetic profiles.** *Proc Natl Acad Sci USA* 1999, **96:**4285–4288.
51. Korbel JO, Doerks T, Jensen LJ, Perez-Iratxeta C, Kaczanowski S, Hooper SD, Andrade MA, Bork P: **Systematic association of genes to phenotypes by genome and literature mining.** *PLoS Biol* 2005, **3:**e134.
52. Koonin EV, Wolf YI: **Genomics of Bacteria and Archaea: the emerging dynamic view of the prokaryotic world.** *Nucleic Acids Res* 2008, **36:**6688–6719.
53. Škunca N, Bošnjak M, Kriško A, Panov P, Džeroski S, Šmuc T, Supek F: **Phyletic profiling with cliques of orthologs is enhanced by signatures of paralogy relationships.** *PLoS Comput Biol* 2013, **9:**e1002852.
54. Retchless AC, Lawrence JG: **Ecological adaptation in bacteria: speciation driven by codon selection.** *Mol Biol Evol* 2012, **29:**3669–3683.
55. Botzman M, Margalit H: **Variation in global codon usage bias among prokaryotic organisms is associated with their lifestyles.** *Genome Biol* 2011, **12:**R109.
56. Gabaldón T: **Comparative genomics-based prediction of protein function.** *Methods Mol Biol* 2008, **439:**387–401.
57. Rui B, Shen T, Zhou H, Liu J, Chen J, Pan X, Liu H, Wu J, Zheng H, Shi Y: **A systematic investigation of Escherichia coli central carbon metabolism in response to superoxide stress.** *BMC Syst Biol* 2010, **4:**122.
58. Singh R, Lemire J, Mailloux RJ, Appanna VD: **A novel strategy involved anti-oxidative defense: the conversion of NADH into NADPH by a metabolic network.** *PLoS ONE* 2008, **3:**e2682.
59. *NCBI Entrez Genome FTP site.* ftp://ftp.ncbi.nih.gov/genomes/Bacteria.
60. Breiman L: **Random forests.** *Mach Learn* 2001, **45:**5–32.
61. NCBI Entrez Microbial Genome Properties: http://www.ncbi.nlm.nih.gov/genome/browse/.
62. Fawcett T: **An introduction to ROC analysis.** *Pattern Recogn Lett* 2006, **27:**861–874.
63. The FastRandomForest Weka Extension: http://fast-random-forest.googlecode.com/.
64. Witten IH, Frank E: *Practical Machine Learning Tools and Techniques, Second Edition.* 2nd edition. San Francisco: Morgan Kaufmann; 2005.
65. Kitagawa M, Ara T, Arifuzzaman M, Ioka-Nakamichi T, Inamoto E, Toyonaga H, Mori H: **Complete Set of ORF clones of Escherichia coli ASKA library (a complete set of E. coli K-12 ORF archive): unique resources for biological research.** *DNA Res* 2006, **12:**291–299.

66. Zhang W, O'Connor K, Wang DIC, Li Z: **Bioreduction with efficient recycling of NADPH by coupled permeabilized microorganisms.** *Appl Environ Microbiol* 2009, **75:**687–694.
67. Rad AM, Janic B, Iskander ASM, Soltanian-Zadeh H, Arbab AS: **Measurement of quantity of iron in magnetically labeled cells: comparison among different UV/VIS spectrometric methods.** *Biotechniques* 2007, **43:**627–628. 630, 632 passim.
68. Markham NR, Zuker M: **UNAFold: software for nucleic acid folding and hybridization.** *Methods Mol Biol* 2008, **453:**3–31.
69. Shultzaberger RK, Bucheimer RE, Rudd KE, Schneider TD: **Anatomy of Escherichia coli ribosome binding sites.** *J Mol Biol* 2001, **313:**215–228.
70. Karlin S, Mrazek J: **Predicted highly expressed genes of diverse prokaryotic genomes.** *J Bacteriol* 2000, **182:**5238–5250.

doi:10.1186/gb-2014-15-3-r44
Cite this article as: Krisko *et al.*: Inferring gene function from evolutionary change in signatures of translation efficiency. *Genome Biology* 2014 **15**:R44.

Supplementary Methods for Kriško *et al*.

Reference sets of highly expressed genes

In each genome, we declare all genes coding for ribosomal proteins to be the 'reference set', assuming them to be very highly expressed, therefore exhibiting a strong influence of translational selection on codon usage. In addition, some other translation-related genes are included in the reference set:

- translation initiation factors:
    - in Bacteria, COGs 0361 (IF-1), 0532 (IF-2) and 0290 (IF-3);
    - in Archaea, COGs 0023 (eIF-1), 1093 (eIF-2α), 1601 (eIF-2β), 5257 (eIF-2γ)
- translation elongation factors:
    - COG 0480 (EF-G), 0264 (EF-Ts), 0050 (EF-Tu), 0231 (EF-P), 2092 (EF-1β), 5256 (EF-1α)
- chaperones:
    - COG 0459 (GroEL / HSP60), 0234 (GroES / HSP10), 0443 (HSP70), 0484 (HSP40), 0576 (GrpE), 0544 (trigger factor), 1832 (prefoldin)

Genes shorter than 80 codons, having internal stop codons or having length (in nucleotides) indivisible by three were excluded from computation; this includes some of the 'reference set' genes. A gene is then represented by a series of codon frequencies for all degenerate codon families, excluding the stop codons and the rare amino acids cysteine and histidine. The frequencies of codons for a single amino acid are adjusted to sum to one, therefore the final data does not reflect amino acid frequencies. If an amino acid is absent from a gene product, the amino acid's codon frequencies are not estimated in any way but are instead represented by a special 'missing value' symbol.

The Random Forest classifier

We employ a supervised machine learning method – the Random Forest (RF) classifier (Breiman 2001) – to detect signatures of selection acting to optimize translational efficiency of genes. The RF implementation we use is FastRandomForest 0.98 (Supek 2011) which integrates into the Weka learning environment (Witten and Frank 2005). RF was our method of choice as it is computationally efficient, robust to noise and handles missing data. The 'forest size' parameter ("-I") was set to 1000; other parameters were left at default values.

The RF algorithm produces an ensemble of decision tree classifiers, where each decision tree is constructed by recursively partitioning the data by attribute value tests (forming 'nodes') so as to reduce the class entropy in the resulting partitions ('branches'). In RF, trees are constructed on bootstrap samples of the entire dataset, and choice of attributes at each node is restricted to introduce variability. The final predictions of a RF model are obtained by averaging over individual trees ('voting').

Detecting selection for translational efficiency in genomes

Nucleotide substitution patterns in DNA molecules influence codon frequencies of genes; additionally, the biases in nucleotide substitution patterns may be region, strand, or chromosome specific (Daubin and Perriere 2003, Rocha 2004). Thus, it is necessary to control for the biases in order to reliably detect influence of translational selection on codon biases.

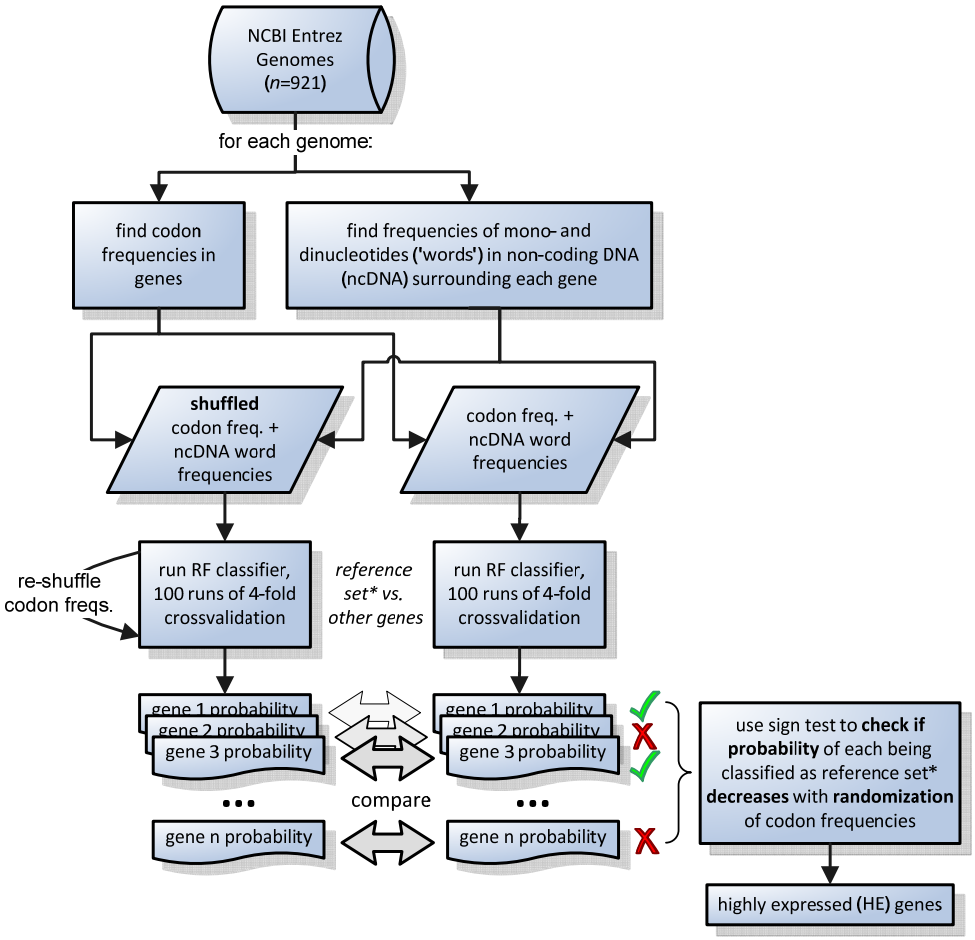
We encode the information about local nucleotide substitution patterns affecting each gene by computing mononucleotide and dinucleotide frequencies in the non-coding regions of DNA neighboring the coding region of the gene. Genes for known functional RNA molecules such as tRNA and rRNA are also treated as coding regions and thus do not contribute toward mono- and di-nucleotide frequencies of intergenic DNA. Additionally, the first 20 nt upstream of each coding region were masked and not taken into consideration when sampling the ncDNA, as this region region of prokaryotic ncDNA was found to be under much stronger selective pressures than the rest of the ncDNA (Molina and van Nimwegen 2008). First 50 available (not masked) non-coding nucleotides upstream of the start codon, together with the first 50 available nucleotides following the stop codon were used as the ncDNA sample for each gene. We also tested wider ±100 and ±200 nt intervals; see below.

To verify that translational selection acts in each genome - a prerequisite for predicting gene expression from codon usage patterns - we employ the following procedure: the RF classifier is first trained to distinguish the reference set genes based on the mono- and di-nucleotide frequencies of genes' neighboring non-coding DNA; the codon frequencies are also supplied to the classifier at this point, but are shuffled between genes, thus rendered uninformative. 100 runs of four-fold crossvalidation are used to estimate the accuracy of the classifier using the area-under-ROC-curve (AUC) (Fawcett 2006) score, and the AUC for each of the 100 runs of crossvalidation is recorded. The codon frequencies are re-shuffled between genes in each run.

The procedure is then repeated for a second time, however now the true (unshuffled) codon frequencies are included in the dataset for the RF classifier training. The AUC scores are again recorded for each of the 100 runs of crossvalidation. To determine if selection for translational efficiency acts on the genome as a whole, the sign test (McDonald 2008) is used to compare 100 AUC scores obtained with shuffled codon frequencies to 100 AUC scores obtained with true codon frequencies, for each genome. If the AUC score exhibits a consistent increase over 100 runs of crossvalidation, the introduction of codon frequencies improves the ability to discriminate the reference set genes, providing evidence that translational selection acts on this specific genome. This was the case at $P \leq 10^{-25}$ (sign test), corresponding to FDR$\leq 9 \cdot 10^{-23}$, for all but one of the genomes, Candidatus *Hodgkinia cicadicola Dsem*, a highly reduced genome (<200 genes) with high GC content (McCutcheon and Moran 2009); here, the difference of AUC scores was still significant at $P \leq 10^{-4}$ (FDR=9.1%). Such a result agrees with the previous findings about selected codon biases being universal among microbial genomes (Supek et al. 2010, Hershberg and Petrov 2009), indicating it is feasible to predict gene expression levels from codon usage in all of the analyzed genomes.

Assigning 'highly expressed' labels to individual genes

In order to check whether each specific gene is affected by translational selection, during the procedure described above which involves two rounds of RF classifier training, the per-gene probabilities of belonging to the 'reference set' are recorded for each of the 100 runs of crossvalidation. The per-gene probabilities are then compared between the two rounds of crossvalidation - the round with the shuffled vs. with the intact codon frequencies. A sign test (McDonald 2008) is used to determine if an increase in probability occurs more frequently than expected by chance; if it does, the gene is labeled as highly expressed. The Supplementary Figure SM1 provides a schematic overview of the computational pipeline.

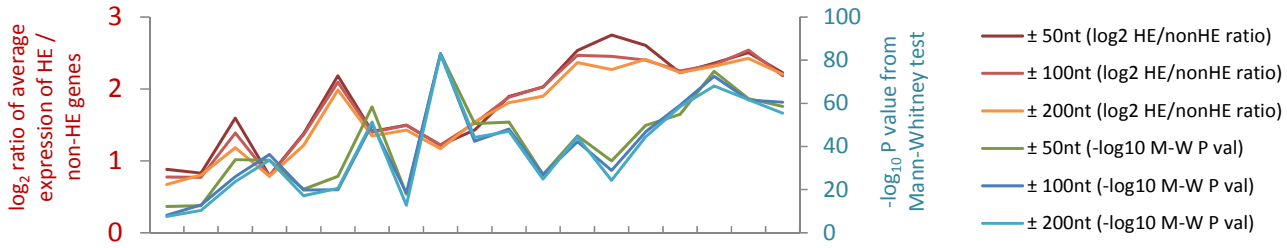


Supplementary Methods Figure SM1. A schematic describing the computational workflow employed to determine highly expressed (HE) genes in genomes.

We have set the sign test threshold *P* value to 10^{-15}; this *P* corresponds to at least 88 out of 100 sign test 'wins' for the dataset with intact codon frequencies. Given that all genomes analyzed had $<10^4$ genes, the used P value threshold corresponds to $<10^{-11}$ false positive HE genes per genome. Also, given the smallest number of HE genes per genome was 13 (for the *Hodgkinia* above), the highest possible FDR for detecting the HE genes using the sign test is $10^{-11}/13 \approx 10^{-12}$.

This sign test *P* value should be regarded as somewhat liberal because the repeated runs of crossvalidation are not independent, being based on repeated sampling from the same set of genes. To obtain an additional, more conservative estimate, we employed a corrected paired *t*-test (Nadeau and Bengio 2003) originally intended for comparison of classification algorithms using repeated runs of crossvalidation. Note that we actually compare RF models derived from a specific dataset, and not the different variants of the underlying RF algorithm itself, and therefore this *P* value likely is pessimistic for our experimental setup (Nadeau and Bengio 2003). For example, using this very conservative *t*-test, the median *P* value for highly expressed genes in the *E. coli* K-12 genome would be $5.6 \cdot 10^{-4}$, while for 95% of the HE genes, $P<0.04$. For the *Bacillus subtilis 168* genome, the median of this conservative *P* value estimate for HE genes would be $5.6 \cdot 10^{-4}$, while for 95% of the HE genes $P<0.07$.

The assignment of the HE/non-HE label to genes depends on a comparison – via the described statistical test – of a gene's codon usage to the dinucleotide composition of neighboring non-coding DNA. As stated above, our analyses were performed with 50 upstream + 50 downstream nucleotides per gene (total 100nt). We also evaluated whether taking neighborhoods of different sizes might have a bearing on the outcome by examining the agreement of the HE labels to experimental gene expression data from NCBI GEO in 19 genomes (sources listed in section below). The results are summarized in the Figure SM2 below: taking wider intervals does not consistently improves either the ratio of average HE/non-HE expression, or the *P* value of a Mann-Whitney test for separation in expression levels between the HE and non-HE groups.



Supplementary Methods Figure SM2. Agreement of the HE/non-HE labels to gene expression data in 19 bacterial genomes, depending on the width of the non-coding DNA neighborhood that is used as a baseline. Order of the genomes same as in Fig 1C.

A comparison to the procedure from Supek *et al.* (PLOS Genetics 2010)

Note that the procedure described above is a derivative of the computational workflow described in reference (Supek et al. 2010) with several changes that resulted in better agreement to microarray measurements of gene expression in 19 phylogenetically diverse bacteria; see below for the list of microarray datasets. The average ratio of expression between the highly expressed genes and the rest of the genome has increased from 2.4x to 3.9x; the *P* value for separation of the expression value in the two classes has improved from $2 \cdot 10^{-12}$ to $7 \cdot 10^{-48}$ (Mann-Whitney test, median over 19 genomes); see Table S1 for the full data. The changes to the algorithm encompass:

- The original algorithm (Supek et al. 2010) uses 50 runs of 4-fold crossvalidation per genome; here, 100 runs of 4-fold crossvalidation are used. This requires more computational time but decreases the variability between repetitions of the pipeline.
- Originally, the first (of the two) round of crossvalidation uses just mono- and di-nucleotide frequencies, while the second uses both these, and the codon frequencies; here, both sources of data are present in both rounds, but codon frequencies are shuffled in the first round. This ensures that all comparisons are made between classification models with an equal number of degrees of freedom. The datasets have 75 features: this includes 55 codon frequencies (64, minus 3 stop codons, 1 Met, 1 Trp, 2 Cys and 2 His codons), 16 dinucleotide and 4 mononucleotide frequencies.
- The original reference set contained only ribosomal protein genes; here, the set is expanded using translation factors and chaperones (see above). These genes were typically used as parts of reference sets in previous studies, see eg. (Karlin and Mrazek 2000), as they are considered to have strong translational codon usage biases in various genomes.
- Originally, all ncDNA in a fixed-size window around each gene is collected (10 kb upstream of the start and 10 kb downstream of the stop codon); here, first 50 non-coding nt upstream of the start and first 50 non-coding nt downstream of the stop are collected, regardless of the distance. This ensures each gene is represented with an equal quantity of ncDNA.
- Originally, all amino acids were considered. Here, the rare amino acids cysteine and histidine are excluded as they were frequently present in very few instances or completely absent from a gene, introducing noise.
- Here, the Random Forest classifier is configured to put a heavier weight on the minority class (this is always the reference set), equal to 1/proportion of reference set genes. Originally, the two classes were treated with equal weights, despite the large imbalance in their sizes.

Gene expression data sources

The datasets with absolute mRNA abundances in prokaryotes were downloaded from the NCBI Gene Expression Omnibus (GEO). The mRNA abundance data was obtained either using single-channel microarrays (commonly Affymetrix) or two-channel microarrays normalized to genomic DNA. We have manually curated to list to include only datasets where growth conditions can be expected to lead to fast growth of the organism, i.e. the conditions where translational selection would be expected to act – exponential growth phase, rich medium, and no physical, chemical or biological stress for the organism. Depending on the growth conditions, the correlation of gene expression to codon usage might be larger or smaller. Finally, we have reduced the list to one dataset per organism, leaving us with 19 datasets with good coverage of various bacteria phyla: Proteobacteria (α, ß, γ and δ), Firmicutes, Actinobacteria and the Thermus/Deinococcus clade. The list of datasets is given below; the set includes four genomes where translational selection was previously not detected in at least 2 out of 3 of the following studies (dos Reis et al. 2004, Sharp et al. 2005, Carbone et al. 2005): *Pseudomonas aeruginosa*, *Streptomyces coelicolor*, *Mycobacterium tuberculosis* and *Nitrosomonas europaea*.

List of mRNA expression datasets; the NCBI GEO Series ID is given ("GSExxxx"), together with the GEO Samples ("GSMxxxx") from the series that were averaged to obtain final expression level. These sets contain absolute mRNA signal intensities, from single-channel microarrays (commonly Affymetrix), except *Streptomyces*, *Desulfovibrio* and *Salmonella*, which are from dual-channel microarrays normalized to genomic DNA.

- *Pseudomonas syringae tomato* DC3000: GSE4848 (GSM109003, 109007, 109009, 109011, 109012, 109014)
- *Mycobacterium tuberculosis* H37Rv: GSE7588 (GSM183531, 183633)
- *Nitrosomonas europaea*: GSE10664 (GSM269650, 269651, 269652)
- *Streptococcus mutans*: GSE6973 (GSM160647 – 160694)
- *Lactobacillus plantarum*: GSE11383 (GSM287549, 287553, 287554, 287555, 287556, 287559)
- *Bacillus subtilis*: GSE11937 (GSM299792, 299793, 299794, 299795)
- *Rhodopseudomonas palustris* CGA009: GSE6221 (GSM143539, 143540, 143541, 143542, 143543, 143544)
- *Thermus thermophilus HB8*: GSE10368 (GSM261559, 261569, 261594, 261560, 261570, 261595)
- *Bradyrhizobium japonicum*: GSE12491 (GSM210242, 210243, 210244, 210245)
- *Desulfovibrio vulgaris Hildenborough*: GSE4447 (GSM101160, 101161, 101162, 101166, 101167, 101168, 101325, 101326, 101330, 101331, 101332, 101333)
- *Haemophilus influenzae*: GSE5061 (GSM114031, 114032, 114033)
- *Listeria monocytogenes*: GSE3247 (GSM73161 – 73185)
- *Rhodobacter sphaeroides 2.4.1*: GSE12269 (GSM308084 - 308107)
- *Staphylococcus aureus* Mu50: GSE2728 (GSM52706, 52707, 52755, 52724, 52725, 52727, 52744, 52745, 52746)
- *Bifidobacterium longum*: GSE5865 (GSM136748, 136749, 136750, 136751, 136752)
- *Streptomyces coelicolor*: GSE7172 (GSM172822, 172823, 172826, 172830)
- *Escherichia coli K12*: GSE13982 (GSM351280, 351282, 351295, 351297)
- *Salmonella typhimurium LT2*: GSE4631 (GSM103442, 103443, 103444, 103445, 103450, 103757, 103758, 103759)
- *Pseudomonas aeruginosa*: GSE4026 (GSM92178, 92179, 92182, 92183, 92186, 92187)

Normalization of experimental results

The measurements resulting from different experimental essays on *E. coli* mutants were normalized to enable genes to be clustered based on their results across different assays. Additionally, the normalization is meant to facilitate interpretability, where 0 means 'no effect' in all essays (same as *wild-type*), and 1 signifies a 'strong effect'. Values <0 or >1 are possible under certain conditions. The individual essays are normalized as follows:

- protein carbonylation increase: optical density (OD) for each mutant was divided by the OD for the *wild-type*, and 1 subtracted from the result
- measuring ROS:
    - DHR-123 increase: fluorescence was divided by the fluorescence of the *wild-type*, and 1 subtracted from the result
    - CellROX increase: same as DHR-123 increase
- measuring Fe:
    - cellular Fe increase: absorbance for each mutant was divided by the absorbance for the *wild-type*, and 1 subtracted from the result
    - dipyridyl rescue: % survival of each mutant after H_2O_2 shock without dipyridyl was subtracted from % survival of the same mutant after H_2O_2 shock with dipyridyl, and the result divided by the % survival of the *wild-type* after H_2O_2 shock without dipyridyl
- measuring NADPH:
    - cellular NADPH decrease: fluorescence for each mutant was divided by the fluorescence for the *wild-type*, and this value subtracted from 1
    - NADPH rescue: same as dipyridyl rescue

Supplementary Methods References

Breiman L (2001) Random forests. Machine Learning 45: 5-32.

Carbone A, Kepes F, Zinovyev A (2005) Codon bias signatures, organization of microorganisms in codon space, and lifestyle. Mol Biol Evol 22: 547-561.

Daubin V, Perriere G (2003) G+C3 structuring along the genome: a common feature in prokaryotes. Mol Biol Evol 20: 471-483.

dos Reis M, Savva R, Wernisch L (2004) Solving the riddle of codon usage preferences: a test for translational selection. Nucleic Acids Res 32: 5036-5044.

Fawcett T (2006) An introduction to ROC analysis. Pattern Recognition Letters 27: 861-874.

Hershberg R, Petrov DA (2009) General rules for optimal codon choice. PLoS Genet 5: e1000556.

Karlin S, Mrazek J (2000) Predicted highly expressed genes of diverse prokaryotic genomes. J Bacteriol 182: 5238-5250.

McCutcheon JP, McDonald BR, Moran NA (2009) Origin of an Alternative Genetic Code in the Extremely Small and GC–Rich Genome of a Bacterial Symbiont. PLoS Genetics 5(7):e1000565.

McDonald JH (2008) Sign test. Handbook of Biological Statistics. Baltimore: Sparky House Publishing. pp. 185-189.

Molina N, van Nimwegen E (2008) Universal patterns of purifying selection at noncoding positions in bacteria. Genome Res 18: 148-160.

Nadeau C, Bengio Y (2003) Inference for the generalization error. Machine Learning 52: 239-281.

Rocha EP (2004) The replication-related organization of bacterial genomes. Microbiology 150: 1609-1627.

Sharp PM, Bailes E, Grocock RJ, Peden JF, Sockett RE (2005) Variation in the strength of selected codon usage bias among bacteria. Nucleic Acids Res 33: 1141-1153.

Supek F (2011) The “Fast Random Forest” software. http://fast-random-forest.googlecode.com/

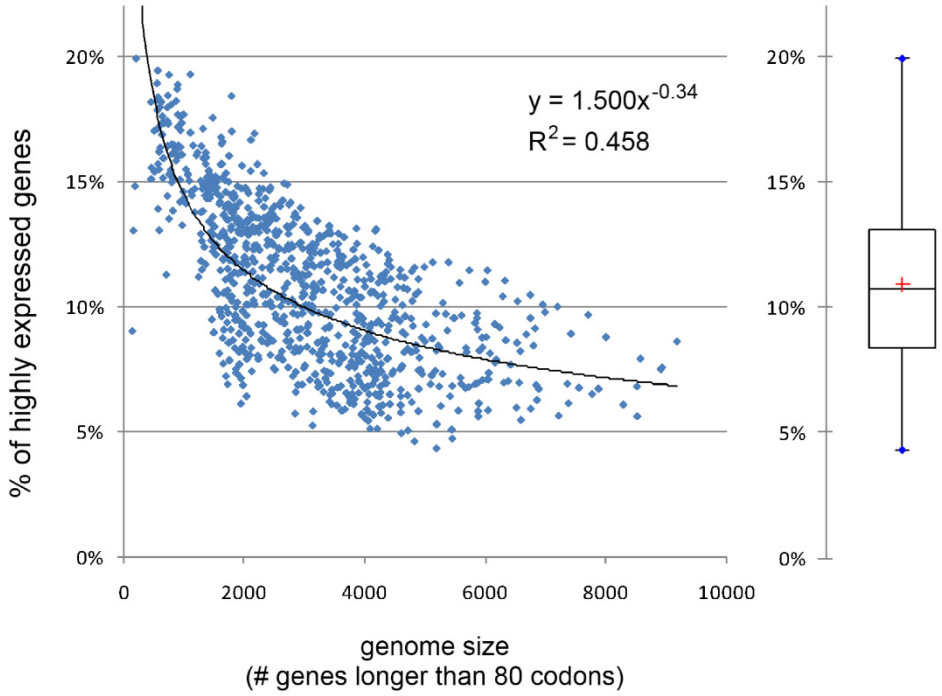
Supek F, Skunca N, Repar J, Vlahovicek K, Smuc T (2010) Translational selection is ubiquitous in prokaryotes. PLoS Genet 6(6): e1001004.

Witten IH, Frank E (2005) Data Mining: Practical machine learning tools and techniques. San Francisco: Morgan Kaufmann.

Additional file 2. Agreement to expression data for the predictions about 'highly expressed' (HE) genes, and a comparison to the original 'OCU' method. ('OCU' stands for optimized codon usage; from Supek et al., PLOS Genetics 2010). *P* values are from a Mann-Whitney test for a difference of microarray signal levels between the HE and non-HE genes, or the OCU and non-OCU genes. The "ratios" are calculated between the average microarray signal of the two groups. The ratio of ribosomal proteins *vs.* whole genome is given for a sense of scale; the ribosomal protein genes are expected to be very highly expressed.

organism	# genes*	% HE	P value	HE/non-HE ratio	% OCU	P value	OCU/non-OCU ratio	ribo.prot/all genes ratio
Pseudomonas syringae tomato DC3000	5280	6.5%	1.2E-13	1.87 x	8.4%	2.2E-01	1.33 x	4.55 x
Mycobacterium tuberculosis H37Rv	3860	11.8%	7.9E-16	1.81 x	17.4%	9.8E-03	1.38 x	4.31 x
Nitrosomonas europaea	2406	12.8%	3.4E-33	2.93 x	15.2%	8.9E-07	1.41 x	6.77 x
Streptococcus mutans	1776	8.8%	5.1E-36	1.75 x	8.1%	1.9E-11	1.42 x	1.94 x
Lactobacillus plantarum	2892	6.3%	1.5E-21	2.74 x	9.8%	1.9E-05	1.68 x	3.22 x
Bacillus subtilis	3809	5.4%	5.1E-27	4.75 x	8.5%	5.4E-03	1.76 x	9.95 x
Rhodopseudomonas palustris CGA009	4658	10.0%	2.6E-57	2.62 x	12.4%	4.1E-24	2.03 x	3.56 x
Thermus thermophilus HB8	2122	14.8%	1.1E-18	2.72 x	13.1%	2.5E-04	2.06 x	6.69 x
Bradyrhizobium japonicum	7972	8.4%	6.1E-88	2.33 x	9.1%	2.8E-33	2.14 x	2.58 x
Desulfovibrio vulgaris Hildenborough	3086	11.5%	7.8E-48	2.67 x	13.5%	4.4E-21	2.31 x	4.72 x
Haemophilus influenzae	1574	9.3%	1.4E-53	3.78 x	10.0%	1.0E-18	2.37 x	4.48 x
Listeria monocytogenes	2734	9.1%	5.3E-35	4.32 x	11.9%	2.7E-05	2.74 x	4.87 x
Rhodobacter sphaeroides 2-4-1	4076	8.7%	1.4E-48	5.74 x	11.8%	1.0E-06	2.84 x	12.09 x
Staphylococcus aureus Mu50	2498	8.0%	7.4E-35	6.76 x	10.0%	1.5E-05	2.87 x	2.31 x
Bifidobacterium longum	1696	10.4%	5.0E-52	6.37 x	13.3%	4.2E-16	3.11 x	6.50 x
Streptomyces coelicolor	7772	6.1%	1.8E-55	4.82 x	6.9%	1.6E-38	3.13 x	15.17 x
Escherichia coli K12	3914	6.9%	1.6E-79	5.33 x	7.5%	4.4E-31	3.18 x	6.81 x
Salmonella typhimurium LT2	4305	5.3%	3.0E-63	5.98 x	7.0%	1.6E-20	3.33 x	7.71 x
Pseudomonas aeruginosa	5426	6.8%	1.4E-60	4.76 x	6.2%	4.1E-30	3.73 x	7.43 x
		average	*median*	*average*	*average*	*median*	*Average*	*Average*
		8.8%	7.8E-48	3.90 x	10.5%	1.9E-11	2.36 x	6.09 x

* number of genes at least 80 codons long



Additional file 3. The relative proportion of highly expressed genes is lower in larger genomes. This correlation was previously explained (Supek *et al*. PLOS Genetics 2010) by different proportions of various gene functional categories in smaller or larger genomes. Many of the functional categories, in turn, tend to have a general preference for higher or lower expression. For instance, larger genomes have a disproportionally increased number of gene regulators, which have a strong tendency of being lowly expressed. Smaller genomes, on the other hand, have a higher proportion of ribosomal proteins, whose absolute number is roughly fixed across genomes regardless of their size.

Additional file 4. Correlations of mRNA 5' end folding free energies and various codon indices to gene expression levels. The free energies are a measure of stability of the structures (more negative = more stable) and are calculated in 42-nucleotide windows on the mRNA sequence using the *hybrid-ss-min* program from UnaFold 3.6 with default parameters, as in Kudla *et al*. (*Science* 2009). The three 42-nt window positions investigated are: [-4,37], found to have a strongest correlation to protein levels in Kudla *et al*.; [-20,21], a window centered over the start codon; and [-30,11], a window centered on the common location of the Shine-Dalgarno sequence at -9 (Shultzaberger *et al*. J Mol Biol 2001). The -10 kcal/mol is the approximate limit for the mRNA folding free energy in 42-nt windows below which more negative values the mRNA folding starts to have a considerable effect on translation efficiency (Supek and Šmuc, *Genetics*, 2010). The mRNA coordinates are given relative to the start codon, where 1 is the A in AUG. The codon indices are: CAI (Sharp and Li, *Nucl Acids Res* 1987), B (Karlin and Mrazek, *J Bact* 2000), and MILC (Supek and Vlahoviček, BMC Bioinformatics 2005). RF is the probability score obtained from a Random Forest classifier (Supek *et al.*, PLOS Genetics, 2010). All codon indices use the same 'reference set' of known highly expressed genes as used in our analyses (Supplementary Methods in Additional file 1).

		Pearson correlation with mRNA levels (from microarray measurements)							% genes in genome with mRNA folding free energy < -10 kcal/mol		
		codon indices for genes				mRNA folding free energies in 42 nucleotide windows					
organism	#genes	CAI	B	MILC	RF	[-4,37]	[-20,21]	[-30,11]	[-4,37]	[-20,21]	[-30,11]
Pseudomonas syringae tomato DC3000	5266	0.21	0.17	0.27	0.33	0.03	0.03	0.05	14%	13%	17%
Mycobacterium tuberculosis H37Rv	3444	0.05	0.13	0.12	0.21	0.10	0.13	0.13	48%	44%	46%
Nitrosomonas europaea	2347	0.28	0.33	0.37	0.43	0.04	0.17	0.13	9%	7%	8%
Streptococcus mutans	1741	0.45	0.29	0.49	0.34	-0.08	-0.02	0.04	2%	1%	1%
Lactobacillus plantarum	2646	0.22	0.20	0.22	0.32	-0.07	-0.03	-0.01	4%	2%	3%
Bacillus subtilis	3635	0.33	0.29	0.38	0.45	-0.02	0.00	-0.02	4%	3%	3%
Rhodopseudomonas palustris CGA009	4632	0.16	0.33	0.32	0.32	0.09	0.07	0.09	31%	27%	34%
Thermus thermophilus HB8	2080	0.20	0.34	0.39	0.45	0.17	0.14	0.11	56%	56%	62%
Bradyrhizobium japonicum	7853	-0.08	0.10	0.02	0.14	0.04	0.05	0.05	32%	31%	35%
Desulfovibrio vulgaris Hildenborough	3024	0.03	0.15	0.16	0.24	0.03	0.04	0.05	24%	22%	28%
Haemophilus influenzae	1573	0.62	0.61	0.71	0.65	-0.11	-0.04	0.01	3%	2%	2%
Listeria monocytogenes	2729	0.35	0.40	0.43	0.43	-0.03	-0.04	0.01	1%	1%	1%
Rhodobacter sphaeroides 2 4 1	3839	0.20	0.40	0.50	0.60	0.08	0.12	0.12	34%	29%	40%
Staphylococcus aureus Mu50	2101	0.43	0.46	0.54	0.56	-0.06	-0.06	-0.02	1%	1%	1%
Bifidobacterium longum	1687	0.41	0.49	0.52	0.56	0.05	0.07	0.08	22%	17%	18%
Streptomyces coelicolor	6984	0.12	0.24	0.30	0.49	0.01	0.03	0.04	48%	42%	45%
Escherichia coli K12	3613	0.57	0.56	0.64	0.59	-0.01	0.03	0.05	9%	4%	6%
Salmonella typhimurium LT2	3982	0.35	0.38	0.44	0.40	0.00	0.04	0.05	8%	5%	7%
Pseudomonas aeruginosa	5404	0.20	0.38	0.50	0.47	0.09	0.08	0.08	23%	19%	25%
median of 19 genomes	***3444***	***0.22***	***0.33***	***0.39***	***0.43***	***0.03***	***0.04***	***0.05***	***14%***	***13%***	***17%***

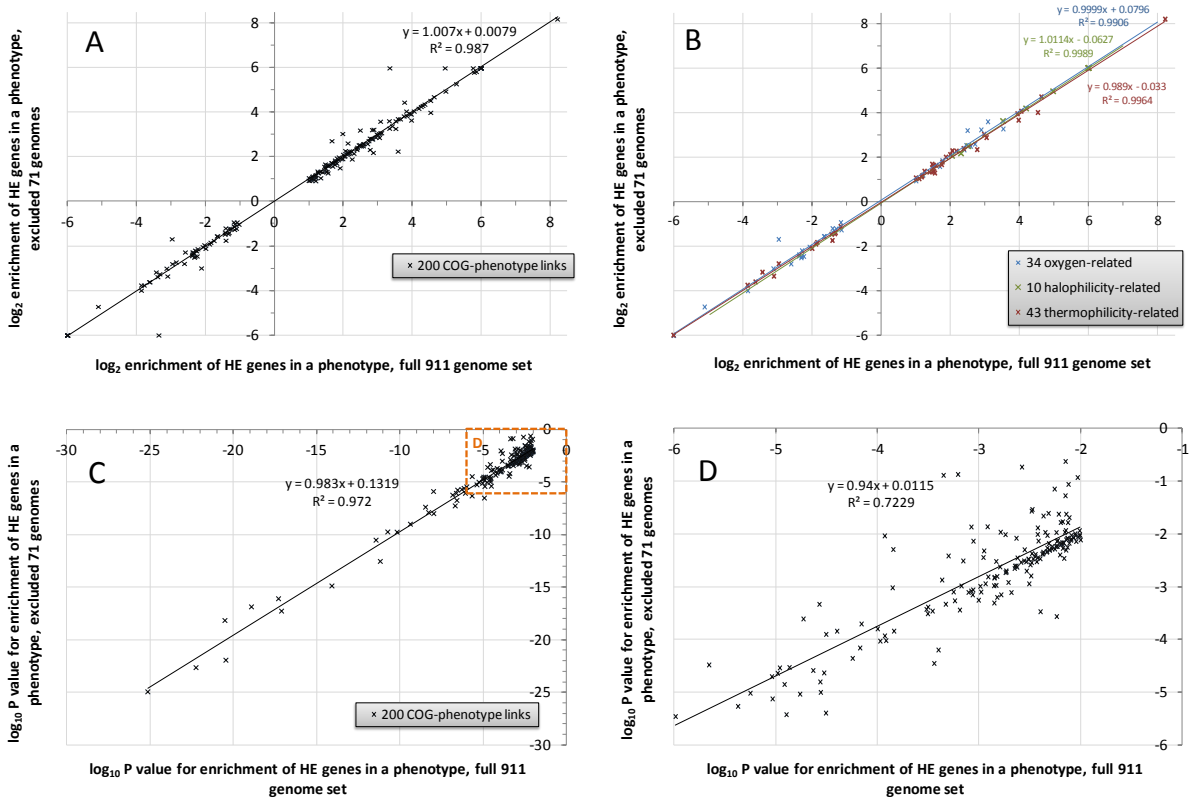
Additional file 5. The 100 features describing each organism which were used in the search for the phenotypes predictive of the changes in translation efficiency within COGs. All features are binary variables, and can be undefined for some organisms. 70 features describing the phylogeny (left/middle columns) and the 6 features describing genome size and GC content (right column, top) are included to ensure that correlations detected with the remaining 24 features (phenotypes, right column) could not be explained by the phylogeny or the genomic size/GC. "#pos" is the number of organisms marked as positive for the feature, and "#neg" as negative.

domain / phylum / class	*#pos*	*#neg*	*order (taxonomy)*	*#pos*	*#neg*	*genomic features*	*#pos*	*#neg*
domain: Archaea	85	826	Actinomycetales	83	801	G+C: above median(>47.5%)	425	387
domain: Bacteria	826	85	Alteromonadales	28	856	G+C: high(>59.9%)	235	576
phylum: Actinobacteria	97	813	Bacillales	46	838	G+C: low(<37.6%)	188	624
phylum: Bacteroidetes	34	876	Bacteroidales	10	874	Genome size: above median(>3.2Mb)	454	443
phylum: Chlamydiae	8	902	Bifidobacteriales	6	878	Genome size: large(>4.8Mb)	226	671
phylum: Chlorobi	10	900	Burkholderiales	47	837	Genome size: small(<2Mb)	195	702
phylum: Chloroflexi	15	895	Campylobacterales	18	866			
phylum: Crenarchaeota	23	887	Chlamydiales	8	876	***phenotypes***	***#pos***	***#neg***
phylum: Cyanobacteria	28	882	Chlorobiales	10	874	Endospores	75	337
phylum: Deinococcus-Thermus	10	900	Chroococcales	20	864	Gram Stain=positive	216	466
phylum: Euryarchaeota	58	852	Clostridiales	37	847	Growth in groups	61	228
phylum: Firmicutes	145	765	Desulfovibrionales	8	876	Habitat=aquatic(pos) *vs.* terrestrial(neg)	167	71
phylum: Proteobacteria	383	527	Enterobacteriales	46	838	Habitat=free-living(pos) *vs.* hostAssociated(neg)	238	222
phylum: Spirochaetes	17	893	Flavobacteriales	14	870	Habitat=multiple(pos) *vs.* single(neg)	190	587
phylum: Tenericutes	28	882	Halobacteriales	12	872	Halophilic	40	140
phylum: Thermotogae	11	899	Lactobacillales	36	848	Motility	379	231
class: Actinobacteria	97	775	Legionellales	3	881	Oxygen = aerotolerant (pos) *vs.* strictly anaerobic (neg)	514	214
class: α-proteobacteria	107	765	Methanococcales	9	875	Oxygen = facultative aerobe (pos) *vs.* strict aerobe (neg)	217	296
class: Bacilli	82	790	Mycoplasmatales	22	862	Psychrophilic	25	760
class: Bacteroidia	10	862	Neisseriales	5	879	Radioresistance	16	887
class: β-proteobacteria	68	804	Pasteurellales	10	874	Shape = coccus (pos) *vs.* rod (neg)	105	506
class: Chlamydiae	8	864	Prochlorales	1	883	Thermophilic	142	643
class: Chlorobia	10	862	Pseudomonadales	18	866	Pathogenic in plants	23	304
class: Clostridia	61	811	Rhizobiales	47	837	Pathogenic in mammals	163	166
class: Deinococci	10	862	Rhodobacterales	12	872	Mammalian pathogen = blood	35	128
class: δ-proteobacteria	33	839	Rickettsiales	25	859	Mammalian pathogen =enteric	31	132
class: ε-proteobacteria	22	850	Spirochaetales	17	867	Mammalian pathogen = heart	8	155
class: Flavobacteria	14	858	Sulfolobales	5	879	Mammalian pathogen = nervous system	17	146
class: γ-proteobacteria	152	720	Thermoanaero-bacterales	20	864	Mammalian pathogen = oportunistic/nosocomial	16	147
class: Halobacteria	12	860	Thermococcales	8	876	Mammalian pathogen = oral cavity	8	155
class: Methanococci	9	863	Thermotogales	11	873	Mammalian pathogen = respiratory	43	120
class: Methanomicrobia	14	858	Thiotrichales	4	880	Mammalian pathogen = skin/soft tissues	24	139
class: Mollicutes	28	844	Vibrionales	9	875			
class: Spirochaetes	17	855	Xanthomonadales	8	876			
class: Thermoprotei	23	849						
class: Thermotogae	11	861						

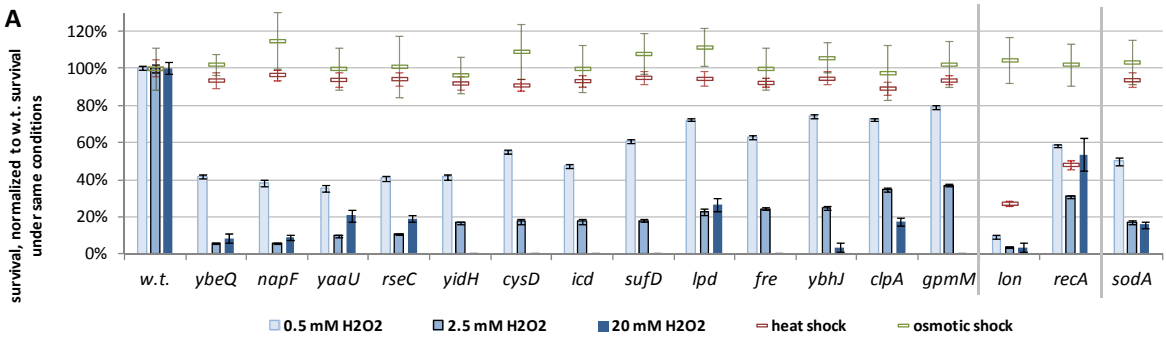
Additional file 6. Genomes where the optimal codons inferred from overrepresentation in HE genes overall do not match the expected optimal codons inferred from the genomic tRNA repertoire. The nine two-fold degenerate amino acids are examined. An optimal codon ("HE" columns) is defined as overrepresented at P<0.001 in a Fisher's exact test on codon counts in HE vs. the non-HE genes; a non-significant result means no codon is optimal. The codons expected to be optimal from tRNAs ("tRNA" columns) are defined in the genomes where tRNA genes with only one of the two possible anticodons are present (found by tRNAscan-SE); then, the codon matching that anticodon by canonical Watson-Crick pairing is considered tRNA-optimal, and the other that uses wobble pairing is considered tRNA-suboptimal. The table shows 71 (of the 911 total) genomes where the optimal and the tRNA-optimal codons disagree in at least 3 of 9 (column "# aa") of the testable amino acids. (In 651/911 genomes there were 0/9 disagreeing amino acids, and 1/9 for further 135 genomes). Thus, in the 71 genomes, the expression level-related codon bias does not, on overall, clearly relate to the tRNA gene repertoire and may possibly not reflect translational selection, but another, unknown factor. We thus excluded the 71 genomes, and re-run the subsequent analyses to verify our findings are robust to inclusion of these genomes (Additional file 7).

		Phe		Tyr		Cys		His		Gln		Asn		Lys		Asp		Glu	
Genome	# aa	HE	tRNA	HE	tRNA	HE	tRNA	HE	tRNA	HE	tRNA	HE	tRNA	HE	tRNA	HE	tRNA	HE	tRNA
Ehrlichia canis Jake	8	TTT	TTC	TAT	TAC	TGT	TGC	CAT	CAC	CAG	CAA	AAT	AAC	AAG	-	GAT	GAC	GAG	GAA
Ehrlichia chaffeensis Arkansas	8	TTT	TTC	TAT	TAC	TGT	TGC	CAT	CAC	CAG	CAA	AAT	AAC	AAG	-	GAT	GAC	GAG	GAA
Borrelia turicatae 91E135	7	TTT	TTC	TAT	TAC	-	TGC	CAT	CAC	CAG	CAA	AAT	AAC	AAG	-	GAT	GAC	GAG	GAA
Cand. Blochmannia floridanus	7	TTT	TTC	TAT	TAC	-	TGC	CAT	CAC	CAG	CAA	AAT	AAC	AAG	-	GAT	GAC	GAG	GAA
Ehrlichia ruminantium Welgevonden	7	TTT	TTC	TAT	TAC	-	TGC	CAT	CAC	CAG	CAA	AAT	AAC	AAG	-	GAT	GAC	GAG	GAA
Nitrosomonas europaea ATCC 19718	7	TTT	TTC	TAT	TAC	TGT	TGC	CAT	CAC	CAA	CAA	AAT	AAC	-	AAA	GAT	GAC	GAG	GAA
Borrelia burgdorferi B31	7	TTT	TTC	TAT	TAC	TGT	TGC	CAT	CAC	CAG	CAA	AAT	AAC	AAG	-	GAT	GAC	GAG	-
Cand. Blochmannia vafer BVAF	6	-	TTC	TAT	TAC	TGT	-	CAT	CAC	CAG	CAA	AAT	AAC	AAG	-	GAT	GAC	GAG	GAA
Borrelia recurrentis A1	6	TTT	TTC	TAT	TAC	-	TGC	-	CAC	CAG	CAA	AAT	AAC	AAG	-	GAT	GAC	GAG	GAA
Borrelia hermsii DAH	6	TTT	TTC	TAT	TAC	-	TGC	CAT	CAC	-	CAA	AAT	AAC	AAG	-	GAT	GAC	GAG	GAA
Cand. Azobacteroides pseudotrichonymphae CFP2	6	TTT	TTC	TAT	TAC	TGT	TGC	-	CAC	CAG	-	AAT	AAC	AAG	-	GAT	GAC	GAG	GAA
Wolbachia endosymbiont of D. melanogaster	6	TTT	TTC	TAT	TAC	-	TGC	CAT	CAC	-	CAA	AAT	AAC	AAG	AAA	GAT	GAC	-	GAA
Acidilobus saccharovorans 345 15	6	TTT	TTC	TAT	TAC	TGT	TGC	CAT	CAC	CAA	-	AAT	AAC	AAA	-	GAT	GAC	GAA	-
Chloroherpeton thalassium ATCC 35110	6	TTT	TTC	TAT	TAC	TGT	TGC	CAT	CAC	CAG	-	AAT	AAC	AAG	-	GAT	GAC	-	-
Desulfobulbus propionicus DSM 2032	6	TTT	TTC	TAT	TAC	TGT	TGC	CAT	CAC	CAA	-	AAT	AAC	AAA	-	GAT	GAC	GAA	-
Geobacter metallireducens GS 15	6	TTT	TTC	TAT	TAC	TGT	TGC	CAT	CAC	CAA	CAA	AAT	AAC	-	AAA	GAT	GAC	GAA	GAA
Geobacter uraniireducens Rf4	6	TTT	TTC	TAT	TAC	TGT	TGC	CAT	CAC	CAA	-	AAT	AAC	AAA	AAA	GAT	GAC	GAA	GAA
Methylacidiphilum infernorum V4	6	TTT	TTC	TAT	TAC	TGT	TGC	CAT	CAC	-	-	AAT	AAC	-	-	GAT	GAC	-	-
Nitrosomonas eutropha C91	6	TTT	TTC	TAT	TAC	TGT	TGC	CAT	CAC	CAA	CAA	AAT	AAC	-	AAA	GAT	GAC	-	GAA
Syntrophus aciditrophicus SB	6	TTT	TTC	TAT	TAC	TGT	TGC	CAT	CAC	-	-	AAT	AAC	-	-	GAT	GAC	GAA	-
Thermobaculum terrenum ATCC BAA 798	6	TTT	TTC	TAT	TAC	TGT	TGC	CAT	CAC	-	-	AAT	AAC	-	-	GAT	GAC	GAA	-
Xylella fastidiosa M12	6	TTT	TTC	TAT	TAC	TGT	TGC	CAT	CAC	CAG	-	AAT	AAC	AAG	-	GAT	GAC	-	-
Neorickettsia risticii Illinois	5	TTT	TTC	-	TAC	-	TGC	-	CAC	CAG	CAA	-	AAC	AAG	AAA	GAT	GAC	GAG	GAA
Borrelia duttonii Ly	5	TTT	TTC	TAT	TAC	-	TGC	-	CAC	-	CAA	AAT	AAC	AAG	-	GAT	GAC	GAG	GAA
Thermofilum pendens Hrk 5	5	TTT	TTC	TAT	TAC	TGT	-	CAT	CAC	CAA	-	AAT	-	AAA	AAG	GAT	-	GAA	GAG
Orientia tsutsugamushi Ikeda	5	TTT	TTC	TAT	TAC	-	TGC	-	CAC	-	CAA	AAT	AAC	AAG	AAA	GAT	GAC	-	GAA
Desulfomicrobium baculatum DSM 4028	5	-	TTC	TAT	TAC	TGT	TGC	CAT	CAC	-	-	AAT	AAC	-	-	GAT	GAC	GAA	-
Wolbachia endosymbiont of Culex quinquefasciatus Pel	5	-	TTC	TAT	TAC	TGT	TGC	CAT	CAC	-	CAA	AAT	AAC	-	AAA	GAT	GAC	-	GAA

Bartonella tribocorum CIP 105476	5	TTT	TTC	TAT	TAC	-	TGC	CAT	CAC	CAG	-	AAT	AAC	AAG	-	GAT	GAC	-	GAA
Nitrosococcus oceani ATCC 19707	5	TTT	TTC	TAT	TAC	-	TGC	CAT	CAC	-	-	AAT	AAC	AAG	-	GAT	GAC	GAG	-
Neorickettsia sennetsu Miyayama	4	-	TTC	-	TAC	-	TGC	-	CAC	CAG	CAA	-	AAC	AAG	AAA	GAT	GAC	GAG	GAA
Anaplasma phagocytophilum HZ	4	-	TTC	-	TAC	TGT	TGC	CAT	CAC	CAG	CAA	-	AAC	AAG	-	-	GAC	GAG	GAA
Wolbachia wRi	4	-	TTC	TAT	TAC	-	TGC	-	CAC	-	CAA	AAT	AAC	AAG	AAA	GAT	GAC	-	GAA
gamma proteobacterium HdN1	4	TTC	TTC	-	TAC	TGT	TGC	CAT	CAC	-	CAA	-	AAC	AAG	AAA	GAT	GAC	-	GAA
Cand. Phytoplasma mali	4	TTT	TTC	TAT	TAC	-	TGC	-	CAC	-	CAA	-	AAC	AAG	AAA	GAT	GAC	-	GAA
Metallosphaera sedula DSM 5348	4	-	TTC	TAT	TAC	-	-	CAT	CAC	CAA	-	AAT	AAC	-	-	GAT	GAC	GAA	-
Cand. Nitrospira defluvii	4	TTT	TTC	-	TAC	-	TGC	CAT	CAC	-	-	AAT	AAC	AAG	-	GAT	GAC	-	-
Sulfolobus acidocaldarius DSM 639	4	TTT	TTC	TAT	-	-	TGC	CAT	CAC	-	-	AAT	AAC	AAA	-	GAT	GAC	GAA	-
Bacteroides helcogenes P 36 108	4	TTT	TTC	-	TAC	TGT	TGC	-	CAC	CAA	-	AAT	AAC	-	-	GAT	GAC	GAA	-
Nitrosococcus halophilus Nc4	4	TTT	TTC	-	TAC	TGT	TGC	-	CAC	CAG	-	AAT	AAC	AAG	-	GAT	GAC	-	-
Borrelia afzelii PKo	4	TTT	TTC	TAT	TAC	-	TGC	-	CAC	-	CAA	AAT	AAC	AAG	-	GAT	GAC	GAG	-
Borrelia garinii PBi	4	TTT	TTC	TAT	TAC	-	TGC	-	CAC	-	CAA	AAT	AAC	AAG	-	GAT	GAC	GAG	-
Sulfolobus islandicus L D 8 5	4	TTT	TTC	TAT	TAC	-	-	-	CAC	-	-	AAT	AAC	-	-	GAT	GAC	GAA	-
Sulfolobus solfataricus P2	4	TTT	TTC	TAT	TAC	-	-	-	CAC	-	-	AAT	AAC	-	-	GAT	GAC	GAA	-
Anaplasma centrale Israel	3	-	TTC	-	TAC	TGT	TGC	-	CAC	CAG	CAA	-	AAC	AAG	-	-	GAC	GAG	GAA
Anaplasma marginale Florida	3	-	TTC	-	TAC	TGT	TGC	-	CAC	CAG	CAA	-	AAC	AAG	-	-	GAC	GAG	GAA
Bartonella clarridgeiae 73	3	TTT	TTC	-	TAC	-	TGC	-	CAC	CAG	CAA	-	AAC	AAG	-	-	GAC	GAG	GAA
Bartonella henselae Houston 1	3	-	TTC	-	TAC	-	TGC	-	CAC	CAG	-	AAT	AAC	AAG	-	GAT	GAC	GAG	GAA
Bartonella quintana Toulouse	3	-	TTC	-	TAC	-	TGC	-	CAC	CAG	-	AAT	AAC	AAG	-	GAT	GAC	GAG	GAA
Caldivirga maquilingensis IC 167	3	-	TTC	TAT	TAC	-	TGC	-	CAC	CAG	-	AAT	AAC	AAG	-	GAT	GAC	GAG	-
Cand. Liberibacter solanacearum CLso ZC1	3	TTT	TTC	TAT	TAC	-	TGC	-	CAC	CAG	-	-	-	AAG	-	GAT	GAC	-	GAA
Cellulophaga algicola DSM 14237	3	TTC	TTC	TAC	TAC	TGT	TGC	CAC	CAC	-	CAA	AAC	AAC	AAG	AAA	GAT	GAC	-	GAA
Chlorobium phaeobacteroides DSM 266	3	-	TTC	-	TAC	TGT	TGC	-	CAC	CAG	-	AAT	AAC	AAG	-	GAT	GAC	GAG	-
Clostridium kluyveri DSM 555	3	-	TTC	TAT	TAC	-	TGC	-	CAC	-	-	AAT	AAC	AAG	-	GAT	GAC	GAA	-
Desulfotalea psychrophila LSv54	3	TTC	TTC	-	TAC	TGT	TGC	-	CAC	-	CAA	-	AAC	AAG	AAA	GAT	GAC	-	-
Dictyoglomus thermophilum H 6 12	3	TTT	TTC	TAT	TAC	-	TGC	-	CAC	-	-	-	AAC	AAG	-	GAT	GAC	GAG	-
Dictyoglomus turgidum DSM 6724	3	-	TTC	TAT	TAC	-	TGC	CAT	CAC	-	-	-	AAC	-	-	GAT	GAC	-	-
Geobacter sulfurreducens PCA	3	TTC	TTC	TAT	TAC	-	TGC	-	CAC	-	CAA	AAT	AAC	-	AAA	GAT	GAC	GAA	GAA
Helicobacter felis ATCC 49179	3	-	TTC	-	TAC	TGT	TGC	-	CAC	-	CAA	AAT	AAC	-	AAA	GAT	GAC	GAA	GAA
Ignisphaera aggregans DSM 17230	3	TTT	TTC	-	TAC	-	-	-	CAC	-	-	AAT	AAC	-	-	GAT	GAC	-	-
Nitrosococcus watsoni C 113	3	TTT	TTC	-	TAC	-	TGC	-	CAC	-	-	AAT	AAC	AAG	-	GAT	GAC	-	-
Pyrobaculum arsenaticum DSM 13514	3	TTT	TTC	-	TAC	TGT	TGC	-	CAC	-	-	-	AAC	AAG	-	GAT	GAC	-	-
Pyrobaculum islandicum DSM 4184	3	TTT	TTC	TAT	TAC	TGT	-	CAT	CAC	-	-	AAT	-	-	-	GAT	-	GAA	-
Rickettsia peacockii Rustic	3	TTT	TTC	-	TAC	-	TGC	-	CAC	CAA	CAA	AAT	AAC	AAA	AAA	GAT	GAC	GAA	GAA
Rickettsia rickettsii Sheila Smith	3	TTT	TTC	-	TAC	-	TGC	-	CAC	-	CAA	AAT	AAC	-	AAA	GAT	GAC	GAA	GAA
Rickettsia typhi Wilmington	3	TTT	TTC	-	TAC	-	TGC	-	CAC	-	CAA	AAT	AAC	-	AAA	GAT	GAC	-	GAA
Spirochaeta thermophila DSM 6192	3	-	TTC	TAT	TAC	-	TGC	-	CAC	CAG	-	AAT	AAC	-	-	GAT	GAC	-	-
Sulfolobus tokodaii 7	3	-	TTC	TAT	TAC	-	TGC	-	CAC	CAA	-	AAT	AAC	AAA	-	GAT	GAC	GAA	-
Thermoproteus neutrophilus V24Sta	3	TTT	TTC	TAT	TAC	-	-	-	CAC	-	CAG	-	-	AAG	AAG	GAT	GAC	GAG	GAG
Thiomicrospira crunogena XCL 2	3	TTC	TTC	-	TAC	TGT	TGC	-	CAC	-	CAA	AAC	AAC	AAG	AAA	GAT	GAC	-	GAA
Wolbachia endosymbiont TRS of Brugia malayi	3	TTT	TTC	TAT	TAC	-	TGC	-	CAC	-	CAA	AAT	AAC	-	AAA	-	GAC	-	GAA



Additional file 7. Robustness of the 200 discovered COG-phenotype links to exclusion of 71 genomes where codon biases were not clearly related to the tRNA gene repertoires. Excluded genomes are listed in Additional file 6. (A) The $\log_2$ enrichment of the 200 COG-phenotype links with the full set of 911 genomes, and after excluding the 71 genomes. (B) Same as A, but limited to the links that we experimentally validated. In the original analysis, a threshold of $\log_2$ enrichment ≥ 1 or ≤ -1 was a requirement for calling the 200 COG-phenotype links; after excluding the 71 genomes, 195 COG-phenotype links still meet this criterion. (C) The $\log_{10}$ P value for significance of the enrichment/depletion (two-tailed Fisher's exact test), again compared between the original and the reduced genome sets. (D) same as C, but only for the COG-phenotype links with $\log_{10}P \geq -6$. In the original analysis, $\log_{10}P \leq -2$ was required for calling the 200 COG-phenotype links; after excluding the 71 genomes, 173/200 links still have $\log_{10}P \leq -2$, and 185/200 still have $\log_{10}P \leq -1.7$ ($P<0.02$).



Additional file 8. Survival of *E. coli* deletion mutants after oxidative stress induced by different hydrogen peroxide concentrations. Survival after heat and osmotic shocks is given for comparison. Deleted genes are on the *x* axis. The *y* axis shows the mutants' survival, normalized to the survival of the *w.t.* under the same conditions, which is 45.6% for the 0.5 mM, 13.8% for the 2.5 mM and 4.2% for the 20 mM H_2O_2, 23.6% for the heat shock and 21.3% for the osmotic shock. *lon* and *recA* mutants are separated for showing a non-specific response, being also sensitive to oxidative and heat stress. *sodA* is a known oxidative stress defense gene, serving as a positive control.

Additional file 9. Complementing *E. coli* deletion mutants with wild-type genes. Survival of *E. coli* deletion mutants in the putative oxidative stress response genes with and without the corresponding genes expressed from a plasmid, at two different H_2O_2 concentrations (2.5 mM and 20 mM). All 20 mM measurements are from at least two experiments in duplicate, and for 2.5 mM, from three experiments in duplicate.

2.5 mM (default)	complementation			no complementation	
strain	**average %survival**	**st. dev.**	**as % of w.t. survival**	**average %survival**	**st. dev.**
wt	N/A	N/A	N/A	13.78	0.34
clpA	13.38	0.64	97%	4.82	0.21
cysD	13.87	0.44	101%	2.42	0.19
fre	13.63	0.46	99%	3.36	0.11
gpmM	13.52	0.45	98%	5.14	0.12
icd	13.62	0.44	99%	2.43	0.28
lon	13.28	0.40	96%	0.50	0.09
lpd	13.20	0.54	96%	3.18	0.32
napF	13.15	0.51	95%	0.83	0.09
recA	13.18	0.50	96%	4.27	0.16
rseC	13.55	0.35	98%	1.50	0.10
soda	13.62	0.50	99%	2.38	0.17
sufD	12.98	0.72	94%	2.48	0.12
yaaU	13.23	0.64	96%	1.38	0.10
ybeQ	13.38	0.33	97%	0.80	0.09
ybhJ	13.47	0.49	98%	3.44	0.16
yidH	13.57	0.51	98%	2.34	0.13

20 mM (high)	complementation			no complementation	
strain	**average %survival**	**st. dev.**	**as % of w.t. survival**	**average %survival**	**st. dev.**
wt	N/A	N/A	N/A	4.16	0.17
clpA	4.00	0.26	96%	0.74	0.11
cysD	3.98	0.15	96%	0.00	0.00
fre	4.10	0.29	99%	0.00	0.00
gpmM	3.98	0.25	96%	0.00	0.00
icd	4.03	0.24	97%	0.00	0.00
lon	3.81	0.09	92%	0.16	0.12
lpd	4.10	0.18	99%	1.11	0.18
napF	4.05	0.35	97%	0.38	0.07
recA	3.93	0.26	94%	2.23	0.46
rseC	4.11	0.18	99%	0.80	0.09
sodA	3.93	0.17	94%	0.66	0.10
sufD	4.10	0.29	99%	0.00	0.00
yaaU	4.03	0.17	97%	0.87	0.17
ybeQ	3.93	0.17	94%	0.36	0.12
ybhJ	3.88	0.10	93%	0.15	0.14
yidH	3.93	0.21	94%	0.00	0.00

Additional file 10. Supporting evidence for putative oxidative stress genes. A survey of the evidence in the literature offering support for the involvement of *sufD*, *clpA*, *icd*, *gpmM*, *lpd* and *cysD* genes in oxidative stress resistance of various organisms.

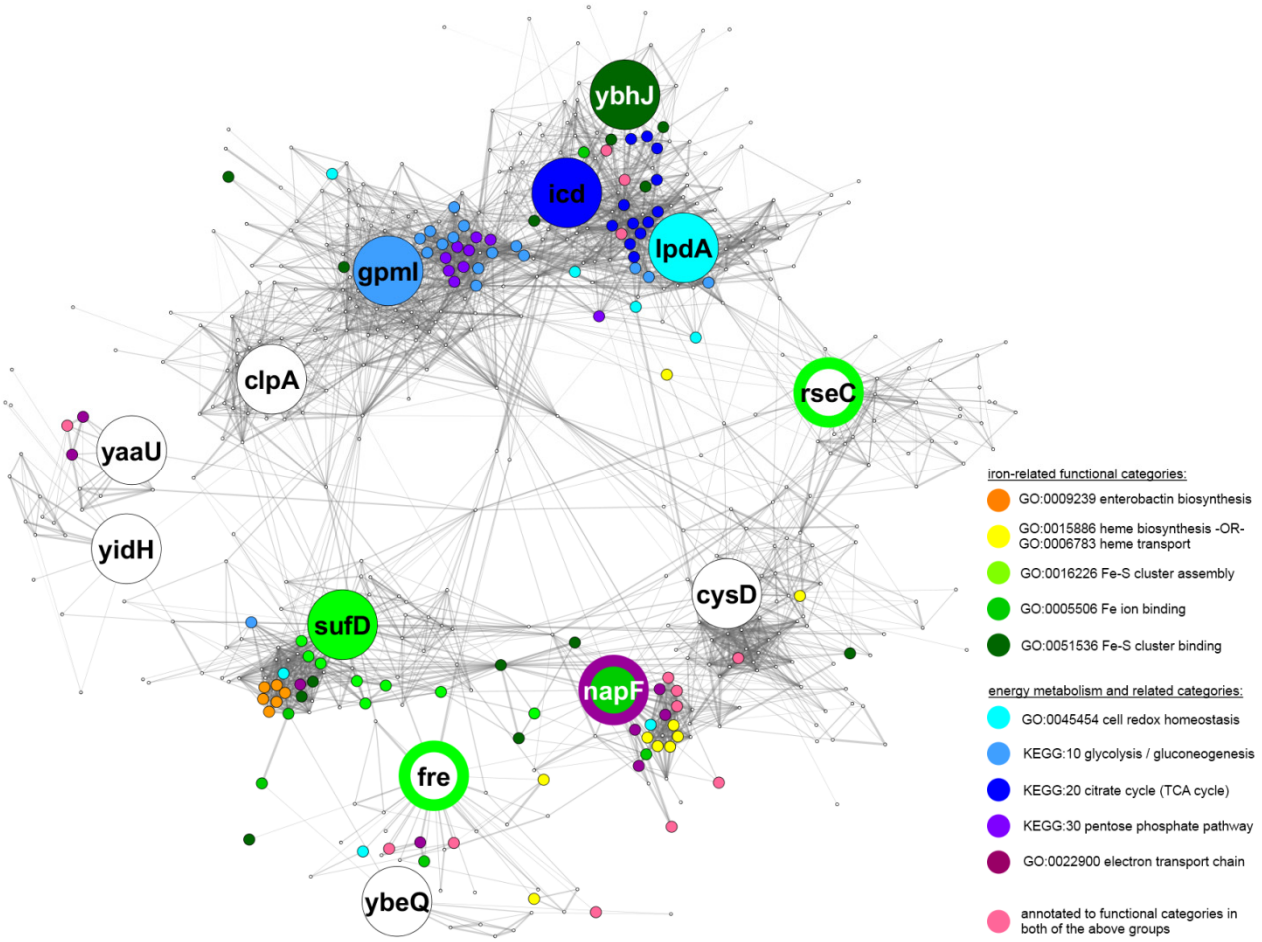
gene	organism	Reference
direct evidence in *E. coli* or homologs in other bacteria		
sufD	*E. coli*	Tokumoto *et al.*, J Biochem 2004 Saini *et al.*, Biochemistry 2010
	Erwinia chrysanthemi	Nachin *et al.*, Mol Microbiol 2001
gpmM	*Mycobacterium tuberculosis*	Chaturvedi *et al.*, JBC 2010
clpA	*Helicobacter pylori*	Loughlin *et al.*, Microb Pathog 2009
	Brucella suis	Ekaza *et al.*, J Bact 2001
direct evidence for homologs in eukaryotes		
icd	mouse	Lee *et al.*, Free Radical Biol Med 2002
gpmM	mouse	Kondoh *et al.*, Cancer Res 2005
evidence of regulation under aerobiosis or oxidative stress		
lpd	E. coli	Cunningham *et al.*, FEMS Microbiol Lett 1998
cysD	*Mycobacterium tuberculosis*	Pinto *et al.*, Microbiology 2004
	Shewanella oneidensis	Brown *et al.*, Mol Cell Proteomics 2006

Additional file 11. Functional interactions with known oxidative stress genes. Predicted functional interactions between 34 COGs we found to have codon adaptation which correlates to the aerobic lifestyle, and 30 COGs encoding known *E. coli* oxidative stress response proteins. The predicted interactions are from the STRING v9.0 database, using exclusively co-expression (top part of table), or exclusively text mining (bottom part) evidence. Only interactions marked as high-confidence by STRING (confidence ≥ 0.7) are shown.

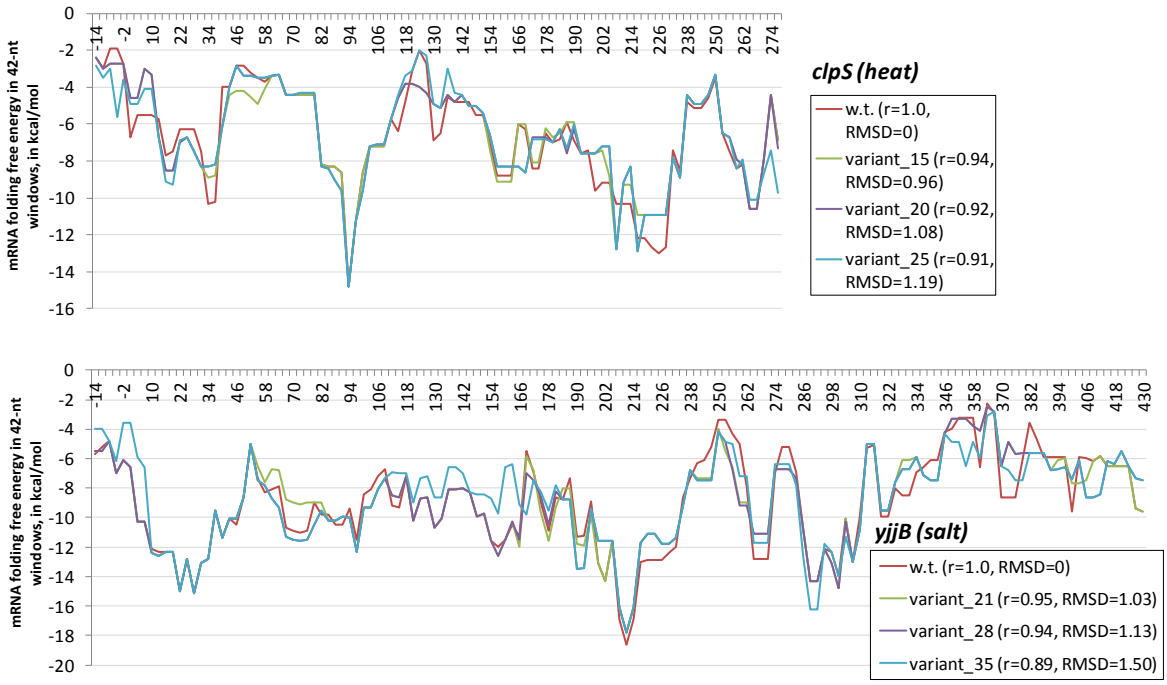
Interactor 1: COG differentially expressed in aerobes *vs*. anaerobes	***E. coli* gene(s)**	**Interactor 2: known *E. coli* oxidative stress response**	***E. coli* gene(s)**	**interaction confidence**
INFERRED FROM CO-EXPRESSION				
COG:0543 2-polyprenylphenol hydroxylase and related flavodoxin oxidoreductases	*fre*	COG:0753 Catalase	*katE*	0.952
COG:0126 3-phosphoglycerate kinase	*pgk*	COG:1225 Peroxiredoxin	*bcp*	0.908
COG:0543 2-polyprenylphenol hydroxylase and related flavodoxin oxidoreductases	*fre*	COG:450 Peroxiredoxin	*ahpC*	0.832
COG:0539 Ribosomal protein S1	*rpsA*	COG:1225 Peroxiredoxin	*bcp*	0.818
COG:0543 2-polyprenylphenol hydroxylase and related flavodoxin oxidoreductases	*fre*	COG:0695 Glutaredoxin and related proteins	*grxA, grxC*	0.814
COG:0126 3-phosphoglycerate kinase	*pgk*	COG:0450 Peroxiredoxin	*ahpC*	0.785
COG:0126 3-phosphoglycerate kinase	*pgk*	COG:0376 Catalase (peroxidase I)	*katG, katP*	0.736
INFERRED FROM TEXT MINING				
COG:1048 Aconitase A	*ybhJ, acnA*	COG:0605 Superoxide dismutase	*sodA, sodB*	0.975
COG:1048 Aconitase A	*ybhJ, acnA*	COG:0753 Catalase	*katE*	0.953
COG:2235 Arginine deiminase	(no genes)	COG:0605 Superoxide dismutase	*sodA, sodB*	0.949
COG:1048 Aconitase A	*ybhJ, acnA*	COG:0276 Protoheme ferro-lyase (ferrochelatase)	*hemH*	0.890
COG:1048 Aconitase A	*ybhJ, acnA*	COG:0386 Glutathione peroxidase	*btuE*	0.880
COG:1048 Aconitase A	*ybhJ, acnA*	COG:0735 Fe2+/Zn2+ uptake regulation proteins	*fur*	0.875
COG:1048 Aconitase A	*ybhJ, acnA*	COG:2032 Cu/Zn superoxide dismutase	*sodC*	0.842
0538 Isocitrate dehydrogenases	*icd*	COG:0753 Catalase	*katE*	0.778
0175 3'-phosphoadenosine 5'-phosphosulfate sulfotransferase (PAPS reductase)/FAD synthetase and related enzymes	*cysD, cysH*	COG:0492 Thioredoxin reductase	*trxB*	0.779
0538 Isocitrate dehydrogenases	*icd*	COG:0605 Superoxide dismutase	*sodA, sodB*	0.810

Additional file 12. Literature data suggesting putative antioxidant mechanism-of-action assignments. Listed for the *sufD*, *fre*, *rseC*, *gpmM*, *lpd* and *icd* genes.

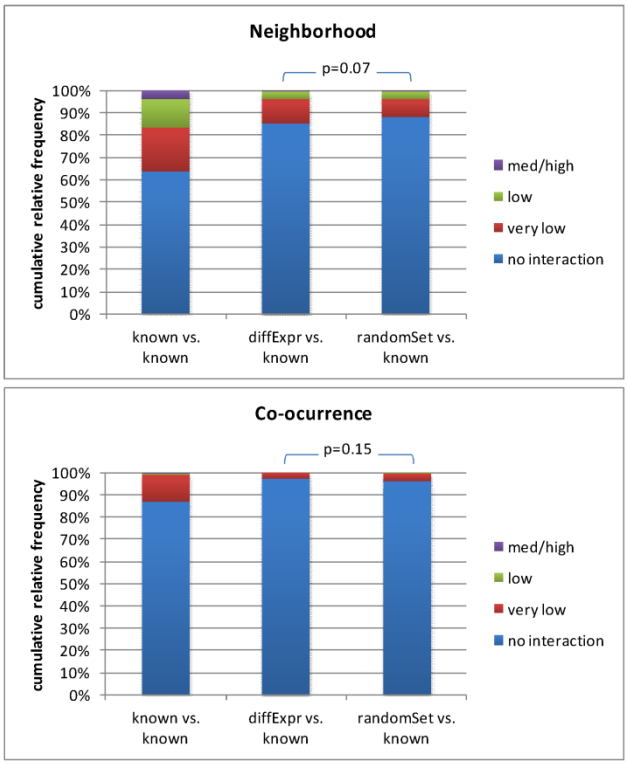
gene	Description	reference
sufD	part of the sufBCD system for assembly of Fe-S clusters under oxidative stress	Jang and Imlay, Mol Microbiol 2010
	required for in vivo iron acquisition, but not during Fe-S cluster assembly or cluster maturation	Saini *et al*, Biochemistry 2010
fre	transfers electrons to reduce an Fe(III) center of ribonucleotide reductase, thereby activating the enzyme	Fontecave *et al*, JBC 1987 Coves *et al*, JBC 1993
	reduces and mobilizes iron from ferrisiderophores	Coves & Fontecave, Eur J Biochem 1993
rseC	reduces the 2Fe-2S cluster in the redox-sensitive transcriptional activator SoxR; oxidised form of SoxR may be unstable	Koo *et al*, EMBO J 2003
gpmM	a glycolysis protein; overexpressing it causes an increased glycolytic flux	Kondoh *et al*, 2005 Cancer Res Kondoh, 2008 Experimental Cell Res
lpd	encodes a subunit of the NADH-producing enzymes pyruvate dehydrogenase and 2-oxoglutarate dehydrogenase	Kim *et al*, J Bact 2008 Bunik & Fernie, Biochem J 2009
	under stress conditions, conversion of NADH into the protective NADPH via NAD-kinase is enhanced	Grose *et al*, PNAS 2006
icd	an NADPH-generating enzyme of the citric acid cycle; also produces alpha-ketoglutarate which can serve as an antioxidant	Mailloux *et al*, PLoS One 2007



Additional file 13. The functional context of the 13 *E. coli* gene representatives of the COGs differentially expressed in aerobic microbes. The genes *recA* and *lon* are not shown as their deletion mutants show nonspecific stress sensitivity (Figure 3). Lines represent the predicted functional interactions from the STRING 9.0 database (medium confidence level, ≥0.4), while dots represent all proteins interacting with at least one of the 13 proteins. A large, highly interconnected set of interacting ribosomal proteins is not shown for clarity. The larger, colored dots are proteins annotated with one of the selected functional categories in *E. coli* (right panel). Hollow circles in *fre* or *rseC* or thick border in *napF* denote putative assignments we inferred for the genes from the literature; all other functional annotations were from the EBI's Uniprot-GOA database. All shown functional categories were found to be enriched among the 13 proteins+interactors at $P<0.05$ (hypergeometric distribution, corrected for multiple testing) using GeneCodis 2.0. Proximity of the discs in the figure roughly corresponds to their functional similarity, as optimized by the "Edge Weighted Spring Embedded" layout in Cytoscape 2.8.1, edge weights being derived from interaction confidence levels in STRING.



Additional file 14. Profiles of folding free energies in 42-nt windows along the *clpS* and *yjjB* gene mRNAs. The *x* axes show the starting coordinate (in nucleotides) of the 42-nt window. The folding free energies were calculated using the *hybrid-ss-min* program from UNAFold 3.6 software with default parameters. Alongside each *E. coli* gene (marked "w.t."), three variants are given with introduced synonymous changes that reduced codon optimality (Fig. 5); the number given after the word "variant" is the number of codons that have been altered, with respect to the wild type. A 14-nt ribosome binding site sequence "AGGAGGUAAAACAU" was prepended before the AUG start codon when determining the folding free energies, as was the case for the actual genes. For each variant, a Pearson's correlation coefficient "r", and the root mean square deviation "RMSD" are given as measures of similarities of their folding free energy profiles to the wild-type sequence.



Additional file 15. Distributions of predicted functional interactions at different confidence levels. Functional interactions were examined between (a) 30 COGs known to have a role in the oxidative stress response, labeled "known *vs.* known"; (b) the "known" group and the 34 COGs found to be differentially expressed between aerotolerant organisms and anaerobes, or between obligate and facultative aerobes, labeled "diffExpr *vs.* known"; (c) the "known" group and a 100 randomly chosen COGs, labeled "randomSet *vs.* known". Two of the 34 COGs were also in the "known" group, and their functional interactions did not count for the "diffExpr" group. In particular: COG0719 (*E. coli sufD* and *sufB* genes) and COG1249 (*E. coli lpd*, *ykgC*, *gor* and *sthA* genes). The predicted functional interactions are from the STRING v9.0 database, http://string-db.org/; the scores vary from 0 to 1, where STRING declares interactions betwen 0.15 and 0.40 to have low confidence, between 0.40 and 0.70 medium, and above 0.70 high confidence. For details on how the scores are computed for each individual source of data, please refer to the papers describing the database (references given at the STRING website). The *P* values were from a χ^2 test.

Additional file 16. Relationships of the aerotolerance phenotype with the presence/absence patterns and with the codon adaptation of the catalase genes. Tables show the count of organisms (not genes) which have the COG absent (first column), present with one or more genes which are all non-HE (second column), or present with one or more genes of which at least one in the genome is HE (third column). The tables below show the same frequencies, but normalized to the total number of aerotolerant or strictly anaerobic organisms. For both COGs, the presence of the catalases in the genome is strongly and significantly correlated with aerobicity (top right panel for each COG). However, the codon adaptation of the catalases is strongly but not significantly correlated with aerobicity (bottom right panel for each COG) due to low numbers of strictly anaerobic genomes that have a catalase gene present.

COG:0376 (*katG*, *katP* in *E. coli*)

# genomes	absent	present, not HE	present, HE
aerotolerant (n=514)	300	189	25
strictly anaerobic (n=214)	185	28	1

absent vs. present
relative risk = 3.07
(95% CI: 2.16 to 4.37)
Fisher's exact test P = 2.5e-14

(normalized per row)	absent	present, not HE	present, HE
aerotolerant (n=514)	58.4%	36.8%	4.9%
strictly anaerobic (n=214)	86.4%	13.1%	0.5%

HE vs. non-HE (when present)
relative risk = 3.39
(95% CI: 0.48 to 24.07)
Fisher's exact test P = 0.33

COG:0753 (*katE* in *E. coli*)

# genomes	absent	present, not HE	present, HE
aerotolerant (*n*=514)	219	254	41
strictly anaerobic (*n*=214)	178	35	1

absent vs. present
relative risk = 3.41
(95% CI: 2.51 to 4.64)
Fisher's exact test P = 4.2e-25

(normalized per row)	absent	present, not HE	present, HE
aerotolerant (*n*=514)	42.6%	49.4%	8.0%
strictly anaerobic (*n*=214)	83.2%	16.4%	0.5%

HE vs. non-HE (when present)
relative risk = 5.00
(95% CI: 0.71 to 35.28)
Fisher's exact test P = 0.064

translation efficiency profile

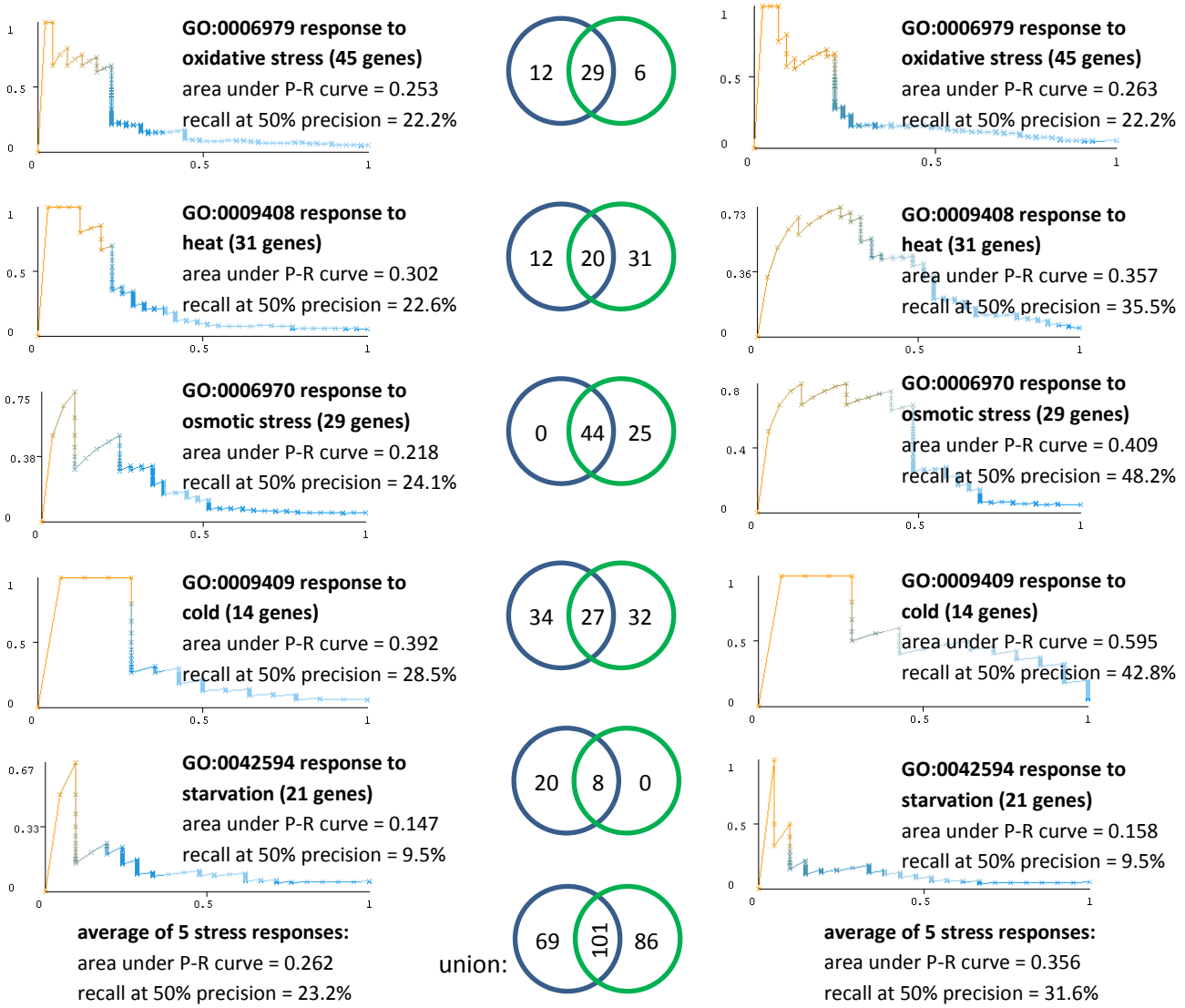
	g1	g2	g3	g4	g5	g6	g7	g8	g9	g10
COG1	0.1		0.8			0.2		0.1		0.9
COG2	0.3					0.3	0.8	0.2		1
COG3	0.2		0.9			0.2	0.9	0.3		0.9
COG4	0.3		1			0.1	0.9		1	1
COG5	0.1		0.9			0.2	0.8	0.2		0.8
COG6		0.8			0.4				0.2	
COG7	0.9			0.3	0.5		0	0.9	0.1	0
COG8		0.9	0.1		0.4	0.9			0.2	
COG9	1	0.7			0.6	1	0.1		0.1	0
COG10		0.9	0.1	0.4	0.5		0	0.9	0.2	

(table for illustration purposes only)

standard phyletic profile

	g1	g2	g3	g4	g5	g6	g7	g8	g9	g10
COG1	1	0	1	0	0	1	0	1	0	1
COG2	1	0	0	0	0	1	1	1	0	1
COG3	1	0	1	0	0	1	1	1	0	1
COG4	1	0	1	0	0	1	1	0	1	1
COG5	1	0	1	0	0	1	1	1	0	1
COG6	0	1	0	0	1	0	0	0	1	0
COG7	1	0	0	1	1	0	1	1	1	1
COG8	0	1	1	0	1	1	0	0	1	0
COG9	1	1	0	0	1	1	1	0	1	1
COG10	0	1	1	1	1	0	1	1	1	0

(table for illustration purposes only)



Additional file 17. A cross-validation test of the ability to retrieve functionally related genes, starting from the translational efficiency profiles of COGs across genomes (left panel), or the gene presence/absence profiles (right panel, equivalent to a standard phyletic profiling approach). The test uses *E. coli* K12 genes that are assigned to a COG and that are annotated with one of the five GO categories above, and compares these genes to a sample of other *E. coli* genes in COGs, but that do not have this GO function assigned. The size of the sample of these 'negative genes' is 19x the number of 'positive' genes, which thus make up 5% of the combined dataset, mimicking a realistic distribution. Then, a Random Forest model is trained to discriminate the two groups of *E. coli* genes, and tested in a *n*-fold crossvalidation scheme (RF in Weka 3.7.9, I=1000, K=30), where *n* is the number of positive genes for that GO. The plots are precision-recall curves: recall is on *x* axis, precision on *y*. Importantly, the "translation efficiency" models (left panel) do not have access to gene presence/absence information and must discriminate the groups only from the codon adaptation of the present genes; absent genes are encoded as missing data. The measure of translation efficiency in the profiles is the difference of classifier probabilities of the intergenic DNA vs. codon usage data (Fig 1A, left *vs*. right). Venn diagrams show the # genes with a newly predicted function when applying the crossvalidated models to the complete *E. coli* genome (3534 genes in COGs with a sufficient phylogenetic representation); left circle = translation efficiency profile, right = phyletic profile; both models were applied at a confidence threshold corresponding to 50% precision.

Additional file 18. An exhaustive list of the inferred clusters of orthologous groups (COGs)-phenotype links.

Available for download from the Genome Biology web site:

http://genomebiology.com/content/supplementary/gb-2014-15-3-r44-s18.xlsx

Additional file 19. A list of *Escherichia coli* strains used.

Available for download from the Genome Biology web site:

http://genomebiology.com/content/supplementary/gb-2014-15-3-r44-s19.xlsx

Additional file 20. Designed variants of *E. coli clpS* and *yjjB* genes, with progressively more optimal codons replaced by suboptimal ones (Fig. 6). The lowercase "a" in the *yjjB* sequences denotes a replacement of the original G with an A to abolish a HsdR site.

ID	DNA
*clpS*_w.t.	ATGGGTAAAACGAACGACTGGCTGGACTTTGATCAACTGGCGGAAGAAAAAGTTCGCGACGCGCTAAAACCGCCATCTATGTATAAAGTGA TATTAGTCAATGATGATTACACTCCGATGGAGTTTGTTATTGACGTGTTACAAAAATTCTTTTCTTATGATGTAGAACGTGCAACGCAATT GATGCTCGCTGTTCACTACCAGGGGAAGGCCATTTGCGGAGTCTTTACCGCCGAGGTTGCAGAAACCAAAGTGGCGATGGTGAACAAGTAC GCGAGGGAGAATGAGCATCCATTGCTGTGTACGCTAGAAAAAGCCTGA
*clpS*_15	ATGGGTAAGACGAACGACTGGCTGGACTTTGATCAACTTGCGGAGGAGAAAGTTCGCGACGCGCTAAAGCCGCCATCTATGTATAAAGTGA TATTAGTCAATGATGATTACACTCCTATGGAGTTTGTTATTGACGTGTTACAAAAATTTTTTAGTTATGATGTAGAACGCGCAACGCAATT GATGCTCGCTGTTCATTATCAGGGGAAGGCCATTTGCGGAGTCTTTACGGCCGAGGTTGCAGAGACCAAAGTGGCGATGGTGAATAAGTAC GCGAGGGAGAATGAGCATCCATTGCTGTGTACGCTAGAGAAAGCCTGA
*clpS*_20	ATGGGTAAGACGAACGACTGGCTGGACTTTGATCAACTTGCGGAGGAGAAAGTTCGCGACGCGCTAAAGCCACCATCTATGTATAAAGTGA TATTAGTCAATGATGATTATACTCCTATGGAGTTTGTTATTGACGTGTTACAAAAATTTTTTAGTTATGATGTAGAACGCGCAACGCAATT GATGCTCGCTGTGCATTATCAGGGGAAGGCCATTTGTGGAGTCTTTACGGCCGAGGTTGCAGAGACCAAAGTGGCGATGGTGAATAAGTAC GCGAGGGAGAATGAGCATCCATTGCTGTGTACGCTAGAGAAGGCCTGA
*clpS*_25	ATGGGTAAGACGAATGACTGGTTAGACTTTGATCAACTTGCGGAGGAGAAAGTTCGCGACGCGCTAAAGCCACCATCTATGTATAAAGTGA TATTAGTCAATGATGATTATACTCCTATGGAGTTTGTTATTGACGTGTTACAAAAGTTTTTTAGTTATGATGTAGAACGCGCAACGCAATT GATGCTCGCTGTGCATTATCAAGGGAAGGCCATTTGTGGAGTCTTTACGGCCGAGGTTGCAGAGACCAAAGTGGCGATGGTGAATAAGTAC GCGAGGGAGAATGAGCATCCATTGCTCTGTACGCTAGAGAAGGCCTGA
*yjjB*_w.t.	ATGGGTGTGATCGAATTTCTGTTAGCGTTGGCGCAGGATATGATCCTCGCCGCCATTCCTGCGGTCGGCTTTGCGATGGTGTTCAACGTTC CCGTaCGGGCGTTACGCTGGTGTGCGCTGCTTGGCTCGATAGGTCATGGTTCCCGAATGATCTTGATGACCAGCGGGTTGAATATTGAGTG GTCAACCTTTATGGCTTCTATGCTGGTCGGTACCATTGGTATTCAATGGTCGCGCTGGTATCTGGCGCATCCGAAAGTGTTTACCGTGGCG GCCGTTATCCCTATGTTCCCGGGCATATCGGCTTATACCGCAATGATTTCGGCGGTAAAAATCAGCCAGTTAGGTTACAGCGAACCGTTGA TGATTACCCTGTTAACCAACTTTCTTACAGCTTCATCGATTGTTGGTGCGTTATCCATCGGTCTTTCCATTCCTGGATTATGGTTGTACCG CAAGCGCCCTCGCGTATAA
*yjjB*_21	ATGGGTGTGATCGAGTTTCTGTTAGCGTTGGCGCAGGATATGATCCTCGCCGCCATTCCTGCGGTCGGCTTTGCGATGGTGTTTAATGTTC CCGTaCGGGCGTTACGCTGGTGTGCGCTCCTTGGCTCGATAGGTCATGGTTCCCGAATGATCTTGATGACCAGCGGGTTGAATATTGAGTG GTCAACCTTTATGGCGAGTATGCTGGTCGGTACCATTGGTATTCAATGGTCGCGCTGGTATTTAGCGCATCCTAAGGTGTTTACAGTGGCG GCCGTCATCCCTATGTTTCCGGGCATATCGGCTTATACCGCAATGATTTCGGCGGTAAAAATTAGCCAATTAGGTTATAGCGAGCCATTGA TGATTACGTTATTAACGAACTTTCTTACAGCTTCATCGATTGTCGGTGCGTTATCCATCGGTCTTTCCATTCCTGGATTATGGTTGTACCG CAAGCGCCCTCGCGTATAA
*yjjB*_28	ATGGGTGTGATCGAGTTTCTGTTAGCGTTGGCGCAGGATATGATCCTCGCCGCCATTCCTGCGGTCGGCTTTGCGATGGTGTTTAATGTCC CCGTaCGGGCGTTACGCTGGTGTGCGCTCCTTGGCTCGATAGGTCATGGTTCCCGAATGATCTTGATGACCAGCGGGTTGAATATTGAGTG GTCAACCTTTATGGCGAGTATGCTTGTCGGTACCATTGGTATTCAATGGTCGCGCTGGTATTTAGCGCATCCTAAGGTGTTTACAGTGGCG GCCGTCATTCCTATGTTTCCGGGCATATCGGCTTATACCGCAATGATTTCGGCGGTAAAGATTAGCCAATTAGGTTATAGCGAGCCATTGA TGATTACGTTATTAACGAACTTTCTTACAGCTTCATCGATTGTCGGTGCGTTATCCATTGGTCTTTCCATTCCTGGATTATGGTTGTATCG CAAGCGCCCTCGCGTATGA
*yjjB*_35	ATGGGTGTGATTGAGTTTTTGTTAGCGTTGGCGCAGGATATGATCCTCGCCGCCATTCCTGCGGTCGGCTTTGCGATGGTGTTTAATGTCC CCGTaCGGGCGTTACGCTGGTGTGCGCTCCTTGGCTCGATAGGTCATGGTTCCCGAATGATTTTGATGACCAGCGGGTTGAATATTGAGTG GTCAACGTTTATGGCGAGTATGCTTGTCGGTACCATTGGTATTCAATGGTCGCGCTGGTATTTAGCGCATCCTAAGGTGTTTACAGTGGCG GCCGTCATTCCTATGTTTCCAGGCATATCGGCCTATACCGCAATGATTTCGGCGGTAAAGATTAGCCAATTAGGTTATAGCGAGCCATTGA TGATTACGTTATTAACGAATTTTCTTACAGCTTCATCGATTGTCGGTGCGTTATCCATTGGTCTTTCCATTCCTGGATTATGGTTGTATCG CAAGCGCCCTCGCGTATGA