


# SCIENTIFIC REPORTS



OPEN

## Highly expressed genes evolve under strong epistasis from a proteome-wide scan in *E. coli*

Pouria Dasmeh<sup>1,2</sup>, Éric Girard<sup>1,2</sup> & Adrian W. R. Serohijos<sup>1,2</sup> 

Epistasis or the non-additivity of mutational effects is a major force in protein evolution, but it has not been systematically quantified at the level of a proteome. Here, we estimated the extent of epistasis for 2,382 genes in *E. coli* using several hundreds of orthologs for each gene within the class *Gammaproteobacteria*. We found that the average epistasis is ~41% across genes in the proteome and that epistasis is stronger among highly expressed genes. This trend is quantitatively explained by the prevailing model of sequence evolution based on minimizing the fitness cost of protein unfolding and aggregation. The genes with the highest epistasis are also functionally involved in the maintenance of proteostasis, translation and central metabolism. In contrast, genes evolving with low epistasis mainly encode for membrane proteins and are involved in transport activity. Our results highlight the coupling between selection and epistasis in the long-term evolution of a proteome.

Resolving the link between genotype and phenotype or the fitness landscape is a central goal in molecular biology and evolution. Knowledge of the structure of the fitness landscape will lead to a better understanding of the evolutionary origin of natural proteins and to solutions to practical evolutionary problems, from rational design of enzymes to the development of new antibiotics<sup>1</sup>. The fitness landscape is complex and a consequence of this complexity is *epistasis* or the dependence of mutational effects on genetic background<sup>2</sup>. The presence of epistasis implies that the effects of multiple mutations are non-additive and that their order of fixation matters. Indeed, epistasis directly affects the potential pathways to explore the fitness landscape<sup>2</sup>. Despite the many experimental and theoretical studies on detecting and elucidating its role in molecular evolution<sup>3,4</sup>, none has investigated the strength of epistasis at a proteome-wide level. Such an analysis can determine correlations between epistasis and genomics properties that could hint at a universal mechanism, if any, for epistasis in proteome evolution. Additionally, a mechanistic understanding of epistasis has practical applications; as yet, it is rarely accounted for in the molecular evolution toolboxes for quantifying from genomic sequences the strength of multiple evolutionary forces—mutation, drift and selection<sup>4,5</sup>.

Epistasis, or the non-additivity of mutational effects, has a direct role on the rate of protein evolution because this implies that the genetic background can attenuate the effect of the mutation and hence affect its likelihood of fixation. Nonetheless, this relationship between epistasis and rate of evolution is complex because of confounding factors that also affect the rate of evolution, most important of which is selection. To estimate the epistasis experienced by genes in long-term evolution, one approach is to compare two rates of amino acid substitutions<sup>6</sup>. These two rates are the average pairwise substitution rate  $R_{dN/dS}$ , which is background-dependent, and the rate of mutational usage  $R_u$ , which is background independent. Both rates are calculated from a multiple sequence alignment (MSA) of orthologs. Specifically, the extent of epistasis is quantified as

$$\varepsilon = 1 - \frac{R_{dN/dS}}{R_u} \quad (1)$$

$R_{dN/dS}$  is the average  $dN/dS$  (the ratio of non-synonymous substitution rate  $dN$  and synonymous substitution rate  $dS$ ) for all pairs of orthologues in an MSA.  $R_{dN/dS}$  is calculated over the entire length of the gene, thus it reflects the co-evolution between sites. This also implies that  $R_{dN/dS}$  accounts for the background- and lineage-specificity of amino acid substitutions. The second rate,

<sup>1</sup>Departement de Biochimie, Université de Montréal, 2900 Édouard-Montpetit, Montréal, Québec, H3T 1J4, Canada. <sup>2</sup>Centre Robert Cedergren en Bioinformatique et Génomique, Université de Montréal, 2900 Édouard-Montpetit, Montréal, Québec, H3T 1J4, Canada. Pouria Dasmeh and Eric Girard contributed equally to this work. Correspondence and requests for materials should be addressed to A.W.R.S. (email: [adrian.serohijos@umontreal.ca](mailto:adrian.serohijos@umontreal.ca))

$$R_u = \left(\frac{1}{L}\right) \sum_i^L \frac{(u_i - 1)}{19}$$

where  $u$  is the mutational usage and is the number of unique amino acids in each site in an MSA.  $L$  is the length of the protein.  $R_u$  represents the ratio between observed accessible amino acid substitutions in a site,  $(u-1)$ , and all possible amino acid substitution assuming no selection, that is,  $(20-1) = 19$ . Unlike  $R_{dN/dS}$ ,  $R_u$  simply counts the number of unique amino acids per site, thus it does not reflect the co-evolution between sites in the protein. Therefore,  $R_u$  is independent of background and lineage. When all mutations are neutral, both  $R_u$  and  $R_{dN/dS}$  are equal to 1. When *random* mutations are not neutral, such as in proteins where they are predominantly destabilizing and deleterious<sup>7</sup>, purifying selection will lead to  $R_u$  and  $R_{dN/dS}$  less than 1. However, the presence of epistasis implies that genetic background further screens substitutions, thus the background dependent  $R_{dN/dS}$  is slower than the background independent  $R_u$ . The expression for epistasis (Eq. 1) estimates how much epistasis or the background specificity of mutational effects slows down the rate of protein evolution. Kondrashov and coworkers<sup>6</sup> applied this method to estimate epistasis in the long-term evolution of 16 mammalian proteins and found epistasis to vary from ~40% to ~80%. This slowing down of evolutionary rates can also arise from the heterogeneity of fitness effects of mutations<sup>8</sup>.

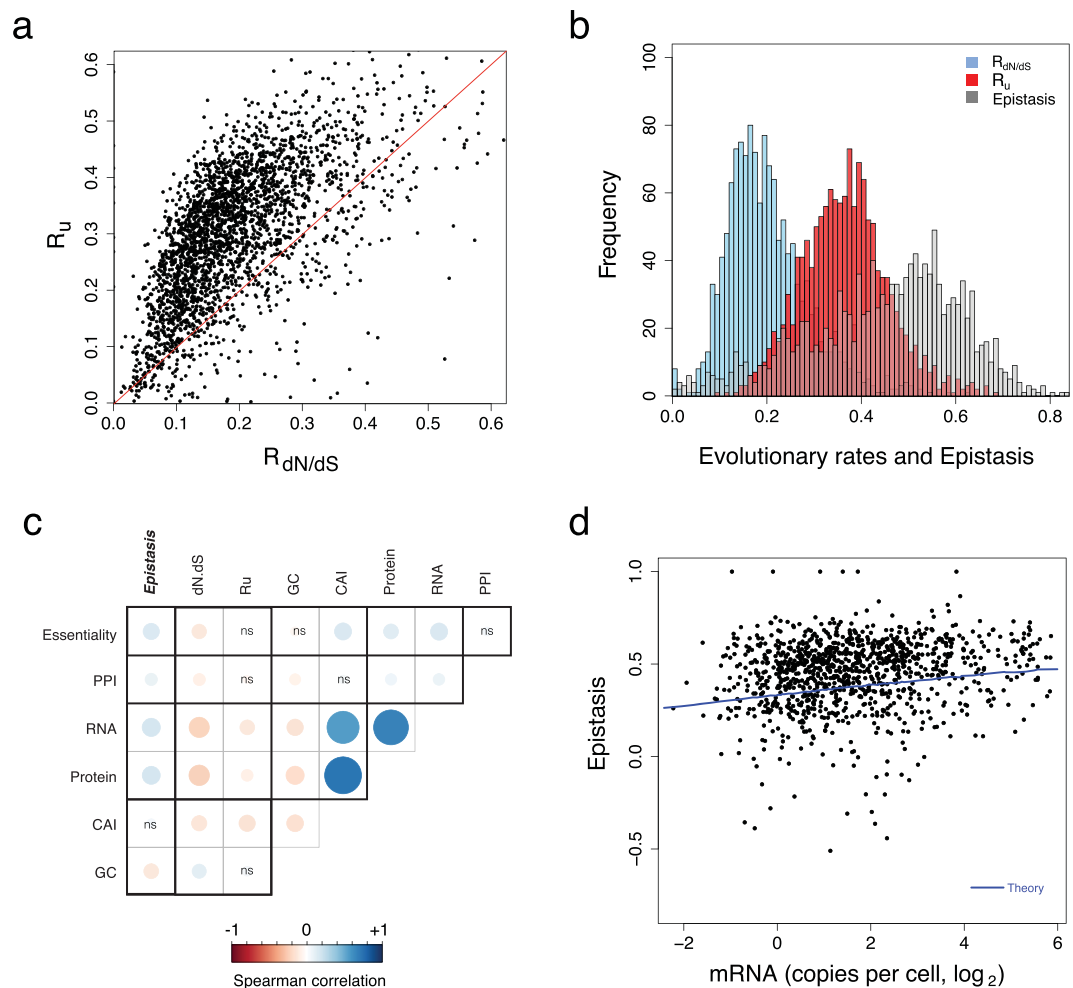
## Results

**Proteome-wide epistasis in *E. coli*.** To determine if epistasis is indeed experienced by all genes in a proteome we estimated epistasis in the evolution of 2,382 genes in *E. coli* using thousands of orthologs within the class *gammaproteobacteria* (see Methods and Supporting Information). We calculated  $R_u$  and  $R_{dN/dS}$  from the multiple sequence alignment of each gene (Methods). The rates  $R_u$  and  $R_{dN/dS}$  and the epistasis for each gene are shown in Fig. 1a and their distributions in Fig. 1b. Since epistasis is expected to slow down the rate of evolution, the lineage-independent rate  $R_u$  is greater than the lineage- and background-dependent  $R_{dN/dS}$  (note the deviation from  $R_u = R_{dN/dS}$  line in Fig. 1a; Wilcoxon signed-rank test, p-value  $< 10^{-16}$ ). The average  $R_u$  and  $R_{dN/dS}$  are  $0.36 \pm 0.09$  and  $0.20 \pm 0.08$ , respectively, which lead to a proteome-wide epistasis of ~41% (Fig. 1b; full data is listed in Table S1). This estimate implies that epistasis and background specificity of mutational effects in proteins slows down the evolutionary rates of proteins in *E. coli*, on average, by ~41%. The magnitude of epistasis is broadly distributed with some genes experiencing epistasis of up to ~80% (Fig. 1b). These estimates over several thousand genes is slightly lower than the value calculated by Breen *et al.*<sup>6</sup> for 16 mammalian proteins.

Next, we checked for the robustness of our results to factors that may influence the calculation of substitution rates and epistasis. First, the evolutionary rates, in particular  $R_u$  that count the number of unique amino acids, are sensitive to the number of orthologs in an MSA. Too few orthologs may lead to undersampling of  $R_u$  and to negative values for epistasis. However, as shown by the plot of epistasis versus the number of orthologs (Fig. S1), this artifact is present only in genes with MSA alignments less than 200 orthologs. Second, a distinction should be made between fixed and non-fixed amino acids. In principle, fixed amino acids are substitutions that are kept in long-term evolution while non-fixed amino acids are segregated in the population at short time scales and are eventually lost. Since Equation 1 estimates epistasis in long-term protein evolution, non-fixed amino acid states or polymorphisms can inflate the mutational usage  $R_u$  and epistasis. To account for the bias due to non-fixed polymorphic states in our amino-acid usage calculation, we used a correction based on the probability of occurrence of non-fixed amino acids at given site in the alignment (Fig. S2 and Supporting Information). The average correction to  $R_u$  due to non-fixed polymorphism is only  $\sim \pm 2\%$  (Fig. S2). Table S2 presents amino acid usage correction along with the probability of observing a non-fixed state as fixed for all genes. Lastly, the calculation of  $R_{dN/dS}$  could be sensitive to the counting method. We control for this effect by using several counting methods (five heuristic and two maximum-likelihood codon-based approaches) for  $dN$ ,  $dS$ , and  $dN/dS$  (Figs S5, S6, S7 and Table S3) and chose the most unbiased (Methods). Altogether, our estimates of epistasis are robust to the number of orthologs, presence of polymorphisms, and approaches for counting substitution rates.

**Relationship of epistasis with genomic properties.** The rate of protein evolution is influenced by several factors ranging from molecular, to cellular, and to population level<sup>9</sup>. We determined if epistasis is also influenced by these factors. Specifically, we calculated the correlation between epistasis and publicly available data on mRNA expression level, protein abundance, gene essentiality, protein-protein interaction (PPI), and codon adaptation index (CAI) (Fig. 1c, Table S4 see Methods). We found that epistasis shows a weak yet significant positive correlation with number of orthologs, PPI, mRNA and protein expression levels, as well as CAI. As shown previously<sup>10</sup>,  $R_{dN/dS}$  negatively correlates with expression level ( $r = -0.24$ , p-value  $< 10^{-16}$ ) implying that highly expressed genes are under strong purifying selection. Since  $R_u$  also reflects selection, it similarly shows negative correlation with expression level ( $r = -0.18$ , p-value  $< 10^{-10}$ ). However, the weaker anti-correlation between  $R_u$  and expression level compared to that of  $R_{dN/dS}$  leads to a positive correlation between epistasis and expression level ( $r = +0.17$ , p-value  $< 10^{-9}$ ) (Fig. 1d). This finding implies that background specificity significantly slows down the rate of evolution among highly expressed genes.

To further check the biological significance of proteins evolving under high epistasis, we performed Gene Ontology (GO) enrichment analyses<sup>11</sup> on the lowest and highest quantiles of epistasis (see Table S5). The lowest quantile corresponds to genes with epistasis less than 18% and lower, and the top quantile with epistasis greater than 53%. Interestingly, essential processes such as amino acid and nucleobase synthesis, ATP and RNA binding and proteostasis were significantly enriched among the genes evolving with high epistasis (Table S6). For example, the genes *glyA* and *prs* encoding the Serine hydroxymethyltransferase and Ribose-phosphate pyrophosphokinase, evolve with ~65% epistasis with ~1000 orthologues and are essential enzymes in the synthesis of amino acids and nucleobases. Other examples are the chaperone protein DnaK with 63% (944 orthologues) and elongation factor 4



**Figure 1.** Proteome-wide estimate of epistasis in *E. coli*. **(a)** Background-dependent evolutionary rate  $R_{dN/dS}$  is significantly slower than the background-independent rate of mutational usage  $R_u$  (Wilcoxon signed-rank test,  $p$ -value  $< 10^{-16}$ ). **(b)** The average epistasis is  $\sim 41 \pm 16\%$  among 2,382 genes in *E. coli*. **(c)** Epistasis is positively correlated with genome-wide factors: mRNA and protein expression levels, essentiality of proteins, number of protein-protein interactions and codon adaptation index (CAI) (see Table S4 for the correlation coefficients and  $p$ -values). Boxes labeled “ns” are not significant ( $p$ -value  $> 0.05$ ). **(d)** Highly expressed genes experience strong epistasis (Spearman  $r = +0.17$ ,  $p$ -value  $< 10^{-9}$ ), which can be explained by a model of sequence evolution based on selection against protein misfolding and aggregation (blue line; see also Figs S3, S4 and S10).

(EF4) with 67% epistasis (953 orthologues). In contrast to highly epistatic genes, those evolving with low epistasis are mainly transmembrane proteins (Table S7). It is well-established that membrane proteins are dramatically less conserved than water soluble proteins and have higher evolutionary rates due to adaptation to the changing environment<sup>12</sup> which could then influence the estimated epistasis using Eq. 1. Altogether, our proteome-wide estimates demonstrate that highly expressed genes not only experience stronger purifying selection, but also greater epistasis in their long-term evolution. This result highlights the coupling of selection and epistasis in proteome evolution<sup>13</sup>.

**Model of sequence evolution based on protein folding explains proteome-wide correlation between epistasis and expression level.** The negative correlation between evolutionary rate  $R_{dN/dS}$  and expression level is well-established<sup>10,14,15</sup>. This observation has been explained by a model of sequence evolution based on selection against protein misfolding due to mistranslation<sup>10</sup> or genetic mutations<sup>14,16</sup>. The biological rationale is that misfolded proteins can form aggregates that are toxic to the cell<sup>17,18</sup>. To determine if the same hypothesis can quantitatively explain the trend between epistasis and mRNA expression level, we combine the population genetic formalism for evolutionary rate with protein folding thermodynamics<sup>14,19–21</sup>. Assuming that cellular fitness  $F$  is *inversely* proportional to the total number of misfolded proteins in the cell, it may be formally written as<sup>10</sup>:

$$F = \exp[-c(\# \text{ of misfolded proteins})] = \exp\left[-c\left(\sum_k^{\Gamma} A_k \frac{1}{1 + \exp(\beta\Delta G_k)}\right)\right] \quad (2)$$

Equation 2 expresses the probability that the protein product of gene  $k$  is unfolded as a function of its stability  $\Delta G_k$ . The energy factor  $\beta = 1/k_b T$  where  $k_b T \sim 0.59$  kcal/mol at room temperature. This probability multiplied by the cellular abundance of the gene  $A_k$  gives the number of misfolded copies (Fig. S10). The summation extends over all genes  $\Gamma$  in the proteome. The parameter  $c$  is the fitness cost of each misfolded protein ( $\sim 10^{-7}$ ) (ref.<sup>18</sup>). As shown previously<sup>10,14</sup> and in our specific dataset (Figs S3 and S4), this fitness function recapitulates the trend between dN/dS and expression level. But to arrive at epistasis, we also need a theoretical estimate for the mutational usage  $R_u$ . In a recent work<sup>19</sup>, we showed that  $R_u$  is the rate of evolution of the most stable sequence in an MSA. By simulating sequence evolution (Supporting Information), we can arrive at a theoretical MSA evolved under the fitness function (Eq. 2) and then calculate  $R_u$  (Fig. S4).

Highly expressed (and more abundant) genes are under strong purifying selection; thus,  $R_u$  and  $R_{dN/dS}$  negatively correlate with mRNA level, both in theory (Fig. S4) and in *E. coli* (Fig. S3). More interestingly, the theoretical dependence of  $R_u$  vs. mRNA is weaker than  $R_{dN/dS}$  vs. mRNA leading to stronger epistasis among highly expressed genes (Fig. S4). Thus, selection against protein misfolding can explain the genomic observation that highly expressed and more abundant genes experience stronger epistasis (Fig. 1c,d). A geometric interpretation of epistasis is the curvature of the fitness landscape; indeed, for the genotype-phenotype relationship based on folding stability (Eq. 2), the landscape exhibits greater curvature at higher expression levels (Fig. S10).

## Discussion

Our study, for the first time, provides proteome-wide estimate of epistasis in *E. coli*. On average, a protein in *E. coli* evolves with  $\sim 41\%$  epistasis. One interpretation of this result is that the rate of protein evolution is reduced by 41% due to background dependence of mutational effects. Moreover, we found that highly expressed proteins evolve with stronger epistasis, which can be explained by selection against protein misfolding. Our results highlight the coupling between selection and epistasis, which has been demonstrated in specific proteins<sup>4,22</sup>, but not in the long-term evolution of a proteome.

We also tested the enrichment of functional groups among genes under high or low epistasis. We found that genes evolving with high epistasis are involved in essential processes such as the maintenance of proteostasis and rRNA and ATP binding. This finding is in line with previous observations that genes with high intergenic pleiotropy in yeast are often involved in more cellular processes than low pleiotropic genes<sup>23</sup>. Here we systematically showed that intragenic epistasis has the same pattern in *E. coli*. We anticipate that future studies on the molecular evolution of proteins evolving with high epistasis could provide a mechanistic understanding of epistasis at the residue level. Furthermore, genes evolving with high epistasis are noteworthy as they tolerate maximum number of novel amino acids and thus are highly evolvable. The methodologies employed in this study can aid in selecting such genes at a genome-wide level. In addition, the coupling between epistasis, abundance and essentiality as described in this work can be used to update substitution matrices and phylogenetic trees of highly expressed proteins.

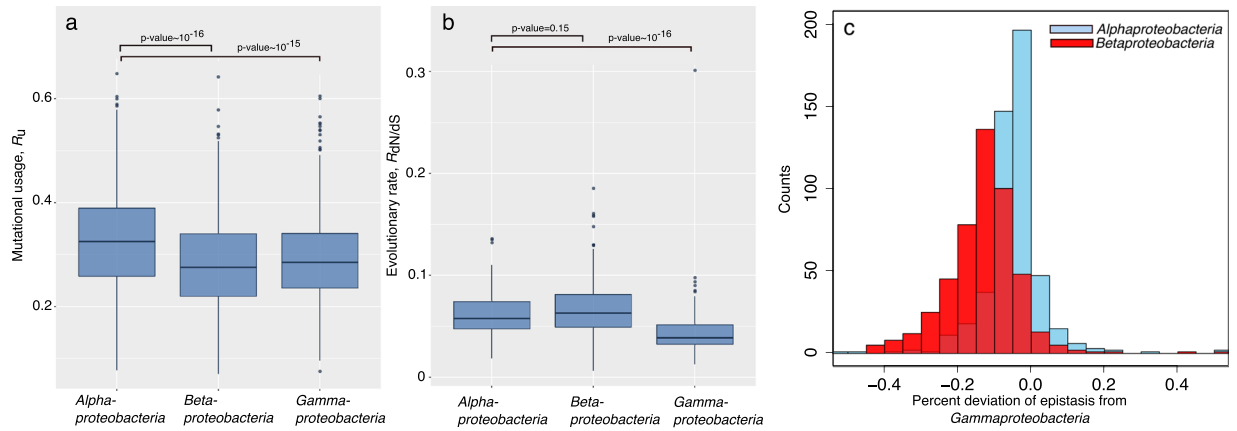
The extent of epistasis as reported in this work depends on amino acid usage and evolutionary rate of proteins, and both quantities were shown to vary in different habitats and can be influenced by environmental conditions<sup>24,25</sup>. To investigate the role of habitat and lifestyle on epistasis, we estimated epistasis in the evolution of *E. coli* orthologous proteins within the two classes of *alpha*- and *betaproteobacteria*. Bacterial species within the classes *alpha* and *betaproteobacteria* generally live in low and high nutrient environments, respectively<sup>26,27</sup>. We retrieved 2098 orthologs for *E. coli* proteins within the class *alphaproteobacteria* and 2174 within the classes *betaproteobacteria*. We then compared epistasis across the three classes, while controlling for the unequal number of orthologues (see Methods).

As shown in Fig. 2a, the average  $R_u$  is slightly higher for orthologs in *alphaproteobacteria* ( $\langle R_u \rangle = 0.33 \pm 0.09$ ) than *betaproteobacteria* ( $\langle R_u \rangle = 0.28 \pm 0.09$ ) and *gammaproteobacteria* ( $\langle R_u \rangle = 0.29 \pm 0.09$ ) with p-values of  $\sim 10^{-16}$  (Wilcoxon signed-rank test). Evolutionary rate,  $R_{dN/dS}$ , when corrected for the level of divergence (dS), is not significantly different (p-value = 0.15) between *alpha* and *betaproteobacteria* with  $\langle R_{dN/dS} \rangle = 0.062 \pm 0.023$  and  $0.068 \pm 0.032$ , respectively (Fig. 2b). This makes the average epistasis  $\sim 0.76$ ,  $0.67$  and  $0.80$  for *alpha*, *beta* and *gammaproteobacteria*. Note that epistasis is overestimated (i.e., compared to previously reported 41%) due to the smaller number of orthologues in this comparison. Therefore, proteins in *alpha* and *betaproteobacteria* evolve with 5% and 16% lower epistasis compared to their orthologs in *gammaproteobacteria* mainly because of differences in mutational usage (Fig. 2c). As elegantly showed by Akashi and Gojobori, mutational usage is significantly different in bacteria with different metabolic profiles<sup>28</sup>. We anticipate that biosynthetic cost minimization, among other factors, may underlie the differences in the extend of mutational usage and hence epistasis in the evolution of proteins within these classes of *proteobacteria* phylum.

This work focused on the *E. coli* proteome; however, it will be interesting to generalize these observations in other well-studied model organisms such as yeast, worm, fly, mouse, and human, where selection (detected by dN/dS) has been shown to strongly correlate with expression level<sup>10</sup>. Demonstrating these results across all kingdoms of life could generalize the finding that selection due to folding stability is a universal mechanism for some of the epistasis experienced in the long-term evolution of a proteome.

## Methods

**Sequences and alignment.** List of genes for *Escherichia coli* K-12 MG1655 was taken from NCBI. From this list (4140 genes), KEGG ids (total of 3059 ids) were used to retrieve functional orthologs within *Gammaproteobacteria* class. Ortholog sequences were available for 2814 of the 3059 genes. To optimize alignment,



**Figure 2.** Epistasis is influenced by bacterial lifestyles. (a)  $R_u$  and (b)  $R_{dN/dS}$  in the evolution of *E. coli* orthologous proteins within the three classes of *alpha*-, *beta*- and *gammaproteobacteria*. In (c) the percent deviation of epistasis in the evolution of *alpha*- and *betaproteobacteria* from *gammaproteobacteria* is calculated. All the p-values are calculated using Wilcoxon rank sum test.

sequences 15% longer or shorter from the reference *E. coli* gene were removed from the set. DNA sequences were converted to protein sequences prior to alignment and calculation of amino-acid usage. For the protein alignments, we used default parameters except for the allowed positions with gaps that were set to half, to allow gaps at positions where less than 50% of sequences had gaps.

**Bioinformatics.** The amino-acid usage measure can be used to obtain an estimation of dN/dS ratio under the assumption of non-epistatic evolution. The amino-acid usage  $\langle u \rangle$  is defined as the number of different amino-acids observed at one site, averaged over all sites in an alignment. We can then estimate non-epistatic dN/dS from  $\langle u \rangle$  using  $(u - 1)/19$  where  $(u - 1)$  is the number of amino-acid states into which the current amino acid can be substituted, divided by 19 amino-acid possibilities. The choice of a proper and unbiased method to estimate dN/dS is crucial in the current work. We thus systematically checked performance of five different heuristic counting approaches and two maximum-likelihood (ML) codon models for 3124 genes in *E. coli* and concluded that the simplest model of Nei and Gojobori<sup>29</sup> gives the most unbiased dS and dN/dS estimates which reasonably fit values from accurate yet computationally expensive ML methods. For the complete analysis check Supplementary methods and Table S4. All the GO-enrichment analyses were done using DAVID bioinformatics resources<sup>30</sup>.

To compare epistasis among the three classes of *alpha*-, *beta*-, and *gammaproteobacteria*, we focused on proteins with more than 300 orthologs within each class. This number was chosen to insure we have more than 500 genes and thus the proteome-wide estimate of epistasis is within 95% confidence level and a margin of error of less than five percent<sup>31</sup>. We then estimated  $R_u$ , dN, dS and  $R_{dN/dS}$  for the exact number of 300 orthologs for each gene. When more orthologs within each class were available, we randomly selected 300 sequences and calculated the average of all quantities for ten repeats. As the number of orthologs are significantly smaller for the classes *alpha*- and *betaproteobacteria*,  $R_u$  is underestimated for proteins in these classes. As a result, epistasis would be overestimated for orthologous proteins in *alpha*- and *betaproteobacteria*. To resolve this issue, we applied the same procedure for *E. coli* proteins within the class *gammaproteobacteria* and calculated evolutionary rates and epistasis.

**Theoretical model.** To calculate the extent of epistasis, we used an expression for substitution rates that takes the stability effects of mutations into account (Equation 1). Fitness is proportional to the number of misfolded copies in the cell which in turn is a product of total abundance and the probability of being in the folded state. This decomposition enables us to utilize the known distribution of mutational effects on protein stability to determine the distribution of fitness effects and calculate evolutionary rates accordingly (see Supplementary information for full analysis). The fitness landscape (Equation 2) contains protein abundance, thus we converted mRNA abundance to protein abundance using their well-established correlation (Supplementary information). The calculation of the rate of mutational usage  $R_u$  and pairwise rate of evolution  $R_{dN/dS}$  based on the fitness landscape of Equation 2 is described in the Supplementary information.

## References

- Romero, P. A. & Arnold, F. H. Exploring protein fitness landscapes by directed evolution. *Nature Reviews Molecular Cell Biology* **10**, 866–876 (2009).
- Phillips, P. C. The language of gene interaction. *Genetics* **149**, 1167–1171 (1998).
- Weinreich, D. M., Lan, Y., Wylie, C. S. & Heckendorn, R. B. Should evolutionary geneticists worry about higher-order epistasis? *Current opinion in genetics & development* **23**, 700–707 (2013).
- Starr, T. N. & Thornton, J. W. Epistasis in protein evolution. *Protein Science* (2016).
- Phillips, P. C. Epistasis—the essential role of gene interactions in the structure and evolution of genetic systems. *Nature Reviews Genetics* **9**, 855–867 (2008).
- Breen, M. S., Kemena, C., Vlasov, P. K., Notredame, C. & Kondrashov, F. A. Epistasis as the primary factor in molecular evolution. *Nature* **490**, 535–538 (2012).



7. Tokuriki, N., Stricher, F., Schymkowitz, J., Serrano, L. & Tawfik, D. S. The stability effects of protein mutations appear to be universally distributed. *Journal of molecular biology* **369**, 1318–1332 (2007).
8. McCandlish, D. M., Rajon, E., Shah, P., Ding, Y. & Plotkin, J. B. The role of epistasis in protein evolution. *Nature* **497**, E1–E2 (2013).
9. Zhang, J. & Yang, J.-R. Determinants of the rate of protein sequence evolution. *Nature Reviews Genetics* **16**, 409–420 (2015).
10. Drummond, D. A. & Wilke, C. O. Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. *Cell* **134**, 341–352 (2008).
11. Ashburner, M. *et al.* Gene Ontology: tool for the unification of biology. *Nature genetics* **25**, 25–29 (2000).
12. Sojo, V., Dessimoz, C., Pomiankowski, A. & Lane, N. Membrane proteins are dramatically less conserved than water-soluble proteins across the tree of life. *Molecular biology and evolution* **33**, 2874–2884 (2016).
13. Gupta, A. & Adami, C. Strong selection significantly increases epistatic interactions in the long-term evolution of a protein. *PLoS Genet* **12**, e1005960 (2016).
14. Serohijos, A. W., Rimas, Z. & Shakhnovich, E. I. Protein biophysics explains why highly abundant proteins evolve slowly. *Cell reports* **2**, 249–256 (2012).
15. Pál, C., Papp, B. & Hurst, L. D. Highly expressed genes in yeast evolve slowly. *Genetics* **158**, 927–931 (2001).
16. Bratulic, S., Gerber, F. & Wagner, A. Mistranslation drives the evolution of robustness in TEM-1  $\beta$ -lactamase. *Proceedings of the National Academy of Sciences* **112**, 12758–12763 (2015).
17. Bershtein, S., Mu, W. & Shakhnovich, E. I. Soluble oligomerization provides a beneficial fitness effect on destabilizing mutations. *Proceedings of the National Academy of Sciences* **109**, 4857–4862 (2012).
18. Geiler-Samerotte, K. A. *et al.* Misfolded proteins impose a dosage-dependent fitness cost and trigger a cytosolic unfolded protein response in yeast. *Proceedings of the National Academy of Sciences* **108**, 680–685 (2011).
19. Dasmeh, P. & Serohijos, A. Estimating The Contribution Of Folding Stability To Non-Specific Epistasis In Protein Evolution. *bioRxiv*, <https://doi.org/10.1101/122259> (2017).
20. Dasmeh, P., Serohijos, A. W., Kepp, K. P. & Shakhnovich, E. I. The influence of selection for protein stability on dN/dS estimations. *Genome biology and evolution* **6**, 2956–2967, <https://doi.org/10.1093/gbe/evu223> (2014).
21. Goldstein, R. A. The evolution and evolutionary consequences of marginal thermostability in proteins. *Proteins: Structure, Function, and Bioinformatics* **79**, 1396–1407 (2011).
22. Bershtein, S., Segal, M., Bekerman, R., Tokuriki, N. & Tawfik, D. S. Robustness–epistasis link shapes the fitness landscape of a randomly drifting protein. *Nature* **444**, 929–932 (2006).
23. He, X. & Zhang, J. Toward a molecular understanding of pleiotropy. *Genetics* **173**, 1885–1891 (2006).
24. Li, S.-J. *et al.* Microbial communities evolve faster in extreme environments. *Scientific reports* **4**, 6205 (2014).
25. Sharp, P. M., Bailes, E., Grocock, R. J., Peden, J. F. & Sockett, R. E. Variation in the strength of selected codon usage bias among bacteria. *Nucleic acids research* **33**, 1141–1153 (2005).
26. Fiebig, A., Herrou, J., Willett, J. & Crosson, S. General stress signaling in the alphaproteobacteria. *Annual review of genetics* **49**, 603–625 (2015).
27. Fierer, N., Bradford, M. A. & Jackson, R. B. Toward an ecological classification of soil bacteria. *Ecology* **88**, 1354–1364 (2007).
28. Akashi, H. & Gojobori, T. Metabolic efficiency and amino acid composition in the proteomes of *Escherichia coli* and *Bacillus subtilis*. *Proceedings of the National Academy of Sciences* **99**, 3695–3700 (2002).
29. Nei, M. & Gojobori, T. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Molecular biology and evolution* **3**, 418–426 (1986).
30. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature protocols* **4**, 44–57 (2009).
31. Krejcie, R. V. & Morgan, D. W. Determining sample size for research activities. *Educational and psychological measurement* **30**, 607–610 (1970).

## Acknowledgements

We thank members of the Serohijos lab for Discussion and Christine Angus-Banaszek for a careful reading of the manuscript. We acknowledge funds from the Université de Montréal, NSERC, and the Merck Foundation.

## Author Contributions

A.S. and P.D. conceived the study. P.D. and E.G. performed the research and analyses. A.S., E.G. and P.D. wrote the manuscript.

## Additional Information

**Supplementary information** accompanies this paper at <https://doi.org/10.1038/s41598-017-16030-z>.

**Competing Interests:** The authors declare that they have no competing interests.

**Publisher's note:** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2017