Classification

Biological Sciences: Evolution, Microbiology

Title

Selection-driven cost-efficiency optimisation of transcript sequences determines the rate of gene sequence evolution in bacteria

Short Title

Selective constraints on bacterial gene evolution

Authors

Emily A. Seward[1], Steven Kelly[1]

Affiliations

[1]Department of Plant Sciences, University of Oxford, South Parks Road, Oxford, OX1 3RB, UK

ORCID iDs

0000-0002-7869-0641, 0000-0001-8583-5362

Corresponding Author

Name: Steven Kelly

Email: steven.kelly@plants.ox.ac.uk

Telephone: +44 (0)1865 275123

Address: Department of Plant Sciences, University of Oxford, South Parks Road, Oxford, OX1 3RB, UK

Keywords

Gene evolution, Synonymous codon use, Codon bias, Translational efficiency, Bacteria, natural selection, Transcript optimisation, molecular evolution

Abstract

Due to genetic redundancy, multiple synonymous codons can code for the same amino acid. However synonymous codons are not used equally and this biased codon use varies between different organisms. Through analysis of 1,320 bacterial genomes we show that bacteria are under genome-wide selection to reduce resource allocation to mRNA production. This is achieved by simultaneously decreasing transcript biosynthetic cost and increasing transcript translational efficiency, with highly expressed genes under the greatest selection. We show that tRNA gene copy number alters the cost-efficiency trade-off of synonymous codons such that for many species it is difficult to both minimise transcript cost and maximise transcript translational efficiency to an equal extent. Finally, we show that genes highly optimised to reduce cost and increase efficiency show reduced rates of synonymous and non-synonymous mutation. This provides a simple mechanistic explanation for variation in evolutionary rate between genes that depends on selection-driven cost-efficiency optimisation of the transcript. These findings reveal how optimisation of resource allocation to mRNA synthesis is a critical factor that determines both the evolution and composition of genes.

Significance statement

Resource limitation limits cell growth. In this work, we show that the simple economic principles of minimising cost and maximising efficiency of gene production are key drivers of molecular sequence evolution in bacteria. We demonstrate that natural selection has optimised the cost and efficiency of transcript sequences directly in proportion to their production demand. Furthermore, we show that selection-driven optimisation of the economics of gene production is a simple unifying mechanism that explains variation in the evolution and composition of genes as well as the correlation between synonymous and non-synonymous mutation rates.

Introduction

Production of proteins is a primary consumer of cell resources (1). It requires allocation of cellular resources to production of RNA sequences as well as allocation of resources to production of nascent polypeptide chains. Whilst a protein's amino acid sequence is functionally constrained, redundancy in the genetic code means that multiple nucleotide sequences can code for the same protein. Since the biosynthetic cost and translational efficiency of synonymous codons varies, biased use of synonymous codons makes it possible to reduce the expenditure of cellular resources on mRNA production without altering the encoded protein sequence. Thus, it is possible to reduce resource allocation to protein synthesis without altering the encoded protein or affecting protein abundance by reducing resource allocation to transcript sequences or by increasing the efficiency with which those transcripts can be translated into protein. Consistent with this, it has been demonstrated that natural selection acts both to reduce biosynthetic cost of RNA sequences (2, 3), and to increase in the efficiency with which those RNA sequences can template the encoded polypeptide chain (4–10). It should be noted here that biased patterns of codon use are also influenced by other factors not associated with cellular resource allocation. These factors include RNA structural constraints to facilitate thermal adaptation (11, 12), RNA sequence constraints to preserve splice sites (13), and translational constraints to ensure accurate protein folding (14, 15). These functional constraints are independent of resource allocation and are specific to individual sites or sets of sites within genes.

Cells employ different strategies to decode synonymous codons (16). These strategies make use of 'wobble' base pairing between the 3rd base of the codon and the 1st base of the anticodon to facilitate translation of all 61 sense codons using a reduced set of tRNAs. As the translational efficiency of a codon is a function of the number of tRNAs that can translate that codon, and as different species encode different subsets of tRNA genes, the same codon is not necessarily equally translationally efficient in all species. In contrast, the

biosynthetic cost of a codon of RNA is determined by the number and type of atoms contained within that codon and the number of high energy phosphate bonds required for their assembly. As translational efficiency varies between species but biosynthetic cost does not, it was hypothesised this must create a corresponding variation in the codon cost-efficiency trade-off between species. For example biosynthetically cheap codons might be translationally efficient in one species but inefficient in another. It was further hypothesised that variation in the codon cost-efficiency trade-off would limit the extent to which a transcript could be optimised to be both biosynthetically inexpensive and translationally efficient.

Here, we show that natural selection acts genome-wide to reduce cellular resource allocation to mRNA synthesis by solving the multi-objective optimisation problem of minimising transcript biosynthetic cost whilst simultaneously maximising transcript translational efficiency. We show that this optimisation is achieved irrespective of the codon cost-efficiency trade-off of a species, and that the extent to which resource allocation is optimised is a function of the production demand of that gene. Finally, we reveal that selection-driven optimisation of resource allocation provides a novel mechanistic explanation for differences in evolutionary rates between genes, and for the previously unexplained correlation in synonymous and non-synonymous mutation rates of genes.

Results

55% of bacteria exhibit a significant trade-off between codon biosynthetic cost and translational efficiency

The biosynthetic cost of a codon is determined by the number and type of atoms contained within the codon, and the number of high energy phosphate bonds required for their assembly. Natural selection acting on biosynthetic cost, either in terms of nitrogen atoms (2) or energetic requirements (3), has been shown to play a role in promoting biased patterns of synonymous codon use. As the energy and nitrogen cost of a codon correlate almost perfectly (Figure 1A), it is not possible to distinguish which factor is responsible for biased

patterns of codon use in the absence of additional information about the biology of the organism in question. However, given the near perfect correlation, analysis of selection acting on overall codon biosynthetic cost can be approximated by analysis of either measure.

tRNA gene copy number varies between species resulting in a corresponding variation in the relative translational efficiency of their associated codons (16, 17). As the biosynthetic cost of a given codon is invariant, the relationship between codon biosynthetic cost and codon translational efficiency (referred to from here on as the codon cost-efficiency trade-off) must therefore vary between species. For example, a hypothetical species encoding a full complement of tRNAs, each present as a single copy, would have a negative correlation between cost and efficiency (Figure 1B). In contrast, a hypothetical species that employed tRNA sparing strategy 1 (no ANN tRNAs) or strategy 2 (no ANN or CNN tRNAs) (16), would show a positive (Figure 1C) or no (Figure 1D) correlation between cost and efficiency. Therefore, a broad range of codon cost-efficiency trade-offs is possible and the gradient of this trade-off is dependent on the tRNA gene copy number of a given species.

Few bacterial species strictly adhere to a single sparing strategy for all synonymous codon groups (e.g. *Escherichia coli* uses strategy 2 for decoding alanine but strategy 1 for decoding glycine), and thus it is anticipated that a continuum of gradients in trade-off between cost and efficiency are possible. To assess this, the codon cost-efficiency trade-off was calculated for 1,320 bacterial species representing 730 different genera (Figure 1E). Approximately 55% of species exhibited a significant codon cost-efficiency trade-off (Figure 1E, $p < 0.05$). Species with a significant negative correlation between cost and efficiency were not observed; instead, the remaining species exhibited non-significant correlations between codon cost and efficiency (Figure 1E, blue dots, ~45% of species). Thus for approximately half of the bacterial species in this analysis there is a significant codon cost-efficiency trade-off whereby resource allocation to mRNA can be reduced by decreasing

biosynthetic cost or increasing translational efficiency but not both simultaneously. This is because the synonymous codons that are most translationally efficient in these species are in general those that consume the most resources for biosynthesis.

Selection acting to minimise biosynthetic cost and maximise translational efficiency of transcript sequences is independent of codon cost-efficiency trade-off

Given that the codon cost-efficiency trade-off varied between species, an analysis was conducted to determine whether this trade-off was associated with concomitant differences in the strength of selection acting on the cost and efficiency of transcript sequences. For each species, genome-wide values for selection on transcript translational efficiency [S_t] and selection on transcript biosynthetic cost [S_c] were inferred using the complete set of open reading frames and tRNAs encoded in that species' genome using the SK model (2) implemented using CodonMuSe (see Methods). There was no significant difference in the mean or range of values for S_c between the groups of species that did or did not exhibit a significant codon cost-efficiency trade-off. Instead, irrespective of codon cost-efficiency trade-off 91% of species had negative S_c values (mean S_c = -0.08), indicating a genome-wide selective pressure to minimise the biosynthetic cost of transcript sequences through synonymous codon use (Fig. 2). This observation is extends previous studies that revealed analogous effects when nitrogen or energy were limited (2, 3). Thus irrespective of codon cost-efficiency trade-off, selection acting on codon biosynthetic cost is an important factor promoting biased patterns of codon use in bacteria.

Similarly, 78% of species had positive values for S_t (mean S_t = 0.1), indicating a genome wide selective pressure to increase the translational efficiency of transcript sequences (Fig. 2). However, species that had a significant codon cost-efficiency trade-off (Fig. 2A) had S_t values that were significantly lower than those that had no trade-off (Fig. 2B, $p < 0.05$, t-test). Moreover, there was an increased variance in the observed values for both S_c and S_t for species that exhibited a significant codon cost-efficiency trade-off compared to those

153 species that did not (Fig. 2A and 2B). Thus, in general the majority of species experience
154 selection to minimise transcript biosynthetic cost while simultaneously maximising transcript
155 translational efficiency. However, the nature of a species' codon cost-efficiency trade-off
156 restricts a transcript's ability to be both cheap and translationally efficient thereby limiting the
157 extent to which resource allocation to transcript sequences can be minimised.

158 **Genes that experience the strongest selection for increased transcript translational**
159 **efficiency are also under the strongest selection to minimise biosynthetic cost**
160 Given that the majority of species exhibited selection to minimise cost and maximise
161 translational efficiency at the genome-wide level, the extent to which this was also seen at
162 the level of an individual gene within species was determined. Here, the strength of selection
163 acting on transcript translational efficiency and strength of selection on transcript
164 biosynthetic cost were inferred for each individual gene in each species. The relationship
165 between these selection coefficients was then compared for each species. For example, in
166 species with no codon cost-efficiency trade-off, such as *Escherichia coli,* there is a
167 significant negative correlation between S_c and S_t (Fig. 3A). Here, the genes that
168 experienced the greatest selection to maximise efficiency are those that experienced the
169 greatest selection to minimise biosynthetic cost. The same phenomenon was also observed
170 for species that exhibit significant codon cost-efficiency trade-offs, such as *Lactobacillus*
171 *amylophilus* (Fig. 3B). Overall, significant correlations between selection coefficients for
172 individual genes were observed for 91% of species ($p < 0.05$, Fig. 3C). Therefore
173 irrespective of the extent of the codon cost-efficiency trade-off, selection is performing multi-
174 objective optimisation of transcript sequences to reduce their biosynthetic cost while
175 increasing their translational efficiency and thereby reducing resource allocation to mRNA
176 production.

177 As the most highly expressed genes in a cell comprise the largest proportion of cellular RNA,
178 the strength of selection experienced by a gene is thought to be dependent on the mRNA

abundance of that gene (18, 19). In agreement with this, evaluation of the relative mRNA abundance of genes in *E. coli* revealed that the most highly expressed genes exhibited the greatest selection to minimise transcript biosynthetic cost (Fig. 4A) whilst also showing the strongest selection to maximise transcript translational efficiency (Fig. 4B). Thus, selection acts in proportion to relative mRNA abundance to perform multi-objective optimisation of codon bias to minimise resource allocation to transcript sequences through production of low cost, high efficiency transcripts.

Sequence optimisation for cost and efficiency constrains molecular evolution rate

Given that codon choice has been shown to provide a selective advantage per codon per generation (20), it was hypothesised that the extent to which a transcript is jointly optimised for codon cost and efficiency would constrain the rate at which the underlying gene sequence can evolve. Specifically, the more highly optimised a transcript is for both biosynthetic cost and translational efficiency, the higher the proportion of spontaneous mutations that would reduce the cost-efficiency optimality of the transcript sequence. Therefore, spontaneous mutations in highly optimised genes would be more likely to be disadvantageous than spontaneous mutations in less optimised genes. As deleterious mutations are lost more rapidly from the population than neutral or advantageous mutations, the more highly optimised a gene sequence is, the lower its apparent evolutionary rate should be.

To test this hypothesis the complete set of gene sequences from *E.coli* was subject to stochastic *in silico* mutagenesis and the proportion of single nucleotide mutations that resulted in reduced transcript cost-efficiency optimality was evaluated. As expected, the proportion of deleterious mutations increased linearly with transcript sequence optimality. This effect was seen for both synonymous (Figure 5A) and non-synonymous mutations (Figure 5B). The effect in non-synonymous mutations is seen because a single base mutation from an optimal codon encoding one amino acid is unlikely to arrive at an equally

205 optimal or better codon encoding any other amino acid. Thus as expected, the more optimal
206 a codon is, the less likely a spontaneous mutation will result in a codon with higher optimality
207 irrespective of whether that codon encodes the same amino acid.

208 The extent to which transcript sequences in *E. coli* were jointly cost-efficiency optimised was
209 compared to the synonymous (K_s) and non-synonymous (K_a) mutation rate of that gene
210 estimated from comparison with *Salmonella enterica*. Consistent with the hypothesis, the
211 rate of both synonymous (K_s Figure 5C) and non-synonymous (K_a Figure 5D) changes were
212 directly proportional to the extent to which the gene sequence had been optimised by natural
213 selection for low biosynthetic cost and high translational efficiency (Figure 5A and B). Thus,
214 the extent to which transcript sequences are jointly optimised for cost and efficiency is
215 sufficient to explain variation in molecular evolutionary rate between genes within a species.
216 Moreover, selection-driven cost-efficiency optimality is sufficient to explain the correlation in
217 rates of synonymous and non-synonymous mutations.

218 ***Discussion***

219 Codon use is biased across the tree-of-life, with patterns of bias varying both between
220 species and between genes within the same species. Here we demonstrate that variation in
221 tRNA content between species creates a corresponding variation in the codon cost-
222 efficiency trade-off whereby codons that cost the least to biosynthesise are not equally
223 translationally efficient in all species. We show that irrespective of the codon cost-efficiency
224 trade-off, natural selection performs multi-objective gene sequence optimisation so that
225 transcript sequences are optimised to be both low cost and high translational efficiency, and
226 that the nature of this trade-off constrains the extent of the solution. We demonstrate that
227 this multi-objective optimisation is dependent on mRNA abundance, such that the transcripts
228 that comprise the largest proportion of cellular mRNA are those that experience the
229 strongest selection to be both low cost and high efficiency. Finally, we show that the extent
230 to which a gene sequence is jointly optimised for reduced transcript cost and enhanced

translational efficiency is sufficient to explain the variation in molecular sequence rate of genes. Furthermore, it is sufficient to explain the previously unexplained phenomenon that the rate of synonymous and non-synonymous mutation for a gene is correlated (21).

Differences in molecular evolution rates between species is thought to be mainly due to differences in organism generation-time (22). However, differences in evolutionary rates between genes in the same species lack a complete mechanistic explanation. Prior to the study presented here it was known that functional constraints of the encoded protein sequence contribute to the constraint of the rate of non-synonymous changes (23). It had also been observed that mRNA abundance and patterns of codon bias correlated with the evolutionary rate of genes (24, 25), and that rates of synonymous and non-synonymous changes were correlated (21). The study presented here unifies these prior observations and provides a novel mechanistic explanation for both variation and correlation in molecular evolution rates of genes. i.e. that stochastic mutations in gene sequences are more likely to result in deleterious alleles in proportion to the extent to which that gene sequence has been jointly optimised by natural selection for reduced transcript biosynthetic cost and enhanced translational efficiency.

The novel mechanism provided here also explains the relationship between mRNA abundance and gene evolutionary rate. Specifically, functional constraints on protein abundance stipulate the quantity of mRNA required to produce that protein. The more mRNA that is required, the larger the percentage of total cellular resources that are invested within that transcript. The mechanism simply entails that the more transcript that is present, the stronger the selective pressure will be to reduce the cellular resources committed to that transcript. Importantly, minimising these resources can be achieved both by using codons that require fewer resources for their biosynthesis, and by utilising translationally efficient codons that increase the protein to transcript ratio and therefore reduce the amount of transcript required to produce the same amount of protein. Overall, this study reveals how

the economics of gene production is a critical factor determining both the evolution and composition of genes.

Methods

Data sources

1,320 bacterial genomes were obtained from the NCBI (www.ncbi.nlm.nih.gov). In order to avoid over-sampling of more frequently sequenced genera, the number of species from each genus was restricted to 5 with a maximum of 1 species for each genus. Therefore, the 1320 species sampled in this study were distributed among 730 different genera. Only genes that were longer than 30 nucleotides, had no in-frame stop codons, and began and ended with start and stop codons respectively were analysed. Each species in this analysis contained a minimum of 500 genes that fit these criteria. Full details of species names, genome accession numbers, strain details and selection coefficients are provided in Supplementary Table 1.

Evaluation of translational efficiency (tAI)

To obtain the number of tRNA genes in each genome, tRNAscan was run on each of the 1,320 bacterial genomes (26). This (current) version of tRNAscan is unable to distinguish between tRNA-Met and tRNA-Ile with the anticodon CAT. Thus tRNA-Ile(CAT), while present, is not detected in any of the genomes. To compensate for this a single copy of tRNA-Ile with the anticodon CAT was added to the tRNA counts for each species if more than one tRNA-Met(CAT) was found. The tRNA adaptation index (tAI) (17), which considers both the tRNA gene copy number and wobble-base pairing when calculating the translational efficiency of a codon, was evaluated using the optimised s_{ij} values for bacteria obtained by Tuller et al (27) and the equation developed by dos Reis et al (28). s_{uu} was set to 0.7 as proposed by Navon et al (29) and s_{uc} was set to 0.95 as U_{34} has been shown to have weak codon-anticodon coupling with cytosine (30). Each species in this analysis was

able to translate all codons, was not missing key tRNAs and did not require unusual tRNA-modifications.

Calculation of relative codon cost and efficiency

Codon biosynthetic cost and translational efficiency were calculated relative to other synonymous codons such that the synonymous codon with the greatest value had a relative cost or efficiency of 1. eg. The nitrogen cost of GCC is 11 atoms. The most expensive synonymous codon is GCG/GCA (13 atoms). Therefore the relative cost of GCC is 11/13 = 0.85. The same evaluation was done to calculate codon translational efficiency.

CodonMuSe: A fast and efficient algorithm for evaluating drivers of codon usage bias

The SK model (2) was used to infer the joint contribution of mutation bias, selection acting on codon biosynthetic cost and selection acting on codon translational efficiency to biased synonymous codon use. To facilitate large scale comparative application of this model a rapid, stand-alone version was implemented in python. The algorithm, instructions for use, and example files are available for download at https://github.com/easeward/CodonMuSe. For each species, the values of M_b, S_c and S_t were inferred using the complete set of protein coding genes and tRNA copy number inferred using tRNAscan. Further details about the algorithm can be found in Supplemental File 1.

Comparing selection acting on codon bias and transcript abundance levels

Transcriptome data for *E. coli* str. K-12 MG1655 were downloaded from NCBI (31). The three biological replicates for the logarithmic growth phase were available, however the third replicate was inconsistent with the first two and so was excluded from this analysis. As each gene had multiple probes, the average probe value for each gene was taken. The three-parameter CodonMuSe model using the value for M_b estimated from a genome-wide analysis was run for each of the 4099 genes in *E. coli* individually, and thus values for S_c

306 and S_t were obtained for each gene. The values for these selection coefficients were plotted
307 against relative mRNA abundance data described above (31).

308 **Calculating the extent to which gene sequences were jointly optimised for cost and**
309 **efficiency**

310 The extent to which transcript sequences were jointly optimised for both biosynthetic cost
311 and translational efficiency was approximated by the distance of a given gene to the cost-
312 efficiency Pareto frontier for that amino acid sequence (d1, Supplementary Figure S1). The
313 cost-efficiency Pareto frontier consists of sequences that are optimised for maximal
314 transcript translational efficiency and minimal transcript biosynthetic cost such that a gene
315 that is 100% optimised lies on the frontier (red frontier, Supplementary Figure S1). Genes
316 that are less than 100% optimised occupy the space between the cost-efficiency Pareto
317 frontier (red frontier) and the opposite frontier (blue frontier, minimising transcript efficiency
318 and maximising cost) for that amino acid sequence (Supplementary Figure S1). The percent
319 optimality of the coding sequence is evaluated as $\left(\frac{d4}{d1+d4}\right) * 100$ (Supplementary Figure S1).

320 **Calculation of molecular evolution rates**

321 Molecular evolutionary rates (K_a and K_s values) were calculated for orthologous genes in *E.*
322 *coli* and *S. enterica*. 2,468 single-copy orthologous genes were identified for *E. coli* and *S.*
323 *enterica* using OrthoFinder v1.1.4 (32). These sequences were aligned at the amino acid-
324 level using MergeAlign (33) and this alignment was then re-threaded with the coding
325 sequences to create codon-level nucleotide alignments. Only aligned sequences longer
326 than 30 nucleotides with less than 10% gaps were used. Gapped regions were removed
327 and KaKs_Calculator 2.0 (34) was run using the GMYN model to evaluate K_a and K_s values
328 for each pair of aligned nucleotide sequences. As the molecular evolution rates represent
329 the average of the mutation rates of the gene-pair since they last shared a common

ancestor, these rates were compared to the average optimality of the same gene-pair in both species.

Acknowledgements

EAS is supported by a BBSRC studentship through BB/J014427/1. SK is a Royal Society University Research Fellow. Work in SK's lab is supported by the European Union's Horizon 2020 research and innovation programme under grant agreement number 637765.

Competing interests

The authors declare that they have no competing interests.

Author contributions

SK and EAS conceived the study, EAS conducted the analysis, EAS and SK wrote the manuscript. Both authors read and approved the final manuscript.

References

1. FARMER IS, JONES CW (1976) The Energetics of Escherichia coli during Aerobic Growth in Continuous Culture. *Eur J Biochem* 67(1):115–122.

2. Seward EA, Kelly S (2016) Dietary nitrogen alters codon bias and genome composition in parasitic microorganisms. *Genome Biol* 17(1):226.

3. Chen W-H, Guanting L, Peer B, Songnian H, Martin J. L (2016) Energy efficiency trade-offs drive nucleotide usage in transcribed regions. *Nat comminications* 7:1–10.

4. Horn D (2008) Codon usage suggests that translational selection has a major impact on protein expression in trypanosomatids. *BMC Genomics* 9:2.

5. Rocha EPC (2004) Codon usage bias from tRNA's point of view: Redundancy, specialization, and efficient decoding for translation optimization. *Genome Res* 14:2279–2286.

6. Sørensen M a, Kurland CG, Pedersen S (1989) Codon usage determines translation rate in Escherichia coli. *J Mol Biol* 207(2):365–377.

7. Hu H, et al. (2013) Codon Optimization Significantly Improves the Expression Level of a Keratinase Gene in Pichia pastoris. *PLoS One* 8(3). doi:10.1371/journal.pone.0058393.

8. Akashi H (1994) Synonymous codon usage in Drosophila melanogaster: Natural selection and translational accuracy. *Genetics* 136(3):927–935.

9. Shah P, Gilchrist M a (2011) Explaining complex codon usage patterns with selection

for translational efficiency, mutation bias, and genetic drift. *Proc Natl Acad Sci U S A* 108(25):10231–10236.

10. Precup J, Parker J (1987) Missense misreading of asparagine codons as a function of codon identity and context. *J Biol Chem* 262(23):11351–11355.

11. Lao PJ, Forsdyke DR (2000) Thermophilic bacteria strictly obey Szybalski's transcription direction rule and politely purine-load RNAs with both adenine and guanine. *Genome Res* 10(2):228–236.

12. Paz A, Mester D, Baca I, Nevo E, Korol A (2004) Adaptive role of increased frequency of polypurine tracts in mRNA sequences of thermophilic prokaryotes. *Proc Natl Acad Sci U S A* 101(9):2951–2956.

13. Eskesen ST, Eskesen FN, Ruvinsky A (2004) Natural selection affects frequencies of AG and GT dinucleotides at the 5′ and 3′ ends of exons. *Genetics* 167(1):543–550.

14. Novoa EM, Ribas de Pouplana L (2012) Speeding with control: Codon usage, tRNAs, and ribosomes. *Trends Genet* 28(11):574–581.

15. Zhang F, Saha S, Shabalina SA, Kashina A (2010) Differential Arginylation of Actin Isoforms Is Regulated by Coding Sequence-Dependent Degradation. *Science (80-)* 329(5998):1534–1537.

16. Grosjean H, de Crécy-Lagard V, Marck C (2010) Deciphering synonymous codons in the three domains of life: Co-evolution with specific tRNA modification enzymes. *FEBS Lett* 584(2):252–264.

17. dos Reis M, Wernisch L, Savva R (2003) Unexpected correlations between gene expression and codon usage bias from microarray data for the whole Escherichia coli K-12 genome. *Nucleic Acids Res* 31(23):6976–6985.

18. Ran W, Higgs PG (2012) Contributions of Speed and Accuracy to Translational Selection in Bacteria. *PLoS One* 7(12). doi:10.1371/journal.pone.0051652.

19. Drummond DA, Wilke CO (2009) The evolutionary consequences of erroneous protein synthesis. *Nat Rev Genet* 10(10):715–724.

20. Brandis G, Hughes D (2016) The Selective Advantage of Synonymous Codon Usage Bias in Salmonella. *PLOS Genet* 12(3):e1005926.

21. Sharp PM (1991) Determinants of DNA sequence divergence between Escherichia coli and Salmonella typhimurium: Codon usage, map position, and concerted evolution. *J Mol Evol* 33(1):23–33.

22. Weller C, Wu M (2015) A generation-time effect on the rate of molecular evolution in bacteria. *Evolution (N Y)* 69(3):643–652.

23. Zuckerkandl E (1976) Evolutionary processes and evolutionary noise at the molecular level. I. Functional Density in Proteins. *J Mol Evol* 7:167–183.

24. Sharp PM, Li WH (1987) The rate of synonymous substitution in enterobacterial genes is inversely related to codon usage bias. *Mol Biol Evol* 4(3):222–230.

399 25. Drummond DA, Raval A, Wilke CO (2006) A single determinant dominates the rate of
400 yeast protein evolution. *Mol Biol Evol* 23(2):327–337.

401 26. Schattner P, Brooks AN, Lowe TM (2005) The tRNAscan-SE, snoscan and snoGPS
402 web servers for the detection of tRNAs and snoRNAs. *Nucleic Acids Res* 33(SUPPL.
403 2):686–689.

404 27. Sabi R, Tuller T (2014) Modelling the Efficiency of Codon – tRNA Interactions Based
405 on Codon Usage Bias. *DNA Res* 21(June):511–525.

406 28. dos Reis M, Savva R, Wernisch L (2004) Solving the riddle of codon usage
407 preferences: A test for translational selection. *Nucleic Acids Res* 32(17):5036–5044.

408 29. Navon S, Pilpel Y (2011) The role of codon selection in regulation of translation
409 efficiency deduced from synthetic libraries. *Genome Biol* 12(2):R12.

410 30. Näsvall SJ, Chen P, Björk GR (2004) The modified wobble nucleoside uridine-5-
411 oxyacetic acid in tRNA Pro cmo 5 UGG promotes reading of all four proline codons in
412 vivo The modified wobble nucleoside uridine-5-oxyacetic acid in tRNA Pro cmo UGG
413 promotes reading of all four proline codons in vi. 1662–1673.

414 31. Cho B-K, et al. (2009) Elucidation of the transcription unit architecture of the
415 Escherichia coli K-12 MG1655 genome. *Nat Biotechnol* 27(11):1043–1049.

416 32. Emms DM, Kelly S (2015) OrthoFinder: solving fundamental biases in whole genome
417 comparisons dramatically improves orthogroup inference accuracy. *Genome Biol*
418 16(1):157.

419 33. Collingridge PW, Kelly S (2012) MergeAlign: improving multiple sequence alignment
420 performance by dynamic reconstruction of consensus multiple sequence alignments.
421 *BMC Bioinformatics* 13:117.

422 34. Wang D, Zhang Y, Zhang Z, Zhu J, Yu J (2010) KaKs_Calculator 2.0: A Toolkit
423 Incorporating Gamma-Series Methods and Sliding Window Strategies. *Genomics,*
424 *Proteomics Bioinforma* 8(1):77–80.

425

426

Figures

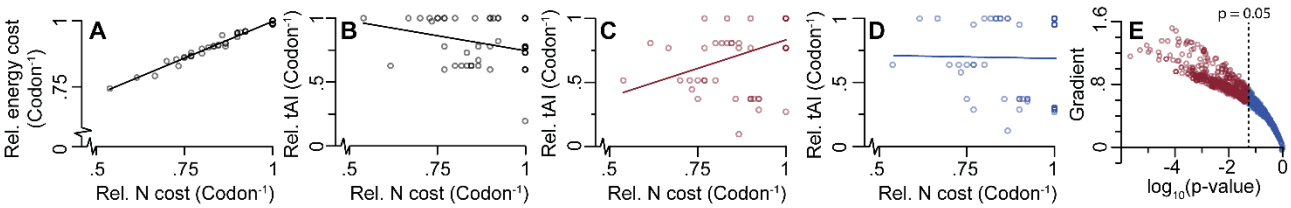


Fig. 1.

Different tRNA sparing strategies alter a species' codon cost-efficiency trade-off. **A)** Codon nitrogen cost (N cost) correlates almost perfectly with codon energetic cost ($p < 0.05$, $y = 0.6x + 0.44$, $R^2 = 0.98$). **B)** A full complement of tRNAs has a negative correlation between codon biosynthetic cost and translational efficiency (tAI) ($p< 0.05$, $y = -0.5x + 1.21$, $R^2 = 0.10$). **C)** tRNA sparing strategy 1 (NNU codons translated by GNN anticodons) has a positive correlation between codon biosynthetic cost and translational efficiency ($p<0.05$, $y = 0.9x – 0.06$, $R^2 = 0.18$). **D)** tRNA sparing strategy 2 (strategy 1 + NNG codons translated by UNN anticodons) has no significant correlation between codon biosynthetic cost and translational efficiency ($p > 0.05$, $y = 0.74$, $R^2 = 0$). **E)** The 1,320 bacterial species in this analysis can be categorised into those with a significant codon cost-efficiency trade-off ($p < 0.05$, red) and those with no trade-off ($p>0.05$, blue). The y-axis is the gradient of the line of best fit between codon biosynthetic cost and translational efficiency.

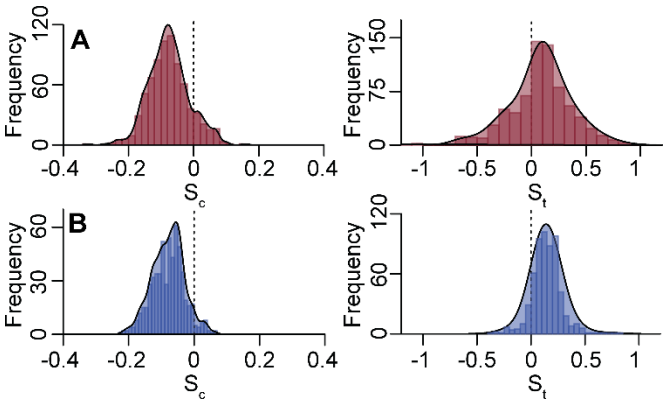


Fig. 2.

Bacterial genomes show selection to minimise nucleotide cost ($-S_c$) and maximise translational efficiency ($+S_t$). **A)** Genome-wide S_c and S_t values for the 726 species with a codon cost-efficiency trade-off. **B)** Genome-wide S_c and S_t values for the 594 species with no codon cost-efficiency trade-off.

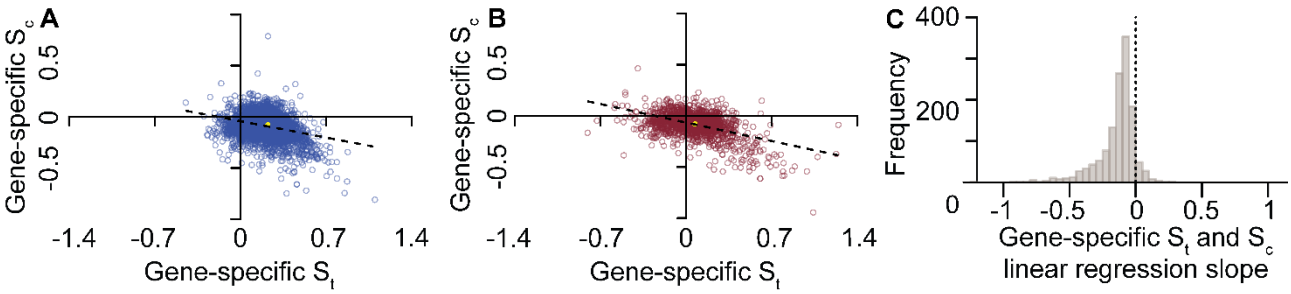


Fig. 3.

The genes under the strongest selection for translational efficiency (+S_t) are also under the strongest selection to minimise nucleotide cost (-S_c). Scatterplots of gene-specific S_t and S_c values for **A)** *Escherichia coli* **B)** *Lactobacillus amylophilus*. In both cases the line of best fit is shown and the yellow dot is the genome-wide best-fit value for each species. **C)** Histogram of the slope between S_c and S_t for individual genes for each of the 1,320 species in this analysis.

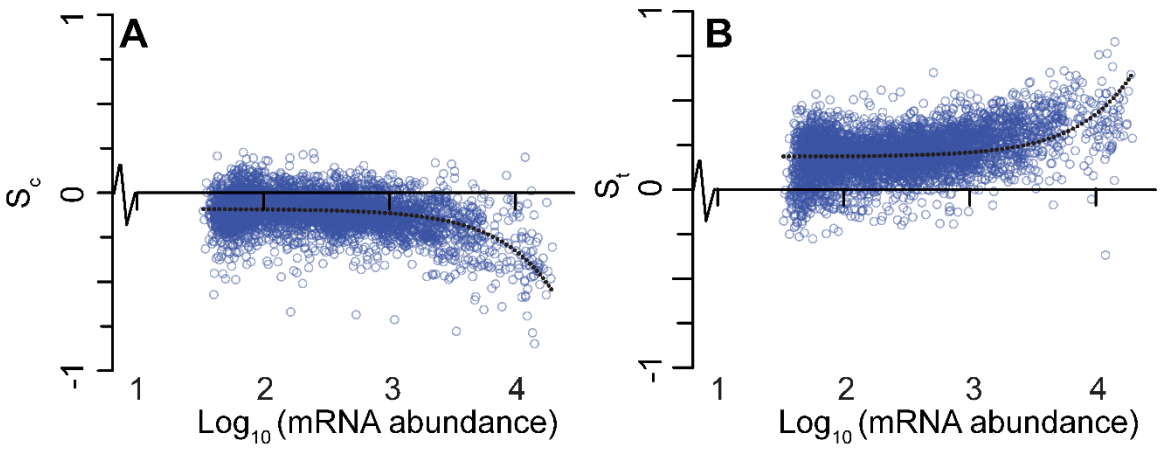
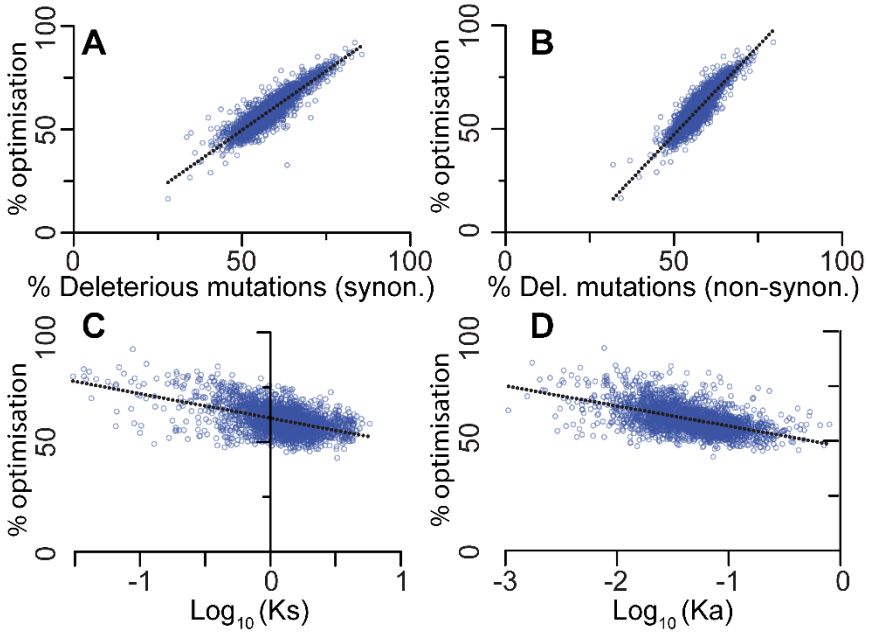


Fig. 4.

Selection acts in proportion to mRNA abundance to decrease codon biosynthetic cost and increase codon translational efficiency in *Escherichia coli*. **A)** There is a negative correlation between selection acting on codon biosynthetic cost (S_c) and mRNA abundance. The line of best fit has an R^2 value of 0.18. **B)** There is a positive correlation between selection acting to increase codon translational efficiency (S_t) and gene expression. The line of best fit has an R^2 value of 0.13.



468

469 **Fig. 5.**

470 Selection-driven optimisation of resource allocation is a critical factor that determines
471 molecular evolution rate. Highly cost-efficiency optimised genes have a higher proportion of
472 deleterious **A)** synonymous (y = 1.15x - 8, R^2 = 0.81) and **B)** non-synonymous (y = 1.71x -
473 38, R_2 = 0.78) mutations. Orthologous genes in *Escherichia coli* and *Salmonella enterica*
474 show a negative correlation between sequence cost-efficiency optimisation and the rate of
475 **C)** synonymous mutations (K_s) (y = -11x + 61, R^2 = 0.26) and **D)** non-synonymous mutation
476 (K_a) (y = -9x + 48, R^2 = 0.28).

477

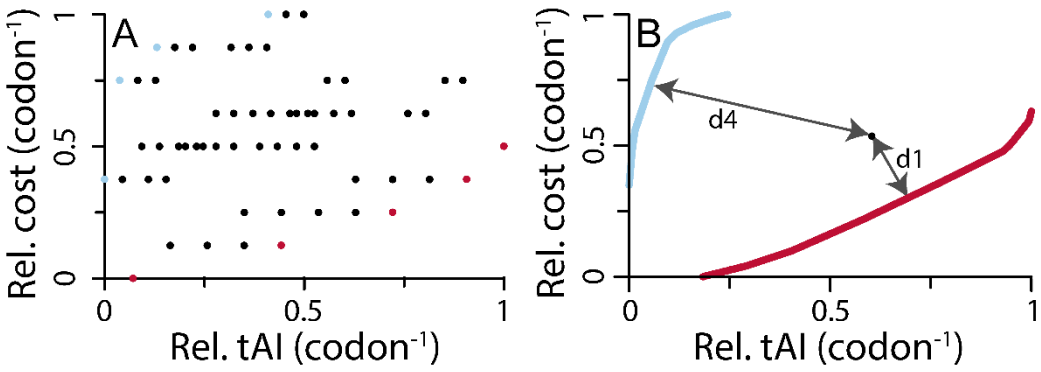


Figure S1

Example cost-efficiency Pareto frontier for a short amino acid sequence. **A)** Scatter plot of the 64 possible coding sequences encoding the amino acid sequence MTGCD. Red dots indicate coding sequences that are positioned on the best cost-efficiency Pareto frontier (the least expensive most translationally efficient sequences possible). Blue dots indicate coding sequences that are positioned on the worst cost-efficiency Pareto frontier (the most expensive least translationally efficient sequences possible). **B)** Evaluating the cost-efficiency optimality of a coding sequence. d1 is the minimum distance between a given coding sequence and the best cost-efficiency Pareto frontier (red) for that amino acid sequence. d4 is the minimum distance of the same gene to the worse cost-efficiency Pareto frontier for that amino acid sequence (blue). The percent optimality of the coding sequence is evaluated as $\left(\frac{d4}{d1+d4}\right) * 100$.

Supplementary Table S1

Full details of species names, genome accession numbers, strain details and selection coefficients.