Efficient translation initiation dictates codon usage at gene start

Kajetan Bentele, Paul Saffert, Robert Rauscher,
Zoya Ignatova and Nils Blüthgen

Supplementary Material

1 Supplementary Figures

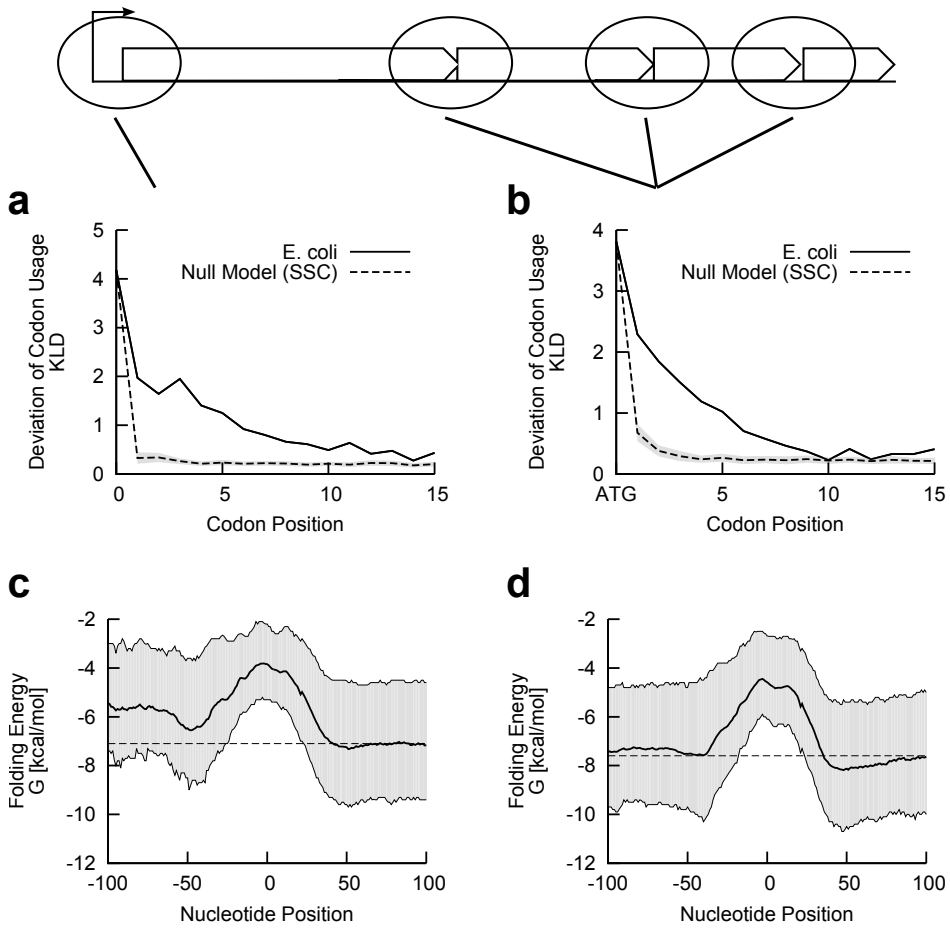

Figure S1: **Deviation of codon usage and suppression of mRNA structure at start of first genes of transcriptional units (TUs) and at start of genes within TUs.** (**a** and **b**) In *E. coli*, the frequency of synonymous codons after translation start in genes at the beginning of TUs (**a**) and in genes within TUs (**b**) deviates from the global codon usage in the genome, as quantified by the Kullback-Leibler divergence (KLD, solid line). (**c** and **d**) Folding energy of *E. coli* mRNA sequences shows a maximum at translation start site for genes at the beginning of TUs (**c**) and genes within TUs (**d**), indicating the suppression of mRNA secondary structure around the start codon. Average folding energy is shown as a solid line, surrounded by the inter-quartile range in gray. The dashed line is a guide to the eye.

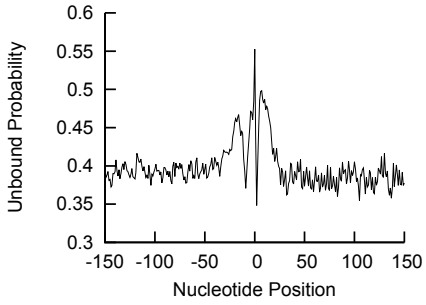

Figure S2: **The average probability to remain unpaired increases around the gene start.** This probability was calculated for each *E. coli* sequence spanning a range from -160 to +160 nucleotides relative to the gene start. These values were averaged among all genes.

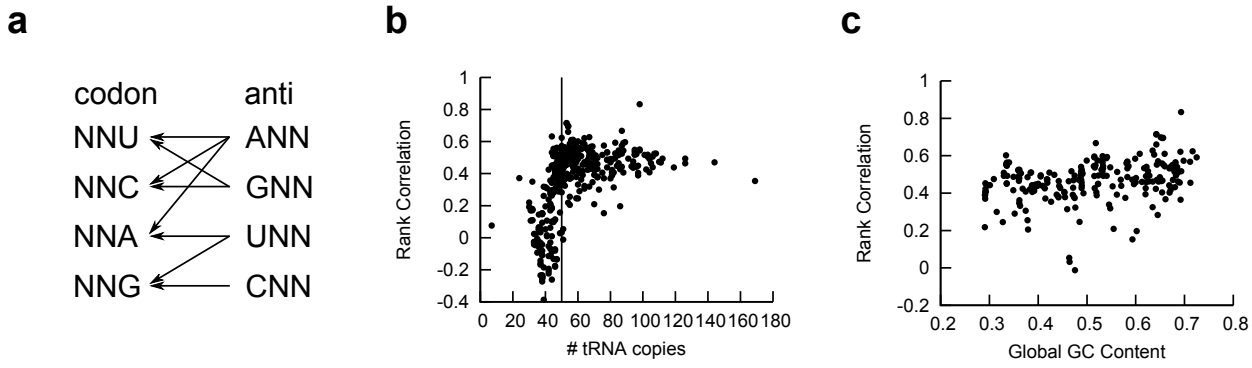

Figure S3: **Translation speed and codon frequency correlate** (**a**) For each genome, we determined the copy number of tRNA genes using tRNAscan-SE (version 1.23) (Lowe and Eddy, 1997). We then calculated the relative adaptivness value (tAI score) for each codon as a measure of elongation speed using the codon/anti-codon scheme as shown and constraints on base parings (dos Reis *et al*, 2004). The scheme is based on Crick's wobble rules (Crick, 1966; dos Reis *et al*, 2004) and the assumption that adenine at the first anticodon position is always converted to inosine, except for the anticodons AAA, ATA, ACA, ATG, ATT, ACT, ATC from which only ATG was observed in one genome. In addition ATA can be read by three different tRNAs (dos Reis *et al*, 2004). Stop codons were disregarded, including TGA codons for selenocysteine. (**b**) The Spearman's rank correlation between global codon frequency and tAI scores as a function of the total number of tRNA genes shows that in genomes with low number of detected tRNA genes the rank correlation is strongly variable, thus we applied a cutoff disregarding genomes with less than 50 tRNA genes. This leaves us with a total number of 243 genomes. (**c**) For genomes with at least a total of 50 tRNA gene copies Spearman's rank correlations between global codon frequency and tAI scores scatter around 0.47 and weakly depends on the global GC-content.

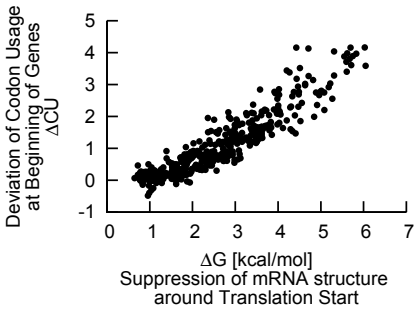

Figure S4: **Correlation of unusual codon usage ΔCU and suppression of mRNA secondary structure around gene start, ΔG, as in Figure 1 A but for non-overlapping genes.** The correlation remains despite this restriction (correlation coefficient $r = 0.92$).

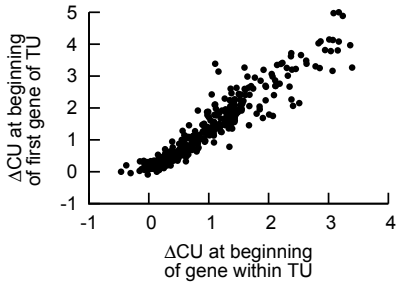

Figure S5: **Codon usage fore the first genes and genes within transcriptional units.** Average deviation of codon usage (ΔCU) of the first 5 codons of genes at the beginning of TUs correlates strongly with average deviation of codon usage ΔCU of genes within TUs, suggesting similar evolutionary pressures.

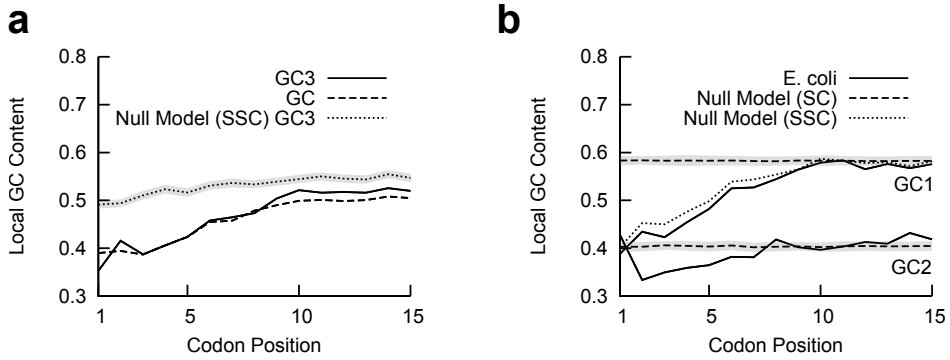

Figure S6: **GC-content at the beginning of first genes in TUs of *E. coli*.** (**a**) GC-content (dashed line) and GC3-content (solid line) of codons decrease at the beginning of first genes in TUs of *E. coli*. Dotted line and gray area shows mean GC3-content ± standard deviation, estimated from the null model (SSC). (**b**) GC1 and GC2-content decrease at gene start in *E. coli* (solid lines) when compared to a null model with shuffled codons (SC). This is primarily due to the choice of amino-acids as GC2-content is fully determined by the amino-acid, and the null model with shuffled synonymous codons (SSC, dotted line) shows only a small deviation for the GC1-content.

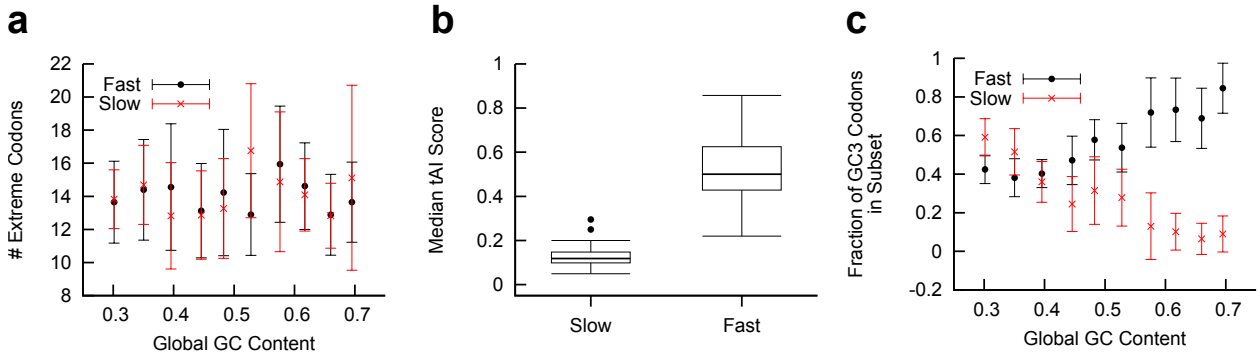

Figure S7: **GC3 composition of slow and fast codons reflects GC-content of the genome** (**a**) We defined sets of slow (small tAI scores) and fast (large tAI scores) codons as the top and bottom 10 codons. As tAI scores are highly degenerated, i.e. many codons exhibit the same value, we expanded the subset such that all other codons with the same tAI score are also included. This results on average in about 12-16 codons in each set. The number of codons within these sets does not correlate with GC-content. (**b**) The sets of fast and slow codons show distinct median tAI score across the analyzed genomes. (**c**) The fraction of GC3 codons in the subset of slow and fast codons is shown as a function of genomic GC-content. We grouped genomes by their global GC-content and plotted mean and standard deviation of the fraction of GC3 codons against the mean global GC-content in each class. With increasing GC-content the number of GC3 codons decreases in the subset of slow, and correspondingly the GC3 increases in the subset of fast codons.

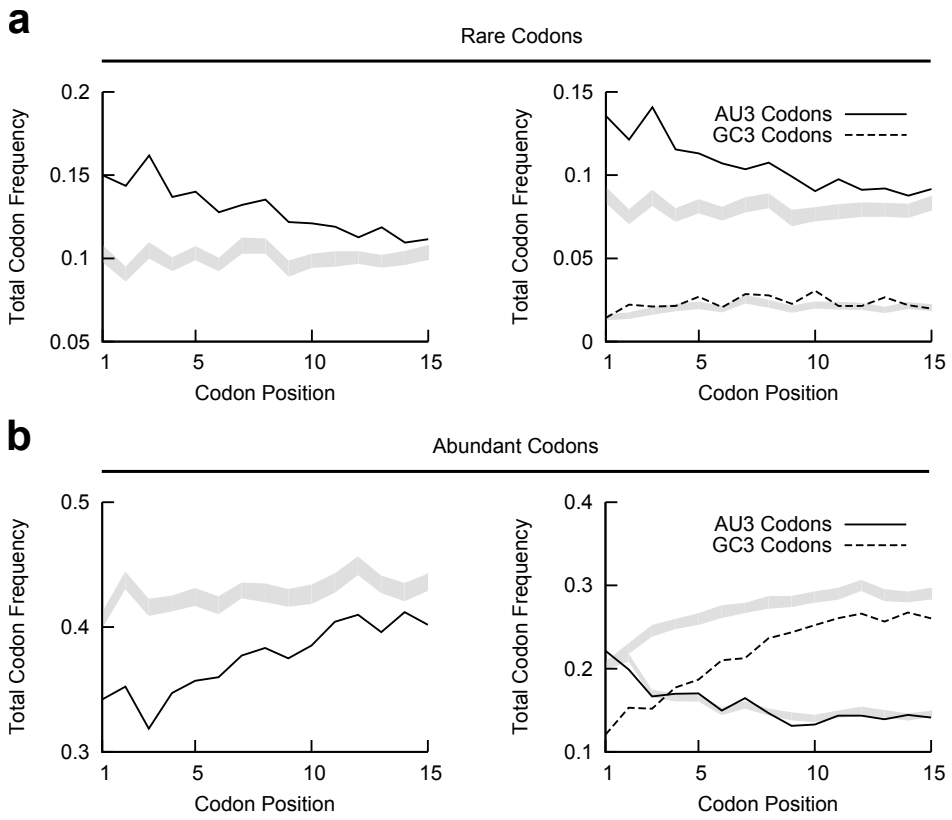

Figure S8: **Enrichment of extreme codons at beginning of first genes in TUs of *E. coli*.** (**a**) Rare codons are enriched at the beginning of first genes in TUs of *E. coli*. Solid line in the left panel shows the total frequency of rare codons. From these only the frequency of AU3 codons exhibit an increase (solid line in right panel) whereas the usage of GC3 codons does not change (dashed line in right panel). In both panels, grayed areas depict the corresponding average total frequency ± standard deviation estimated from the null model SSC. (**b**) Same plot as in (**a**) but for abundant codons. Frequency of abundant GC3 codons is strongly reduced.

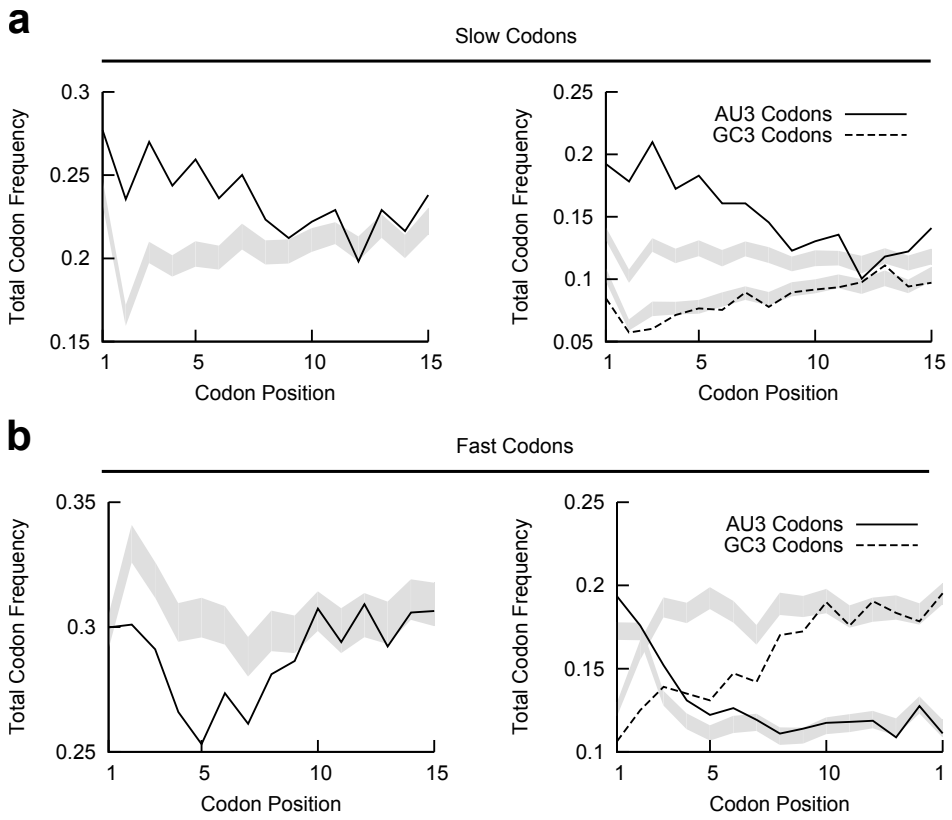

Figure S9: **Enrichment of slow and fast codons at beginning of genes within TUs of *E. coli*.** (**a**) The total frequency of slow codons (small tAI score, solid line in left panel) increases at the beginning of genes in *E. coli*. But as for rare codons, only AU3 codons from this subsets (solid line in right panel) are enriched, whereas frequency of slow GC3 codons (dashed line in right panel) is slightly decreased. Grayed areas show the corresponding average total frequency ± standard deviation estimated from the null model SSC. (**b**) Same plot as in (**a**) but for fast codons (large tAI score). Frequency of fast GC3 codons is strongly decreased, whereas fast AU3 codons are enriched.

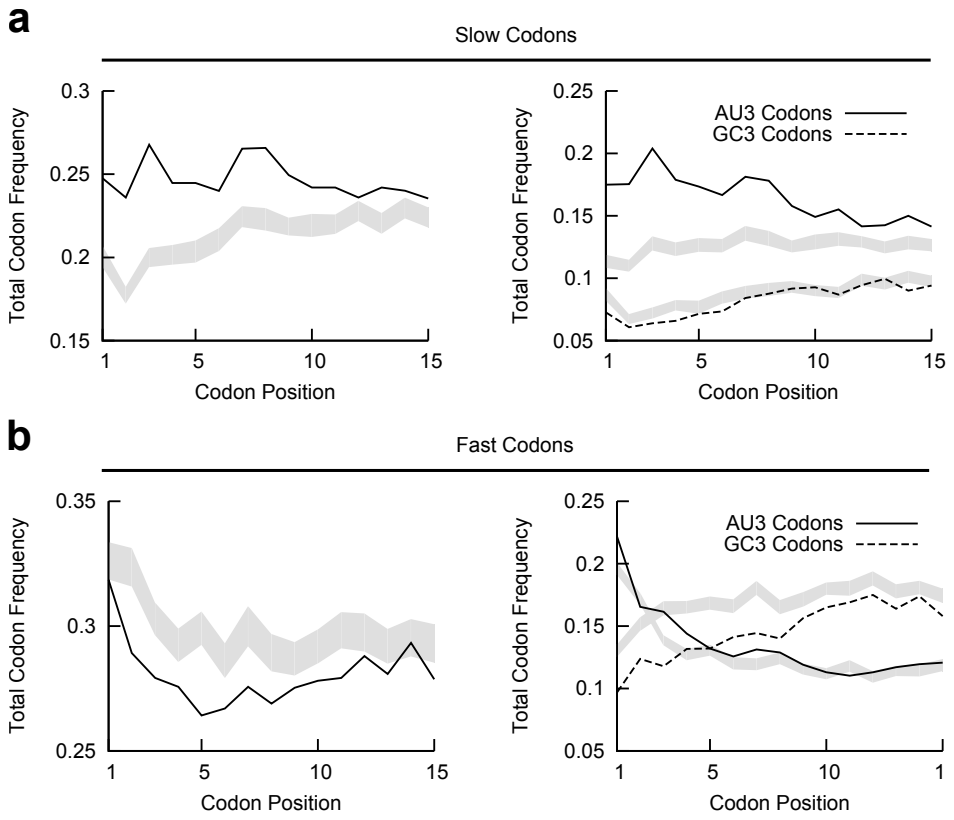

Figure S10: **Enrichment of slow and fast codons at beginning of first genes in TUs of *E. coli*.** Same plot as in Fig. S9 but for first genes in TUs. The same asymmetry is observed: AU3 codons are enriched, whereas GC3 codons are depleted irrespective of their elongation speed.

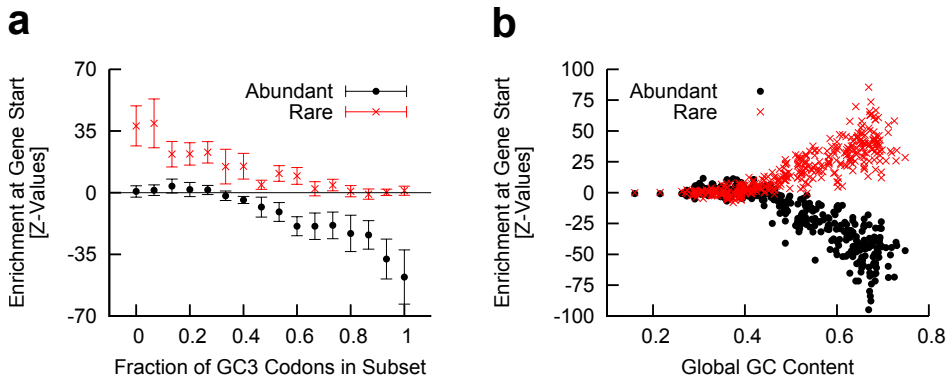

Figure S11: **Enrichment of extreme codons and deviation of GC3-content at start of first genes in TUs of bacterial genomes.** Enrichment of codons was assessed by calculating Z-values of fold change for codon frequency at the beginning of genes for rare (red crosses) and abundant (black dots) codons compared to the null model (SSC). (**a**) Genomes were grouped according to the fraction of GC3 codons in the subset of rare and abundant codons, and mean enrichment $\pm$ standard deviation is shown for these groups. Genomes with GC3-rich abundant codons show a depletion of abundant codons, and genomes with AU3-rich rare codons show an enrichment of rare codons at gene start. (**b**) Enrichment of extreme codons shown as a function of GC-content. Rare codons are only enriched and abundant codons depleted in genomes with GC-content larger than about 0.5.

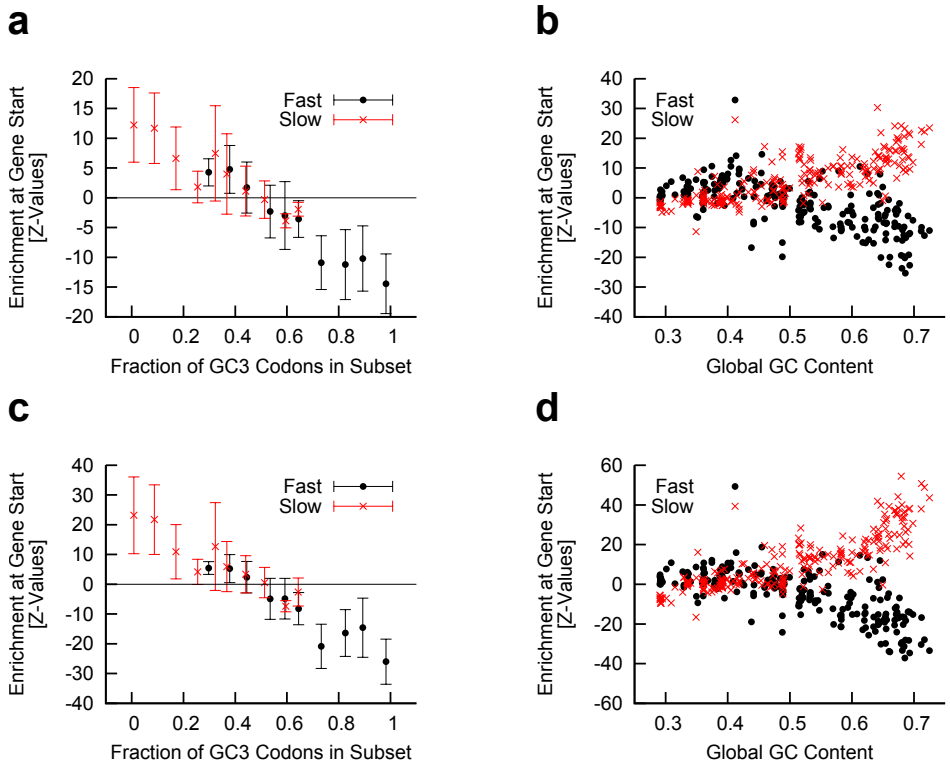

Figure S12: **Enrichment of slow and fast codons reflects GC-content of the genome.** (**a**) We grouped genomes according to the GC3-content of their fast and slow codons, and quantified enrichment of the of slow and fast codons at start of genes within TUs by calculating Z-values using the null model with synonymous shuffled codons (SSC). Symbols show the mean, and error bars show standard deviation. Slow codons are only enriched if the majority of slow codons are AU3 codons. (**b**) We observed a gradual increase of slow and a corresponding decrease of fast codons at the beginning of genes with increasing global GC-content. (**c**) and (**d**) Same plot as in (**a**) and (**b**) but for the first genes in TUs. Effects are even more pronounced.

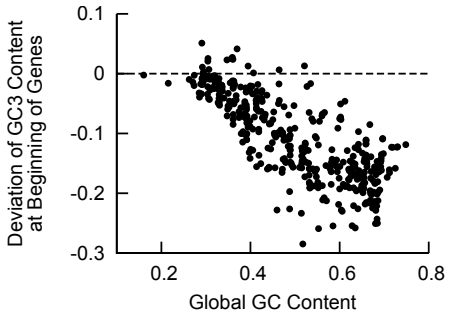

Figure S13: **Deviation of GC3-content at start of first genes in TUs.** The average deviation from the genomic GC3-content for codons 1 to 5 depends on the global GC-content. Virtually all genomes show a reduction in GC3-content at the gene start, and genomes with higher genomic GC-content typically show a stronger reduction (correlation coefficient $r = -0.79$).

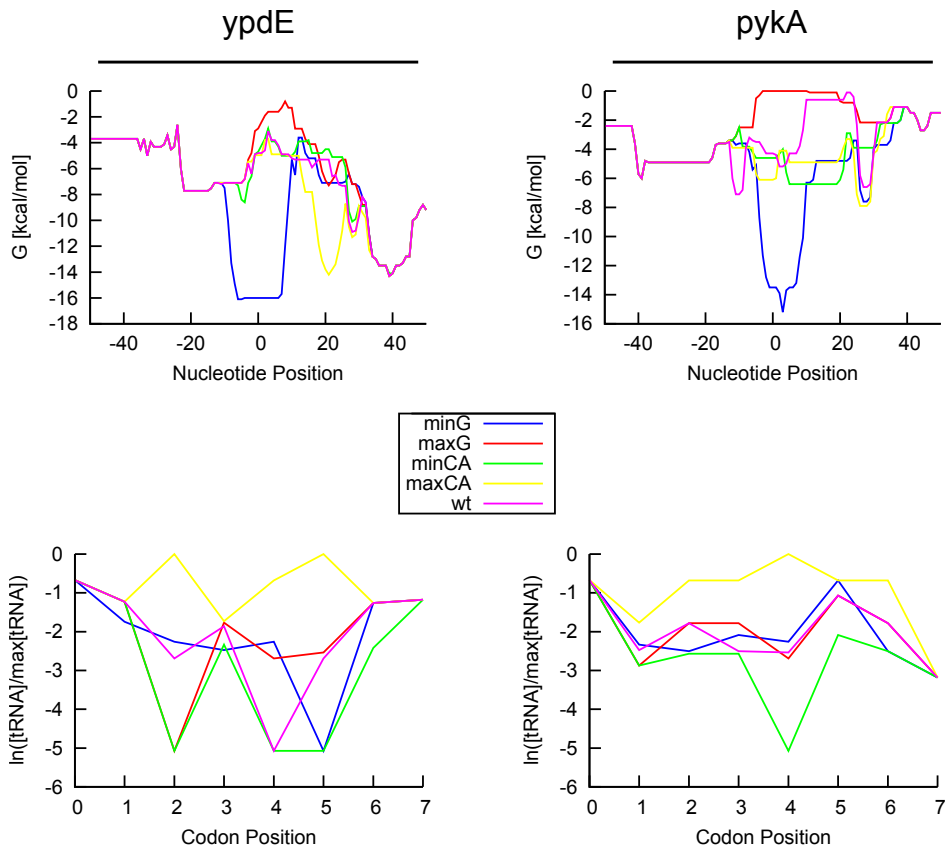

Figure S14: **Folding energy and codon usage profiles for different constructs.** The constructs *minG* and *maxG* exhibit different degree of structure at the gene start, but have on average the same codon usage as the *wt* sequence. In contrast, mRNA folding energies of *minCA* and *maxCA* do not differ much from the *wt* construct, but show opposite codon adaptation.

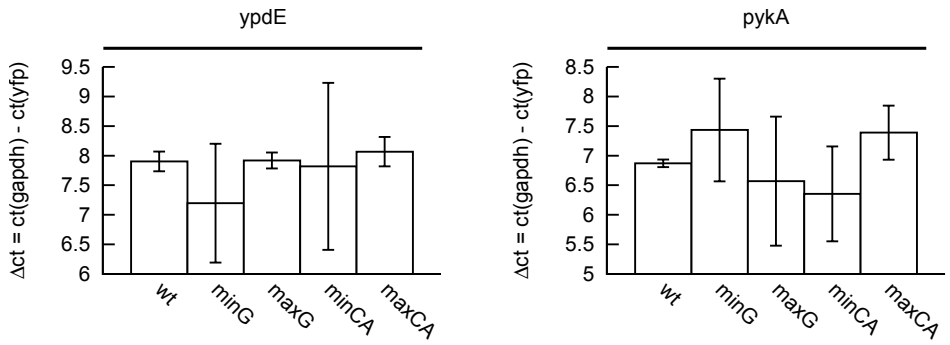

Figure S15: **Results of qRT-PCR measurements for different constructs after induction.** Within the experimental error there are no differences between constructs derived from the respective *E. coli* genes.

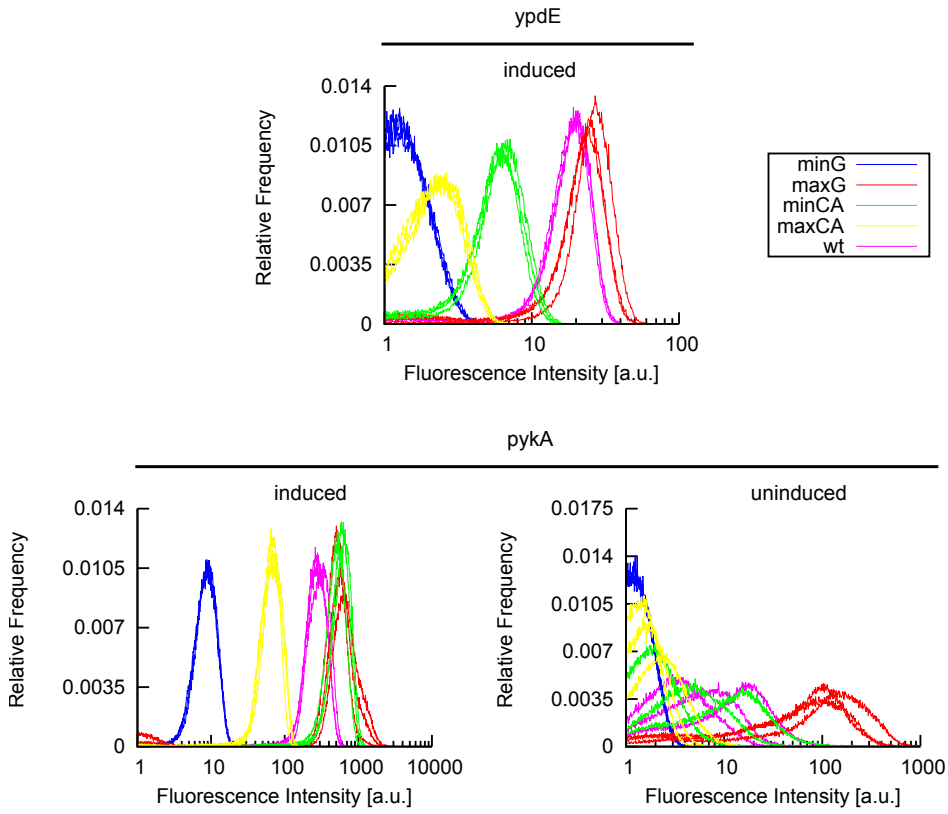

Figure S16: **Distribution of fluorescence levels of different constructs as measured by flow cytometry.** Values are background corrected.

2 Supplementary Tables

Constructs

Table S1: Sequences derived from *ypdE*. *wt*, *minG*, *maxG*, *minCA* and *maxCA* indicate the corresponding altered sequences.

ypdE	nucleotide sequence
5' UTR and start codon	GAAGTGCTCTACGCCAAGCCGAAAACAGTGTTGCTCACGG GAGAGGCATAATG
wt	GAT TTA TCG CTA TTA AAA
minG	GAC CTC TCC CTC CTA AAA
maxG	GAT CTA TCT TTA CTT AAA
minCA	GAT CTA AGT CTA CTA AAG
maxCA	GAT CTG AGC TTG CTG AAA
3' sequence and glycine linker	GCGTTGAGCGAGGCAGATGCGATCGCCTCCTCGGAACAGG AAGTGCGGCAGATCCTGCTGGAAGAAGCGGATGGCGGCGGA

Table S2: Sequences derived from *pykA*. *wt*, *minG*, *maxG*, *minCA* and *maxCA* indicate the corresponding altered sequences.

pykA	nucleotide sequence
5' UTR and start codon	GGATTTCAAGTTCAAGCAACACCTGGTTGTTTCAGTCAAC GGAGTATTACATG
wt	TCC AGA AGG CTT CGC AGA
minG	AGT AGG CGG CTC CGT AGG
maxG	TCA AGA AGA TTA CGC AGA
minCA	TCA CGA CGA CTA CGG AGG
maxCA	TCT CGT CGT CTG CGT CGT
3' sequence and glycine linker	ACAAAAATCGTTACCACGTTAGGCCCAGCAACAGATCGCG ATAATAATCTTGAAAAAGTTATCGCGGCGGGTGGCGGCGGA

References

Crick FH (1966) Codon–anticodon pairing: the wobble hypothesis. *Journal of molecular biology* **19**: 548–55

dos Reis M, Savva R, Wernisch L (2004) Solving the riddle of codon usage preferences: a test for translational selection. *Nucleic acids research* **32**: 5036–44

Lowe TM, Eddy SR (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic acids research* **25**: 955–64