Local genic base composition impacts protein production and cellular fitness

Erik M. Quandt, Charles C. Traverse and Howard Ochman

Department of Integrative Biology, University of Texas at Austin, Austin, TX, United States of America

ABSTRACT

The maintenance of a G + C content that is higher than the mutational input to a genome provides support for the view that selection serves to increase G + C contents in bacteria. Recent experimental evidence from *Escherichia coli* demonstrated that selection for increasing G + C content operates at the level of translation, but the precise mechanism by which this occurs is unknown. To determine the substrate of selection, we asked whether selection on G + C content acts across all sites within a gene or is confined to particular genic regions or nucleotide positions. We systematically altered the G + C contents of the GFP gene and assayed its effects on the fitness of strains harboring each variant. Fitness differences were attributable to the base compositional variation in the terminal portion of the gene, suggesting a connection to the folding of a specific protein feature. Variants containing sequence features that are thought to result in rapid translation, such as low G + C content and high levels of codon adaptation, displayed highly reduced growth rates. Taken together, our results show that purifying selection acting against A and T mutations most likely results from their tendency to increase the rate of translation, which can perturb the dynamics of protein folding.

INTRODUCTION

Bacterial species exhibit a wide range of genomic base compositions, ranging from 13% to 75% G + C (*Thomas et al., 2008*; *McCutcheon & Moran, 2010*). Because mutations in bacteria are universally biased towards A and T, the maintenance of a G + C content that is higher than the mutational input to a genome implies a role of natural selection in shaping genomic base composition (*Hershberg & Petrov, 2010*; *Hildebrand, Meyer & Eyre-Walker, 2010*). However, it has been difficult to establish the specific traits or circumstances for which higher G + C contents might be advantageous. Many hypotheses have been suggested to explain the differences in genomic G + C contents among organisms, most commonly citing an association and advantage of a particular base composition with an environmental variable (*Singer & Ames, 1970*; *Kagawa et al., 1984*; *McEwan, Gatherer & McEwan, 1998*; *Foerstner et al., 2005*; *Rocha & Feil, 2010*). However, these correlations are rarely robust across taxa (*Hurst & Merchant, 2001*) or are complicated by other factors (*Rocha & Feil, 2010*) leaving the actual basis of the variation unknown. Moreover, the selection coefficients necessary to favor individual nucleotide substitutions that alter genomic base composition seem unrealistically small, even given the large effective population sizes of bacteria.

Prior research on *E. coli* has demonstrated that strains expressing genes of lower G + C contents had slower doubling times, even when the sequences of the encoded proteins were identical (*Raghavan, Kelkar & Ochman, 2012*). Furthermore, this GC-effect was dependent on translation such that isogenic constructs lacking ribosome binding sites did not recapitulate the fitness defect (*Raghavan, Kelkar & Ochman, 2012*). These findings suggest that selection acts at the level of individual genes and not on genomic base composition as a whole, and that the translational process is somehow affected by changes in base composition.

Shifting the focus towards the translation of individual genes mitigates some of the concerns about the efficacy of selection needed to favor compositional changes at individual sites but raises questions about how cellular fitness could be linked to the G + C content of a gene. One potential explanation is that genes of different base composition differ in their patterns of ribosome turnover such that ribosomes become limiting during rapid growth (*Andersson & Kurland, 1990*; *Plotkin & Kudla, 2011*; *Weiße et al., 2015*; *Gorochowski et al., 2016*). Selection for translational efficiency is reflected in the sequences of highly expressed genes, which are biased towards using codons that can be translated rapidly due to high concentrations of their cognate tRNAs (*Ikemura, 1985*; *Andersson & Kurland, 1990*; *Dong, Nilsson & Kurland, 1996*). It is possible that differences in the speed at which A/T or G/C bases are translated, or the tendency to form sequence motifs that affect decoding, result in a similar form of selective pressure on base composition. Alternatively, differences in base composition might alter the rate of translation and affect the folding of nascent polypeptides. An inappropriate rate of translation could cause proteins to misfold and aggregate, and be detrimental to the cell.

To distinguish the specific substrate of G + C selection, we asked if selection for G + C content is acting over the entire gene or is confined to certain motifs or nucleotide positions. To accomplish this, we synthesized a large set of gene constructs comprising variants of mosaic base compositions and tested for any growth defects associated with each variant when expressed under a variety of promoters. We found that the fitness benefit incurred from increased G + C contents varied with genic location and was associated with a reduced rate of co-translational protein folding. Based on these findings, selection for higher G + C contents serves to counter the increased rate of translation caused by A and T mutations, which disrupt protein folding dynamics by decreasing the stability of mRNA secondary structures.

MATERIALS AND METHODS

Bacterial strains and growth conditions

Growth and fluorescence assays were performed with *E. coli* strain BW25113 (*F*-, Δ*(araD-araB)567*, Δ*lacZ4787*(::rrnB-3), λ⁻, *rph-1*, Δ*(rhaD-rhaB)568*, *hsdR514*) (*Datsenko & Wanner, 2000*; *Baba et al., 2006*). Cells were grown in Lysogeny Broth (LB) of the Lennox variety (5 g/L NaCl). When appropriate, antibiotics were supplemented at the following concentrations: ampicillin (Amp) (100 μg/mL), kanamycin (Kan) (30 μg/mL), streptomycin (Strep/Sm) (100 μg/mL).

Design of recoded GFP-L, M and H genes

The 239 amino-acid superfolder GFP (sfGFP) protein-coding sequence served as the basis for GFP designs. This sequence has been shown to be a superior reporter for gene expression due to its fast folding kinetics and improved stability (*Pédelacq et al., 2006*). The sfGFP DNA sequence was recoded at synonymous sites using the EuGene genetic optimization software (*Gaspar et al., 2012*). To create the GFP-L, M, and H variants, the sfGFP gene was recoded three times, with settings in EuGene set to target G + C contents of 40%, 50%, and 60% and with the CAI of each variant optimized to the *E. coli* MG1655 genome. The first 51 base pairs (52.9% G + C) and last 54 base pairs (48.1% G + C) were held constant in all gene variants as these regions are known to alter gene expression (*Grunberg-Manago, 1999*; *Kudla et al., 2009*; *Goodman, Church & Kosuri, 2013*; *Umu et al., 2016*). The CA- and GT-enriched variants of the GFP-L terminal fragment were recoded manually using EuGene by changing each codon to a CA- or GT-rich version wherever possible. Sequences of all gene variants used in this study are presented in Table S2.

CAI and mRNA folding energy calculations

Codon adaptation index (CAI) (*Sharp & Li, 1987*) values for each GFP variant were calculated using the default options of the 'cai' function in Emboss 6.6.0 (*Rice, Longden & Bleasby, 2000*) based on a set of 40 highly expressed genes that included 37 ribosomal proteins and 3 elongation factors (*Sharp et al., 2005*; *Hilterbrand, Saelens & Putonti, 2012*). The folding energy of each GFP variant was calculated using RNALfold (*Hofacker, Priwitzer & Stadler, 2004*) with default options. These values are presented in Table S1.

Gene and plasmid construction

The GFP-L, M, and H gene sequences were each subdivided into three non-overlapping, roughly equal-sized sequence fragments: 5′-distal, proximal, and terminal. The 5′-distal fragment spanned base pair positions 1–255 (encoding amino acids 1–68), the proximal fragment spanned base pair positions 256–459 (encoding amino acids 69–153), and the terminal-fragment spanned base pair positions 460–717 (encoding amino acids 154–239). These nine sequence fragments were synthesized as dsDNA gBLOCK gene fragments (IDT).

To create the full-length GFP-L, M and H genes, as well as the compositional mosaics of high, medium and low-GC fragments, the selected fragments were assembled using either the Ligase Cycling Reaction (LCR) method (*De Kok et al., 2014*) or a modified Gibson assembly method (*Paetzold et al., 2013*), which employed bridging oligos that overlapped with ≈30 bp of each fragment to mediate assembly. Full-length genes were assembled into pUC19 and then subcloned by Gibson assembly (*Gibson et al., 2009*) into the pFAB-series expression vectors—pFAB3701, pFAB3845, pFAB3857, pFAB3833, pFAB3665 and pFAB3689 (*Mutalik et al., 2013*)—in place of the previously encoded RFP gene. The pFAB-series expression vectors contain well-characterized constitutive promoters of varying strength as well as a bicistronic design that has been shown to normalize translation initiation strength regardless of the sequence in the translation initiation region of the expressed gene. By using these vectors, we aimed to eliminate fluctuations in expression due to differential efficiencies of translation initiation caused by the coding sequence and to instead focus on effects arising from the elongation phase. Gene and promoter sequences of all constructs were verified by Sanger sequencing.

Cell growth and fluorescence measurements

Strains were grown overnight at 37 °C with shaking at 200 rpm in LB supplemented with the appropriate antibiotic for plasmid maintenance. Overnight cultures were used to inoculate assay media at a starting $OD_{600} = 0.05$. 100 μL of the inoculated assay media were added to a well of a 96-well flat-bottom plate (Corning, Inc., Corning, NY, USA), and plates were incubated at 37 °C with shaking at 600 rpm. OD_{600} readings were taken on a VICTOR X3 (Perkin Elmer) plate reader. Growth rates were calculated using the Gompertz equation for bacterial cell growth (*Zwietering et al., 1990*) using GraphPad Prism Version 6.0. For fluorescence measurements, 10 μL of culture media were added to 90 μL of ddH$_2$0 in a black, clear bottom 96-well plate (Corning, Inc., Corning, NY, USA) and assayed at an excitation/emission wavelength of 485/535 nm in the plate reader. Simultaneously, an OD_{600} measurement was taken, and fluorescence was calculated as $RFU_{485/535}/OD_{600}$.

Strain construction

The P90Q mutation was introduced into the *rpsL* gene of *E. coli* strain BW25113 using the oligo-mediated λ red 'recombineering' method (*Ellis et al., 2001*). Briefly, strain BW25113 was transformed with plasmid pKD46 (*Datsenko & Wanner, 2000*), which contains the λ red recombinase genes under the control of the AraC arabinose-inducible promoter. Cells were grown in LB to an $OD_{600} \approx 0.4$, then treated with 0.1% L-Arabinose for 1 hr to induce expression of the λ red recombinase genes. Cells were made electrocompetent and then electroporated with 1 μg of oligo EQ_rpsL_P90Q (A*A*GCG CAC CAC GTA CGG TGT GGT AAC GAA CAC CC TGG AGG TCT TTA ACA CGA CCG CCA CGG ATC AGG A*T*C), allowed to recover, and plated on LB-strep agar. The *rpsL* gene of the resulting Sm^R strain was PCR amplified and sequenced to confirm the introduction of the P90Q mutation.

RESULTS

The base composition of the GFP gene was systematically modified from the whole-gene to the individual-nucleotide levels to pinpoint the specific features responsible for altering *E. coli* growth rates. We initially assayed the fitness effects and expression levels of GFP genes recoded across synonymous sites to span a range of G + C contents. We then synthesized GFP genes that were mosaics of high- and low-GC fragments to localize the source of fitness defects to a particular region of the gene. This guided our focus towards specific codon positions at which synonymous rare codon substitutions were capable of restoring fitness to wild-type levels.

Overall base composition of GFP genes modulates cellular fitness and gene expression levels

To test the effects of genic G + C content on cellular fitness, we recoded the nucleotide sequence of the GFP gene to produce coding sequences with a consistently low (L; 41%), medium (M; 50%), or high (H; 59%) G + C content across the entire gene (Table S2). This was achieved by recoding genes with synonymous codons that differ in G + C content, which ensured that the protein sequence encoded by the various gene constructs remained identical.

The three GFP genes, GFP-L, GFP-M, and GFP-H, were each tested for expression and growth-rate characteristics in a series of pFAB vectors containing promoters of different strengths (*Mutalik et al., 2013*). Overall levels of expression, as measured by cellular GFP fluorescence, depended on the intrinsic strength of the pFAB promoter; however, in each of the expression vectors in which all three GFP constructs could be cloned, we observed the identical trend: the GFP gene of low G + C contents (GFP-L) expressed at higher levels than the gene of high G + C contents (GFP-H), and the gene of intermediate G + C contents (GFP-M) expressed at the lowest level (Fig. 1A).

Of particular note is that the GFP-L gene, which typically expresses at the highest level, could not be stably cloned into the three expression vectors possessing the strongest promoters (Fig. 1A). Transformations of these constructs yielded dramatically reduced numbers of recombinant clones and were usually not fluorescent (but see later section). Those few clones that were recovered were found to possess mutations in the coding or promoter sequences, suggesting that high-level expression of the GFP-L gene is lethal, leading to selection for mutations that alleviate its toxicity.

The toxicity of the GFP-L gene was similarly apparent from the growth rates of strains that were capable of expressing this gene from weaker promoters: growth rates of strains expressing the GFP-L gene decreased sharply with increasing expression level (Fig. 1B). In contrast, expression of the GFP-M or GFP-H genes did not adversely affect growth rates, even at the highest levels of expression. These results agree with the previous findings that recoded GFP genes of low G + C contents impose a fitness cost when expressed in *E. coli* (*Raghavan, Kelkar & Ochman, 2012*; *Kelkar, Phillips & Ochman, 2015*) and show that this fitness decrement is dose-dependent with protein expression levels (Fig. S1).

Compositional mosaics reveal sequence features underlying cellular toxicity

To determine whether a specific sequence motif or the G + C content of a particular genic region was responsible for the observed differences in gene expression and cellular toxicity, we shuffled portions of the previously synthesized GFP genes to produce compositional mosaics. Each of the three compositional variants of the GFP from above was subdivided into three gene fragments representing the 5′-distal, proximal, and terminal regions. These fragments were reassembled to produce mosaic genes corresponding to all 27 possible full-length sequence combinations. The compositional mosaics were assayed in the pFAB3857 vector, which was selected on account of its intermediate level of expression (Fig. 1A) and because differences in cellular growth rate resulting from the G + C content of expressed GFP genes were explicit (Fig. 1B).

The 27 compositional mosaics in this library differed by up to 5-fold in GFP fluorescence (Fig. 2A). There was a negative association between fluorescence and overall genic G + C content ($r^2 = 0.141$, $p = 0.0006$) as well as with overall genic CAI ($r^2 = 0.378$, $p < 0.0001$) (Fig. S2). However, these trends were due, in large part, to the genes containing a L-fragment in the 5′-distal position, which together form a separate group (Fig. S3) independent of overall genic G + C content or overall genic CAI (Fig. S4). The genes with the 5′-distal L fragments displayed significantly higher levels of fluorescence than those with the 5′-distal
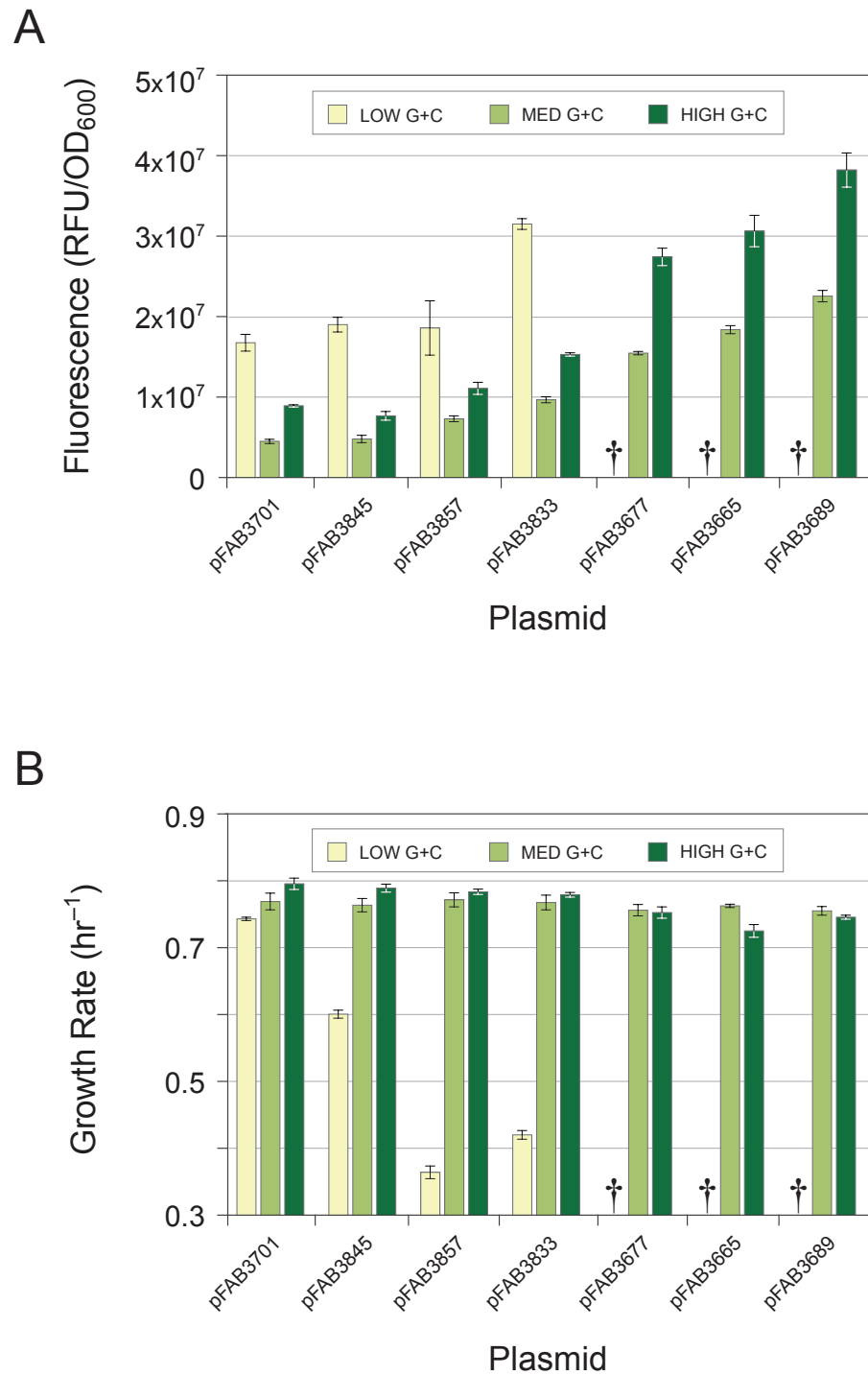
A

B

Figure 1 **Effects of promoter strength and base composition on expression level and cellular growth rate.** GFP genes were recoded at synonymous sites to possess a consistently low (yellow; 41%), medium (light green; 50%) or high (dark green; 59%) G + C content over the entire gene. (A) GFP constructs of different G + C contents tested for expression levels in multiple pFAB vectors, displayed in order of increasing promoter strength. Expression levels determined by intensity of cell fluorescence. Daggers (†) denote constructs refractory to cloning, all of which were of low G + C contents. (B) Fitness measurements of strains expressing GFP constructs from (A) determined by exponential-phase growth rates. Daggers (†) denote constructs not assayed. Bars represent the mean ± standard deviation of three biological replicates.
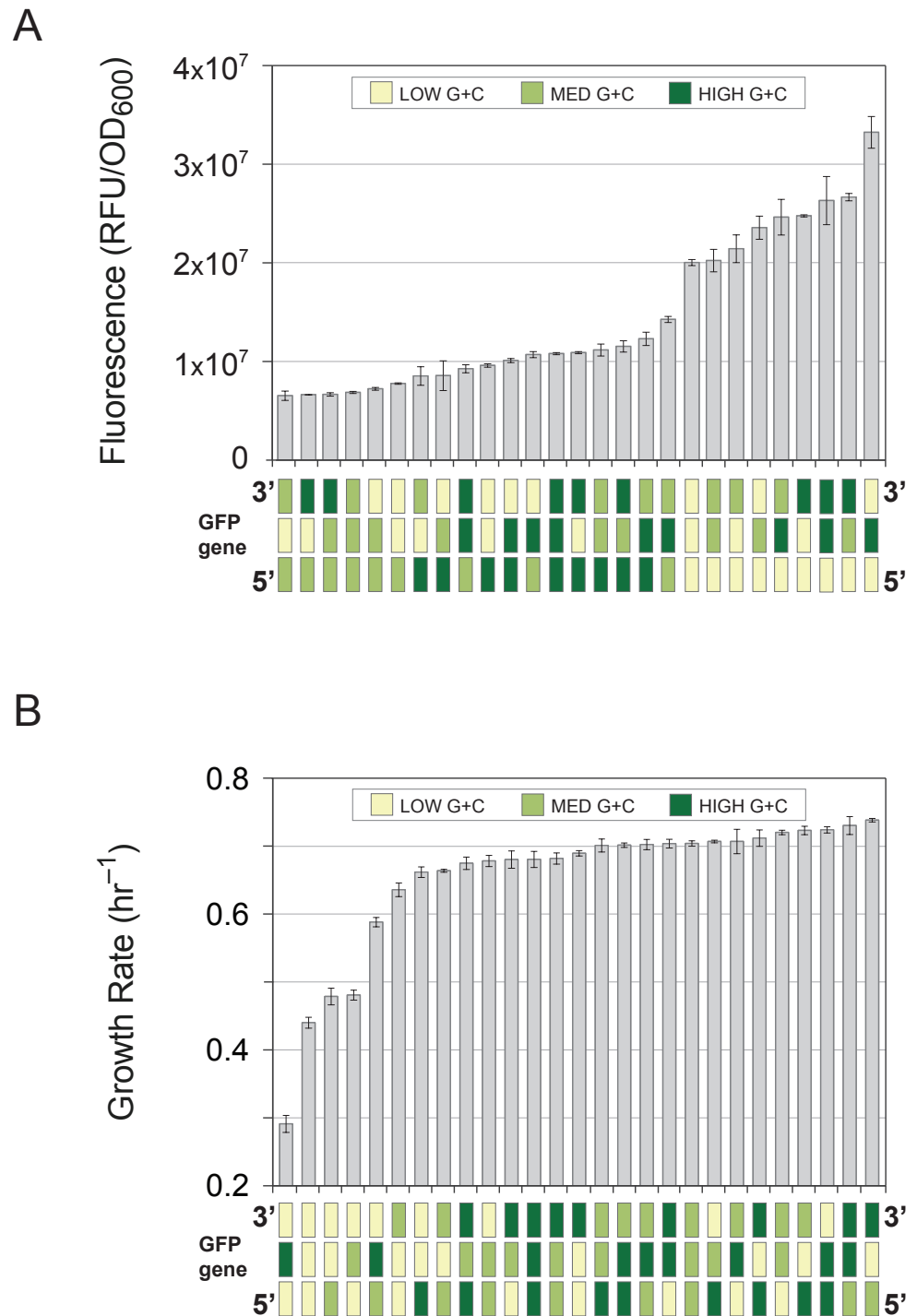
Full-size 🖼 DOI: 10.7717/peerj.4286/fig-1

Figure 2 **Gene expression and cellular growth rate of compositional mosaics.** Recoded GFP genes (as shown in Fig. 1) were each subdivided into three fragments and shuffled to create all 27 possible full-length gene combinations. Each of the shuffled GFP genes encodes the identical protein sequence, and all constructs were cloned and expressed in the pFAB3857 vector. Gene (continued on next page...)

Full-size ⬚ DOI: 10.7717/peerj.4286/fig-2

Figure 2 (…continued)
diagrams on the *x*-axis show the mosaic structure of each of the tested genes, shaded according to the GC-content of each fragment: yellow; 41% G + C; light green; 50% G + C; dark green; 59% G + C. (A) Gene expression levels of mosaics GFP genes determined by intensity of cell fluorescence, with constructs arranged in order of increasing expression. Bars represent the mean ± standard deviation of three biological replicates. (B) Growth rates of *E. coli* strains expressing mosaic GFP genes. Fitness measurements the 27 GFP variants determined by exponential-phase growth rate, with constructs arranged in order of increasing growth rates. Bars represent the mean ± standard deviation of three biological replicates.

M fragment ($p < 0.0001$) or the 5′-distal H fragment ($p < 0.0001$) (Fig. S5). These results indicate that some property associated with the G + C content of the 5′-distal fragment modulates expression levels. This finding aligns with prior studies of the translation initiation region that have found that low G + C sequences near the 5′-end tend to increase expression levels (*Kudla et al., 2009*; *Goodman, Church & Kosuri, 2013*). However, there has been some disagreement as to the underlying mechanism. The leading hypothesis is that lower levels of G + C content in the initiation region encourage ribosome loading due to a reduction in the strength of mRNA secondary structures (*Kudla et al., 2009*; *Goodman, Church & Kosuri, 2013*). However, it has also been argued that the effect is attributable to the low CAI values associated with many low G + C codons. Under this model, slower translation at the start of the transcript, as caused by low CAI codons, reduces ribosome collisions that are detrimental to expression (*Tuller et al., 2010*).

We observe a slightly stronger correlation between fluorescence and the G + C content of the 5′-distal fragment ($r^2 = 0.6219$, $p < 0.0001$) than with CAI ($r^2 = 0.4793$, $p < 0.0001$) (Fig. S6). It is curious that our constructs displayed this effect since all constructs were identical for the first 51 bp—a length thought to mitigate any sequence-specific effects caused by this region, as had been shown for native *E. coli* gene sequences (*Goodman, Church & Kosuri, 2013*). Additionally, we employed a bicistronic expression system, which has been shown to normalize translation initiation regardless of gene sequence (*Mutalik et al., 2013*). These results suggest that the size of the region affecting the efficiency of translation initiation is longer than previously established and that its effect cannot be wholly eliminated with the bicistronic expression format. In support of this view, another recent study has also shown that a more significant portion of the coding sequence may play a role in defining rates of translation initiation (*Burkhardt et al., 2017*).

When assayed for their effects on cellular fitness, most of the 27 compositional mosaics resulted in only minor reductions in growth rates; however, those displaying severely reduced growth rates all contained an L-fragment in the terminal position (Fig. 2B). The magnitude of the fitness defect caused by the terminal L-fragment was associated with expression level: those strains expressing higher levels of the GFP gene, largely due to the G + C content of the 5′-distal fragment, displayed greater degrees of growth rate inhibition ($r^2 = 0.70$, $p < 0.0001$) (Fig. 3).

Sequence features causing cellular toxicity
Having established that the G + C content of the terminal-fragment of the GFP gene produces the greatest effect on cellular fitness, we sought to determine whether the overall G + C content of this region or, alternatively, if some specific sequence feature within
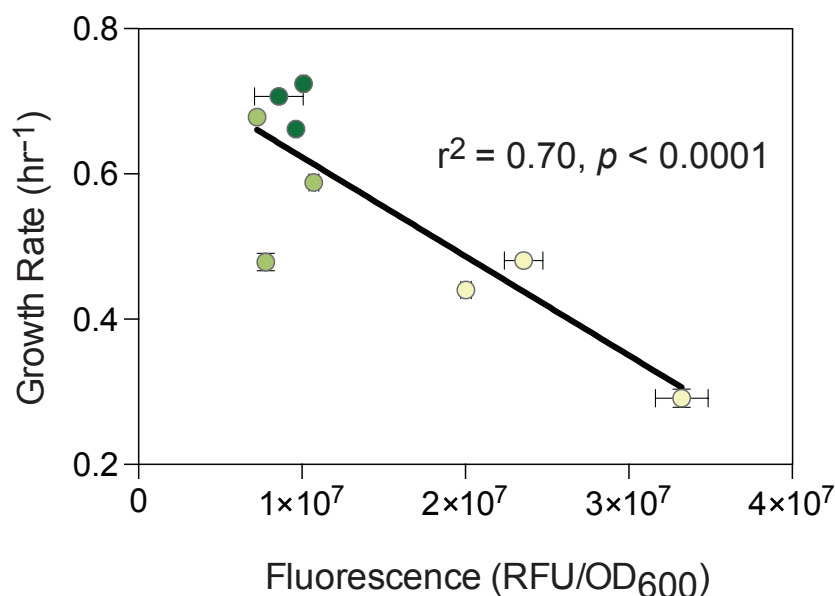
Figure 3 **Association between expression level and cellular growth rates in *E. coli* harboring mosaic GFP genes that each contain the L-fragment in the terminal position.** Shading denotes the GC-content of the 5′-distal fragment: yellow; 41% G + C; light green; 50% G + C; dark green; 59% G + C. Points represent the mean ± standard deviation of three biological replicates.

Full-size ◪ DOI: 10.7717/peerj.4286/fig-3

this region dictated this effect. To distinguish between these hypotheses, we created two additional sets of terminal fragments:

In the first, we created fragments of the same overall G + C contents as the "canonical" L-fragment ($\approx$41%) but recoded to contain a biased nucleotide content in the mRNA coding strand. Two such constructs were created: one in which the coding strand was enriched for cytosine and adenine bases ('CA↑') and another in which the coding strand was enriched for guanine and thymine bases ('GT↑'). Importantly, these nucleotide-enriched constructs were not normalized for CAI values, and as a result, the CA↑ terminal fragment had a lower CAI value than the canonical-L and GT↑ terminal fragments (Table S1). This difference was largely driven by the presence of the rare CTA codon—the leucine codon most enriched in C and A—at all six leucine positions within the region.

The second set consisted of chimeric terminal fragments that were designed to dissect the precise location of any discrete sequence motif(s) responsible for the fitness detriment associated with a terminal L-fragment. Two chimeric terminal fragments were created, one containing half of the canonical L sequence followed by half of the canonical H-sequence ('L/H'), and another containing half of the canonical H sequence followed by half of the canonical L sequence ('H/L'). Both of these chimeras were assembled with the canonical low G + C L-5′-distal and L-proximal fragments to create full-length genes, and assayed in the pFAB3857 expression vector.

The terminal gene-fragment variants in the first set ('CA↑' and 'GT↑') affected cellular fitness in different ways: although both constructs had the same overall base composition,

the CA↑ fragment did not produce the fitness defect normally associated with the canonical L fragment. In contrast, the GT↑ terminal-fragment induced growth defects greater than that produced by the canonical L-fragment (Fig. 4). Note that both variants resulted in lower levels of GFP fluorescence relative to the canonical construct but that only the GT↑ fragment was associated with the fitness defect (Fig. 4).

The chimeric L/H and H/L variants both displayed about the same growth rate and expression level as the variant with canonical H-sequence throughout the terminal fragment (Fig. 4). These two sets of constructs, one of which altered base composition but not the overall %G + C of the terminal fragment and the other which swapped the positions of the GC-rich region in the terminal fragment, indicated that the toxicity associated with the canonical L-terminal fragment is not simply a function of %G + C but is mediated by a specific combination of sequence features within the mRNA.

Attenuation of toxicity by rare codons

The toxicity of the canonical GFP-L gene variant made it impossible to recover clones when expressed by the three strongest promoters (Fig. 1B). Fortuitously, in an attempt to clone this fragment into the highly-expressing pFAB3665 vector, a single viable and fluorescent colony was recovered. This clone possessed a point mutation within the terminal-fragment of the GFP-L gene, converting G to A at nucleotide position 534. This mutation resulted in a synonymous substitution in the leucine codon at amino acid position 178, changing it from the preferred CTG codon to the rarely used CTA codon (*Ikemura, 1985*; *Sharp & Li, 1987*). This mutation, which restored cellular growth, was especially surprising given that it reduced the G + C content of the gene.

To characterize the effects of this mutation on cellular fitness and fluorescence properties, we cloned the GFP-L L178L sequence variant (Leu-1) into a lower-expressing pFAB3857 vector, which is viable and allows comparisons with the canonical GFP-L sequence. This L178L sequence variant exhibited growth rates similar to those expressing the GFP-M or H genes and near wild-type levels of fluorescence (Fig. 4). Therefore, this single synonymous mutation, which imparts a *lower* G + C content, suppresses the toxicity normally associated with the L-terminal fragment, but it does not operate through the reduction of gene expression.

This result led us to wonder if the suppression induced by the L178L mutation was due to a reduced rate of translation, as can occur at rare codons (*Sørensen, Kurland & Pedersen, 1989*; *Gardin et al., 2014*). Such attenuation sites have previously been shown to be involved in the co-translational folding of nascent proteins (*Kimchi-Sarfaty et al., 2007*; *Komar, 2009*; *Zhang, Hubalewska & Ignatova, 2009*) and may affect the fidelity of translation (*Kramer & Farabaugh, 2007*), either of which could have consequences for cellular fitness (*Drummond & Wilke, 2009*). To test if the L178L mutation is associated with translational attenuation—and if the specific location of this putative attenuation site is critical for suppression—we designed two variants of the canonical GFP-L terminal sequence in which the CTA leucine codon was substituted at various sites downstream of leucine 178. In one construct (Leu-2), we converted the CTG leucine codons at sites 194 and 195 to rare leucine CTA codons, and in the other (Leu-3), we converted the CTG
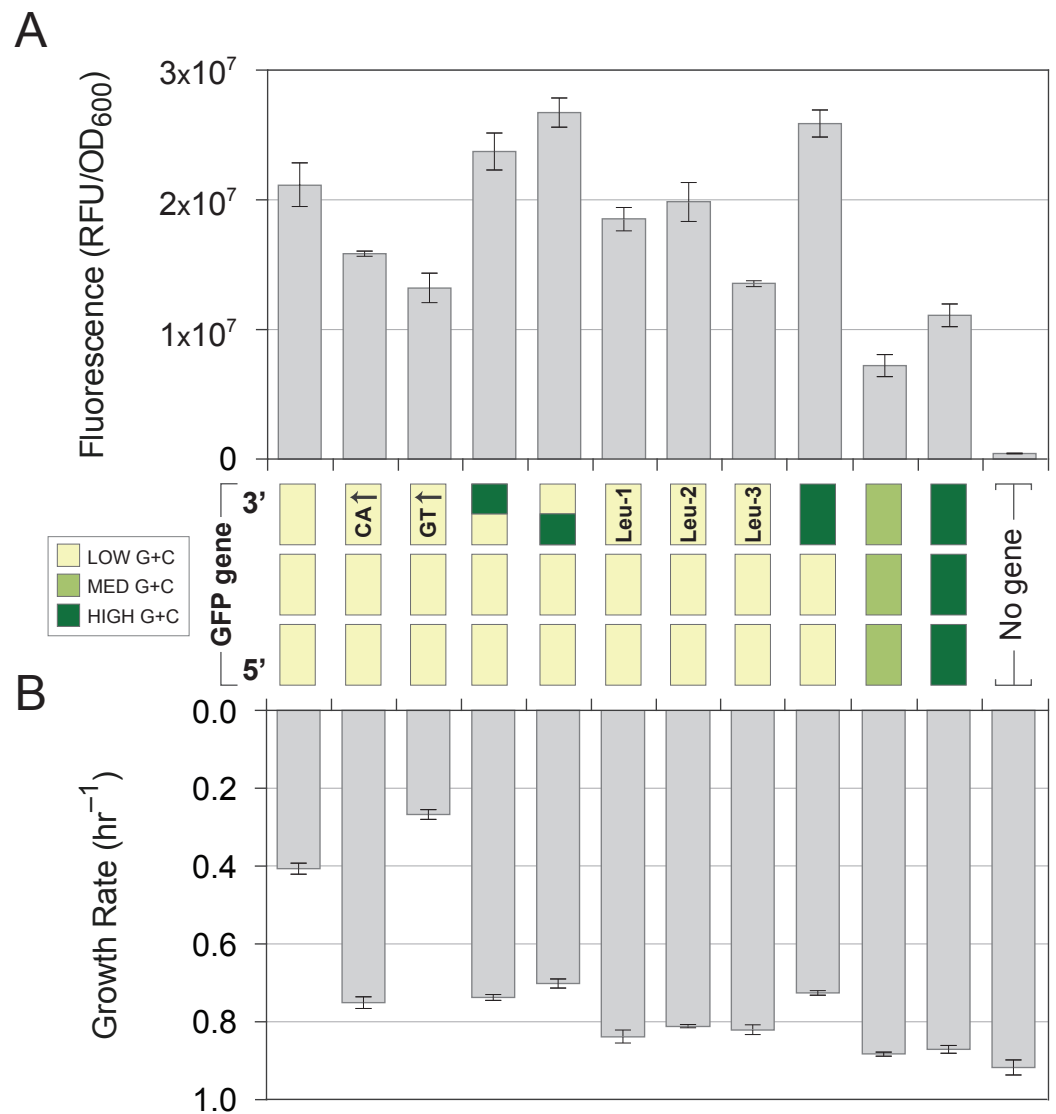
Figure 4 **Synonymous nucleotide substitutions in the GFP terminal-fragment influence cellular growth rates and gene expression levels.** Terminal-fragments of GFP genes (nucleotide positions 460–717) were modified to contain different sequence features, all of which preserved the GFP protein sequence, and assembled into full-length genes along with canonical low (yellow; 41% G + C) 5′-distal and proximal fragments. All constructs were cloned and expressed in plasmid pFAB3857. Notation of terminal-fragments is as follows: yellow unlabeled, the canonical low terminal-fragment of 41% G + C; CA↑, a low G + C fragment enriched in cytosine and adenine nucleotides in the coding strand; GT↑, a low G + C fragment enriched guanine and thymidine nucleotides in the coding strand; yellow-green chimera, a sequence that contains half of the canonical low G + C sequence followed half of the canonical high G + C sequence; green-yellow chimera, a sequence that contains half of the canonical high G + C sequence followed half of the canonical low G + C sequence; Leu-1, the canonical low G + C terminal-fragment possessing a synonymous mutation that altered the leucine codon at amino acid position 178 from CTG to CTA; Leu-2, the canonical low G + C terminal-fragment possessing synonymous mutations that altered the leucine codons at amino acid positions 194 and 195 from CTG to CTA; Leu-3, the canonical low G + C terminal-fragment possessing synonymous mutations that altered the leucine codons at amino acid positions 220 and 221 from CTG to CTA; dark (continued on next page...)

leucine codons at sites 220 and 221 to CTA codons. These new terminal-fragments were assembled with the canonical low G + C L-5′-distal and L-proximal fragments to create full-length genes, and cloned into the pFAB3857 expression vector.

Each of the three rare-codon variants was able to rescue the fitness defect previously observed in the canonical L terminal-fragment, and like the original L178L mutation, each displayed only a minimal reduction in GFP fluorescence (Fig. 4). These results strengthen the hypothesis that the toxic effect of the canonical L terminal-fragment is due to a too-rapid rate of translation. Furthermore, each of these rare-codon constructs suppresses translational toxicity to a similar degree although they spanned a region of 42 amino acids.

Reductions in global translation rates can alleviate cellular toxicity

Our finding that insertion of rare codons into the L terminal-fragment was sufficient to ameliorate the fitness defect associated with its expression suggested that the mechanism of suppression involved a reduction in translational rate through this region. It has previously been observed that eukaryotic proteins (such as GFP) are prone to misfolding when expressed in bacteria (*Fukuda, Arai & Kuwajima, 2000*; *Chang et al., 2005*). This propensity for misfolding has been attributed, in part, to the faster overall rate of translation elongation in bacteria, which can interfere with proper co-translational protein folding (*Siller et al., 2010*). We reasoned that if this faster rate of translation was responsible for the fitness defect observed from expression of GFP genes containing the L terminal-fragment then reducing the overall rate of translation in this context would have an alleviating effect on fitness. For these experiments, we took advantage of the fact that the mutations in the *rpsL* 30S ribosomal subunit S12 gene that confer resistance to the antibiotic streptomycin (Sm^R) do so, in part, by reducing the overall rate of translation elongation (*Kurland, Hughes & Ehrenberg, 1996*).

We introduced the P90Q mutation into the chromosomal *rpsL* sequence, producing a Sm^R strain with an error-restrictive phenotype (*Holberger & Hayes, 2009*) associated with reduced elongation rates and hyper-accurate translation (*Kurland, Hughes & Ehrenberg, 1996*). We then compared growth rates and fluorescence properties associated with the canonical GFP-L gene in wild-type and isogenic Sm^R strains using vectors of different expression strengths (pFAB3845 and pFAB3857). With both vectors, levels of growth rate inhibition were significantly lower in the Sm^R strain than in the wild-type strain ($p = 0.0018$ for pFAB3845; $p < 0.0001$ for pFAB3857, unpaired Student's t-test), supporting the hypothesis that translation rates underlie the fitness defect associated with canonical L terminal-fragment (Fig. 5A). The level of GFP fluorescence was not found to be significantly different between both strains when the GFP-L gene was expressed from the pFAB3845 vector ($p = 0.068$, unpaired Student's t-test), but was higher for the Sm^R strain when
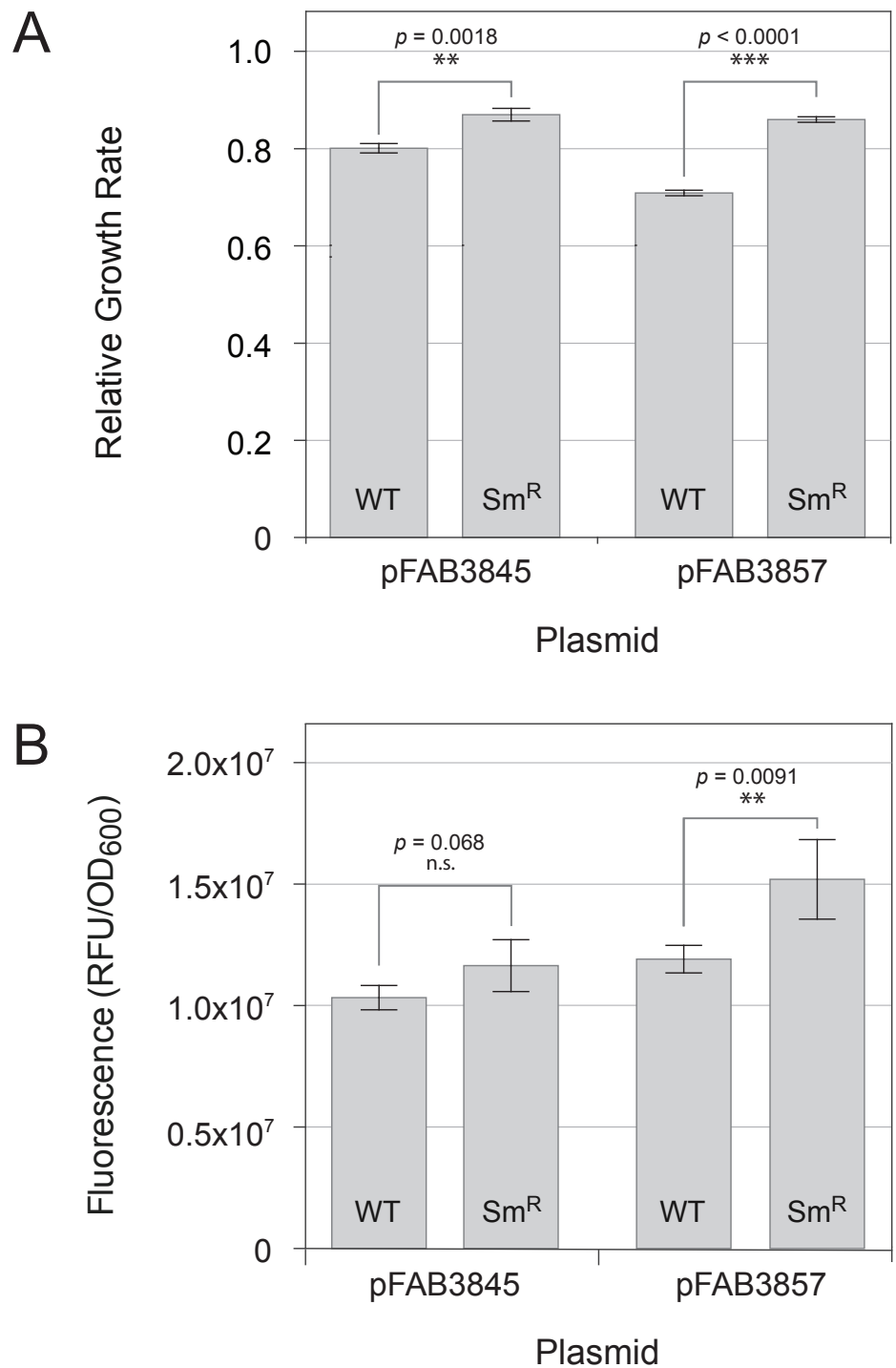
Figure 5 Effect of reduced translation rate on cellular fitness and gene expression. The canonical GFP-L gene, which possesses a consistently low G + C content over the entire gene sequence, was cloned and expressed in plasmid vectors with either a weak (pFAB3845) or strong (pFAB3857) promoter. Cloned constructs were assayed in the wild-type *E. coli* (WT, strain BW25113) (continued on next page...)

Full-size 🖾 DOI: 10.7717/peerj.4286/fig-5

GFP-L was expressed from the pFAB3857 vector ($p = 0.0091$, unpaired Student's t-test) (Fig. 5B). This result rules out the possibility that the growth rate enhancement of the Sm^R strains is a consequence of reduced protein expression. Taken together, these results suggest that expression of the GFP-L gene is toxic to wild-type cells because of a fitness cost associated with rapid translation.

DISCUSSION

We show that the fitness costs associated with the expression of GFP sequences of low $G + C$ content are due almost entirely to the sequence within a very limited region near the 3′ terminus of the gene, indicating that selection does not act uniformly on base composition across the entire gene sequence. This finding indicates that the GC-effect on fitness is not caused by collective differences in the rate of translation at G/C *vs.* A/T sites but instead results from local sequence features that affect translation. Systematic dissection of the region associated with the fitness decrement revealed that introduction of either a high G + C sequence or a small number of rare synonymous codons were sufficient to restore fitness to normal levels, suggesting that both affected the same process.

A feature common to these alterations is that both the presence of certain rare codons and increases to the G + C content of mRNA can reduce rates of translation. Rare synonymous codons, like those used in our experiments, decrease rates of translation in *E. coli* on account of the limited concentrations of their isoaccepting tRNAs (*Pedersen, 1984*; *Varenne et al., 1984*; *Sørensen, Kurland & Pedersen, 1989*; *Dong, Nilsson & Kurland, 1996*; *Chaney & Clark, 2015*). Similarly, higher G + C contents tend to increase the stability of mRNA secondary structures, which can impede ribosome progression along the transcript (*Somogyi et al., 1993*; *Wen et al., 2008*; *Chen et al., 2013*). To confirm that translation speed is responsible for the observed fitness differences, we determined that expression of low G + C GFP variants are less deleterious when global rates of translation are reduced. Collectively, these findings all indicate that the primary source of the fitness defect caused by low G + C contents results from a regional rate of translation that is too fast.

Because the GFP gene variants differed in their overall base compositions, we attempted to control for differences in translation rate that might arise from this compositional variation by designing constructs with similar overall levels of adaptive codon usage bias. However, our results demonstrate that the G + C context in which a particular codon exists can alter the rate of translation in a way that impacts cellular fitness, even in genes, such as GFP, that have no functional relevance in *E. coli*. Such factors are not considered in current measures of codon adaptation, which are typically based solely on the frequency

of each codon in the most highly expressed genes in a genome (*Sharp & Li, 1987*) or on the relative concentrations of isoaccepting tRNAs (*Dos Reis, Savva & Wernisch, 2004*). As *Gorochowski et al. (2015)* suggest, the failure to account for G + C content and its effect on the mRNA structural context of a given codon may be one reason why ribosomal profiling often fails to confirm the expected correlation between measures of codon adaptation and translational rate (*Li, Oh & Weissman, 2012*; *Qian et al., 2012*; *Charneski & Hurst, 2013*; *Pop et al., 2014*; *Gorochowski et al., 2015*).

That slowing the rate of translation imparts fitness benefits seems to contradict the view that a fast translation rate is generally beneficial because its ensures that the concentration of ribosomes does not becomes limiting during periods of rapid growth (*Andersson & Kurland, 1990*; *Plotkin & Kudla, 2011*). However, a fast rate of translation is not always ideal. For example, since proteins begin to fold as they emerge from the ribosome, the rate at which the sequence is translated will affect the conformational space that the protein can explore as it folds (*Tsai et al., 2008*; *Komar, 2009*; *O'Brien et al., 2014*). Attenuation sites often serve to slow or pause translation so that certain portions of the protein can adopt a particular configuration, and the absence of such sites can cause protein misfolding and result in altered functionality (*Kimchi-Sarfaty et al., 2007*) or loss of solubility and aggregation (*Zhang, Hubalewska & Ignatova, 2009*). The production of misfolded proteins is energetically costly and can cause toxic effects due to their altered structures (*Bucciantini et al., 2002*; *Drummond & Wilke, 2008*; *Drummond & Wilke, 2009*). And aside from directly affecting protein folding, a fast rate of translation can also negatively impact the fidelity of translation (*Tubulekas & Hughes, 1993*; *Johansson, Zhang & Ehrenberg, 2012*; *Rodnina, 2012*; *Yang, Chen & Zhang, 2014*) resulting in mutated or nonfunctional proteins that affect fitness (*Drummond & Wilke, 2009*). Additionally, this process explains why we observe higher fitness levels when low G + C variants were expressed in a Sm^R strain with hyper-accurate translation. Future experiments that directly measure rates of translation elongation and levels of misfolded or mistranslated proteins in relation to genic G + C content would strengthen support for these hypotheses.

If local levels of G + C content are under selection for the translational tuning of each protein, we might expect heterogeneity in the base composition of genes within a genome. But despite their differences in amino acid composition, the majority of genes within a bacterial genome are of similar G + C contents (*Sueoka, 1962*; *Muto & Osawa, 1987*; *Karlin, Campbell & Mrázek, 1998*), indicating that other factors that play a role in shaping base composition. For instance, translation rates can also be modulated by codon and amino acid usage (*Ingolia et al., 2009*; *Pavlov et al., 2009*; *Gingold & Pilpel, 2011*; *Charneski & Hurst, 2013*), the presence of anti-Shine-Dalgarno sequences (*Li, Oh & Weissman, 2012*; *Fluman et al., 2014*; *Vasquez et al., 2016*), and interactions between the mRNA transcript and other RNAs in the cell (*Umu et al., 2016*). The combined influence of these factors makes selection on base composition highly dependent on sequence context, potentially having a homogenizing effect on overall base composition.

Although low GC-variants of GFP cause fitness defects, it is not known where the GFP protein falls on the spectrum of protein-folding robustness in relation to its base composition. Because GFP is known to have some co-translational folding requirements

(*Kelkar et al., 2012*) and originates from an eukaryote, which have markedly slower translation rates than bacteria (*Siller et al., 2010*), it may be prone to misfolding when translated at speeds typical of *E. coli*. Furthermore, the GFP protein is comprised mainly of ß-strands, which are thought to favor slower rates of translation, as evidenced by the preferential usage of poorly adapted codons in their coding sequences (*Thanaraj & Argos, 1996*). Assaying genes that are native to *E. coli* will help determine if resident sequences experience patterns of selection for base composition similar to those observed for GFP.

Horizontally acquired coding sequences are often of lower G + C content than genes native to a genome (*Lawrence & Ochman, 1997*; *Daubin, Lerat & Perrière, 2003*). In enteric bacteria, these acquired sequences are silenced by the H-NS protein, which targets and prevents the expression of sequences with low G + C contents (*Lucchini et al., 2006*; *Navarre et al., 2006*). The silencing of foreign genes has conventionally been viewed as a way to counteract the costs associated with the unregulated expression of superfluous genes. We have shown that low G + C sequences are deleterious at the level of translation, thereby providing an additional reason why such sequences need to be suppressed by H-NS.

The extent to which selection on genic G + C contents operates in bacteria other than *E. coli* is largely unknown. It might be expected that selection would be strongest in genomes with extreme GC-contents; however, the effect was observed in only one of the two high G + C organisms tested (*Kelkar, Phillips & Ochman, 2015*). Despite selection for higher GC-contents, bacterial species display base compositions as low as 13% G + C (*McCutcheon & Moran, 2010*). This variation in base composition among species could be ascribed to differences in the efficacy of selection caused by population-level processes or to other mechanisms that can compensate for detriments arising from suboptimal rates of translation. It is notable that endosymbiotic bacteria, which possess the most AT-biased genomes, typically overproduce the chaperone GroEL (*Dale & Moran, 2006*). This chaperone assists in folding the highly degraded protein sequences that accumulate in these genomes (*Fares et al., 2002*). Selection for higher GC-contents is likely to be too weak in these genomes to counteract the fixation of deleterious AT-mutations by drift, such that GroEL might also have a role in stabilizing proteins that fold incorrectly due to an improper rate of translation.

CONCLUSIONS

The maintenance of a translation rate conducive to the production of properly folded proteins places selective pressure both on base composition and on codon usage, which both interact to determine local rates of translation. Since small changes in base composition can alter the rate at which key codons are decoded, it implies that selection can act on individual mutations that affect base composition. Thus, protein folding requirements for translational speed can serve a role in shaping base composition.

ACKNOWLEDGEMENTS

ADDITIONAL INFORMATION AND DECLARATIONS

Competing Interests

The authors declare there are no competing interests.

Author Contributions

- Erik M. Quandt and Charles C. Traverse conceived and designed the experiments, performed the experiments, analyzed the data, wrote the paper, prepared figures and/or tables, reviewed drafts of the paper.
- Howard Ochman conceived and designed the experiments, analyzed the data, wrote the paper, prepared figures and/or tables, reviewed drafts of the paper.

Data Availability

The following information was supplied regarding data availability:
The raw data has been provided as Supplemental File.

Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj.4286#supplemental-information.

REFERENCES

Andersson SG, Kurland CG. 1990. Codon preferences in free-living microorganisms. *Microbiological Reviews* **54**:198–210.

Baba T, Ara T, Hasegawa M, Takai Y, Okumura Y, Baba M, Datsenko KA, Tomita M, Wanner BL, Mori H. 2006. Construction of *Escherichia coli* K-12 in-frame, single-gene knockout mutants: the Keio collection. *Molecular Systems Biology* **2**:2006.0008 DOI 10.1038/msb4100050.

Bucciantini M, Giannoni E, Chiti F, Baroni F, Formigli L, Zurdo J, Taddei N, Ramponi G, Dobson CM, Stefani M. 2002. Inherent toxicity of aggregates implies a common mechanism for protein misfolding diseases. *Nature* **416**:507–511 DOI 10.1038/416507a.

Burkhardt DH, Rouskin S, Zhang Y, Li GW, Weissman JS, Gross CA. 2017. Operon mRNAs are organized into ORF-centric structures that predict translation efficiency. *eLife* **6**:e22037 DOI 10.7554/eLife.22037.

Chaney JL, Clark PL. 2015. Roles for synonymous codon usage in protein biogenesis. *Annual Review of Biophysics* **44**:143–166 DOI 10.1146/annurev-biophys-060414-034333.

Chang HC, Kaiser CM, Hartl FU, Barral JM. 2005. *De novo* folding of GFP fusion proteins: high efficiency in eukaryotes but not in bacteria. *Journal of Molecular Biology* **353**:397–409 DOI 10.1016/j.jmb.2005.08.052.

Charneski CA, Hurst LD. 2013. Positively charged residues are the major determinants of ribosomal velocity. *PLOS Biology* **11**:e1001508 DOI 10.1371/journal.pbio.1001508.

Chen C, Zhang H, Broitman SL, Reiche M, Farrell I, Cooperman BS, Goldman YE. 2013. Dynamics of translation by single ribosomes through mRNA secondary structures. *Nature Structural & Molecular Biology* **20**:582–588 DOI 10.1038/nsmb.2544.

Dale C, Moran NA. 2006. Molecular interactions between bacterial symbionts and their hosts. *Cell* **126**:453–465 DOI 10.1016/j.cell.2006.07.014.

Datsenko KA, Wanner BL. 2000. One-step inactivation of chromosomal genes in *Escherichia coli* K-12 using PCR products. *Proceedings of the National Academy of Sciences of the United States of America* **97**:6640–6645 DOI 10.1073/pnas.120163297.

Daubin V, Lerat E, Perrière G. 2003. The source of laterally transferred genes in bacterial genomes. *Genome Biology* **4**:R57 DOI 10.1186/gb-2003-4-9-r57.

De Kok S, Stanton LH, Slaby T, Durot M, Holmes VF, Patel KG, Platt D, Shapland EB, Serber Z, Dean J, Newman JD, Chandran SS. 2014. Rapid and reliable DNA assembly via ligase cycling reaction. *ACS Synthetic Biology* **3**:97–106 DOI 10.1021/sb4001992.

Dong H, Nilsson L, Kurland CG. 1996. Co-variation of tRNA abundance and codon usage in *Escherichia coli* at different growth rates. *Journal of Molecular Biology* **260**:649–663 DOI 10.1006/jmbi.1996.0428.

Dos Reis M, Savva R, Wernisch L. 2004. Solving the riddle of codon usage preferences: a test for translational selection. *Nucleic Acids Research* **32**:5036–5044 DOI 10.1093/nar/gkh834.

Drummond DA, Wilke CO. 2008. Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. *Cell* **134**:341–352 DOI 10.1016/j.cell.2008.05.042.

Drummond DA, Wilke CO. 2009. The evolutionary consequences of erroneous protein synthesis. *Nature Reviews Genetics* **10**:715–724 DOI 10.1038/nrg2662.

Ellis HM, Yu D, DiTizio T, Court DL. 2001. High efficiency mutagenesis, repair, and engineering of chromosomal DNA using single-stranded oligonucleotides. *Proceedings of the National Academy of Sciences of the United States of America* **98**:6742–6746 DOI 10.1073/pnas.121164898.

Fares MA, Ruiz-González MX, Moya A, Elena SF, Barrio E. 2002. Endosymbiotic bacteria: GroEL buffers against deleterious mutations. *Nature* **417**:398 DOI 10.1038/417398a.

Fluman N, Navon S, Bibi E, Pilpel Y. 2014. mRNA-programmed translation pauses in the targeting of *E. coli* membrane proteins. *eLife* **3**:e03440 DOI 10.7554/eLife.03440.

Foerstner KU, Von Mering C, Hooper SD, Bork P. 2005. Environments shape the nucleotide composition of genomes. *EMBO Reports* **6**:1208–1213 DOI 10.1038/sj.embor.7400538.

Fukuda H, Arai M, Kuwajima K. 2000. Folding of green fluorescent protein and the Cycle3 mutant. *Biochemistry* **39**:12025–12032 DOI 10.1021/bi000543l.

Gardin J, Yeasmin R, Yurovsky A, Cai Y, Skiena S, Futcher B. 2014. Measurement of average decoding rates of the 61 sense codons *in vivo*. *eLife* **3**:e03735 DOI 10.7554/eLife.03735.

Gaspar P, Oliveira JL, Frommlet J, Santos MAS, Moura G. 2012. EuGene: maximizing synthetic gene design for heterologous expression. *Bioinformatics* **28**:2683–2684 DOI 10.1093/bioinformatics/bts465.

Gibson DG, Young L, Chuang RY, Venter JC, Hutchison CA, Smith HO, Hutchison 3rd CA. 2009. Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nature Methods* **6**:343–345 DOI 10.1038/nmeth.1318.

Gingold H, Pilpel Y. 2011. Determinants of translation efficiency and accuracy. *Molecular Systems Biology* **7**:481 DOI 10.1038/msb.2011.14.

Goodman DB, Church GM, Kosuri S. 2013. Causes and effects of N-terminal codon bias in bacterial genes. *Science* **342**:475–479 DOI 10.1126/science.1241934.

Gorochowski TE, Avcilar-Kucukgoze I, Bovenberg RAL, Roubos JA, Ignatova Z. 2016. A minimal model of ribosome allocation dynamics captures trade-offs in expression between endogenous and synthetic genes. *ACS Synthetic Biology* **5**:710–720 DOI 10.1021/acssynbio.6b00040.

Gorochowski TE, Ignatova Z, Bovenberg RAL, Roubos JA. 2015. Trade-offs between tRNA abundance and mRNA secondary structure support smoothing of translation elongation rate. *Nucleic Acids Research* **43**:3022–3032 DOI 10.1093/nar/gkv199.

Grunberg-Manago M. 1999. Messenger RNA stability and its role in control of gene expression in bacteria and phages. *Annual Review of Genetics* **33**:193–227 DOI 10.1146/annurev.genet.33.1.193.

Hershberg R, Petrov DA. 2010. Evidence that mutation is universally biased towards AT in bacteria. *PLOS Genetics* **6**:e1001115 DOI 10.1371/journal.pgen.1001115.

Hildebrand F, Meyer A, Eyre-Walker A. 2010. Evidence of selection upon genomic GC-content in bacteria. *PLOS Genetics* **6**:e1001107 DOI 10.1371/journal.pgen.1001107.

Hilterbrand A, Saelens J, Putonti C. 2012. CBDB: the codon bias database. *BMC Bioinformatics* **13**:62 DOI 10.1186/1471-2105-13-62.

Hofacker IL, Priwitzer B, Stadler PF. 2004. Prediction of locally stable RNA secondary structures for genome-wide surveys. *Bioinformatics* **20**:186–190 DOI 10.1093/bioinformatics/btg388.

Holberger LE, Hayes CS. 2009. Ribosomal protein S12 and aminoglycoside antibiotics modulate A-site mRNA cleavage and transfer-messenger RNA activity in *Escherichia coli*. *Journal of Biological Chemistry* **284**:32188–32200 DOI 10.1074/jbc.M109.062745.

Hurst LD, Merchant AR. 2001. High guanine-cytosine content is not an adaptation to high temperature: a comparative analysis amongst prokaryotes. *Proceedings of the Royal Society. B. Biological Sciences* **268**:493–497 DOI 10.1098/rspb.2000.1397.

Ikemura T. 1985. Codon usage and tRNA content in unicellular and multicellular organisms. *Molecular Biology and Evolution* **2**:13–34.

Ingolia NT, Ghaemmaghami S, Newman JRS, Weissman JS. 2009. Genome-wide analysis *in vivo* of translation with nucleotide resolution using ribosome profiling. *Science* **324**:218–223 DOI 10.1126/science.1168978.

Johansson M, Zhang J, Ehrenberg M. 2012. Genetic code translation displays a linear trade-off between efficiency and accuracy of tRNA selection. *Proceedings of the National Academy of Sciences of the United States of America* **109**:131–136 DOI 10.1073/pnas.1116480109.

Kagawa Y, Nojima H, Nukiwa N, Ishizuka M, Nakajima T, Yasuhara T, Tanaka T, Oshima T. 1984. High guanine plus cytosine content in the third letter of codons of an extreme thermophile. DNA sequence of the isopropylmalate dehydrogenase of *Thermus thermophilus*. *Journal of Biological Chemistry* **259**:2956–2960.

Karlin S, Campbell AM, Mrázek J. 1998. Comparative DNA analysis across diverse genomes. *Annual Review of Genetics* **32**:185–225 DOI 10.1146/annurev.genet.32.1.185.

Kelkar DA, Khushoo A, Yang Z, Skach WR. 2012. Kinetic analysis of ribosome-bound fluorescent proteins reveals an early, stable, cotranslational folding intermediate. *Journal of Biological Chemistry* **287**:2568–2578 DOI 10.1074/jbc.M111.318766.

Kelkar YD, Phillips DS, Ochman H. 2015. Effects of genic base composition on growth rate in G + C-rich genomes. *G3* **5**:1247–1252 DOI 10.1534/g3.115.016824.

Kimchi-Sarfaty C, Oh JM, Kim I-W, Sauna ZE, Calcagno AM, Ambudkar SV, Gottesman MM. 2007. A "silent" polymorphism in the MDR1 gene changes substrate specificity. *Science* **315**:525–528 DOI 10.1126/science.1135308.

Komar AA. 2009. A pause for thought along the co-translational folding pathway. *Trends in Biochemical Sciences* **34**:16–24 DOI 10.1016/j.tibs.2008.10.002.

Kramer EB, Farabaugh PJ. 2007. The frequency of translational misreading errors in *E. coli* is largely determined by tRNA competition. *RNA* **13**:87–96 DOI 10.1261/rna.294907.

Kudla G, Murray AW, Tollervey D, Plotkin JB. 2009. Coding-sequence determinants of gene expression in *Escherichia coli*. *Science* **324**:255–258 DOI 10.1126/science.1170160.

Kurland CG, Hughes D, Ehrenberg M. 1996. Limitations of translational accuracy. In: Neidhardt FC, Curtis III R, Ingraham JL, eds. *Escherichia coli and Salmonella: cellular and molecular biology*. Washington, D.C.: American Society for Microbiology Press, 979–1004.

Lawrence JG, Ochman H. 1997. Amelioration of bacterial genomes: rates of change and exchange. *Journal of Molecular Evolution* **44**:383–397 DOI 10.1007/PL00006158.

Li G-W, Oh E, Weissman JS. 2012. The anti-Shine–Dalgarno sequence drives translational pausing and codon choice in bacteria. *Nature* **484**:538–541 DOI 10.1038/nature10965.

Lucchini S, Rowley G, Goldberg MD, Hurd D, Harrison M, Hinton JCD. 2006. H-NS mediates the silencing of laterally acquired genes in bacteria. *PLOS Pathogens* **2**:e81 DOI 10.1371/journal.ppat.0020081.

McCutcheon JP, Moran NA. 2010. Functional convergence in reduced genomes of bacterial symbionts spanning 200 My of evolution. *Genome Biology and Evolution* **2**:708–718 DOI 10.1093/gbe/evq055.

McEwan CE, Gatherer D, McEwan NR. 1998. Nitrogen-fixing aerobic bacteria have higher genomic GC content than non-fixing species within the same genus. *Hereditas* **128**:173–178 DOI 10.1111/j.1601-5223.1998.00173.x.

Mutalik VK, Guimaraes JC, Cambray G, Lam C, Christoffersen MJ, Mai Q-A, Tran AB, Paull M, Keasling JD, Arkin AP, Endy D. 2013. Precise and reliable gene expression via standard transcription and translation initiation elements. *Nature Methods* **10**:354–360 DOI 10.1038/nmeth.2404.

Muto A, Osawa S. 1987. The guanine and cytosine content of genomic DNA and bacterial evolution. *Proceedings of the National Academy of Sciences of the United States of America* **84**:166–169 DOI 10.1073/pnas.84.1.166.

Navarre WW, Porwollik S, Wang Y, McClelland M, Rosen H, Libby SJ, Fang FC. 2006. Selective silencing of foreign DNA with low GC content by the H-NS protein in *Salmonella*. *Science* **313**:236–238 DOI 10.1126/science.1128794.

O'Brien EP, Ciryam P, Vendruscolo M, Dobson CM. 2014. Understanding the influence of codon translation rates on cotranslational protein folding. *Accounts of Chemical Research* **47**:1536–1544 DOI 10.1021/ar5000117.

Paetzold B, Carolis C, Ferrar T, Serrano L, Lluch-Senar M. 2013. *In situ* overlap and sequence synthesis during DNA assembly. *ACS Synthetic Biology* **2**:750–755 DOI 10.1021/sb400067v.

Pavlov MY, Watts RE, Tan Z, Cornish VW, Ehrenberg M, Forster AC. 2009. Slow peptide bond formation by proline and other N-alkylamino acids in translation. *Proceedings of the National Academy of Sciences of the United States of America* **106**:50–54 DOI 10.1073/pnas.0809211106.

Pédelacq J-D, Cabantous S, Tran T, Terwilliger TC, Waldo GS. 2006. Engineering and characterization of a superfolder green fluorescent protein. *Nature Biotechnology* **24**:79–88 DOI 10.1038/nbt1172.

Pedersen S. 1984. *Escherichia coli* ribosomes translate *in vivo* with variable rate. *The EMBO Journal* **3**:2895–2898.

Plotkin JB, Kudla G. 2011. Synonymous but not the same: the causes and consequences of codon bias. *Nature Reviews Genetics* **12**:32–42 DOI 10.1038/nrg2899.

Pop C, Rouskin S, Ingolia NT, Han L, Phizicky EM, Weissman JS, Koller D. 2014. Causal signals between codon bias, mRNA structure, and the efficiency of translation and elongation. *Molecular Systems Biology* **10**:770 DOI 10.15252/msb.20145524.

Qian W, Yang JR, Pearson NM, Maclean C, Zhang J. 2012. Balanced codon usage optimizes eukaryotic translational efficiency. *PLOS Genetics* **8**:e1002603 DOI 10.1371/journal.pgen.1002603.

Raghavan R, Kelkar YD, Ochman H. 2012. A selective force favoring increased G + C content in bacterial genes. *Proceedings of the National Academy of Sciences of the United States of America* **109**:14504–14507 DOI 10.1073/pnas.1205683109.

Rice P, Longden I, Bleasby A. 2000. EMBOSS: the European Molecular Biology Open Software Suite. *Trends in Genetics* **16**:276–277 DOI 10.1016/j.cocis.2008.07.002.

Rocha EPC, Feil EJ. 2010. Mutational patterns cannot explain genome composition: are there any neutral sites in the genomes of bacteria? *PLOS Genetics* **6**:e1001104 DOI 10.1371/journal.pgen.1001104.

Rodnina MV. 2012. Quality control of mRNA decoding on the bacterial ribosome. *Advances in Protein Chemistry and Structural Biology* **86**:95–128 DOI 10.1016/B978-0-12-386497-0.00003-7.

Sharp PM, Bailes E, Grocock RJ, Peden JF, Sockett RE. 2005. Variation in the strength of selected codon usage bias among bacteria. *Nucleic Acids Research* **33**:1141–1153 DOI 10.1093/nar/gki242.

Sharp PM, Li WH. 1987. The codon adaptation index—a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Research* **15**:1281–1295 DOI 10.1093/nar/15.3.1281.

Siller E, DeZwaan DC, Anderson JF, Freeman BC, Barral JM. 2010. Slowing bacterial translation speed enhances eukaryotic protein folding efficiency. *Journal of Molecular Biology* **396**:1310–1318 DOI 10.1016/j.jmb.2009.12.042.

Singer CE, Ames BN. 1970. Sunlight ultraviolet and bacterial DNA base ratios. *Science* **170**:822–826 DOI 10.1126/science.170.3960.822.

Somogyi P, Jenner AJ, Brierley I, Inglis SC. 1993. Ribosomal pausing during translation of an RNA pseudoknot. *Molecular and Cellular Biology* **13**:6931–6940 DOI 10.1128/MCB.13.11.6931.

Sørensen MA, Kurland CG, Pedersen S. 1989. Codon usage determines translation rate in *Escherichia coli*. *Journal of Molecular Biology* **207**:365–377 DOI 10.1016/0022-2836(89)90260-X.

Sueoka N. 1962. On the genetic basis of variation and heterogeneity of DNA base composition. *Proceedings of the National Academy of Sciences of the United States of America* **48**:582–592 DOI 10.1073/pnas.48.4.582.

Thanaraj TA, Argos P. 1996. Protein secondary structural types are differentially coded on messenger RNA. *Protein Science* **5**:1973–1983 DOI 10.1002/pro.5560051003.

Thomas SH, Wagner RD, Arakaki AK, Skolnick J, Kirby JR, Shimkets LJ, Sanford RA, Löffler FE. 2008. The mosaic genome of *Anaeromyxobacter dehalogenans* strain 2CP-C suggests an aerobic common ancestor to the delta-proteobacteria. *PLOS ONE* **3**:e2103 DOI 10.1371/journal.pone.0002103.

Tsai CJ, Sauna ZE, Kimchi-Sarfaty C, Ambudkar SV, Gottesman MM, Nussinov R. 2008. Synonymous mutations and ribosome stalling can lead to altered folding pathways and distinct minima. *Journal of Molecular Biology* **383**:281–291 DOI 10.1016/j.jmb.2008.08.012.

Tubulekas I, Hughes D. 1993. Suppression of *rpsL* phenotypes by *tuf* mutations reveals a unique relationship between translation elongation and growth rate. *Molecular Microbiology* **7**:275–284 DOI 10.1111/j.1365-2958.1993.tb01118.x.

Tuller T, Carmi A, Vestsigian K, Navon S, Dorfan Y, Zaborske J, Pan T, Dahan O, Furman I, Pilpel Y. 2010. An evolutionarily conserved mechanism for controlling the efficiency of protein translation. *Cell* **141**:344–354 DOI 10.1016/j.cell.2010.03.031.

Umu SU, Poole AM, Dobson RC, Gardner PP. 2016. Avoidance of stochastic RNA interactions can be harnessed to control protein expression levels in bacteria and archaea. *eLife* **5**:e13479 DOI 10.7554/eLife.13479.

Varenne S, Buc J, Lloubes R, Lazdunski C. 1984. Translation is a non-uniform process. Effect of tRNA availability on the rate of elongation of nascent polypeptide chains. *Journal of Molecular Biology* **180**:549–576 DOI 10.1016/0022-2836(84)90027-5.

Vasquez KA, Hatridge TA, Curtis NC, Contreras LM. 2016. Slowing translation between protein domains by increasing affinity between mRNAs and the ribosomal anti-Shine-Dalgarno sequence improves solubility. *ACS Synthetic Biology* **5**:133–145 DOI 10.1021/acssynbio.5b00193.

Weiße AY, Oyarzún DA, Danos V, Swain PS. 2015. Mechanistic links between cellular trade-offs, gene expression, and growth. *Proceedings of the National Academy of Sciences of the United States of America* **112**:E1038–E1047 DOI 10.1073/pnas.1416533112.

Wen J-D, Lancaster L, Hodges C, Zeri A-C, Yoshimura SH, Noller HF, Bustamante C, Tinoco I. 2008. Following translation by single ribosomes one codon at a time. *Nature* **452**:598–603 DOI 10.1038/nature06716.

Yang JR, Chen X, Zhang J. 2014. Codon-by-codon modulation of translational speed and accuracy via mRNA folding. *PLOS Biology* **12**:e1001910 DOI 10.1371/journal.pbio.1001910.

Zhang G, Hubalewska M, Ignatova Z. 2009. Transient ribosomal attenuation coordinates protein synthesis and co-translational folding. *Nature Structural & Molecular Biology* **16**:274–280 DOI 10.1038/nsmb.1554.

Zwietering MH, Jongenburger I, Rombouts FM, Van't Riet K. 1990. Modeling of the bacterial growth curve. *Applied and Environmental Microbiology* **56**:1875–1881.